UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Mathematics “Tullio Levi-Civita”
Doctoral Program in Mathemathics
Curriculum: Computational Mathematics
(Ciclo XXXII)

Ph.D. Thesis

Numerical methods for Shallow Water Equations on regular surfaces

Coordinator
Prof. Martino Bardi

Supervisor
Prof. Mario Putti

Candidate
Elena Bachini

ABSTRACT

Shallow water models of geophysical flows must be adapted to geometric characteristics in the presence of a general bottom topography with non-negligible slopes and curvatures, such as a mountain landscape. In this thesis we derive an intrinsic shallow water model starting from the Navier-Stokes equations defined on a local reference frame anchored on the bottom surface. The resulting equations are characterized by non-autonomous flux functions and source terms embodying only geometric information. We show that the proposed model is rotational invariant, admits a conserved energy, is well-balanced, and it is formally a second order approximation of the Navier-Stokes equations with respect to a geometry-based order parameter.

We then derive numerical discretization schemes compatibles with the intrinsic setting of the formulation, starting from studying a first order upwind Godunov Finite Volume scheme intrinsically defined on the bottom surface. We analyze convergence properties of the resulting scheme both theoretically and numerically. Simulations on several synthetic test cases are used to validate the theoretical results as well as more experimental properties of the solver. The results show the importance of taking into full consideration the bottom geometry even for relatively mild and slowly varying curvatures.

The low-order discretization method is subsequently extended to the Discontinuous Galerkin framework. We implement a linear version of the DG scheme defined intrinsically on the surface and we start from the resolution of the scalar transport equation. We test the scheme for convergence and then we move towards the intrinsic shallow water model. Simulations on synthetic test cases are reported and the improvement with respect to the first order finite volume discretization is clearly visible.

Finally, we consider a finite element method for advection-diffusion-reaction equations on surfaces. Unlike many previous techniques, this approach is based on the geometrically intrinsic formulation and the resulting finite element method is fully intrinsic to the surface. In the last part of this work, we lay out in detail the formulation and compare it to a well-established finite element scheme for surface PDEs. We then evaluate the method for several steady and transient problems involving both diffusion and advection-dominated regime. The experimental results show the theoretically expected convergence rates and good performance of the established finite element methods.

SOMMARIO

I modelli di equazioni Shallow Water (SWE) per flussi geofisici devono essere adattati alle caratteristiche geometriche del fondo su cui scorrono in presenza di pendenze e curvature non trascurabili, come ad esempio nei pendii montani. Questa tesi si occupa di derivare un modello SWE intrinseco, partendo dalle equazioni di Navier-Stokes definite su un sistema di riferimento locale ancorato alla superficie del fondo. Le equazioni risultanti sono caratterizzate da flussi non autonomi e termini sorgente che contengono solo informazioni geometriche. Si dimostra che il modello proposto è invariante per rotazioni, conserva l'energia ed è well-balance. Inoltre, il modello è formalmente un'approssimazione del secondo ordine delle equazioni di Navier-Stokes rispetto ad un parametro dipendente dalla geometria per piccole altezze del pelo libero.

In seguito, si propone la soluzione numerical del modello studiato con metodi di discretizzazione compatibili con il setting intrinseco della formulazione. Un approccio numerico completamente intrinseco offre diversi vantaggi sia teorici che pratici. Per prima cosa la forma originale delle equazioni viene mantenuta in termini di leggi di conservazione (o bilancio) che rimangono descritte come tali in un quadro discreto. Questo permette di sfruttare al meglio le proprietà dello schema di discretizzazione anche per equazioni definite su superfici. Inoltre, l'analisi dello schema risulta facilitata perché un'estensione dell'analisi nel caso standard. Infine, l'uso di un sistema di riferimento locale garantisce migliori caratteristiche di ben-condizionamento dello schema numerico anche in casi complessi.

Viene quindi sviluppato uno schema ai volumi finiti, upwind alla Godunov, definito intrinsicamente sulla superficie del fondo. Sono analizzate le proprietà di convergenza dello schema risultante sia dal punto di vista teorico che numerico. Diverse simulazioni su casi test sintetici sono usate per validare i risultati teorici, così come proprietà sperimentali del solutore. I risultati mostrano l'importanza di tenere in considerazione la geometria del fondale anche per curvature relativamente piccole e quasi costanti.

Lo schema ai volumi finiti del primo ordine è successivamente esteso al framework dei metodi di Galerkin discontinui. Viene implementata la versione con funzioni di base lineari di uno schema DG definito intrinsecamente sulla superficie. Dapprima viene verificata la convergenza dello schema per la risoluzione numerica di equazioni di trasporto scalari e in seguito il metodo è applicato alla soluzione del modello intrinseco delle equazioni di Shallow-Water. Sono riportate simulazioni su alcuni casi test sintetici: è possibile apprezzare il miglioramento rispetto allo schema FV del primo ordine precedentemente considerato. Si rileva infatti una più accurata rappresentazione dei treni d'onda a monte del fronte e una drastica diminuzione della viscosità numerica necessaria per la stabilità dello schema in presenza di soluzioni caratterizzate da fronti ripidi o discontinuità.

Infine, viene proposto un metodo ad elementi finiti per equazioni di convezione-diffusione-reazione su superfici. Diversamente da molte tecniche precedenti, questo approccio si basa su una formulazione geometrica intrinseca e di conseguenza anche il metodo ad elementi finiti risultante è completamente intrinseco alla superficie. Nell'ultima parte di questo lavoro si riporta il dettaglio della formulazione e il confronto con uno schema agli elementi finiti per PDE su superfici preso dalla letteratura. In seguito il metodo è verificato su diversi problemi sia stazionari che transitori in regime diffusivo e convettivo. I risultati confermano le stime di convergenza attese teoricamente e la buona performance dello schema proposto.

Contents

List of Figures

List of Tables

1

Introduction

In many practical geophysical applications, such as large-scale modeling of fluid motion in oceans and rivers [49], river morphodynamics and debris flow [75, 53, 51], avalanche dynamics [40], and atmospheric circulation [50], the Shallow Water (SW) approximation is used as an effective tool to derive appropriate reduced dimensionality models for quantitative simulations of such phenomena. Generally, the typically accepted hypothesis is that the flow develops preferentially along one direction, e.g., horizontal, or, equivalently, that one component of the fluid velocity is negligible with respect to the other two. This is the so-called SW assumption. In most cases, the negligible component of the flow velocity is perpendicular to the surface over which motion develops. Then dimensionality reduction proceeds via depth integration of the three dimensional governing equations, e.g., the Navier-Stokes (NS) equations, to arrive at a two-dimensional system with specific flow rate and water depth as unknowns [see e.g. 31].

For slowly varying bottom topographies fluid depth is evaluated along the vertical direction as an approximation of the bottom normal. This approach is generally used in modeling large scale ocean dynamics or atmospheric flows, where the bottom boundary is the geo-sphere and the normal direction coincides with the direction of gravitational forces [61]. Also at smaller scales, typical of models of river hydraulics or granular and snow avalanches, the vertical direction is ordinarily utilized [52]. However, this approximation is valid only for relatively small bottom slopes, commonly estimated at about six degrees with respect to horizontal [22], and for negligible curvatures of the bottom relief. To improve accuracy, ad hoc pressure corrections are often devised to take into account deviations of the vertical pressure profile from the hydrostatic behavior due to bathimetry variability [49]. More recent attempts look at extensions of the Boussinesq scaling approach to evaluate these corrective terms [34], employing sufficiently low order Green-Naghdi polynomial expansions of the vertical pressure profile to combine accuracy and computational efficiency of the resulting model [74]. Another recent non-hydrostatic pressure solver for the nonlinear shallow water equations is proposed in Aricò and Lo Re [7].

The presence of a curved bottom topography plays an important role, increasing the geometric complexity of the fluid streamlines. It is then difficult to accurately identify the negligible velocity component under the SW hypothesis, as the average flow field drastically departs from a rectilinear behavior. To address this problem, Savage and Hutter [66, 67] developed a formulation of the SW model in local curvilinear coordinates based on depth integration along the normal to the topography. Their approach is valid only for small and essentially one-dimensional bottom curvatures and in practice assumes that the fluid surface is parallel to the bottom. This strategy was extended by Bouchut et al. [17], Bouchut and Westdickenberg [16] to consider two-dimensional settings and less restrictive bottom geometries. Starting from the NS equations, the flow velocity component perpendicular to the bottom is considered negligible and a hydrostatic pressure distribution is assumed along local normals. The resulting SW equations are derived by depth integration along the normal direction, under the further hypotheses of a linear velocity distribution (equivalent to assuming constant depth-averaged velocity) and of a fluid depth sufficiently small to guarantee the invertibility of the coordinate transformation. In addition, Bouchut and Westdickenberg [16] proved that their SW system (i) admits a conservative energy equation, (ii) preserves the steady-state of a lake-at-rest, and (iii) is an approximation of order ϵ^2, where ϵ is the aspect ratio between the depth of the fluid layer and the characteristic length along which the phenomenon develops (the SW hypothesis states that $\epsilon \ll 1$). Applications of this model are described in Fernández-Nieto et al. [38], Pelanti et al. [62], Bouchut and Boyaval [15], Moretti et al. [56]. A more intrinsic approach was recently proposed by Fent et al. [37], who suggests to perform depth integration following the so-called "cross-flow" path, along which the tangential component of the fluid velocity is zero. Unfortunately the definition of the "cross-flow" paths is implicit, as it requires the knowledge of the unknown NS velocity field. For this reason, Fent et al. [37] approximate the "cross-flow" path with the direction normal to the bottom starting from a NS system defined on a curvilinear coordinate frame defined on the bottom geometry. The system of SW Equations (SWE) resulting from depth integration turns out to be closely related to the model of Bouchut and Westdickenberg [16], and shares similar approximations and limitations in terms of the geometry of bed topography.

In this work, we propose a new geometrically intrinsic formulation of the SWE on general topography and study its mathematical structure and numerical solution. Our developments take inspiration from the works of Bouchut and Westdickenberg [16] and of Fent et al. [37], both of whom include the effects of the bottom geometry on the SW system. In the former, the derived model includes the bottom geometry using a three-dimensional Cartesian coordinate system. As a consequence, the SWEs contain non-conservative terms that need to be properly handled in the numerical solution phase. In the latter approach, depth integration proceeds using a local coordinate system defined on the bottom surface. Again, non-conservative terms arise in the covariant form of the equations. Similarly, in our work we describe the SW model on a local reference frame. However, differently from previous work, by careful use of contravariant and covariant vectors we are able to arrive at a system that is completely intrinsic to the bottom geometry, with a source term that contains only bottom slope and curvature information. The resulting set of equations is characterized by spatially varying flux functions and bottom-related sources. We study the mathematical structure of the proposed approximation, proving order of accuracy with respect to a "geometric" aspect ratio parameter ϵ_G that includes information on local curvatures and slope of the bottom surface. We then study the hyperbolic structure of the proposed system using bottom-intrinsic differential operators and show that it is invariant under rotational transformations, satisfies the lake-at-rest condition, and admits a conserved energy in absence of bed resistance.

The intrinsic nature of the developed SWE allows the formulation of an intrinsic Finite Volume (FV) discretization, with some complications due to the presence of non-autonomous fluxes and space-varying source terms. Rossmanith et al. [63] were among the first to study the numerical solution of a hyperbolic system on a general manifold by means of a FV scheme defined on a quadrilateral grid. However, the discretization of geometric quantities based on the surface fundamental forms by quadrilateral meshes turns out to be non-consistent [57]. For this reason, in this work we use triangular grids to derive a first order Godunov type FV method. Surface interpolation of the geometric quantities and the existence of an intrinsic divergence theorem provide the necessary tools to produce a bottom-intrinsic discretization. We prove that approximated geometric quantities converge over subsequent refinements of the surface mesh with second order with respect to the mesh parameter h, and consequently that the discrete divergence theorem is exact up to second order. Using the rotation invariance of the SW equations, we define a geometrically adapted one-dimensional Riemann problem on the curvilinear triangle edges. The exact Riemann solver, proposed by Toro [70] for dam-break problems in classical SW system, is directly applied in solving the Riemann problem after carefully assigning the left and right states at the triangle edges in the corresponding local reference frames. The same considerations are implemented in the HLL and HLLC schemes used to address problems with more general wave patterns. Particular attention is required in order to maintain the well-balance property in the discrete setting. We follow the approach proposed by Audusse et al. [8], Bouchut [14] who introduce a general strategy based on a local hydrostatic reconstruction that ensures well-balance and preserves the non-negativity of the water height. We extend the same idea to our intrinsic setting, obtaining a quadrature rule for the curvilinear source integrals that is consistent with the order of the scheme and is exact in a steady-state configuration.

A number of test cases performed over slowly varying bottom topography are used to show the effectiveness of the numerical approach and to verify the importance of considering the geometric features of the bed topography in the equations. The bottom surfaces are defined by explicit formulas, thus allowing the exact calculation of the metric tensor. The approximation of these quantities starting from real digital elevation maps requires the use of computational geometry tools [see e.g. 59, 58], which is beyond the scope of this study.

The promising results obtained with a low order discretization scheme suggested the formulation of our approach using higher-order methods. For this purpose we chose to work within the framework of the Discontinuous Galerkin (DG) approach for its similarity to the FV method. For example, [3] describe a DG scheme for the standard SWEs that generalizes and extends the Godunov FV method [21] in several ways. The choice of DG methods was also dictated because of other reasons. First, the use of higher-order polynomials is naturally built into the method. Instead of computing higher-order terms through some sort of ad hoc post-processing procedure, such as in the reconstruction phase in ENO-WENO approaches in FV [see e.g. 55], they are defined directly in the variational formulation. Second, diffusive terms can be incorporated into the method using the local DG methods as described in [3].

In addition, the DG method allows the use of non-conforming grids, i.e., grids with hanging nodes without the use of mortar spaces. This feature could be very useful in dealing with complicated geometries and adaptive meshes.

DG methods, associated with Runge-Kutta discretization in time, were first analyzed by Cockburn and Shu in a series of papers [25, 27, 28, 29, 26]. Extensions to the Local DG (LDG) method were initially provided in [24]. Stability for nonlinear advection-diffusion systems and a priori error estimates for constant coefficient problems were studied in [20]. Finally, Cockburn and Dawson [23] generalized these results to multi-dimensional equations with spatially varying coefficients.

Our application of the DG method to our intrinsic framework starts from the simpler scalar transport equation and then moves to the ISWE model. Because of the inherent limitations of surface approximations, we confine the developments to linear basis functions. Again, the approach remains completely intrinsic by a proper formulation of the linear basis functions and geometric interpolations to define quadrature points starting from surface information on triangulation nodes.

Building on this work on DG, we extend the intrinsic approach at linear basis functions on triangles to address convection-diffusion problems on surfaces with conforming Galerkin Finite Element (FE) methods. Starting from the recent review work of Dziuk and Elliott [36] and the more recent work of Olshanskii et al. [60], we develop linear FEM in a completely intrinsic framework for the convection-diffusion-reaction equation on surfaces. The development is carried out in such a way that the convergence results reported in Dziuk and Elliott [36] can be immediately extended in the intrinsic approach. Numerical examples taken from the literature are used to reaffirm the previous assertion experimentally.

The successful development of the intrinsic FEM approach suggested the possibility to extend to surface equations the Entropy-Viscosity (EV) approach recently proposed by [45, 41, 44, 47]. Entropy viscosity methods are appealing for several reasons. At a fundamental level, the notion of driving numerical dissipation via the physical dissipation dictated through an entropy inequality has sound grounding in the mathematical theory for hyperbolic conservation laws [43]. More practically, entropy viscosity methods are valid for complex geometries with unstructured meshes and bring many desirable features of finite volume methods to a continuous Galerkin finite element framework. That is, they can be advanced with explicit time integration and are maximum principle preserving (or invariant domain preserving in the case of systems) [46, 44]. Then it is completely natural to try to extend this approach to our intrinsic setting to handle the numerical solution of PDEs on surfaces. In this thesis, only a preliminary work mostly dedicated to the development of the formulation is reported, while implementation and testing are left as future work.

The thesis is organized as follows. Chapter 2 contains all the definitions and results related to differential geometry that are needed for the development of the thesis work. Chapter 3 addresses the derivation of the intrinsic shallow water equations from the Navier-Stokes system. We start from the formal definition of a local reference frame, covariant differential operators, and progresses by performing the depth-integration. The chapter ends with the analysis of the properties of the model. Chapter 4 develops the first order finite volume method. First the approximations of the geometrical quantities are described and analyzed. Subsequently, the full model is derived and its properties are studied. Finally, numerical results ranging from simple domains to complex three-dimensional surfaces are reported. All the theoretical results are verified experimentally. Chapter 5 deals with the extension of the FV approach to the discontinuous Galerkin method. After an initial development part on the scalar transport equation, the extension to ISWE is reported, and the chapter concludes with numerical examples. In chapter 6 we address the intrinsic conforming finite element method for the advection-diffusion-reaction equation. Building on the developments of the DG method, the intrinsic scheme is described and tested on literature examples.

2

Geometric Setting

In this chapter we summarize the basic geometric notions that will be needed to treat partial differential equations on surfaces. We will take advantage of the concepts presented mainly in two classical differential geometry books: Abate and Tovena [2] and Do Carmo [33], reporting here standard results for the most parts without proofs. We will take care of replacing the proofs by proper citations. We start by the definition of regular surfaces and tangent planes. Then, we present some geometric properties, as the first and second fundamental forms, the notion of curvatures and geodesics. We conclude this chapter by applying these concepts to the field of PDEs and we introduce the setting of surface PDEs, with some examples.

2.1 Regular surfaces

We start this section with the definition of regular surface in $\mathbb{R}^3$. Roughly speaking a regular surface is a special subsets of $\mathbb{R}^3$ that in a certain sense is two-dimensional and smooth enough so that the usual notion of calculus can be extended to it. The proper definition is stated as:

Definition 2.1.1. *A subset $\mathcal{S} \subset \mathbb{R}^3$ is said to be a C^k* -regular surface *if, for each $\mathbf{p} \in \mathcal{S}$, there exists a neighborhood $V \subset \mathbb{R}^3$ and a map $\phi_{\mathbf{p}} : U \to V \cap \mathcal{S}$ of an open set*

$U \subseteq \mathbb{R}^2$ *onto* $V \cap \mathcal{S} \subset \mathbb{R}^3$ *such that:*

i) $\phi_{\mathbf{p}}(U) \subseteq \mathcal{S}$ *is an open neighborhood of* $\mathbf{p} \in \mathcal{S}$ *(i.e., there exists an open neighborhood of* $\mathbf{p}$*,* $V \subset \mathbb{R}^3$*, such that* $\phi_{\mathbf{p}}(U) = V \cap \mathcal{S}$*);*

ii) $\phi_{\mathbf{p}}$ *is a homeomorphism with its image ;*

iii) the differential $\mathrm{d}\phi_{\mathbf{q}} : \mathbb{R}^2 \to \mathbb{R}^3$ *is injective for all* $\mathbf{q} \in U$ *(i.e., it has maximum rank).*

We call $\phi_{\mathbf{p}}$ the *parametrization* or the *system of (local) coordinates* in $\mathbf{p}$. The inverse map $\phi_{\mathbf{p}}^{-1}$ is called the *local chart* in $\mathbf{p}$. The neighborhood $V \cap \mathcal{S}$ of $\mathbf{p}$ in $\mathcal{S}$ is called a *coordinate neighborhood*, and the coordinates $(x^1_{\mathbf{p}}, x^2_{\mathbf{p}}) = \phi_{\mathbf{p}}^{-1}(\mathbf{p})$ are called *local coordinates* of $\mathbf{p}$. The curve $\lambda \mapsto \phi_{\mathbf{p}}(\mathbf{o} + \lambda \mathbf{e}_j)$ is the j-th *coordinate curve* through $\mathbf{p} = \phi_{\mathbf{p}}(\mathbf{o})$. Given two points $\mathbf{p}, \mathbf{q} \in \mathcal{S}$ and their local parametrizations $\phi_{\mathbf{p}}, \phi_{\mathbf{q}}$, with $U_{\mathbf{p}} \cap U_{\mathbf{q}} \neq \emptyset$, if the *transition* map $\phi_{\mathbf{p}} \circ \phi_{\mathbf{q}}^{-1}$ is a C^k-diffeomorphism we say that the local parametrizations are *compatible*. A family $\{\phi_\alpha\}$ of compatible local parametrizations $\phi_\alpha : U_\alpha \to \mathcal{S}$ such that $\mathcal{S} = \bigcup_\alpha \phi_\alpha(U_\alpha)$ is called an *atlas* for a surface $\mathcal{S} \subset \mathbb{R}^3$. We will neglect the subscript $\mathbf{p}$ when indicating the parametrization symbol ϕ in cases when the central point $\mathbf{p}$ is obvious by context.

We give now some examples of regular surfaces.

Example 2.1. Given $\mathbf{p} \in \mathbb{R}^3$ and two linearly independent vectors $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3$, the plane $\mathcal{S} \subseteq \mathbb{R}^3$ through $\mathbf{p}$ and parallel to the two vectors is a regular surface with a single local parametrization $\phi : \mathbb{R}^2 \to \mathbb{R}^3$, given by $\phi(\mathbf{x}) = \mathbf{p} + x^1\mathbf{v}_1 + x^2\mathbf{v}_2$. In this case we have obviously a global parametrization for $\mathcal{S}$.

Example 2.2. Let U be is an open subset of $\mathbb{R}^2$ and $f : U \to \mathbb{R}$ an arbitrary smooth function. Then, the graph of f,

$$\mathrm{Graph}(f) := \{(x^1, x^2, f(x^1, x^2)) \mid (x^1, x^2) \in U\} ,$$

is a regular surface in $\mathbb{R}^3$. The map $\phi : U \to \mathbb{R}^3$ given by $\phi(\mathbf{x}) = (x^1, x^2, f(x^1, x^2))$ is a single local parametrization. We can check if it satisfies the conditions in the definition. Item i) is clearly satisfied. The restriction to Graph(f) of the projection on the first two coordinates is the inverse of ϕ, and it is continuous, so item ii) is satisfied. For the third condition, we have:

$$\mathbf{J}\phi(\mathbf{x}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial f}{\partial x^1}(\mathbf{x}) & \frac{\partial f}{\partial x^2}(\mathbf{x}) \end{bmatrix} ,$$

that has rank 2 everywhere.

We can give an alternative definition of a regular surface with a constructive flavor. We start from the following definition.

Definition 2.1.2. *Let* $V \subset \mathbb{R}^3$ *be an open set, and* $f : V \to \mathbb{R}$ *a* C^∞ *function. If* $\mathrm{d}f : \mathbb{R}^3 \to \mathbb{R}$ *is surjective we say that* $\mathbf{p}$ *is a* regular point *for* f*, otherwise we call it a* critical point *for* f*. If* $\mathbf{p} \in V$ *is a critical point, then we say that* $f(\mathbf{p}) \in \mathbb{R}$ *is a* critical value*. A point* $q \in f(V)$ *that is not a critical value is a* regular value*.*

With this definition in mind, we can state the following fact, whose proof comes from a special application of the implicit function theorem.

Proposition 2.1.3. *Let* $V \subseteq \mathbb{R}^3$ *be an open set, and* $f \in C^\infty(V)$*. If* $q \in \mathbb{R}$ *is a regular value of* f*, then every connected component of the level set* $f^{-1}(q) = \{\mathbf{p} \in V \mid f(\mathbf{p}) = q\}$ *is a regular surface.*

We present now a local converse of example 2.2, that says that every regular surface is locally the graph of a differentiable function, [2, prop. 3.1.29].

Proposition 2.1.4. *If* $\mathcal{S} \subset \mathbb{R}^3$ *is a regular surface and* $\mathbf{p} \in \mathcal{S}$*, then there exists a local parametrization* $\phi : U \to \mathcal{S}$ *in* $\mathbf{p}$ *that takes one of the following forms:*

$$\phi(x^1, x^2) = \begin{cases} (x^1, x^2, f(x^1, x^2)), & \textit{or} \\ (x^1, f(x^1, x^2), x^2), & \textit{or} \\ (f(x^1, x^2), x^1, x^2), & \end{cases}$$

for a certain function $f \in C^\infty(U)$*. In particular, there is an open neighborhood* $\mathcal{N}_\mathcal{S} \subseteq \mathbb{R}^3$ *of* $\mathcal{S}$ *such that* $\mathcal{S}$ *is closed in* $\mathcal{N}_\mathcal{S}$*.*

As introduced before, the main idea behind the definition of a regular surface is to define a set that is two-dimensional and regular enough so that the notions of calculus can be extended to it. Let us see how to use local parametrizations to define functions over a surface, and in particular how to extend the notions of continuity and differentiability.

Definition 2.1.5. *Let* $\mathcal{S} \subset \mathbb{R}^3$ *be a regular surface and consider* $\mathbf{p} \in \mathcal{S}$*. A function* $f : \mathcal{S} \to \mathbb{R}$ *is of class* C^∞*, or* smooth*, at* $\mathbf{p}$ *if there exists a local parametrization* $\phi : U \to \mathcal{S}$ *at* $\mathbf{p}$ *such that* $f \circ \phi : U \to \mathbb{R}$ *is of class* C^∞ *in a neighborhood of* $\phi^{-1}(\mathbf{p}) \subset \mathbb{R}^2$*.*

A smooth function is automatically continuous. Moreover, note that being smooth on a regular surface is a property of the function, and does not depend on the local parametrization, as the following theorem states [2]:

Theorem 2.1.6. *Let $\mathcal{S}$ be a surface, and let $\phi : U \to \mathcal{S}, \psi : V \to \mathcal{S}$ be two local parametrizations, with $W = \phi(U) \cap \psi(V) \neq \emptyset$. Then, the map $\phi^{-1} \circ \psi\Big|_{\psi^{-1}(W)} : \psi^{-1}(W) \to \phi^{-1}(W)$ is a diffeomorphism.*

2.1.1 The tangent space

We want to describe now some of the properties and objects that arise naturally when dealing with surfaces. We start from the notion of tangent space and tangent bundle. Consider a curve σ of class C^∞ lying on a regular surface $\mathcal{S}$, i.e., a differentiable function:

$$\sigma : (-\varepsilon, \varepsilon) \to \mathcal{S}\,,$$

and define $\mathbf{p} := \sigma(0)$. A *tangent vector* to $\mathcal{S}$ at $\mathbf{p}$ is a vector of the form $\sigma'(0)$ (σ' is called *velocity* vector for obvious analogy). The collection of the velocity vectors of all the curves passing through $\mathbf{p}$ as $t = 0$ is the *tangent cone* of $\mathcal{S}$ at $\mathbf{p}$:

$$T_\mathbf{p}\mathcal{S} := \{\sigma'(0) \mid \sigma : (-\varepsilon, \varepsilon) \to \mathcal{S} \text{ and } \sigma(0) = \mathbf{p}\}\ .$$

Since $\mathcal{S}$ is a regular surface defined by a smooth parametrization, we have that for every $\mathbf{p} \in \mathcal{S}$ the tangent cone $T_\mathbf{p}\mathcal{S}$ is a plane and does not degenerate into a straight line or a point. Let us show what happens in the simplest case of the plane.

Example 2.3. Let $U \subseteq \mathbb{R}^2$ be an open set, and $\mathbf{p} \in U$. Every curve contained in U is planar, and so the tangent vectors to U at $\mathbf{p}$ lie in $\mathbb{R}^2$. Conversely, if $\mathbf{v} \in \mathbb{R}^2$, the curve $\sigma : (-\varepsilon, \varepsilon) \to \mathcal{S}$ given by $\sigma(\lambda) = \mathbf{p} + \lambda\mathbf{v}$ has its support within U for ε small enough, and $\mathbf{v}$ is the tangent vector. Thus, we proved that $T_\mathbf{p}U = \mathbb{R}^2$.

If we consider a local parametrization for the surface, we can describe the relation between subsets of the plane and the surface (see fig. 2.1), obtaining the following:

Proposition 2.1.7. *Let $\mathcal{S} \subset \mathbb{R}^3$ be a surface, $\mathbf{p} \in \mathcal{S}$, and $\phi : U \to \mathcal{S}$ a local parametrization at $\mathbf{p}$ with $\phi(\mathbf{x}_0) = \mathbf{p}$, $\mathbf{x}_0 \in U$. Then, $\mathrm{d}\phi_{\mathbf{x}_0}$ is an isomorphism between $\mathbb{R}^2$ and $T_\mathbf{p}\mathcal{S}$.*

In particular, $T_\mathbf{p}\mathcal{S} = \mathrm{d}\phi_{\mathbf{x}_0}(\mathbb{R}^2)$ is always a 2-dimensional vector space, and $\mathrm{d}\phi_{\mathbf{x}_0}(\mathbb{R}^2)$ does not depend on ϕ but only on $\mathcal{S}$ and $\mathbf{p}$.

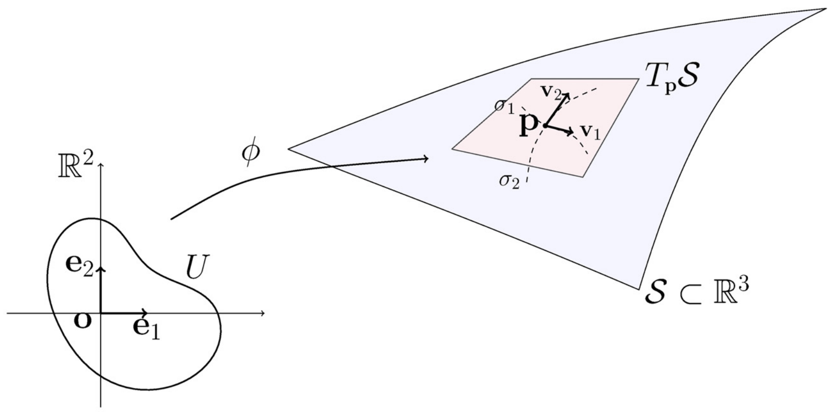

Figure 2.1: Example of local parametrization and affine tangent plane to the surface $\mathcal{S}$ at point $\mathbf{p}$.

Definition 2.1.8. *The vector space* $T_{\mathbf{p}}\mathcal{S} \subset \mathbb{R}^3$ *is the* tangent plane *to* $\mathcal{S}$ *at* $\mathbf{p}$. *The disjoint union of the tangent planes of* $\mathcal{S}$ *is called the* tangent bundle *of* $\mathcal{S}$, *i.e.,* $\mathrm{T}\,\mathcal{S} = \bigsqcup_{\mathbf{p}\in\mathcal{S}} T_{\mathbf{p}}\mathcal{S}$.

Thus, a *tangent vector field* on a surface $\mathcal{S}$ is a map X from $\mathcal{S}$ to $\mathbb{R}^3$ such that the vector $X_{\mathbf{p}} := X(\mathbf{p}) \in T_{\mathbf{p}}\mathcal{S}$, for each $\mathbf{p} \in \mathcal{S}$.

Remark 2.1. According to the definition, the tangent plane is a vector space in $\mathbb{R}^3$, so it passes through the origin. Our idea of the tangent plane attached to the surface $\mathcal{S}$ at point $\mathbf{p}$ is the *affine tangent plane*, i.e., the plane $\mathbf{p} + T_{\mathbf{p}}\mathcal{S}$ (see fig. 2.1). It can be proved that the affine tangent plane is the plane best approximating the surface $\mathcal{S}$ at $\mathbf{p}$.

In the case of a surface defined as level set of a function g, we are able to tell how to find the tangent plane from which the best approximating property is apparent. In fact, we have following proposition [2, prop. 3.3.11]:

Proposition 2.1.9. *Let* $V \subseteq \mathbb{R}^3$ *an open set, and* $q \in \mathbb{R}$ *a regular value of a function* $g \in C^\infty(V)$. *If* $\mathcal{S}$ *is a connected component of* $g^{-1}(q)$ *and* $\mathbf{p} \in \mathcal{S}$, *the tangent plane* $T_{\mathbf{p}}\mathcal{S}$ *is the subspace of* $\mathbb{R}^3$ *orthogonal to* $\nabla g(\mathbf{p})$.

If $\mathcal{S} = \mathrm{Graph}(f)$ with $\phi : U \to \mathcal{S}$ given by $\phi(\mathbf{x}) = (x^1, x^2, f(x^1, x^2))$ and $\mathbf{p} = \phi(\mathbf{o})$, we can define a level function $g = x^3 - f(x^1, x^2)$ and $\mathcal{S} = g^{-1}(0)$. Then, we can see the affine tangent plane as the linear approximation of the regular surface, by considering a Taylor expansion, i.e. given $\mathbf{q} \in \phi(U)$:

$$g(\mathbf{q}) = g(\mathbf{p}) + (\mathbf{q} - \mathbf{p}) \cdot \nabla g(\mathbf{p}) + \mathcal{O}\Big((\mathbf{q} - \mathbf{p})^2\Big) .$$

Let $W \subset \phi(U)$ be a neighborhood of $\mathbf{p}$ in $\mathcal{S}$, then we can formally write

$$W = T_{\mathbf{p}}\mathcal{S} + \mathcal{O}\Big(\mathrm{diam}(W)^2\Big) \; . \tag{2.1}$$

The isomorphism given by proposition 2.1.7 allows us to define a special basis for the tangent plane. Let $\mathcal{S} \subset \mathbb{R}^3$ be a regular surface, $\mathbf{p} \in \mathcal{S}$ and $\phi : U \to \mathcal{S}$ a local parametrization centered in $\mathbf{p}$ (i.e., $\phi(\mathbf{o}) = \mathbf{p}$). Consider $\{\mathbf{e}_1, \mathbf{e}_2\}$ the canonical basis of $\mathbb{R}^2$, then we can define:

$$\partial_i|_{\mathbf{p}} = \left.\frac{\partial}{\partial x^i}\right|_{\mathbf{p}} = \mathrm{d}\phi_{\mathbf{o}}(\mathbf{e}_i) = \frac{\partial \phi}{\partial x^i}(\mathbf{o}) \qquad i = 1, 2 \; .$$

These are vectors lying on the tangent plane in $\mathbf{p}$. The set $\{\partial_1|_{\mathbf{p}}, \partial_2|_{\mathbf{p}}\}$ forms a basis for $T_{\mathbf{p}}\mathcal{S}$ and is called the basis *induced* by the local parametrization.

Example 2.4. Let $\mathrm{Graph}(f) \subset \mathbb{R}^3$ be the graph of a function $f \in C^\infty(U)$. Consider the local parametrization from example 2.2. The basis induced by ϕ at $\mathbf{p} = (x^1, x^2, f(x^1, x^2)) \in \mathrm{Graph}(f)$ is defined by:

$$\left.\frac{\partial}{\partial x^1}\right|_{\mathbf{p}} = \begin{bmatrix} 1 \\ 0 \\ \frac{\partial f}{\partial x^1}(x^1, x^2) \end{bmatrix} \; , \quad \left.\frac{\partial}{\partial x^2}\right|_{\mathbf{p}} = \begin{bmatrix} 0 \\ 1 \\ \frac{\partial f}{\partial x^2}(x^1, x^2) \end{bmatrix} \; .$$

Remark 2.2. It is useful to remind what a coordinate system and a coordinate frame are. Intuitively, for a regular surface $\mathcal{S} \subset \mathbb{R}^3$ we can always use the “straight” coordinate system of $\mathbb{R}^3$ to describe the surface. This is classically denoted by (x^1, x^2, x^3), and in this case the canonical basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the coordinate frame at each point. The basis does not change with the point, and we will call this the *Global Coordinate System* (GCS). In general, we are interested in different coordinate systems, which can describe the surface in a more intrinsic way. An example of what is called “curvilinear” coordinate system is the *Local Coordinate System* (LCS) in the coordinates $(x^1_{\mathbf{p}}, x^2_{\mathbf{p}})$ derived from a local parametrization of the surface. In this case the definition strictly depends on the point. The induced basis, as previously defined, is the coordinate (or reference) frame for $T_{\mathbf{p}}\mathcal{S}$.

Remark 2.3. It is important to underline that each local parametrization $\phi : U \to \mathcal{S}$ defines a coordinate system on $\mathcal{S}$ starting from the coordinate system on U. Different local parametrizations provide different local coordinate systems.

We give now the definition of the so-called *Lie bracket*, a more algebraic concept related to the notion of vector fields. Given two vector fields $X, Y \in \mathrm{T}\mathcal{S}$, the tangent

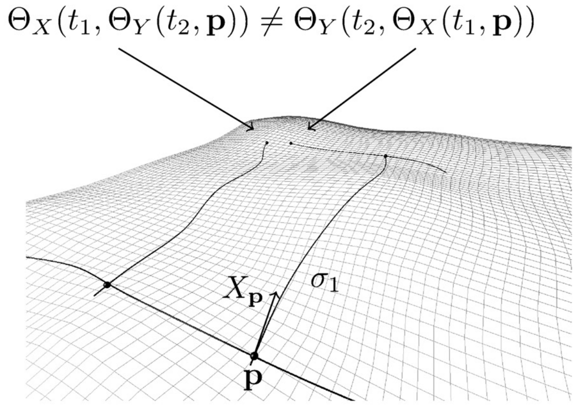

Figure 2.2: Example of non-commuting vector fields

bundle of $\mathcal{S}$, we can define the $\mathbb{R}$-bilinear operation $[\cdot, \cdot]$ as:

$$[X, Y] = XY - YX\,.$$

Explicitly, if we express the vector fields in local coordinates as $X = \sum_{i=1}^{2} X^i \partial_i$ and $Y = \sum_{j=1}^{2} Y^j \partial_j$, with $X^i, Y^j \in C^\infty(\mathcal{S})$, the Lie bracket of X, Y can be written as:

$$[X, Y] = \sum_{i,j=1}^{2} \left(X^i \, \partial_i Y^j - Y^i \, \partial_i X^j \right) \partial_j\,.$$

Consider now a point $\mathbf{p} \in \mathcal{S}$ and an open neighborhood $W \subset \mathcal{S}$ of $\mathbf{p}$. With reference to fig. 2.2, there exist an interval $I_\varepsilon = (-\varepsilon, \varepsilon)$ and a curve $\sigma_1 : I_\varepsilon \to \mathcal{S}$ such that the ODE

$$\frac{d\sigma_1(t)}{dt} = X_{\sigma_1(t)}\,,$$

with initial condition $\sigma_1(0) = \mathbf{p}$, has a unique solution. Following a re-statement of the existence theorem for ODE [13], we can define the flows $\Theta_X, \Theta_Y : I_\varepsilon \times W \to \mathcal{S}$ associated with the vector fields X, Y as:

$$\frac{d\Theta_X}{dt}(t, \mathbf{p}) = X_{\Theta_X(t,\mathbf{p})}\,, \qquad \frac{d\Theta_Y}{dt}(t, \mathbf{p}) = Y_{\Theta_Y(t,\mathbf{p})}\,,$$

with initial conditions $\Theta_X(0, \mathbf{p}) = \mathbf{p}$, $\Theta_Y(0, \mathbf{p}) = \mathbf{p}$. We call *commutativity condition* the property of the flows such that starting from a point $\mathbf{p}$ we can follow the flow Θ_X for a time t_1 and then the flow Θ_Y for another time t_2, and this is equivalent to inverting the order and follow Θ_Y for t_2 and then Θ_X for t_1. Figure 2.2 shows an example of non commuting flows. It is possible to prove that the condition $[X, Y] = 0$ is equivalent to the commutativity condition of the flows [1], and commutativity is a necessary condition to build a well-defined coordinate system given a coordinate frame. In particular, in every coordinate system, the induced local basis vectors ∂_1, ∂_2 commute. It is also possible to prove that this is also a sufficient condition in the construction of a basis reference frame for the tangent plane, as stated in the following theorem [1, thm. 3.7.4]:

Theorem 2.1.10. *Given two linearly independent vector fields X_1, X_2 of $\mathrm{T}\,\mathcal{S}$, then the following are equivalent:*

- *i) for every point $\mathbf{p} \in \mathcal{S}$ there exists an open subset V of $\mathcal{S}$ containing $\mathbf{p}$ and a coordinate system defined on V such that*

$$\frac{\partial}{\partial x^i} = X_i\Big|_V \qquad \text{for } i = 1, 2\,,$$

- *ii) $[X_i, X_j] = 0 \qquad$ for $i, j = 1, 2$.*

A coordinate system satisfying item i) is called a flat coordinate system.

2.2 Intrinsic and extrinsic quantities

Now that we have the basic structure of a regular surface, we focus on its geometric properties. In particular, we want to underline the differences between intrinsic and extrinsic quantities related to the surface. Geometry is first born from an extrinsic point of view by watching objects embedded in some space and considering their characteristics in relation to the surrounding space. Examples of extrinsic quantities are the tangent vectors seen as vectors of $\mathbb{R}^3$, or any quantity expressed in the GCS.

The intrinsic point of view allows to consider features of an object without an explicit reference to the space in which the object might be immersed. The concept of *metric* is the basic intrinsic quantity that allows all the computations to be made without "leaving" the surface. The first fundamental form is an example of such a metric in the case of regular surfaces. Because of the main theme of this work, we will present these quantities with a direct focus on the intrinsic geometry of the surface.

2.2.1 The first fundamental form

We start this section by introducing the first fundamental form of a regular surface, a positive definite quadratic form defined on each tangent plane that allows us to compute lengths of tangent vectors to the surface, as well as areas of regions of the surface. Let $\mathcal{S} \subset \mathbb{R}^3$ be a regular surface. For all $\mathbf{p} \in \mathcal{S}$ we denote by $\langle \cdot, \cdot \rangle_{\mathbf{p}}$ the positive definite scalar product on the tangent plane $T_{\mathbf{p}}\mathcal{S}$ induced by the canonical scalar product of $\mathbb{R}^3$.

Definition 2.2.1. *The* first fundamental form $\mathrm{I}_{\mathbf{p}}$ *is the (positive definite) quadratic form associated with the scalar product:*

$$\mathrm{I}_{\mathbf{p}} : T_{\mathbf{p}}\mathcal{S} \to \mathbb{R} \, , \quad \mathrm{I}_{\mathbf{p}}(\mathbf{v}) = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{p}} \geq 0 \, .$$

Remark 2.4. As introduced before, the term intrinsic refers to properties that do not depend on the immersion of the surface in $\mathbb{R}^3$. In particular, the properties that only depends on the first fundamental forms are called *intrinsic.*

The knowledge of the first fundamental form $\mathrm{I}_{\mathbf{p}}$ is equivalent to the knowledge of the scalar product $\langle \cdot, \cdot \rangle_{\mathbf{p}}$:

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{p}} = \frac{1}{2} \left[\mathrm{I}_{\mathbf{p}}(\mathbf{v} + \mathbf{w}) - \mathrm{I}_{\mathbf{p}}(\mathbf{v}) - \mathrm{I}_{\mathbf{p}}(\mathbf{w}) \right] = \frac{1}{4} \left[\mathrm{I}_{\mathbf{p}}(\mathbf{v} + \mathbf{w}) - \mathrm{I}_{\mathbf{p}}(\mathbf{v} - \mathbf{w}) \right] \, .$$

In particular, we can also give a definition of it by considering a given local parametrization $\phi : U \to \mathcal{S}$ at $\mathbf{p} \in \mathcal{S}$, and $\{\partial_1, \partial_2\}$ the basis of $T_{\mathbf{p}}\mathcal{S}$ induced by ϕ. We can write any tangent vector as a linear combination of basis vectors, i.e., for every $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$, $\mathbf{v} = v^1\partial_1 + v^2\partial_2$. Then, we can express the scalar product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{p}}$ between two vectors $\mathbf{v}, \mathbf{w} \in T_{\mathbf{p}}\mathcal{S}$ in coordinates as:

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{p}} = v^1 w^1 \langle \partial_1, \partial_1 \rangle_{\mathbf{p}} + (v^1 w^2 + v^2 w^1) \langle \partial_1, \partial_2 \rangle_{\mathbf{p}} + v^2 w^2 \langle \partial_2, \partial_2 \rangle_{\mathbf{p}} \, .$$

Then we can define the metric coefficients E, F and G as the scalar products of the basis vectors of $T_{\mathbf{p}}\mathcal{S}$ as in the following definition.

Definition 2.2.2. *The* metric coefficients *of* $\mathcal{S}$ *with respect to* ϕ *are the functions* $E, F, G : U \to \mathbb{R}$ *given by*

$$E(\mathbf{x}) = \langle \partial_1, \partial_1 \rangle_{\mathbf{p}} \, , \quad F(\mathbf{x}) = \langle \partial_1, \partial_2 \rangle_{\mathbf{p}} \, , \quad G(\mathbf{x}) = \langle \partial_2, \partial_2 \rangle_{\mathbf{p}} \, ,$$

for all $\mathbf{x} \in U$*.*

We can relate the first notation introduced by Gauss in the early 19th century to the modern notation of differential geometry (and Einstein summation rule) whereby repeated lower/upper indices are implicitly summed, and write:

$$E = g_{11} \, , \quad F = g_{12} = g_{21} \, , \quad G = g_{22} \, ,$$
$$\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{p}} = g_{ij}(\mathbf{p}) v^i w^j =: \langle \mathbf{v}, \mathbf{w} \rangle_{\mathcal{G}_{\mathbf{p}}} \, ,$$

where $\mathcal{G}_{\mathbf{p}} = (g_{ij})_{ij}$ is called the *metric tensor* and we replace the operator $\langle \cdot, \cdot \rangle_{\mathbf{p}}$ by the symbol $\langle \cdot, \cdot \rangle_{\mathcal{G}_{\mathbf{p}}}$ when the vectors are expressed with respect to the basis vector of $T_{\mathbf{p}}\mathcal{S}$.

We will drop the subscript $\mathbf{p}$ when the dependence on the point $\mathbf{p}$ is clear from the context.

For a regular surface the metric coefficients are functions in $C^\infty(U)$ and they completely determine the first fundamental form and the scalar product, so we can write:

$$\begin{aligned}\mathrm{I}_{\mathbf{p}}(\mathbf{v}) &= E(\mathbf{x})(v^1)^2 + 2F(\mathbf{x})v^1v^2 + G(\mathbf{x})(v^2)^2 = \\ &= \begin{bmatrix} v^1 & v^2 \end{bmatrix} \begin{bmatrix} E(\mathbf{x}) & F(\mathbf{x}) \\ F(\mathbf{x}) & G(\mathbf{x}) \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = \mathbf{v}^{\mathrm{T}}\mathcal{G}\mathbf{v} = \langle \mathbf{v}, \mathbf{v} \rangle_{\mathcal{G}} \ ,\end{aligned}$$

for all $\mathbf{p} = \phi(\mathbf{x}) \in \phi(U)$ and $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$. Note that, the metric coefficients strongly depend on the chosen local parametrization, even though the first fundamental form is related to the intrinsic metric properties of the surface. For the applications we have in mind, the importance of $\mathrm{I}_{\mathbf{p}}$ comes from the fact that we can treat metric questions on a regular surface without further references to the ambient space $\mathbb{R}^3$.

Length of a curve. The first fundamental form allows us to compute lengths of curves. Given $\sigma : I \to \mathcal{S}$ a parametrization of a curve lying on the surface, easy calculations yield:

$$\ell_\sigma(\lambda) = \int_0^\lambda |\sigma'(\tau)| \,\mathrm{d}\tau = \int_0^\lambda \sqrt{\mathrm{I}_{\sigma(\tau)}(\sigma'(\tau))}\,\mathrm{d}\tau \,.$$

Conversely, if we can compute the length of curves with support on the surface $\mathcal{S}$, we can define the first fundamental form. Indeed, given $\mathbf{p} \in \mathcal{S}$ and $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$, let $\sigma : (-\varepsilon, \varepsilon) \to \mathcal{S}$ be a curve in the parameter $\lambda \in (-\varepsilon, \varepsilon)$, with $\sigma(0) = \mathbf{p}$ and $\sigma'(0) = \mathbf{v}$, with length ℓ_σ. Then, taking the derivative of ℓ_σ from the first equality of the above equation we obtain immediately:

$$\mathrm{I}_{\mathbf{p}}(\mathbf{v}) = \left| \left(\frac{d\ell_\sigma}{d\lambda}(0) \right)^2 \right| .$$

Moreover, if the support of the curve $\sigma(\lambda) = \phi(x^1(\lambda), x^2(\lambda))$ is contained in a coordinate neighborhood corresponding to the parametrization $\phi(x^1, x^2)$, we can compute also the angle θ of the coordinate curves of the parametrization:

$$\cos\theta = \frac{\langle \partial_1, \partial_2 \rangle}{\|\partial_1\| \, \|\partial_2\|} = \frac{F}{\sqrt{EG}} \ ,$$

noting that, the coordinate curves of a parametrization are orthogonal if and only if $F = 0$ for all the points in the neighborhood. In this case, the parametrization is called *orthogonal*, and as a consequence the metric tensor $\mathcal{G}$ is diagonal.

Example 2.5. Let $\mathcal{S} \subset \mathbb{R}^3$ be a plane passing through $\mathbf{p} \in \mathbb{R}^3$ parallel to $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3$ two linearly independent vectors, and let $\phi(\mathbf{x}) = \mathbf{p} + x^1\mathbf{v}_1 + x^2\mathbf{v}_2$ be its local parametrization introduced in example 2.1. For each $\mathbf{p} \in T_{\mathbf{p}}\mathcal{S}$, the induced basis of the tangent plane is given by $\partial_1 = \mathbf{v}_1, \partial_2 = \mathbf{v}_2$, and the metric coefficients are:

$$E = \langle \mathbf{v}_1, \mathbf{v}_1 \rangle \ , \quad F = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle \ , \quad G = \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \ .$$

If the two vectors are orthonormal, then $E = G = 1$, and $F = 0$.

Example 2.6. Let $U \subseteq \mathbb{R}^2$ be an open set, $f \in C^\infty(U)$, and ϕ the local parametrization of the graph of f, as in example 2.2. In this case, the metric coefficients are:

$$E = 1 + \left|\frac{\partial f}{\partial x^1}\right|^2 \ , \quad F = \frac{\partial f}{\partial x^1}\frac{\partial f}{\partial x^2} \quad G = 1 + \left|\frac{\partial f}{\partial x^2}\right|^2 \ .$$

Remark 2.5. As apparent from example 2.5, an identity metric tensor implies a flat (planar) surface. Indeed, it can be shown that if $E = G = 1$ and $F = 0$, then the surface is locally isometric to a plane. If this happens at all points of the surface, then the surface is essentially flat (a plane).

Area of a regular region. The first fundamental form allows also to compute the area of bounded regions of a surface. First, we need to define what a region of a surface is:

Definition 2.2.3. *A* regular region $R \subseteq \mathcal{S}$ *is a connected compact subset of* $\mathcal{S}$ *obtained as the closure of its interior* R° *and whose boundary is parametrized by finitely many curvilinear polygons with disjoint supports. If* $\mathcal{S}$ *is compact, then* $R = \mathcal{S}$ *is a regular region without boundary.*

Definition 2.2.4. *Let* $R \subseteq \mathcal{S}$ *be a regular region of a surface* $\mathcal{S}$*. A* partition *of* R *is a finite family* $\mathcal{R} = \{R_1, \dots, R_n\}$ *of regular regions contained in* R*, such that* $R = \bigcup_i R_i$ *and* $R_i \cap R_j \subseteq \partial R_i \cap \partial R_j$*, for* $i, j = 1, \dots, n$ *and* $i \neq j$*. The* diameter $\operatorname{diam} R$ *of a partition is the maximum of the diameters of the elements of* $\mathcal{R}$*.*
A pointed partition *of* R *is a pair* $(\mathcal{R}, \mathbf{P})$ *given by a partition* $\mathcal{R}$ *of* R *and a n-tuple* $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ *of points of* R *such that* $\mathbf{p}_i \in R_i$*,* $i = 1, \dots, n$*.*

The notion of partition of a region will be useful also later in the numerical sections, where we will describe the discretization of the domain by introducing the definition of triangulation.

Definition 2.2.5. *Let* $R \subseteq \mathcal{S}$ *be a regular partition of a regular surface* $\mathcal{S}$ *and* $(\mathcal{R}, \mathbf{P})$ *a pointed partition of* R*. For all* $R_i \in \mathcal{R}$*, denote by* $\pi_i(R_i)$ *the orthogonal projection of* R_i *on the affine tangent plane* $\mathbf{p}_i + T_{\mathbf{p}_i}\mathcal{S}$*. The* area of the pointed partition *is defined as:*

$$\mathrm{Area}(\mathcal{R}, \mathbf{P}) = \sum_i \mathrm{Area}(\pi_i(R_i)) \; .$$

The region R *is* rectifiable *if the limit*

$$\mathcal{A}_R = \lim_{\mathrm{diam}\, R \to 0} \mathrm{Area}(\mathcal{R}, \mathbf{P})$$

exists and is finite. If it is the case, the limit is the area of R.

Note that the concept of rectifiability of a region is the direct extension of the idea of a rectifiable curve. Intuitively, a rectifiable region is a finite union of piece-wise smooth surfaces, and as such it has the desirable properties of smooth surfaces.

The following result [2, thm. 4.2.5], derived from the classical theorem of Change of Variables for multiple integrals, states how to compute the area of a region:

Theorem 2.2.6. *Let* $R \subseteq \mathcal{S}$ *be a regular region contained in the image of a local parametrization* $\phi : U \to \mathcal{S}$ *of a surface* $\mathcal{S}$*. Then,* R *is rectifiable and its area is*

$$\mathcal{A}_R = \int_{\phi^{-1}(R)} \sqrt{EG - F^2}\, \mathrm{d}\mathbf{x} \; .$$

Moreover we have that:

Lemma 2.2.7. *Given* $\phi : U \to \mathcal{S}$ *a local parametrization of a surface* $\mathcal{S}$*, then:*

$$\|\partial_1 \wedge \partial_2\| = \sqrt{EG - F^2} \; ,$$

where the symbol $\wedge$ *denotes the vector product in* $\mathbb{R}^3$*. Moreover, if* $\psi : V \to \mathcal{S}$ *is another local parametrization with* $W = \psi(V) \cap \phi(U) \neq \emptyset$*, and* $f = \psi^{-1} \circ \phi|_{\phi(U)^{-1}}$*, then*

$$\partial_1 \wedge \partial_2|_{\phi(\mathbf{x})} = \det(\mathbf{J}f)(\mathbf{x})\, \tilde{\partial}_1 \wedge \tilde{\partial}_2|_{\psi \circ f(\mathbf{x})}$$

for all $\mathbf{x} \in \phi^{-1}(W)$*, where* $\{\tilde{\partial}_1, \tilde{\partial}_2\}$ *is the basis induced by* ψ*.*

Integral of a function and Stokes theorems. A direct consequence of the previous lemma is the fact that the integral does not depend on the local parametrization, and thus we can define the integral of a function over a surface as follows:

Definition 2.2.8. *Let* $R \subseteq \mathcal{S}$ *be a regular region contained in the image of a local parametrization* $\phi : U \to \mathcal{S}$ *of a regular surface* $\mathcal{S}$*, and* $f : R \to \mathbb{R}$ *a continuous function. The* integral of f on R *is given by*

$$\int_R f = \int_{\phi^{-1}(R)} (f \circ \phi) \sqrt{EG - F^2}\, \mathrm{d}\mathbf{x} \; .$$

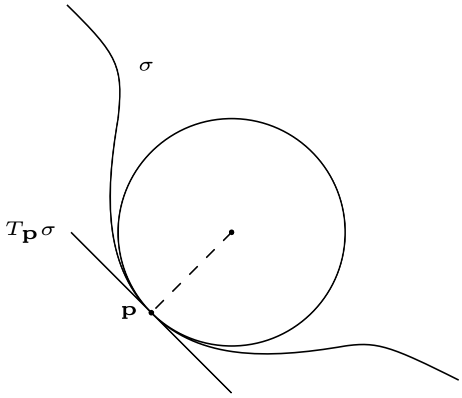

Figure 2.3: Example of the osculating circle at the point $\mathbf{p}$ of a planar curve σ.

Moreover, we are interested in applying Stokes theorems, and to be able to speak about them we need to introduce the concept of 1-form. Here we just quickly recall the results we need, forgoing some precision for the sake of brevity. We refer to differential geometry books, such as [13, 1], for a rigorous and detailed description of this topic.

Let $f, g \in C^{\infty}(\mathcal{S})$ be two smooth scalar functions, then a general 1-form is given by $\omega = f\,\mathrm{d}x^1 + g\,\mathrm{d}x^2$. The differential of ω is given by $\mathrm{d}\omega = (\frac{\partial g}{\partial x^1} - \frac{\partial f}{\partial x^2})\,\mathrm{d}x^1 \wedge \mathrm{d}x^2$, where now $\wedge$ denotes the wedge product and $\mathrm{d}x^1 \wedge \mathrm{d}x^2$ is the volume form for $\mathcal{S}$. Then, Stokes theorem can be written as [1]:

Theorem 2.2.9. *Let $\mathcal{S} \subset \mathbb{R}^3$ be a surface with smooth boundary $\partial\mathcal{S}$ and ω a 1-differential form with compact support on $\mathcal{S}$. Then:*

$$\int_{\partial\mathcal{S}} \omega = \int_{\mathcal{S}} \mathrm{d}\omega\,.$$

Using the above expressions for ω and $\mathrm{d}\omega$, we can write

$$\int_{\partial\mathcal{S}} f\,\mathrm{d}x^1 + g\,\mathrm{d}x^2 = \int_{\mathcal{S}} \left(\frac{\partial g}{\partial x^1} - \frac{\partial f}{\partial x^2}\right) \mathrm{d}x^1 \wedge \mathrm{d}x^2\,, \tag{2.2}$$

that is the classical Gauss-Green theorem.

2.2.2 Second fundamental form and curvatures

In this section we move towards the notion of curvature of a surface. The idea forms its basis in the definition of the curvature of a curve in $\mathbb{R}^2$, which intuitively can be defined by the rate of change of the tangent line to the curve. The direct extension of this concept to regular surfaces suggests the idea of trying to measure how rapidly a surface $\mathcal{S}$ pulls away from its tangent plane $T_{\mathbf{p}}\mathcal{S}$ in a neighborhood of a point $\mathbf{p}$, or, in other words, we want to measure the rate of change of the normal to the surface.

Another equivalent intuitive definition for the curvature of a planar curve is to relate it to the reciprocal of the radius of the "osculating circle", which is the unique circle tangent to the curve (fig. 2.3). In this case, the extension to two dimensions is not the radius of the tangent sphere, but we would need to consider the principal radii of a ellipsoid tangent to the surface. This is because in principle the surface can have different curvatures along different directions.

We start our discussion from the definition of the normal vector field and of the map that relates a point on the surface to the outer normal to the surface.

Definition 2.2.10. *A* normal vector field *on a surface* $\mathcal{S} \subset \mathbb{R}^3$ *is a map* $\mathbf{N} : \mathcal{S} \to \mathbb{R}^3$ *of class* C^∞*, such that* $\mathbf{N}(\mathbf{p})$ *is orthogonal to the tangent plane* $T_\mathbf{p}\mathcal{S}$ *for all* $\mathbf{p} \in \mathcal{S}$*. If* $\|\mathbf{N}\| = 1$*, i.e.,* $\mathbf{N} \in S^2$*,* $\mathbf{N}$ *is called the* normal versor field *to* $\mathcal{S}$ *and the map* $\mathbf{N} : \mathcal{S} \to S^2$ *is called the* Gauss map*.*

Not every surface has a normal vector field globally, but given a local parametrization $\phi : U \to \mathcal{S}$ at a point $\mathbf{p}$, then the Gauss map of $\phi(U)$ always exists and is defined by:

$$\mathbf{N}(\mathbf{p}) = \frac{\partial_1 \wedge \partial_2}{\|\partial_1 \wedge \partial_2\|}(\mathbf{p})\,.$$

Remark 2.6. The existence of the normal vector field is related to the notion of *orientability.* Following [33], we give the subsequent definition of orientability: a regular surface $\mathcal{S}$ is called *orientable* if it is possible to cover it with a family of coordinate neighborhoods in such a way that if a point $\mathbf{p}$ belongs to two neighborhoods of the family, then the change of coordinates has a positive Jacobian at $\mathbf{p}$. Note that this definition of orientation is purely intrinsic.

The Gauss map determines uniquely the tangent planes to the surface, since $\mathbf{N}(\mathbf{p})$ is orthogonal to $T_\mathbf{p}\mathcal{S}$. Thus, its variation measures how the tangent planes change over the surface. In particular, the Gauss map is differentiable and its differential is a vector of the tangent plane $T_\mathbf{p}\mathcal{S}$ related to the rate of change of $\mathbf{N}$, that is, how close is a surface from being planar. As example, recall the parametrization of the plane in example 2.1, the normal versor is simply the vector resulting from the vector product of $\mathbf{v}_1, \mathbf{v}_2$, that is constant along the plane, and thus, the differential is zero.

We state here an important property of the Gauss map:

Proposition 2.2.11. *Given a surface* $\mathcal{S} \subset \mathbb{R}^3$ *with Gauss map* $\mathbf{N} : \mathcal{S} \to S^2$*, the differential* $\mathrm{d}\mathbf{N}_\mathbf{p}$ *is an endomorphism of* $T_\mathbf{p}\mathcal{S}$*, symmetric with respect to the scalar product* $\langle \cdot, \cdot \rangle_\mathbf{p}$*, for all* $\mathbf{p} \in \mathcal{S}$*.*

Remark 2.7. The map $\mathcal{W} : \mathrm{T}\,\mathcal{S} \to \mathrm{T}\,\mathcal{S}$, that for each point $\mathbf{p} \in \mathcal{S}$ is defined as $\mathcal{W}_{\mathbf{p}} = \mathrm{d}\mathbf{N}_{\mathbf{p}} : T_{\mathbf{p}}\mathcal{S} \to T_{\mathbf{p}}\mathcal{S}$, the endomoprhism of the tangent plane, is called the *Weingarten map.*

We are now able to give the following definitions:

Definition 2.2.12. *The* second fundamental form *of $\mathcal{S}$ is the quadratic form $\mathrm{II}_{\mathbf{p}} : T_{\mathbf{p}}\mathcal{S} \to \mathbb{R}$ given by*

$$\mathrm{II}_{\mathbf{p}}(\mathbf{v}) = -\left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\mathbf{v}), \mathbf{v} \right\rangle_{\mathbf{p}} , \qquad \forall\, \mathbf{v} \in T_{\mathbf{p}}\mathcal{S} .$$

Given $\phi : U \to \mathcal{S}$ a local parametrization of the surface centered at $\mathbf{p}$, the form coefficients *of $\mathcal{S}$ with respect to ϕ are the functions e, f and $g : U \to \mathbb{R}$ defined by:*

$$e(\mathbf{x}) = -\left\langle \mathrm{d}\mathbf{N}_{\phi(\mathbf{x})}(\partial_1), \partial_1 \right\rangle_{\mathbf{p}} , \quad f(\mathbf{x}) = -\left\langle \mathrm{d}\mathbf{N}_{\phi(\mathbf{x})}(\partial_1), \partial_2 \right\rangle_{\mathbf{p}} , \quad g(\mathbf{x}) = -\left\langle \mathrm{d}\mathbf{N}_{\phi(\mathbf{x})}(\partial_2), \partial_2 \right\rangle_{\mathbf{p}} ,$$

for all $\mathbf{x} \in U$.

As in the case of the metric coefficients, the form coefficients completely determine the second fundamental form. For all $\mathbf{p} = \phi(\mathbf{x})$, $\mathbf{x} \in U$, and $\mathbf{v} = v^1\partial_1 + v^2\partial_2 \in T_{\mathbf{p}}\mathcal{S}$, we get:

$$\mathrm{II}_{\mathbf{p}}(v^1\partial_1 + v^2\partial_2) = e(\mathbf{x})(v^1)^2 + 2f(\mathbf{x})v^1v^2 + g(\mathbf{x})(v^2)^2 .$$

And, again as for the metric coefficients, they depend on the chosen local parametrization.

Example 2.7. Let $\mathrm{Graph}(f) \subset \mathbb{R}^3$ be the graph of $f : U \to \mathbb{R}$, and ϕ the usual parametrization (see example 2.2). The Gauss map exists globally and is given by:

$$\mathbf{N} \circ \phi = \frac{\partial_1 \wedge \partial_2}{\|\partial_1 \wedge \partial_2\|} = \frac{1}{\sqrt{1 + \|\nabla f\|^2}} \begin{bmatrix} -\dfrac{\partial f}{\partial x^1} \\ -\dfrac{\partial f}{\partial x^2} \\ 1 \end{bmatrix} .$$

If we compute now the differential of $\mathbf{N}$ in $\mathbf{p}$, we obtain:

$$\begin{aligned} \mathrm{d}\mathbf{N}_{\mathbf{p}}(\partial_j) &= \frac{\partial \mathbf{N} \circ \phi}{\partial x^j}(\mathbf{x}) = \\ &= \frac{1}{(1 + \|\nabla f\|^2)^{3/2}} \left[\frac{\partial f}{\partial x^1}\frac{\partial f}{\partial x^2}\frac{\partial^2 f}{\partial x^j \partial x^2} - \left(1 + \left(\frac{\partial f}{\partial x^2}\right)^2\right) \frac{\partial^2 f}{\partial x^j \partial x^1} \right] \partial_1 \\ &\quad + \frac{1}{(1 + \|\nabla f\|^2)^{3/2}} \left[\frac{\partial f}{\partial x^1}\frac{\partial f}{\partial x^2}\frac{\partial^2 f}{\partial x^j \partial x^1} - \left(1 + \left(\frac{\partial f}{\partial x^1}\right)^2\right) \frac{\partial^2 f}{\partial x^j \partial x^2} \right] \partial_2 . \end{aligned}$$

And finally, we can compute the second fundamental form:

$$\begin{aligned}\mathrm{II}_{\mathbf{p}}(\mathbf{v}) &= -\left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\mathbf{v}), \mathbf{v}\right\rangle_{\mathbf{p}} = \\ &= -\left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\partial_1), \partial_1\right\rangle_{\mathbf{p}} (v^1)^2 - 2\left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\partial_1), \partial_2\right\rangle_{\mathbf{p}} v^1 v^2 - \left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\partial_2), \partial_2\right\rangle_{\mathbf{p}} (v^2)^2 = \\ &= \frac{1}{\sqrt{1+\|\nabla f\|^2}}\left[\frac{\partial^2 f}{\partial (x^1)^2}(v^1)^2 + 2\frac{\partial^2 f}{\partial x^1 \partial x^2} v^1 v^2 + \frac{\partial^2 f}{\partial (x^2)^2}(v^2)^2\right] ,\end{aligned}$$

that can be written in matrix form with respect to the basis $\{\partial_1, \partial_2\}$ of $T_{\mathbf{p}}\mathcal{S}$ as $\mathrm{II}_{\mathbf{p}}(\mathbf{v}) = \left(1+\|\nabla f\|^2\right)^{-1/2}\mathbf{H}(f)$, where $\mathbf{H}(f)$ is the Hessian matrix of f. In particular the form coefficients are:

$$e = \frac{1}{\sqrt{1+\|\nabla f\|^2}}\frac{\partial^2 f}{\partial (x^1)^2} , \quad f = \frac{1}{\sqrt{1+\|\nabla f\|^2}}\frac{\partial^2 f}{\partial x^1 \partial x^2} , \quad g = \frac{1}{\sqrt{1+\|\nabla f\|^2}}\frac{\partial^2 f}{\partial (x^2)^2} .$$

The fact that $\mathrm{d}\mathbf{N}_{\mathbf{p}}$ (see proposition 2.2.11) is a symmetric endomorphism helps us to define new quantities. The first definition follows from the fact that $\mathrm{d}\mathbf{N}_{\mathbf{p}}$ is diagonalizable:

Definition 2.2.13. *A* principal direction *of $\mathcal{S}$ at $\mathbf{p}$ is an eigenvector of $\mathrm{d}\mathbf{N}_{\mathbf{p}}$ of length one, and the corresponding eigenvalue with the sign changed is a* principal curvature*.*

If $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$ is a principal direction with principal curvature k, we have the relation:

$$\mathrm{II}_{\mathbf{p}}(\mathbf{v}) = -\left\langle \mathrm{d}\mathbf{N}_{\mathbf{p}}(\mathbf{v}), \mathbf{v}\right\rangle_{\mathbf{p}} = -\left\langle -k\mathbf{v}, \mathbf{v}\right\rangle_{\mathbf{p}} = k .$$

Moreover, we can give a name to the trace and determinant of $\mathrm{II}_{\mathbf{p}}$:

Definition 2.2.14. *The* Gaussian curvature *of $\mathcal{S}$ is the function $K : \mathcal{S} \to \mathbb{R}$ given by:*

$$K(\mathbf{p}) = \det(\mathrm{d}\mathbf{N}_{\mathbf{p}}) , \quad \forall\, \mathbf{p} \in \mathcal{S} .$$

The function $H : \mathcal{S} \to \mathbb{R}$, defined as:

$$H(\mathbf{p}) = -\frac{1}{2}\operatorname{tr}(\mathrm{d}\mathbf{N}_{\mathbf{p}}) , \quad \forall\, \mathbf{p} \in \mathcal{S} ,$$

is called the mean curvature *of $\mathcal{S}$.*

Since these quantities are directly computed from the Gauss map, the Gaussian curvature and the absolute value of the mean curvature do not depend on the parametrization of the surface. If we denote the principal curvatures of $\mathcal{S}$ at $\mathbf{p}$ by k_1, k_2, then $K(\mathbf{p}) = k_1 k_2$ and $H(\mathbf{p}) = (k_1 + k_2)/2$. Moreover, it is interesting to note that the Gaussian curvature can be interpreted in terms of ratios of areas. Given $\phi : U \to \mathcal{S}$ a local parametrization of a surface $\mathcal{S}$ at $\mathbf{p}$, if $B_\varepsilon \subset \mathbb{R}^2$ is a ball centered in $\mathbf{p}$ and with radius $\varepsilon > 0$, then if $K(\mathbf{p}) \neq 0$ we have:

$$|K(\mathbf{p})| = \lim_{\varepsilon \to 0} \frac{\mathrm{Area}(\mathbf{N} \circ \phi(B_\varepsilon))}{\mathrm{Area}(\phi(B_\varepsilon))} \,.$$

We can now relate the coefficients of the metric tensor and the second fundamental form. Let $M \in \mathbb{R}^{2,2}$ be the matrix representing the endomorphism $\mathrm{d}\mathbf{N}_\mathbf{p}$ with respect to the induced basis $\{\partial_1, \partial_2\}$ of the tangent plane in $\mathbf{p}$. For every $\mathbf{v} = v^1\partial_1 + v^2\partial_2, \mathbf{w} = w^1\partial_1 + w^2\partial_2 \in T_\mathbf{p}\mathcal{S}$ we have:

$$\begin{aligned} -\left\langle \mathrm{d}\mathbf{N}_\mathbf{p}(\mathbf{v}), \mathbf{w} \right\rangle_\mathbf{p} &= \begin{bmatrix} w^1 & w^2 \end{bmatrix} \begin{bmatrix} e & f \\ f & g \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = \\ &= -\begin{bmatrix} w^1 & w^2 \end{bmatrix} \begin{bmatrix} E & F \\ F & G \end{bmatrix} M \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} \,, \end{aligned}$$

and we can directly compute the values of M:

$$M = -\frac{1}{EG - F^2} \begin{bmatrix} eG - fF & fG - gF \\ fE - eF & gE - fF \end{bmatrix} \,.$$

In particular, the Gaussian curvature is given by

$$K = \det(M) = \frac{eg - f^2}{EG - F^2} \,,$$

the mean curvature is

$$H = -\frac{1}{2}\,\mathrm{tr}(M) = \frac{1}{2} \frac{eG - 2fF + gE}{EG - F^2} \,,$$

and the principal curvatures are

$$k_{1,2} = H \pm \sqrt{H^2 - K} \,.$$

2.2.3 Geodesics

The goal of this section is to extend to the surface the role that straight lines have in the plane. There are different ways of defining lines: a global geometric characterization that sees a line segment as the shortest curve between its endpoints; and a local analytic one, defining a line as the curve with a constant tangent vector. The extension on a surface of the global characterization presents some issues, such as the fact that the curve may not exists or not be unique; while, the local characterization seems to be more consistent with the intrinsic approach we are pursuing, but needs some additional definitions. We start here by the introduction of the definition of *covariant derivative*, which helps in the definition of constant vectors on a surface.

Recall that a tangent vector field in an open set $W \subset \mathcal{S}$ of a regular surface $\mathcal{S}$ is a map X that assigns at each point $\mathbf{p} \in W$ a vector $X(\mathbf{p}) \in T_{\mathbf{p}}\mathcal{S}$. In particular, we can consider the vector field along a curve $\sigma : I \to W$ of class C^∞, defined as:

$$X : I \to \mathbb{R}^3 \qquad \text{such that} \qquad X(\lambda) \in T_{\sigma(\lambda)}\mathcal{S} \;, \qquad \forall\, \lambda \in I \;.$$

The tangent vector $\sigma'(\lambda)$ of a curve σ is a typical example of vector field along a curve.

Definition 2.2.15. *Let $X \in \mathrm{T}\,\mathcal{S}$ be a vector field along a curve $\sigma : I \to \mathcal{S}$ of class C^∞, and fix $\mathbf{p} = \sigma(\bar{\lambda})$ and $Y = \sigma'(\bar{\lambda})$ for some $\bar{\lambda} \in I$. The vector obtained by the orthogonal projection of $\dfrac{dX}{d\lambda}(\bar{\lambda})$ onto the plane $T_{\mathbf{p}}\mathcal{S}$ is called the* covariant derivative *at $\mathbf{p}$ of the vector field X relative to the vector Y and is defined as:*

$$D_Y X = \pi_{\mathbf{p}}\left(\frac{dX}{d\lambda}(\bar{\lambda})\right) \;,$$

where $\pi_{\mathbf{p}}(\cdot) : \mathbb{R}^3 \to T_{\mathbf{p}}\mathcal{S}$ is the orthogonal projection on the tangent plane to $\mathcal{S}$ at $\mathbf{p}$.

Remark 2.8. If we do not fix the parameter $\bar{\lambda} \in I$, we can define the vector field $DX : I \to \mathbb{R}^3$ as $DX(\lambda) = D_{\sigma'(\lambda)}X \in T_{\sigma(\lambda)}\mathcal{S}$. This is called the covariant derivative of the vector field X along a curve σ.

The above definition uses a particular curve $\sigma(\lambda)$, tangent to Y at $\mathbf{p}$. However, the covariant derivative is a purely intrinsic notion that depends on the first fundamental form of $\mathcal{S}$. To show this fact let us express $D_Y X$ in local coordinates. If $\phi : U \to \mathcal{S}$ be a local parametrization whose image contains the support of a curve $\sigma : I \to \mathcal{S}$, then we have $\sigma(\lambda) = \phi(\sigma_1(\lambda), \sigma_2(\lambda))$. A vector field X along σ can be written in components with respect to the local induced basis as $X(\lambda) = X^1\partial_1 + X^2\partial_2$, then we can compute the covariant derivative and obtain:

$$D_Y X = \sum_{k=1}^{2}\left[\frac{dX^k}{d\lambda} + \sum_{i,j=1}^{2}\left(\Gamma^k_{ij} \circ \sigma\right)\sigma'_i X^j\right]\partial_k \;. \tag{2.3}$$

In particular, if the curve is aligned to one of the coordinate curve, i.e., $Y = \partial_j$, and the vector field is also tangent to a coordinate curve, $X = \partial_i$, then:

$$D_Y X = \sum_{k=1}^{2}\Gamma^k_{ji}\partial_k \;.$$

Definition 2.2.16. *The functions* Γ_{ij}^k *are called* Christoffel symbols *with respect to the parametrization* ϕ*, and they can be computed by the following formula:*

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^{2} g^{kl}\left(\frac{\partial g_{il}}{\partial x^j} + \frac{\partial g_{lj}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^l}\right) ,$$

where g_{ij} *are the coefficients of the first fundamental form, and* g^{ij} *of its inverse.*

Definition 2.2.17. *A vector field* X *on a parametrized curve* $\sigma : I \to \mathcal{S}$ *is said to be* parallel *if* $DX(\lambda) = 0$ *for every* $\lambda \in I$*.*

Definition 2.2.18. *A* geodesic *on a surface* $\mathcal{S}$ *is a curve* $\sigma : I \to \mathcal{S}$ *of class* C^∞ *such that* $\sigma' \in \mathrm{T}\,\mathcal{S}$ *is parallel, i.e., such that* $D\sigma' = 0$*.*

As previously said, a geodesic curve on a surface plays the role of a straight line on the plane, in term of distances and connection of points. We need to show that geodesics exist. Given a curve σ on a surface $\mathcal{S}$ and a local parametrization $\phi : U \to \mathcal{S}$ that contains the curve, we can say from eq. (2.3) that the curve $\sigma = \phi(x^1, x^2)$ is a geodesic if and only if satisfies the following system of differential equations:

$$\sigma_k'' + \sum_{i,j=1}^{2}\left(\Gamma_{ij}^k \circ \sigma\right)\sigma_i''\sigma_j'' = 0 , \qquad k = 1, 2 . \tag{2.4}$$

The previous system of equations is called *geodesic equation.*

Proposition 2.2.19. *For all* $\mathbf{p} \in \mathcal{S}$ *and* $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$ *there exists a geodesic* $\sigma : I \to \mathcal{S}$ *such that* $0 \in I$*,* $\sigma(0) = \mathbf{p}$ *and* $\sigma'(0) = \mathbf{v}$*. Moreover, if* $\tilde{\sigma} : \tilde{I} \to \mathcal{S}$ *is another geodesic satisfying the same conditions, then* σ *and* $\tilde{\sigma}$ *coincide in* $I \cap \tilde{I}$*. In particular, for all* $\mathbf{p} \in \mathcal{S}$ *and* $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$ *there exists a maximal open interval* $I_{\mathbf{v}}$ *and unique geodesic* $\sigma_{\mathbf{v}} : I_{\mathbf{v}} \to \mathcal{S}$ *such that* $\sigma_{\mathbf{v}}(0) = \mathbf{p}$ *and* $\sigma_{\mathbf{v}}'(0) = \mathbf{v}$*.*

Thus, geodesics exist on regular surfaces and can be computed solving the geodesic equation. However, the explicit computation of geodesics is very difficult even for simple surfaces. Let us show this in the case of the graph of a function $f \in C^\infty(U)$, for $U \subseteq \mathbb{R}^2$. Recalling the definition of the Christoffel symbols and eq. (2.4), the system to solve is:

$$\sigma_k'' + \frac{1}{1 + \|\nabla f \circ \sigma\|^2}\left(\frac{\partial f}{\partial x^k} \circ \sigma\right)\langle(\mathbf{H}(f) \circ \sigma) \cdot \sigma', \sigma'\rangle = 0 , \qquad k = 1, 2 .$$

Moreover, for parallel vector fields we have that:

Lemma 2.2.20. *Given* $\sigma : I \to \mathcal{S}$ *with* $0 \in I$ *and* $\mathbf{p} = \sigma(0)$*, then for all* $\mathbf{v} \in T_{\mathbf{p}}\mathcal{S}$ *there exists a unique parallel vector field* X *such that* $X(0) = \mathbf{v}$*.*

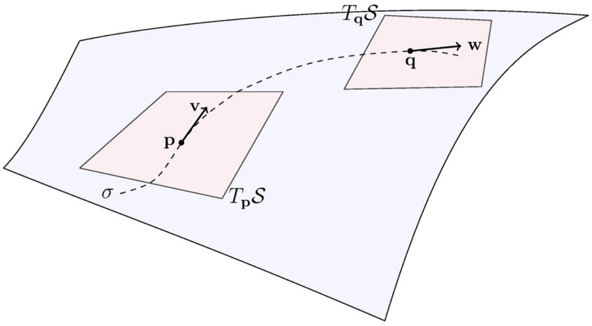

Figure 2.4: Example of parallel transport of a vector $\mathbf{v}$ along the curve σ.

Then this lemma allows us to talk about parallel transport of a vector along a parametrized curve, see fig. 2.4:

Definition 2.2.21. *Given a curve σ on $\mathcal{S}$ defined on an open interval I, $\mathbf{p} := \sigma(\lambda_0), \mathbf{q} := \sigma(\lambda_1) \in \sigma(I)$ and a vector $\mathbf{v}$ of $T_{\mathbf{p}}\mathcal{S}$, the* parallel transport *of $\mathbf{v}$ in $\mathbf{q}$ is the parallel vector field X along σ such that $X(\lambda_0) = \mathbf{v}$ and $X(\lambda_1) = \mathbf{w}$.*

2.3 PDEs on surfaces

In many practical applications the proper definition of partial differential equations over a manifold can be crucial. We are interested in particular in equations over surfaces in $\mathbb{R}^3$. The already high level of challenge that characterizes the analysis of PDEs is even higher in these cases, due to the fact that the domain in which the equations hold is curved. A geometric setting that takes into account all the principal geometric properties is needed. To complete our description we will need to express points in $\mathbb{R}^3$ with respect to both global and local reference frames. The coordinates of the global coordinate system will be identified with (x^1, x^2, x^3), while the local coordinates will be (s^1, s^2, s^3). We introduce here the last definitions needed to set the framework in which we are able to define our equations of interest.

2.3.1 Physical, contravariant and covariant components

We need to generalize our theoretical physical knowledge to give an invariant structure to physical laws under a curvilinear coordinate system. In particular, we need to distinguish physical, covariant, and contravariant versions of vectors and tensors, to ensure coordinate invariance.

Let us consider the standard basis of $\mathbb{R}^3$ formed by unit vectors $\mathbf{e}_i$, $i = 1, 2, 3$, in the classical Global Cartesian Coordinate System (GCS) (x^1, x^2, x^3). For every $\vec{u} \in \mathbb{R}^3$

there is a unique set of coefficients, $u_{(i)}$, such that

$$\vec{u} = \sum_i u_{(i)} \mathbf{e}_i \ .$$

The values $u_{(i)}$ are called the *physical* components of $\vec{u}$ relative to the standard basis set, and we denote them by subscripts surrounded with parentheses. Consider now a regular surface in $\mathbb{R}^3$ and a local parametrization $\phi : U \to \mathcal{S}$ of $\mathcal{S}$ centered at $\mathbf{p}$, with local coordinates $(s^1_{\mathbf{p}}, s^2_{\mathbf{p}})$. We can compute the induced reference basis vectors $\partial\phi/\partial s^1, \partial\phi/\partial s^2 \in T_{\mathbf{p}}\mathcal{S}$, that we will denote respectively by $\mathbf{t}_1$ and $\mathbf{t}_2$ for simplicity of notation. The associated metric $\mathcal{G}_{\mathcal{S}}$, namely the first fundamental form of $\mathcal{S}$ at $\mathbf{p}$, is defined by the following:

$$\mathcal{G}_{\mathcal{S}} := \begin{pmatrix} \langle \mathbf{t}_1, \mathbf{t}_1 \rangle & \langle \mathbf{t}_1, \mathbf{t}_2 \rangle \\ \langle \mathbf{t}_2, \mathbf{t}_1 \rangle & \langle \mathbf{t}_2, \mathbf{t}_2 \rangle \end{pmatrix} \ .$$

Note that, we have used the subscript $\mathcal{S}$ to relate the first fundamental form to the surface $\mathcal{S}$ and distinguish the associated 2×2 metric tensor from the 3×3 metric tensor necessary in the $\mathbb{R}^3$ local reference frame that will be defined in the next paragraphs.

In view of the equations and quantities we are going to study later, we need to define a three-dimensional local system of curvilinear coordinates, i.e., a coordinate system spanning a neighborhood $\mathcal{N}_{\mathbf{p}} \subset \mathbb{R}^3$ of a point $\mathbf{p}$ belonging to the surface. We have the following proposition [33]:

Proposition 2.3.1. *Let $\mathcal{S}$ be a regular surface and $\phi : U \to \mathcal{S}$ a local parametrization centered at $\mathbf{p} \in \mathcal{S}$. Then there exists a neighborhood $W \subset \phi(U)$ of $\mathbf{p}$ in $\mathcal{S}$ and a number $\varepsilon > 0$ such that the segments of the normal lines passing through points $\mathbf{q} \in W$, centered at $\mathbf{q}$ and with length 2ε, are disjoint.*

The union of all the segments of normal lines of length 2ε passing through $\mathbf{q}$ is called the *tubular neighborhood* $\mathcal{N}_{\mathbf{p}}$ of W. The previous proposition allows the extension of the local coordinate system to a tubular neighborhood of each point $\mathbf{p}$. We can define a $\mathbf{t}_3$ that extends the local basis $\{\mathbf{t}_1, \mathbf{t}_2\}$ to a triplet of basis vectors attached to $\mathbf{p} \in \mathcal{S}$ that can be used to describe all other points in $\mathcal{N}_{\mathbf{p}}$. This three-dimensional reference frame will be called the "Local Curvilinear coordinate System" (LCS), and it will be associated with the local coordinates (s^1, s^2, s^3), with metric tensor $\mathcal{G}$. Because of proposition 2.3.1, the three-dimensional coordinate transformation $\Phi_{\mathbf{p}}$ of each point $\mathbf{p} \in \mathcal{N}_{\mathbf{p}}$ from the global to the local coordinate is a differentiable map whose inverse is

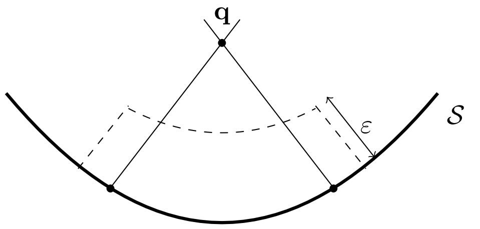

Figure 2.5: Example of normal intersection for a point $\mathbf{q}$ outside the tubular neighborhood.

differentiable (it is a diffeomorphism). We can identify these transformations as:

$$\begin{aligned} \Phi_{\mathbf{p}} &: \mathbb{R}^3 \to \mathbb{R}^3 & \qquad \Psi_{\mathbf{p}} &:= \Phi_{\mathbf{p}}^{-1} : \mathbb{R}^3 \to \mathbb{R}^3 \ , \\ \mathbf{x_p} &\mapsto \mathbf{s_p} & \mathbf{s_p} &\mapsto \mathbf{x_p} \\ \text{GCS} &\to \text{LCS} & \text{LCS} &\to \text{GCS} \end{aligned} \tag{2.5}$$

where $\mathbf{x_p} = (x^1_{\mathbf{p}}, x^2_{\mathbf{p}}, x^3_{\mathbf{p}})$ and $\mathbf{s_p} = (s^1_{\mathbf{p}}, s^2_{\mathbf{p}}, s^3_{\mathbf{p}})$ are the coordinates of $\mathbf{p}$ with respect to the GCS and the LCS, respectively. Note that, if we restrict the domain of $\Phi_{\mathbf{p}}$ to points on the surface, this is exactly the local parametrization ϕ.

Remark 2.9. Given a neighborhood $\mathcal{N}_{\mathbf{p}}$ of $\mathbf{p}$, every point $\mathbf{q} \in \mathcal{N}_{\mathbf{p}}$ can be expressed in the LCS as follows. Let $\mathbf{q} \in \mathcal{N}_{\mathbf{p}}$ be given in the GCS by $\mathbf{q} = (x^1_{\mathbf{q}}, x^2_{\mathbf{q}}, x^3_{\mathbf{q}})$. Consider the line passing through $\mathbf{q}$ and parallel to $\mathbf{t}_3$, which can be given the following parametric form:

$$\gamma(\lambda) : \lambda \mapsto (x^1(\mathbf{q}), x^2(\mathbf{q}), x^3(\mathbf{q})) - \mathbf{t}_3\lambda, \tag{2.6}$$

with $r = \gamma(\bar{\lambda})$ the intersection between the coordinate line γ and $\mathcal{S}$. Thus, the tuple:

$$(s^1(\mathbf{q}), s^2(\mathbf{q}), s^3(\mathbf{q})) = \left(x^1(r), x^2(r), \bar{\lambda}\right) \ ,$$

consists of the associated local coordinates of the point $\mathbf{q}$.

Remark 2.10. The LCS thus defined is not a global bijection. In fact, singular points may arise, for example, at the intersection of normal vectors leaving $\mathcal{S}$, fig. 2.5. However, proposition 2.3.1 ensures that this LCS is a diffeomorphism in the tubular ε-neighborhood of $\mathbf{p}$, where $\Phi_{\mathbf{q}}$ and $\Phi_{\mathbf{q}}^{-1}$ exist for each point $\mathbf{q}$ in this neighborhood and are continuous, i.e., the transformation of coordinates is a diffeomorphism.

We have now two coordinate systems, (x^1, x^2, x^3) and (s^1, s^2, s^3) for the neighborhood $\mathcal{N}_{\mathbf{p}} \subseteq \mathbb{R}^3$, for all $\mathbf{p} \in \mathcal{S}$, where x^1, x^2, x^3 are the standard Cartesian coordinates, while s^1, s^2, s^3 are the local curvilinear ones, and we assume to have their transformation relations $\Phi_{\mathbf{p}}$ and $\Psi_{\mathbf{p}}$. As usual, the Jacobian matrix of the coordinate transformation is the matrix of the change of coordinates between GCS and LCS. Going back to consider the "physical" vector $\vec{u}$, we assume it attached to the surface $\mathcal{S}$. Then, the vector field $\vec{u}$ can be written as $\vec{u} = \sum_i u^i \mathbf{t}_i$, where u^i are the components with respect to the induced local basis and are called "contravariant" components. Following standard notation, contravariant components are identified by means of superscripts. Note that the LCS basis $\mathbf{t}_i$ is formed by "covariant" vectors, and we use subscripts for their identification. If we consider an orthogonal reference frame, the relation between physical and contravariant components can be shown directly as:

$$u_{(i)} = \sum_i \sqrt{g_{ii}} u^i \ .$$

2.3.2 Differential operators

We start here by the definition of the intrinsic differential operators on a regular surface. We present the following proposition that we state without proof. The proof is immediate.

Proposition 2.3.2. *Let* (s^1, s^2) *be the curvilinear coordinates on* $\mathcal{S}$ *and* $\mathcal{G}_\mathcal{S}$ *the associated metric tensor. Let* $f : \mathcal{S} \to \mathbb{R}$ *a scalar differentiable function on* $\mathcal{S}$, $X : \mathcal{S} \to \mathbb{R}^2$ *a vector field on* $\mathcal{S}$ *and* $\mathbb{T} : \Omega \to \mathbb{R}^{3\times 3}$ *a rank-2 contravariant tensor given by* $\mathbb{T} = \{\tau^{ij}\}$. *Then, the* intrinsic *differential operators on* $\mathcal{S}$ *expressed in the local curvilinear coordinate system are given by the following expressions:*

- *the* intrinsic *gradient of* f *is:*

$$\nabla_\mathcal{G} f = \mathcal{G}_\mathcal{S}^{-1} \nabla f = g^{ij} \frac{\partial f}{\partial s^i} \ ; \tag{2.7}$$

- *the* intrinsic *divergence of* X *is:*

$$\nabla_\mathcal{G} \cdot X = \frac{1}{\sqrt{\det(\mathcal{G}_\mathcal{S})}} \nabla \cdot \left(\sqrt{\det(\mathcal{G}_\mathcal{S})} X \right) = \frac{\partial X^i}{\partial s^i} + \Gamma^i_{ij} X^j \ ; \tag{2.8}$$

- *the* j*-th component of the divergence of* $\mathbb{T}$ *is:*

$$(\nabla_\mathcal{G} \cdot \mathbb{T})^j = \nabla_{\mathcal{G}i} \tau^{ij} = \frac{1}{\sqrt{\det(\mathcal{G}_\mathcal{S})}} \frac{\partial}{\partial s^i} \left(\sqrt{\det(\mathcal{G}_\mathcal{S})} \tau^{ij} \right) + \Gamma^j_{ik} \tau^{ik} \ , \tag{2.9}$$

  where $\nabla_\mathcal{G} \cdot \tau^{(\cdot j)}$ *identifies the divergence of the* j*-th column of* $\mathbb{T}$, *and* Γ^k_{ij} *denote the Christoffel symbols;*

- *the* intrinsic *Laplace-Beltrami operator of f is:*

$$\Delta_{\mathcal{G}} f = \nabla_{\mathcal{G}} \cdot \nabla_{\mathcal{G}} f = \frac{1}{\sqrt{\det(\mathcal{G}_{\mathcal{S}})}} \frac{\partial}{\partial s^i} \left(\sqrt{\det(\mathcal{G}_{\mathcal{S}})} g^{ij} \frac{\partial f}{\partial s^j} \right) . \tag{2.10}$$

We state now some classical results useful in the study of partial differential equations that are still valid in the framework of surfaces. Starting from Stokes theorem 2.2.9 and Gauss-Green theorem given in eq. (2.2), with reference to the definition of the intrinsic differential operators, it is possible to write the divergence theorem on surfaces in the following intrinsic form:

Lemma 2.3.3. *Let $\mathcal{S} \subset \mathbb{R}^3$ be a surface with smooth boundary $\partial\mathcal{S}$ and X be continuously differentiable vector field. Then:*

$$\int_{\mathcal{S}} \nabla_{\mathcal{G}} \cdot X \, \mathrm{ds} = \int_{\partial\mathcal{S}} \langle X, \mu \rangle_{\mathcal{G}} \, \mathrm{d}\sigma$$

where $\mu : \mathcal{S} \to \mathbb{R}^2$ denotes the vector tangent to $\mathcal{S}$ and normal to $\partial\mathcal{S}$ with components written with respect to the local reference frame (i.e., $\mu = \mu^1\partial_1 + \mu^2\partial_2$), and ds *and* $\mathrm{d}\sigma$ *are the surface area measure and the curve length measure, respectively.*

Another consequence of theorem 2.2.9 is the intrinsic Green's formula (or integration by parts):

Lemma 2.3.4. *Let $\mathcal{S} \subset \mathbb{R}^3$ be a surface with smooth boundary $\partial\mathcal{S}$ and $f, g \in C^2(\bar{\mathcal{S}})$ be continuously differentiable functions over $\bar{\mathcal{S}}$. Then:*

$$\int_{\mathcal{S}} \langle \nabla_{\mathcal{G}} f, \nabla_{\mathcal{G}} g \rangle_{\mathcal{G}} \, \mathrm{ds} = - \int_{\mathcal{S}} \Delta_{\mathcal{G}} f \; g \, \mathrm{ds} + \int_{\partial\mathcal{S}} \langle \nabla_{\mathcal{G}} f, \mu \rangle_{\mathcal{G}} \, g \, \mathrm{d}\sigma \; , \tag{2.11}$$

where $\mu : \mathcal{S} \to \mathbb{R}^2$ denotes the vector tangent to $\mathcal{S}$ and normal to $\partial\mathcal{S}$ with components written with respect to the local reference frame, and ds *and* $\mathrm{d}\sigma$ *are the surface area measure and the curve length measure, respectively.*

2.3.3 Examples

Let us introduce now some classical examples of partial differential equations defined on a surface $\mathcal{S}$. We consider $\Gamma \subseteq \mathcal{S}$ a compact subset of the surface $\mathcal{S}$, to be able to handle boundary conditions.

Laplace equation. Laplace equation is one of the simplest example of elliptic partial differential equation. Let $u : \Gamma \to \mathbb{R}$ be a scalar function defined on the surface, then

Laplace equation is a second-order PDE and it can be defined on a surface considering the definition of the Laplace-Beltrami operator defined in eq. (2.10):

$$\Delta_{\mathcal{G}} u = \nabla_{\mathcal{G}} \cdot \nabla_{\mathcal{G}}\, u = 0\,.$$

This can be directly extended considering a variable diffusion tensor $\mathbf{K}$ and the equation:

$$\nabla_{\mathcal{G}} \cdot (\mathbf{K}\, \nabla_{\mathcal{G}}\, u) = 0\,.$$

Note that the tensor $\mathbf{K}$ is physically an intrinsic quantity that measures the diffusivity on the surface. Hence, in our setting the equation is well-defined and we do not need to proceed with further adaptation, like projections of quantities on the surface (like the approach presented in [36]).

Transport equation. Consider now a time-dependent case. The transport equation is a first order conservation law for the quantity $u : \Gamma \times [0, t_f] \to \mathbb{R}$. Assume Γ be fixed in time, and $\mathbf{w} : \Gamma \times [0, t_f] \to \mathbb{R}^2$ a given velocity field everywhere tangent to the surface. Then the linear transport equation reads:

$$\frac{\partial u}{\partial t} + \nabla_{\mathcal{G}} \cdot (\mathbf{w}\, u) = 0\,. \tag{2.12}$$

The equation can be written as:

$$\frac{\partial u}{\partial t} + \langle \mathbf{w}, \nabla_{\mathcal{G}}\, u \rangle_{\mathcal{G}} = 0\,,$$

if we assume a divergence-free velocity vector, i.e., $\nabla_{\mathcal{G}} \cdot \mathbf{w} = 0$.

Advection-diffusion-reaction equation. The advection-diffusion-reaction equation is a combination of the diffusion and advection equations used to describe a combination of physical phenomena. The general equation reads:

$$\frac{\partial u}{\partial t} - \nabla_{\mathcal{G}} \cdot (\mathbf{K}\, \nabla_{\mathcal{G}}\, u) + \langle \mathbf{w}, \nabla_{\mathcal{G}}\, u \rangle_{\mathcal{G}} + c\, u = f\,, \tag{2.13}$$

where $u : \Gamma \times [0, t_f] \to \mathbb{R}$ is the variable of interest and is a scalar function defined on the surface, $\mathbf{w} : \Gamma \times [0, t_f] \to \mathbb{R}^2$ is a given divergence-free velocity field everywhere tangential to the surface, $\mathbf{K}$ is the diffusion tensor, and we assume $c \geq 0$ and $f \in L^2(\Gamma)$. We also assume Γ to be fixed in time.

Navier-Stokes equations. We conclude by considering the Navier-Stokes equations for incompressible fluids. This system is defined in a three-dimensional domain, which we denote by $\Omega \subset \mathbb{R}^3$, and we assume that $\Omega \subseteq \mathcal{N}_{\mathcal{S}}$, the tubular neighborhood of a regular surface. In other words, we assume that the fluid moves on the surface $\mathcal{S}$. Let $\vec{u} : \Omega \times [0, t_f] \to \mathbb{R}^3$ be the velocity of the fluid, ρ the density, $p : \Omega \times [0, t_f] \to \mathbb{R}$ is the pressure and ϵ the kinematic viscosity. The following nonlinear equations govern the movement of an incompressible viscous fluid:

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla_{\mathcal{G}} \vec{u} = -\frac{1}{\rho} \nabla_{\mathcal{G}} p + \epsilon \Delta_{\mathcal{G}} \vec{u} \tag{2.14}$$

$$\nabla_{\mathcal{G}} \cdot \vec{u} = 0\,. \tag{2.15}$$

These equations are written with respect to an LCS anchored on the surface $\mathcal{S}$, which is considered as the bed over which the fluid moves. The proper and complete definition of the LCS will be described in the next chapter. This is the more complex and more interesting example we presented, and gives the basis of the study of Shallow Water Equations that are the focus of this work. We will see in the next chapter how this system of equations enters into account.

3

Intrinsic Shallow Water Equations

Shallow Water models are two-dimensional models for fluid dynamics based on the assumption that the fluid moves in waves with amplitude that is negligible with respect to wave length. These equations are derived from the three-dimensional Navier-Stokes equations after a process of depth average along a specific direction. Their importance is due to the fact that depth integration allows the reduction from three to two dimensions enabling efficient numerical computations, decreasing the burden and the complexity of a three dimensional model at large scales. Shallow Water models arise in a wide variety of applications, from meteorologic and atmospheric models to oceanography. Recently, applications for modeling avalanches, debris flows, landslides and other phenomena have been proposed. Although defined at smaller spatial scales than the previously mentioned models, these applications must consider a general topography, such as mountain landscapes, and this fact requires a rigorous derivation of the equations that takes into account the geometric setting.

Savage and Hutter [66, 67] were the first to develop a formulation of the SW model in local curvilinear coordinates based on depth integration along the normal to the topography. Their approach is valid only for small and essentially one-dimensional bottom curvatures and in practice assumes that the fluid surface is parallel to the bottom. This strategy was then extended by Bouchut et al. [17] and Bouchut and

Westdickenberg [16] to consider two-dimensional settings and less restrictive bottom geometries. Starting from the NS equations, the flow velocity component perpendicular to the bottom is considered negligible and a hydrostatic pressure distribution is assumed along local normals. A more intrinsic approach was recently proposed by Fent et al. [37], who approximate the optimal integration path, named the “cross-flow” path, with the direction normal to the bottom starting from a NS system defined on a curvilinear reference frame anchored to the bottom surface.

The aim of this chapter is to derive a new geometrically intrinsic formulation of the SWE on general topography, which we call the Intrinsic Shallow Water Equations (ISWE), from the Navier-Stokes system and prove some of their main properties. Our developments take inspiration from the works of Bouchut and Westdickenberg [16] and of Fent et al. [37], both of whom include the effects of the bottom geometry on the SW system. In the former, the derived model includes the bottom geometry using a three-dimensional Cartesian coordinate system. As a consequence, the SWEs contain non-conservative terms that need to be properly handled. In the latter approach, depth integration proceeds using a local reference frame defined on the bottom surface. Again, non-conservative terms arise in the covariant form of the equations. Similarly, in our work we describe the SW model on a local reference frame. However, differently from previous work, by careful use of contravariant and covariant vectors we are able to arrive at a system that is completely intrinsic to the bottom geometry, with a source term that contains only bottom slope and curvature information. The resulting set of equations is characterized by spatially varying flux functions and bottom-related sources. We study the mathematical structure of the proposed approximation, proving order of accuracy with respect to a “geometric” aspect ratio parameter $\epsilon_{\mathcal{G}}$ that includes information on local curvatures and slope of the bottom surface. We then study the hyperbolic structure of the proposed system using bottom-intrinsic differential operators and show that it is invariant under rotational transformations, satisfies the lake-at-rest condition, and admits a conserved energy in absence of bed resistance.

3.1 Incompressible Navier-Stokes equations

The incompressible Navier-Stokes equations are a system of four partial differential equations that fully describe the dynamics of a viscous fluid, in which the stresses are assumed proportional to the gradient of velocity. They are an important example of hyperbolic partial differential equations, directly derived from the equations for the conservation of mass and momentum. Consider the incompressible Navier-Stokes equations on an open domain $\Omega \subset \mathbb{R}^3$ written as:

$$\nabla \cdot \vec{u} = 0 \ , \tag{3.1a}$$

$$\frac{\partial \vec{u}}{\partial t} + \nabla \cdot (\vec{u} \otimes \vec{u}) = -\frac{1}{\rho} \nabla p + \frac{1}{\rho} \nabla \cdot \mathbb{T} + \vec{g} \ , \tag{3.1b}$$

where $\vec{u} : \Omega \times [0, t_f] \to \mathbb{R}^3$ is the fluid velocity, ρ its density, assumed constant, $p : \Omega \times [0, t_f] \to \mathbb{R}$ is the fluid pressure, $\mathbb{T} : \Omega \to \mathbb{R}^{3\times 3}$ the deviatoric stress tensor, and $\vec{g}$ the gravity acceleration. Note that we have used the product rule of differentiation and the incompressibility condition (3.1a) to write the convective term in conservative form. Explicitly, we exploited the relation: $\nabla \cdot \vec{u} \otimes \vec{u} = \vec{u} \cdot \nabla \vec{u} + \vec{u} \nabla \cdot \vec{u}$. Moreover, using the standard Reynolds averaging in time and the Boussinesq simplification it is possible to relate the term $\nabla \cdot \mathbb{T}$ with the more usual $\epsilon \Delta \vec{u}$ [69], to recover the more classical formulation as seen in eq. (2.14). We assume that the domain boundary $\partial\Omega$ is smooth and formed by the union of the bottom surface ($\mathcal{S}_{\mathcal{B}}$), the free surface ($\mathcal{S}_{\mathcal{F}}$), and the lateral surface. Smoothness is detailed by the hypothesis that all these surfaces are regular and can be identified by the graph of some function. Thus, the bottom surface will be given by the graph of the function $\mathcal{B} : U \times \mathbb{R} \to \mathbb{R}$, $U \subset \mathbb{R}^2$ open, i.e., in a global (Cartesian) coordinate system x^1, x^2, x^3 (GCS) with x^3 assumed aligned with the action of gravity but in the opposite direction:

$$\mathcal{S}_{\mathcal{B}} := \left\{ (x^1, x^2, x^3, t) \in \mathbb{R}^3 \times \mathbb{R} \text{ such that } x^3 = \mathcal{B}(x^1, x^2) \right\} \ .$$

Equivalently, $\mathcal{S}_{\mathcal{B}}$ can be defined as $\overset{-1}{F_{\mathcal{B}}}(0)$, where $F_{\mathcal{B}}(x^1, x^2, x^3) := x^3 - \mathcal{B}(x^1, x^2)$. Similarly, we can define the fluid free surface $\hat{\mathcal{F}} : U \times [0, t_f] \to \mathbb{R}$ as:

$$\mathcal{S}_{\hat{\mathcal{F}}} := \left\{ (x^1, x^2, x^3, t) \in \mathbb{R}^3 \times \mathbb{R} \text{ such that } x^3 = \hat{\mathcal{F}}(x^1, x^2, t) \right\} = \overset{-1}{F_{\hat{\mathcal{F}}}}(0) \ ,$$

where $F_{\hat{\mathcal{F}}}(x^1, x^2, x^3, t) := x^3 - \hat{\mathcal{F}}(x^1, x^2, t)$. The lateral boundary is fixed and independent of time and does not enter our discussion. Next, we want to move our geometric description to a new coordinate system s^1, s^2, s^3 that locally follows the bottom surface.

3.1.1 Working on arbitrary topography

Following Fent et al. [37] we define a local curvilinear reference system (LCS) positioned on the surface representing the topography of the bottom. All the developments, including depth integration, will be carried out with respect to this local reference system. We would like to describe the motion of a fluid particle using a coordinate system that satisfies the following two main conditions:

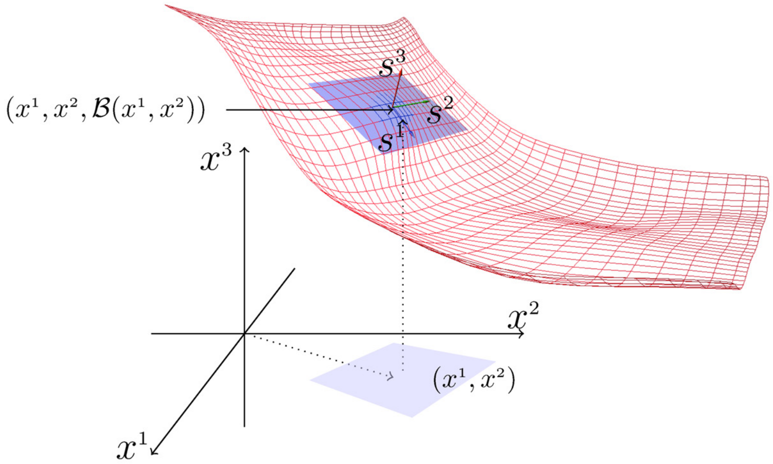

Figure 3.1: Example of bottom surface and LCS coordinate system.

i) the first two coordinates run along the bottom surface $\mathcal{S}_{\mathcal{B}}$, their tangent vectors belonging at each point $\mathbf{p} \in \mathcal{S}_{\mathcal{B}}$ to the tangent plane $T_{\mathbf{p}}\mathcal{S}_{\mathcal{B}}$;

ii) the third coordinate crosses the surface orthogonally so a vector tangent to $\mathcal{S}_{\mathcal{B}}$ is everywhere orthogonal to $\mathbf{N}$, the surface normal vector (fig. 3.1).

We start with a rigorous definition of the tangent plane on the bottom surface and proceed with the definition of a local coordinate system. This local coordinate system for the surface will be later extended to be a coordinate system in $\mathbb{R}^3$ for a tubular neighborhood of the surface.

Monge Coordinate System. Consider the bottom surface $\mathcal{S}_{\mathcal{B}}$. As explained in the previous chapter, we can locally characterize a surface by writing its local parametrization and express quantities in the local coordinates. In the case of a surface given by the graph of a scalar function (in our case the bottom surface $\mathcal{S}_{\mathcal{B}}$ can be described using the function $\mathcal{B}$), we can consider a particular example of parametrization, called the *Monge parametrization*, which is defined by:

$$\begin{aligned}\phi :& U \subseteq \mathbb{R}^2 \to \mathbb{R}^3 \\ &(s^1, s^2) \mapsto (x^1(s^1, s^2), x^2(s^1, s^2), x^3(s^1, s^2)) := (s^1, s^2, \mathcal{B}(s^1, s^2))\,.\end{aligned}$$

As previously seen in example 2.4, we can compute the induced basis for the tangent plane as:

$$\frac{\partial \phi}{\partial s^1} = \left(\frac{\partial x^1(s^1, s^2)}{\partial s^1}, \frac{\partial x^2(s^1, s^2)}{\partial s^1}, \frac{\partial x^3(s^1, s^2)}{\partial s^1}\right) = (1, 0, \mathcal{B}_{s^1})\ ,$$

and

$$\frac{\partial \phi}{\partial s^2} = \left(\frac{\partial x^1(s^1, s^2)}{\partial s^2}, \frac{\partial x^2(s^1, s^2)}{\partial s^2}, \frac{\partial x^3(s^1, s^2)}{\partial s^2} \right) = (0, 1, \mathcal{B}_{s^2}) \ ,$$

where $\mathcal{B}_{s^i} = \partial \mathcal{B}/\partial s^i$, for $i = 1, 2$. We need to underline that in general the induced reference frame is not orthonormal, and not even orthogonal. In fact:

$$\begin{aligned} \langle \partial_1 \phi, \partial_1 \phi \rangle &= \langle (1, 0, \mathcal{B}_{s^1}), (1, 0, \mathcal{B}_{s^1}) \rangle = 1 + \mathcal{B}_{s^1}^2 \ , \\ \langle \partial_1 \phi, \partial_2 \phi \rangle &= \langle (1, 0, \mathcal{B}_{s^1}), (0, 1, \mathcal{B}_{s^2}) \rangle = \mathcal{B}_{s^1} \mathcal{B}_{s^2} \ , \\ \langle \partial_2 \phi, \partial_2 \phi \rangle &= \langle (0, 1, \mathcal{B}_{s^2}), (0, 1, \mathcal{B}_{s^2}) \rangle = 1 + \mathcal{B}_{s^2}^2 \ . \end{aligned}$$

With reference to the examples related to the graph of a scalar function in chapter 2, it is easy to compute all the other important geometric quantities.

To ensure numerical stability we are interested in a reference frame with orthogonal vectors. We present here two theoretical results following [2].

Theorem 3.1.1. *Given two differentiable and linearly independent vector fields* X^1, X^2 *on a surface* $\mathcal{S}$, *then there exists a local parametrization* $\phi : U \to \mathcal{S}$ *such that* $\partial_1 \phi$ *is proportional to* X^1 *and* $\partial_2 \phi$ *is proportional to* X^2. *Moreover, the support of the integral lines coincide with the support of the coordinate lines.*

Corollary 3.1.2. *There exists a local parametrization at each point of* $\mathcal{S}$ *such that* $\partial_1 \phi$ *and* $\partial_2 \phi$ *are orthogonal.*

Proof. To prove this corollary it is enough to apply theorem 3.1.1 to the two vector fields of an existing frame, the second of which has been orthogonalized with respect to the first one using Gram-Schmidt. □

As a consequence, we are allowed to select orthogonal vector fields and find a flat coordinate system for them. In our case, a coordinate frame can be formed by taking $\partial_1 \phi$ and applying Gram-Schmidt to orthogonalize $\partial_2 \phi$ with respect to $\partial_1 \phi$. Then, the resulting coordinate system is well-defined, as summarized in the following proposition:

Proposition 3.1.3. *Given a regular surface* $\mathcal{S}_{\mathcal{B}} \subset \mathbb{R}^3$ *defined by the graph of a differentiable function* $\mathcal{B} \in C^\infty(U)$, U *open subset of* $\mathbb{R}^2$. *Let* $\phi : (s^1, s^2) \to (s^1, s^2, \mathcal{B}(s^1, s^2))$ *be the Monge parametrization for* $\mathcal{S}_{\mathcal{B}}$, ∂_1, ∂_2 *be the induced basis for the tangent plane, and define:*

$$\mathbf{t}_1 := a \, \partial_1 \qquad \textit{and} \qquad \mathbf{t}_2 := b \left(\partial_2 - \frac{\partial_2 \cdot \partial_1}{\partial_1 \cdot \partial_1} \partial_1 \right)$$

for some function $a, b \in C^\infty(\mathcal{S}_{\mathcal{B}})$. *Then, around each point* $\mathbf{p} \in \mathcal{S}_{\mathcal{B}}$ *there exists a neighborhood* $\tilde{V}$ *and a parametrization* $\tilde{\phi} : \tilde{U} \to \tilde{V}$ *such that the two partial derivatives* $\tilde{\partial}_1$ *and* $\tilde{\partial}_2$ *coincide with* $\mathbf{t}_1$ *and* $\mathbf{t}_2$ *on* $\tilde{V}$.

Moreover, the two vectors $\mathbf{t}_1$ *and* $\mathbf{t}_2$ *form an orthogonal flat coordinate system, i.e.:*

$$\begin{cases} \langle \mathbf{t}_1, \mathbf{t}_2 \rangle = 0 \\ [\mathbf{t}_1, \mathbf{t}_2] = 0 \end{cases} .$$

Proof. Given the parametrization $\phi : U \to \mathcal{S}_{\mathcal{B}}$, $\phi(s^1, s^2) = (s^1, s^2, \mathcal{B}(s^1, s^2))$ and the natural coordinate frame $\{\partial_1, \partial_2\}$, let X_1, X_2 be the vector fields given by $X_1 = \partial_1$, $X_2 = \partial_2 - \frac{\partial_2 \cdot \partial_1}{\partial_1 \cdot \partial_1} \partial_1$. From corollary 3.1.2, there exists a parametrization $\tilde{\phi} : \tilde{U} \to \mathcal{S}_{\mathcal{B}}$, with coordinates $\tilde{s^1}, \tilde{s^2} \in \tilde{U}$, such that:

$$\frac{\partial \tilde{\phi}}{\partial \tilde{s^1}} \propto X_1 \,, \qquad \frac{\partial \tilde{\phi}}{\partial \tilde{s^2}} \propto X_2 \,,$$

or equivalently that there exist $a, b \in C^\infty(\mathcal{S}_{\mathcal{B}})$ such that $\tilde{\partial}_1 = a\, X_1$ and $\tilde{\partial}_2 = b\, X_2$.

Using the equivalence stated in theorem 2.1.10, we conclude that $\tilde{\partial}_1$ and $\tilde{\partial}_2$ satisfy:

$$\begin{cases} \langle a\, X_1, b\, X_2 \rangle = 0 \\ [a\, X_1, b\, X_2] = 0 \end{cases} ,$$

where the first equation is satisfied by construction. □

In the next paragraph we take care of giving an explicit expression for our vectors, after the extension of the local coordinate system to the three-dimensional tubular neighborhood.

Local curvilinear system (LCS). Recall that we are interested in a local curvilinear system anchored to the bottom surface that can characterize all the fluid domain $\Omega \subset \mathbb{R}^3$ and that satisfies the two main conditions in items i) and ii). With reference to section 2.3.1, we know that we can extend the local coordinate system on $\mathcal{S}_{\mathcal{B}}$ to a three-dimensional coordinate system for the tubular neighborhood of the surface. In particular, see eq. (2.5) for the definition of the three-dimensional coordinate transformation Φ. Thus, regarding the ensuing reference frame, the previous requests amount to asking that the three vector fields $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ in $\mathbb{R}^3$ are such that:

$$\mathbf{t}_1(\mathbf{p}), \mathbf{t}_2(\mathbf{p}) \in T_{\mathbf{p}}\mathcal{S}_{\mathcal{B}} \quad \forall\, \mathbf{p} \in \mathcal{S}_{\mathcal{B}} \,,$$

and we define $\mathbf{t}_3(\mathbf{p})$ in order to be orthogonal to the other two frame vectors and such that the right-hand rule is satisfied. This is equivalent to define an orientation to the surface (see remark 2.6). Moreover, as previously mentioned, we need that $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ commute in all $\mathbb{R}^3$ and to be pairwise orthogonal. The practical definition of the LCS proceeds as follows. First, we calculate the two tangent vectors $\hat{\mathbf{t}}_1(\mathbf{p})$ and $\hat{\mathbf{t}}_2(\mathbf{p})$, which is equivalent to compute the differential of $\Phi_{\mathbf{p}}$ applied to the canonical basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, of the GCS, or, equivalently, as the derivatives of the coordinate transformation with respect to the LCS variables:

$$\hat{\mathbf{t}}_i(\mathbf{p}) = \mathrm{d}\Phi_{\mathbf{p}}(\mathbf{e}_i) = \left(\frac{\partial x^1}{\partial s^i}, \frac{\partial x^2}{\partial s^i}, \frac{\partial x^3}{\partial s^i}\right), \qquad i = 1, 2 \;,$$

where $\mathrm{d}\Phi_{\mathbf{p}}$ is the Jacobian matrix of the coordinate transformation. Then, vector $\hat{\mathbf{t}}_2$ is orthogonalized with respect to $\hat{\mathbf{t}}_1$ via Gram-Schmidt, yielding the desired orthogonal frame $\mathbf{t}_1, \mathbf{t}_2$ on $T_{\mathbf{p}}\mathcal{S}_{\mathcal{B}}$. The frame-completing vector $\mathbf{t}_3$ is chosen to be orthogonal to the previous two and unitary, i.e. $\|\mathbf{t}_3(\mathbf{p})\| = 1$. Note that normalization of the other two basis vectors cannot be done, as this would amount to assume a zero curvature of $\mathcal{S}_{\mathcal{B}}$ at $\mathbf{p}$, loosing all the geometric information we would like to preserve in our LCS. The explicit expression of the LCS frame vectors at $\mathbf{p} \in \mathcal{S}_{\mathcal{B}}$ is:

$$\mathbf{t}_1(\mathbf{p}) = (1, 0, \mathcal{B}_{s^1}) \;, \tag{3.2a}$$

$$\mathbf{t}_2(\mathbf{p}) = \left(-\frac{\mathcal{B}_{s^1}\mathcal{B}_{s^2}}{1 + (\mathcal{B}_{s^1})^2}, 1, \frac{\mathcal{B}_{s^2}}{1 + (\mathcal{B}_{s^1})^2}\right) \;, \tag{3.2b}$$

$$\mathbf{t}_3(\mathbf{p}) = \mathbf{N}(\mathbf{p}) = \frac{\mathbf{t}_1(\mathbf{p}) \wedge \mathbf{t}_2(\mathbf{p})}{\|\mathbf{t}_1(\mathbf{p})\| \, \|\mathbf{t}_2(\mathbf{p})\|} = \frac{(-\mathcal{B}_{s^1}, -\mathcal{B}_{s^2}, 1)}{\|\mathbf{t}_1(\mathbf{p})\| \, \|\mathbf{t}_2(\mathbf{p})\|} \;. \tag{3.2c}$$

The associated metric tensor, as a consequence of the orthogonality property, becomes the diagonal matrix given by:

$$\mathcal{G} := \begin{pmatrix} \|\mathbf{t}_1(\mathbf{p})\|^2 & 0 & 0 \\ 0 & \|\mathbf{t}_2(\mathbf{p})\|^2 & 0 \\ 0 & 0 & \|\mathbf{t}_3(\mathbf{p})\|^2 \end{pmatrix} = \begin{pmatrix} h_{(1)}^2 & 0 & 0 \\ 0 & h_{(2)}^2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \;. \tag{3.3}$$

It is important to underline that this particular definition of the tangent vectors has been made for our convenience, but any local coordinate system that satisfies items i) and ii) can be used. What follows is independent on the parametrization of the surface and the definition of the LCS.

Remark 3.1. To simplify the exposition, we do not fully employ the classical tensor calculus notation and do not use Einstein summation convention. However we need to distinguish physical, covariant, and contravariant versions of vectors and tensors, to ensure coordinate invariance. As previously described we denote the physical components of vectors by subscripts surrounded with parentheses. In the particular case of our LCS s^1, s^2, s^3, equipped with the associated diagonal metric $\mathcal{G}$ and the reference basis vectors $\mathbf{t}_i$, the physical vector components need to be scaled with $\sqrt{\mathcal{G}}$, or $\sqrt{g_{ii}} = h_{(i)}$. Hence, a vector field $\vec{u}$ can be written as $\vec{u} = \sum_i u^i \mathbf{t}_i$, where the contravariant components u^i are related to the physical components by the relation:

$$u_{(i)} = h_{(i)} u^i \, .$$

We recall that, following standard notation, contravariant components are identified by means of superscripts, while covariant components are denoted by subscripts.

Next, we need to adapt to the LCS the expressions of the differential operators that appear in the Navier-Stokes equations, i.e., the gradient of a scalar function, the divergence of a vector field, and the divergence of a tensor field, as stated in proposition 2.3.2. We report here the particular formulation we obtain considering the LCS:

- the gradient of a scalar function f is:

$$\nabla_{\mathcal{G}} f = \left(\frac{1}{h^2_{(1)}} \frac{\partial f}{\partial s^1}, \; \frac{1}{h^2_{(2)}} \frac{\partial f}{\partial s^2}, \; \frac{\partial f}{\partial s^3} \right) ; \tag{3.4}$$

- the divergence of a contravariant vector field $\vec{u} = u^1 \mathbf{t}_1 + u^2 \mathbf{t}_2 + u^3 \mathbf{t}_3$ is:

$$\nabla_{\mathcal{G}} \cdot \vec{u} = \frac{1}{h_{(1)} h_{(2)}} \left(\frac{\partial \left(h_{(1)} h_{(2)} u^1 \right)}{\partial s^1} + \frac{\partial \left(h_{(1)} h_{(2)} u^2 \right)}{\partial s^2} + \frac{\partial \left(h_{(1)} h_{(2)} u^3 \right)}{\partial s^3} \right) ; \tag{3.5}$$

- the j-th component of the divergence of a 3×3 rank-2 contravariant tensor $\mathbb{T} = \{\tau^{ij}\}$ is:

$$\begin{aligned} \left(\nabla_{\mathcal{G}} \cdot \mathbb{T} \right)^j = \nabla_{\mathcal{G}} \cdot \tau^{(\cdot j)} &+ \frac{1}{h_{(j)}} \left(2 \tau^{1j} \frac{\partial h_{(j)}}{\partial s^1} - \tau^{11} \frac{h_{(1)}}{h_{(j)}} \frac{\partial h_{(1)}}{\partial s^j} \right) \\ &+ \frac{1}{h_{(j)}} \left(2 \tau^{2j} \frac{\partial h_{(j)}}{\partial s^2} - \tau^{22} \frac{h_{(2)}}{h_{(j)}} \frac{\partial h_{(2)}}{\partial s^j} \right) , \end{aligned} \tag{3.6}$$

  where $\nabla_{\mathcal{G}} \cdot \tau^{(\cdot j)}$ identifies the divergence of the j-th column of $\mathbb{T}$.

Remark 3.2. In the next sections, we will reduce our system to a two-dimensional tangent-following local system describing only points of the bottom surface by means of the coordinates s^1, s^2. In this case, the metric tensor reduces to the sub-tensor containing only the information related to those two directions, identified previously with $\mathcal{G}_{S_{\mathcal{B}}}$. For simplicity, we will use the same symbols, equations, and operators in compact form independently of the spatial dimension, and the context will provide the appropriate definition.

Curvilinear Navier-Stokes equations. Using eqs. (3.4) to (3.6), the Navier-Stokes equations given in eq. (3.1) can be written in the LCS as:

$$\nabla_{\mathcal{G}} \cdot \vec{u} = 0 \; , \tag{3.7a}$$

$$\frac{\partial \vec{u}}{\partial t} + \nabla_{\mathcal{G}} \cdot (\vec{u} \otimes \vec{u}) = -\frac{1}{\rho} \nabla_{\mathcal{G}} p + \frac{1}{\rho} \nabla_{\mathcal{G}} \cdot \mathbb{T} + \vec{g} \; . \tag{3.7b}$$

3.2 Derivation of the intrinsic SW model

The derivation of the SWE starts from the formulation of Navier-Stokes equations in the local coordinate system eq. (3.7). Next, we perform depth integration along s^3, the direction locally normal to the terrain surface running between the bottom and the free surfaces. We start this task by looking at the boundary conditions on these surfaces.

Boundary conditions

Using the LCS, the bottom and free surfaces are given by:

$$\mathcal{S}_{\mathcal{B}} := \Big\{ (s^1, s^2, s^3) \in \mathbb{R}^3 \text{ such that } s^3 = \mathcal{B}(s^1, s^2) \equiv 0 \Big\} \; ,$$
$$\mathcal{S}_{\mathcal{F}} := \Big\{ (s^1, s^2, s^3, t) \in \mathbb{R}^3 \times [0, t_f] \text{ such that } s^3 = \mathcal{F}(s^1, s^2, t) \equiv \eta(s^1, s^2, t) \Big\} \; ,$$

where $\eta(s^1, s^2, t) = \mathcal{F}(s^1, s^2, t) - \mathcal{B}(s^1, s^2)$ denotes the fluid depth. We assume that the bottom is not eroding and thus maintains a fixed geometry, while the fluid surface is a function of time. The kinematic conditions postulate that the free surface moves with the fluid and that the bottom is impermeable. Thus we can write:

$$\frac{dF_{\mathcal{M}}}{dt} = \frac{\partial F_{\mathcal{M}}}{\partial t} + \vec{u} \cdot \nabla_{\mathcal{G}} F_{\mathcal{M}} \Big|_{\mathcal{M}} = 0 \; ,$$

where $\mathcal{M} = \mathcal{B}$ or $\mathcal{F}$. Since $F_{\mathcal{B}} = s^3 - \mathcal{B}(s^1, s^2)$ and $F_{\mathcal{F}} = s^3 - \mathcal{F}(s^1, s^2, t)$, we obtain immediately:

$$\frac{dF_{\mathcal{B}}}{dt} = \vec{u} \Big|_{\mathcal{B}} \cdot \nabla_{\mathcal{G}} F_{\mathcal{B}} = \vec{u} \cdot \nabla_{\mathcal{G}} F_{\mathcal{B}} \Big|_{s^3=0} = 0 \; , \tag{3.8a}$$

$$\frac{dF_{\mathcal{F}}}{dt} = -\frac{\partial \eta}{\partial t} + \vec{u} \cdot \nabla_{\mathcal{G}} F_{\eta} \Big|_{s^3=\eta} = -\frac{\partial \eta}{\partial t} - \left(\frac{u^1}{h_{(1)}} \frac{\partial \eta}{\partial s^1} + \frac{u^2}{h_{(2)}} \frac{\partial \eta}{\partial s^2} - u^3 \right) \Big|_{s^3=\eta} = 0 \; . \tag{3.8b}$$

Moreover, assuming that the external actions on the fluid surface are negligible, the dynamic condition at the fluid-air interface translates into a zero-stress boundary equation:

$$\mathbb{T}_{\mathcal{F}} \cdot \mathbf{N}_{\mathcal{F}} = 0 \; , \qquad \mathbf{N}_{\mathcal{F}} = \frac{\nabla \mathcal{F}}{||\nabla \mathcal{F}||} \; , \tag{3.9}$$

where $\mathbf{N}_{\mathcal{F}}$ is the unit normal vector on the free surface $\mathcal{F}$. The bed boundary condition imposes the value of the shear stress:

$$\mathbb{T}_{\mathcal{B}} \cdot \mathbf{N}_{\mathcal{B}} = \mathbf{f}_{\mathcal{B}} = \tau_b^1 \mathbf{t}_1 + \tau_b^2 \mathbf{t}_2 + p_{\mathcal{B}} \mathbf{t}_3 \,, \tag{3.10}$$

where $p_{\mathcal{B}}$ indicates the bottom pressure. The values for τ_b^i are expressed by means of classical steady-state empirical friction laws (e.g., Chezy, Manning, or Gauckler-Strickler equations).

Depth integration of the Navier-Stokes equations

Starting from the Navier-Stokes equations written in the local curvilinear coordinate system as given in eq. (3.7), we perform depth integration along the normal direction s^3 from $s^3 = \mathcal{B}(s^1, s^2) \equiv 0$ to $s^3 = \mathcal{F}(s^1, s^2, t) \equiv \eta(s^1, s^2, t)$. In the interest of space, we will omit the measure symbol $\mathrm{d}s^3$ in all the integrals.

The continuity equation (eq. (3.7a)). Applying Leibniz rule and recalling the kinematic boundary conditions given by eq. (3.8), we obtain:

$$\begin{aligned}
\int_0^\eta \nabla_{\mathcal{G}} \cdot \vec{u} = & \int_0^\eta \frac{1}{h_{(1)}h_{(2)}} \left(\frac{\partial\left(h_{(1)}h_{(2)}u^1\right)}{\partial s^1} + \frac{\partial\left(h_{(1)}h_{(2)}u^2\right)}{\partial s^2} + \frac{\partial\left(h_{(1)}h_{(2)}u^3\right)}{\partial s^3} \right) = \\
= & \frac{1}{h_{(1)}h_{(2)}} \frac{\partial}{\partial s^1} \int_0^\eta h_{(1)}h_{(2)}u^1 + \frac{1}{h_{(1)}h_{(2)}} \frac{\partial}{\partial s^2} \int_0^\eta h_{(1)}h_{(2)}u^2 \\
& + u^3\Big|_{s^3=\eta} - \frac{u^1}{h_{(1)}} \frac{\partial \mathcal{F}}{\partial s^1}\Big|_{s^3=\eta} - \frac{u^2}{h_{(2)}} \frac{\partial \mathcal{F}}{\partial s^2}\Big|_{s^3=\eta} \\
& - u^3\Big|_{s^3=0} + \frac{u^1}{h_{(1)}} \frac{\partial \mathcal{B}}{\partial s^1}\Big|_{s^3=0} + \frac{u^2}{h_{(2)}} \frac{\partial \mathcal{B}}{\partial s^2}\Big|_{s^3=0} = \\
= & \frac{\partial \eta}{\partial t} + \nabla_{\mathcal{G}} \cdot \int_0^\eta \underline{\vec{u}} \,,
\end{aligned}$$

where $\underline{\vec{u}} := [u^1, u^2]^{\mathrm{T}}$ and the curvilinear divergence operator $\nabla_{\mathcal{G}} \cdot$ is adapted to the two-dimensional setting (see remark 3.2). Recall that application of Leibniz rule requires enough regularity of both bottom and free surfaces as well as the velocity vector $\vec{u}$.

The momentum equation (eq. (3.7b)). Integration along the s^3-direction yields:

$$\int_0^\eta \frac{\partial \vec{u}}{\partial t} + \int_0^\eta \nabla_{\mathcal{G}} \cdot (\vec{u} \otimes \vec{u}) = -\frac{1}{\rho} \int_0^\eta \nabla_{\mathcal{G}}\, p - g \int_0^\eta \nabla_{\mathcal{G}}\, x^3 + \frac{1}{\rho} \int_0^\eta \nabla_{\mathcal{G}} \cdot \mathbb{T} \,.$$

Employing Leibniz rule and the kinematic BC, the left-hand-side can be written as:

$$\begin{aligned}
& \frac{\partial}{\partial t} \int_0^\eta \vec{u} - \vec{u} \frac{\partial \eta}{\partial t} + \nabla_{\mathcal{G}} \cdot \int_0^\eta \vec{u} \otimes \vec{u} - (\vec{u} \otimes \vec{u})\, \nabla_{\mathcal{G}}\, \mathcal{F}\Big|_{s^3=\eta} + (\vec{u} \otimes \vec{u})\, \nabla_{\mathcal{G}}\, \mathcal{B}\Big|_{s^3=0} = \\
& = \frac{\partial}{\partial t} \int_0^\eta \vec{u} + \nabla_{\mathcal{G}} \cdot \int_0^\eta \vec{u} \otimes \vec{u} - \vec{u} \left(\frac{\partial \eta}{\partial t} + \vec{u} \cdot \nabla_{\mathcal{G}}\, \mathcal{F}\Big|_{s^3=\eta} \right) = \frac{\partial}{\partial t} \int_0^\eta \vec{u} + \nabla_{\mathcal{G}} \cdot \int_0^\eta \vec{u} \otimes \vec{u} \,,
\end{aligned}$$

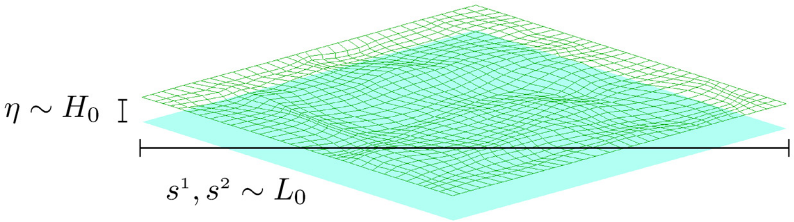

Figure 3.2: Example of a thin and wide layer of fluid over an horizontal bottom surface.

Similar computations for the right-hand-side produce:

$$
\begin{aligned}
-\frac{1}{\rho}\int_0^\eta \nabla_{\mathcal{G}}\, p - g\int_0^\eta \nabla_{\mathcal{G}}\, x^3 + \frac{1}{\rho}\nabla_{\mathcal{G}}\cdot\int_0^\eta \mathbb{T} - \frac{\mathbb{T}}{\rho}\nabla_{\mathcal{G}}\,\mathcal{F}\Big|_{s^3=\eta} + \frac{\mathbb{T}}{\rho}\nabla_{\mathcal{G}}\,\mathcal{B}\Big|_{s^3=0} &= \\
= -\frac{1}{\rho}\int_0^\eta \nabla_{\mathcal{G}}\, p - g\int_0^\eta \nabla_{\mathcal{G}}\, x^3 + \frac{1}{\rho}\nabla_{\mathcal{G}}\cdot\int_0^\eta \mathbb{T} + \frac{1}{\rho}\mathbb{T}_{\mathcal{B}}\cdot\mathbf{N}_{\mathcal{B}}\ ,
\end{aligned}
$$

where we made use here of the dynamic boundary conditions given in eqs. (3.9) and (3.10).

The normally integrated Navier-Stokes equations. Putting all the above calculations together we obtain the final system of normally integrated Navier-Stokes equations:

$$
\frac{\partial\eta}{\partial t} + \nabla_{\mathcal{G}}\cdot\int_0^\eta \underline{\vec{u}} = 0\ , \tag{3.11a}
$$

$$
\frac{\partial}{\partial t}\int_0^\eta \vec{u} + \nabla_{\mathcal{G}}\cdot\int_0^\eta \vec{u}\otimes\vec{u} = -\frac{1}{\rho}\int_0^\eta \nabla_{\mathcal{G}}\, p - g\int_0^\eta \nabla_{\mathcal{G}}\, x^3 + \frac{1}{\rho}\nabla_{\mathcal{G}}\cdot\int_0^\eta \mathbb{T} + \frac{1}{\rho}\mathbb{T}_{\mathcal{B}}\cdot\mathbf{N}_{\mathcal{B}}\ . \tag{3.11b}
$$

Here eq. (3.11a) is a scalar equation while eq. (3.11b) is a three-component system. In the following section, we derive the reduced two-component equations governing momentum balance by doing a formal order expansion and invoking the SW hypothesis.

Length scales and physical quantities

The classical SW hypothesis states that the characteristic depth of the fluid is smaller than the characteristic wavelength. In the present context, this statement is equivalent to our assumption of small normal velocity. To see this, assume a setting with a relatively thin and wide fluid moving on the terrain surface (fig. 3.2). Denote with L_0 the length scale in a direction tangential to the bottom and with H_0 the length scale of the fluid depth measured along the normal. The shallow water scaling assumes that

$H_0/L_0 = \epsilon \ll 1$. We would like to connect this idea with the order of approximation of the model in our curvilinear setting.

Denote by V_0 the scale of the contravariant tangential velocity components u^1 and u^2, and by W_0 the scale of the contravariant normal component u^3. From the hypothesis of regularity of the bottom surface, the metric coefficients can be considered of order $\mathcal{O}(1)$, and represent just the length scale of the coordinate transformation between the GCS and the LCS. However, we do not have any *a priori* information on the order of magnitude of the derivatives of these metric coefficients. Formal application of the chain rule of differentiation to the continuity equation (3.7a) yields:

$$\begin{aligned} \nabla_{\mathcal{G}} \cdot \vec{u} &= \frac{1}{h_{(1)}h_{(2)}}\left(\frac{\partial}{\partial s^1}(h_{(1)}h_{(2)}u^1) + \frac{\partial}{\partial s^2}(h_{(1)}h_{(2)}u^2) + \frac{\partial}{\partial s^3}(h_{(1)}h_{(2)}u^3)\right) = \\ &= \underbrace{\frac{\partial u^1}{\partial s^1}}_{\mathcal{O}\left(\frac{V_0}{L_0}\right)} + \underbrace{\frac{\partial u^2}{\partial s^2}}_{\mathcal{O}\left(\frac{V_0}{L_0}\right)} + \underbrace{\frac{\partial u^3}{\partial s^3}}_{\mathcal{O}\left(\frac{W_0}{H_0}\right)} + \underbrace{\frac{u^1}{h_{(1)}}}_{\mathcal{O}(V_0)}\frac{\partial h_{(1)}}{\partial s^1} + \underbrace{\frac{u^1}{h_{(2)}}}_{\mathcal{O}(V_0)}\frac{\partial h_{(2)}}{\partial s^1} + \underbrace{\frac{u^2}{h_{(1)}}}_{\mathcal{O}(V_0)}\frac{\partial h_{(1)}}{\partial s^2} + \underbrace{\frac{u^2}{h_{(2)}}}_{\mathcal{O}(V_0)}\frac{\partial h_{(2)}}{\partial s^2} = 0\,. \end{aligned}$$

Multiplying by H_0, we have that:

$$\frac{H_0}{L_0}V_0 + W_0 + H_0V_0 \max\left\{\frac{\partial h_{(1)}}{\partial s^1}, \frac{\partial h_{(2)}}{\partial s^1}, \frac{\partial h_{(1)}}{\partial s^2}, \frac{\partial h_{(2)}}{\partial s^2}\right\} \sim 0\,,$$

from which we can estimate the scaling of the s^3-velocity as:

$$W_0 \sim \max\left\{\epsilon, H_0\frac{\partial h_{(1)}}{\partial s^1}, H_0\frac{\partial h_{(2)}}{\partial s^1}, H_0\frac{\partial h_{(1)}}{\partial s^2}, H_0\frac{\partial h_{(2)}}{\partial s^2}\right\} V_0 = \epsilon_{\mathcal{G}}V_0\,. \tag{3.12}$$

This defines a "geometric" aspect ratio $\epsilon_{\mathcal{G}}$ given by::

$$\epsilon_{\mathcal{G}} := \max\left\{\epsilon, H_0\frac{\partial h_{(1)}}{\partial s^1}, H_0\frac{\partial h_{(2)}}{\partial s^1}, H_0\frac{\partial h_{(1)}}{\partial s^2}, H_0\frac{\partial h_{(2)}}{\partial s^2}\right\}\,, \tag{3.13}$$

that connects local curvatures information to the global length scale parameter ϵ. Hence, the SW approximation can be restated by the assumption $\epsilon_{\mathcal{G}} \ll 1$, which effectively adds a restriction on the shape of the bottom surface that ensures that the derivatives of the metric coefficients are of the order of $1/L_0$.

Formal order of approximation and the SW hypothesis

Starting from the integrated Navier-Stokes equations in the curvilinear coordinate system shown in eq. (3.11), we proceed by applying the SW hypothesis. Thus we postulate that the fluid motion is much faster in the tangential directions than in the normal direction, i.e., $u^3 = \epsilon_{\mathcal{G}} u^i$, $i = 1,2$, $\epsilon_{\mathcal{G}} \ll 1$, and we expand the components of the velocity vector in powers of the parameter $\epsilon_{\mathcal{G}}$:

$$u^i = u^i_{(0)} + \epsilon_{\mathcal{G}} u^i_{(1)} + \epsilon^2_{\mathcal{G}} u^i_{(2)} + \mathcal{O}\Big(\epsilon^3_{\mathcal{G}}\Big) \qquad i = 1,2 \ , \tag{3.14a}$$

$$u^3 = \epsilon_{\mathcal{G}} u^3_{(1)} + \epsilon^2_{\mathcal{G}} u^3_{(2)} + \mathcal{O}\Big(\epsilon^3_{\mathcal{G}}\Big) \ . \tag{3.14b}$$

We also expand the general stress tensor components using the same parameter $\epsilon_{\mathcal{G}}$ to obtain:

$$\tau^{ij} = \tau^{ij}_{(0)} + \epsilon_{\mathcal{G}} \tau^{ij}_{(1)} + \epsilon^2_{\mathcal{G}} \tau^{ij}_{(2)} + \mathcal{O}\Big(\epsilon^3_{\mathcal{G}}\Big) \qquad i,j = 1,2,3. \tag{3.15}$$

Again, we assume that the terms $\tau^{i3} = \tau^{3i}$, $i = 1,2,3$, which contain u^3, can be expanded as:

$$\tau^{3i} = \epsilon_{\mathcal{G}} \tau^{3i}_{(1)} + \epsilon^2_{\mathcal{G}} \tau^{3i}_{(2)} + \mathcal{O}\Big(\epsilon^3_{\mathcal{G}}\Big) \quad \text{for } i = 1,2, \quad \text{and} \quad \tau^{33} = \epsilon^2_{\mathcal{G}} \tau^{33}_{(2)} + \mathcal{O}\Big(\epsilon^3_{\mathcal{G}}\Big) \ .$$

We split the velocity vector and stress tensor as the sum of the corresponding average values $\vec{U}$ and $\mathbf{T}$ and fluctuations $\tilde{u}$ and $\tilde{\tau}$ around the mean. Thus we write:

$$\vec{u} = \vec{U} + \tilde{u}, \ \text{where } \vec{U}(s^1, s^2, t) = \frac{1}{\eta}\int_0^\eta \vec{u}(\mathbf{s}, t) \ \mathrm{d}s^3, \quad \int_0^\eta \tilde{u}(\mathbf{s}, t) \ \mathrm{d}s^3 = 0 \ , \tag{3.16}$$

$$\mathbb{T} = \mathbf{T} + \tilde{\tau}, \ \text{where } \mathbf{T}(s^1, s^2, t) = \frac{1}{\eta}\int_0^\eta \mathbb{T}(\mathbf{s}, t) \ \mathrm{d}s^3, \quad \int_0^\eta \tilde{\tau}(\mathbf{s}, t) \ \mathrm{d}s^3 = 0 \ . \tag{3.17}$$

Expansions in powers of $\epsilon_{\mathcal{G}}$ are readily written for all these quantities. Rearranging terms, we can write the velocity vector as its average value times a perturbation of the identity $\mathbb{1} = [1,1,1]^T$:

$$\vec{u} = \begin{bmatrix} U^1 + \tilde{u}^1 \\ U^2 + \tilde{u}^2 \\ U^3 + \tilde{u}^3 \end{bmatrix} = \begin{bmatrix} U^1(1 + \tilde{u}^1/U^1) \\ U^2(1 + \tilde{u}^2/U^2) \\ U^3(1 + \tilde{u}^3/U^2) \end{bmatrix} = \vec{U} \circ (\mathbb{1} + \tilde{\mathbb{u}}) \ ,$$

where the symbol $\circ$ denotes the component-wise product (Hadamard product) and

$$\mathbb{1} + \tilde{\mathbb{u}} = \begin{bmatrix} 1 + \tilde{\mathbb{u}}^1 \\ 1 + \tilde{\mathbb{u}}^2 \\ 1 + \tilde{\mathbb{u}}^3 \end{bmatrix} = \begin{bmatrix} 1 + \tilde{u}^1/U^1 \\ 1 + \tilde{u}^2/U^2 \\ 1 + \tilde{u}^3/U^3 \end{bmatrix} \ .$$

Using eqs. (3.16) and (3.17), depth integration of $\vec{u}$ and $\mathbb{T}$ yields:

$$\int_0^\eta \vec{u} = \vec{U} \circ \int_0^\eta (\mathbb{1} + \tilde{\mathbb{u}}) = \eta \vec{U} \ , \qquad \int_0^\eta \mathbb{T} = \int_0^\eta (\mathbf{T} + \tilde{\tau}) = \eta \mathbf{T} \ .$$

The averaged tensor continues to display velocity fluctuations and can be expressed as:

$$\int_0^\eta \vec{u} \otimes \vec{u} = \int_0^\eta \begin{bmatrix} \left(U^1(1 + \tilde{\mathbb{u}}^1)\right)^2 & U^1 U^2 (1 + \tilde{\mathbb{u}}^1)(1 + \tilde{\mathbb{u}}^2) \\ U^1 U^2 (1 + \tilde{\mathbb{u}}^1)(1 + \tilde{\mathbb{u}}^2) & \left(U^2(1 + \tilde{\mathbb{u}}^2)\right)^2 \end{bmatrix} =$$

$$= \vec{U} \otimes \vec{U} \circ \int_0^\eta (\mathbb{1} + \tilde{\mathbb{u}}) \otimes (\mathbb{1} + \tilde{\mathbb{u}}) = \vec{U} \otimes \vec{U} \circ \int_0^\eta \mathbb{I} + \tilde{\mathbb{u}} \otimes \tilde{\mathbb{u}} = \eta \vec{U} \otimes \vec{U} \circ \mathbb{A} \ ,$$

where the "reduced fluctuation" tensor associated with $\vec{u}$ is defined as:

$$\mathbb{A} = \frac{1}{\eta}\int_0^\eta \mathbb{I} + \tilde{\mathbb{u}} \otimes \tilde{\mathbb{u}} \ . \tag{3.18}$$

The intrinsic shallow water equations. Using the formal expansions in powers of $\epsilon_{\mathcal{G}}$ in the normally integrated NS system eq. (3.11) we obtain our reduced formulation, which we name Intrinsic Shallow Water Equations (ISWE), as given in the next theorem. We use the following notation. The couple (s^1, s^2) indicates the curvilinear coordinate system associated with the LCS defined in eq. (3.2) with the ensuing metric tensor $\mathcal{G}_{sw}$ given by the principal 2-minor of eq. (3.3). The vector $\vec{q} = [\eta U^1, \eta U^2]^{\mathrm{T}}$ denotes the depth-averaged velocity vector, while the tensors

$$\mathbb{A}_{sw} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{T}_{sw} = \eta \begin{bmatrix} \mathbf{T}^{11} & \mathbf{T}^{12} \\ \mathbf{T}^{21} & \mathbf{T}^{22} \end{bmatrix}$$

are the principal 2-minors of $\mathbb{A}$ and $\mathbf{T}$, respectively. Vector $\mathbf{f}_{\mathcal{B}} = [\tau_b^1, \tau_b^2]^T$ is the vector field accounting for bed friction. The we can state the following theorem.

Theorem 3.2.1. *The intrinsic shallow water equations, written with respect to the LCS (eqs.* (3.2) *and* (3.3)*), are given by:*

$$\frac{\partial \eta}{\partial t} + \nabla_{\mathcal{G}} \cdot \vec{q} = 0 \ , \tag{3.19a}$$

$$\begin{aligned} &\frac{\partial \vec{q}}{\partial t} + \nabla_{\mathcal{G}} \cdot \left(\frac{1}{\eta} (\vec{q} \otimes \vec{q}) \circ \mathbb{A}_{sw} + \left(\frac{g\eta^2}{2} \frac{\partial x^3}{\partial s^3} \right) \mathcal{G}_{sw}^{-1} \right) \\ &\qquad + \frac{g\eta^2}{2} \nabla_{\mathcal{G}} \left(\frac{\partial x^3}{\partial s^3} \right) + g\eta \, \nabla_{\mathcal{G}}(x^3) - \frac{1}{\rho} \left(\nabla_{\mathcal{G}} \cdot \mathbf{T}_{sw} + \mathbf{f}_{\mathcal{B}} \right) = 0 \ . \end{aligned} \tag{3.19b}$$

They provide an approximation of order $\mathcal{O}(\epsilon_{\mathcal{G}}^2)$ of the Navier-Stokes equations, under the assumption of thin fluid layer, $\eta = \mathcal{O}(\epsilon_{\mathcal{G}})$.

Proof. We start by re-writing eq. (3.11b) component-wise. Using the operators defined in eqs. (3.4) to (3.6) and recalling that the terms $\partial h_{(1)}/\partial s^3, \partial h_{(2)}/\partial s^3$ vanish, the third equation, describing momentum conservation along s^3, takes on the form:

$$\begin{aligned} &\frac{\partial}{\partial t}\int_0^\eta u^3 + \frac{1}{h_{(1)}h_{(2)}} \left(\frac{\partial}{\partial s^1} \int_0^\eta h_{(1)}h_{(2)} u^1 u^3 + \frac{\partial}{\partial s^2} \int_0^\eta h_{(1)}h_{(2)} u^2 u^3 + \frac{\partial}{\partial s^3} \int_0^\eta h_{(1)}h_{(2)} (u^3)^2 \right) = \\ &\quad = -\frac{1}{\rho}\int_0^\eta \frac{\partial p}{\partial s^3} - g\int_0^\eta \frac{\partial x^3}{\partial s^3} + \frac{1}{\rho h_{(1)}h_{(2)}} \left(\frac{\partial}{\partial s^1} \int_0^\eta h_{(1)}h_{(2)} \tau^{31} + \frac{\partial}{\partial s^2} \int_0^\eta h_{(1)}h_{(2)} \tau^{32} \right. \\ &\qquad\qquad \left. + \frac{\partial}{\partial s^3} \int_0^\eta h_{(1)}h_{(2)} \tau^{33} \right) + \frac{p_{\mathcal{B}}}{\rho} \ , \end{aligned}$$

where $p_{\mathcal{B}}$ is the bed pressure as given in eq. (3.10). Introducing the expanded velocity, eq. (3.14), and tensor components, eq. (3.15), we obtain

$$\begin{aligned}
&\frac{1}{\rho}\int_0^\eta \frac{\partial p}{\partial s^3} + g\int_0^\eta \frac{\partial x^3}{\partial s^3} \\
&+ \epsilon_{\mathcal{G}}\left[\frac{\partial}{\partial t}\left(\eta U^3_{(1)}\right) + \frac{1}{h_{(1)}h_{(2)}}\left(\frac{\partial}{\partial s^1}\left(\eta U^1_{(0)}U^3_{(1)}\alpha_{13}h_{(1)}h_{(2)}\right) + \frac{\partial}{\partial s^2}\left(\eta U^2_{(0)}U^3_{(1)}\alpha_{23}h_{(1)}h_{(2)}\right)\right)\right. \\
&\left. - \frac{1}{\rho h_{(1)}h_{(2)}}\left(\frac{\partial}{\partial s^1}\int_0^\eta \tau^{31}_{(1)}h_{(1)}h_{(2)} + \frac{\partial}{\partial s^2}\int_0^\eta \tau^{32}_{(1)}h_{(1)}h_{(2)}\right) + \frac{p_{\mathcal{B},(1)}}{\rho}\right] + \mathcal{O}\left(\epsilon^2_{\mathcal{G}}\right) = 0\,, \qquad (3.20)
\end{aligned}$$

where the term $p_{\mathcal{B},(1)}$, the s^3-component of the shear stress $\mathbb{T}_{\mathcal{B}} \cdot \mathbf{N}_{\mathcal{B}}$, is a first order approximation (as evidenced by the (1) subscript) and thus is assumed to be proportional to $\epsilon_{\mathcal{G}}$. Looking at the zero-order terms, i.e., the terms proportional to $\epsilon^0_{\mathcal{G}}$, we can write:

$$\frac{1}{\rho}\int_0^\eta \frac{\partial p}{\partial s^3} + g\int_0^\eta \frac{\partial x^3}{\partial s^3} = \mathcal{O}(\epsilon_{\mathcal{G}})$$

that corresponds to the hydrostatic pressure condition along the normal direction to the bottom surface. Since we neglect the effects of surface tension and wind on the free surface, we can set $p\Big|_{s^3=\eta} = p_{atm} = 0$, to obtain:

$$p\Big|_0 = \rho g \eta \frac{\partial x^3}{\partial s^3} + \mathcal{O}(\epsilon_{\mathcal{G}})\ . \qquad (3.21)$$

Recall that the term $\frac{\partial x^3}{\partial s^3}$ does not depend on s^3, since we are assuming the direction s^3 to be rectilinear. This condition states, up to terms of order $\epsilon_{\mathcal{G}}$, that the fluid pressure varies linearly along the s^3 direction. Substitution of this algebraic expression for the pressure ensures the actual reduction of the system of PDEs from four to three equations, as in the classical SWE derivation.

We turn now our attention to the other two components of eq. (3.11b). Focusing only on the s^1-equation, the other being analogous, we can write:

$$\begin{aligned}
&\frac{\partial}{\partial t}\int_0^\eta u^1 + \frac{1}{h_{(1)}h_{(2)}}\left[\frac{\partial}{\partial s^1}\int_0^\eta h_{(1)}h_{(2)}(u^1)^2 + \frac{\partial}{\partial s^2}\int_0^\eta h_{(1)}h_{(2)}u^1u^2 + \frac{\partial}{\partial s^3}\int_0^\eta h_{(1)}h_{(2)}u^3u^1\right] \\
&\qquad + \int_0^\eta \frac{(u^1)^2}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^1} + 2\int_0^\eta \frac{u^1u^2}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^2} - \int_0^\eta (u^2)^2\frac{h_{(2)}}{h^2_{(1)}}\frac{\partial h_{(2)}}{\partial s^1} = \\
&= -\frac{1}{\rho}\int_0^\eta \frac{1}{h^2_{(1)}}\frac{\partial p}{\partial s^1} - g\int_0^\eta \frac{1}{h^2_{(1)}}\frac{\partial x^3}{\partial s^1} + \frac{1}{\rho h_{(1)}h_{(2)}}\left[\frac{\partial}{\partial s^1}\int_0^\eta h_{(1)}h_{(2)}\tau^{11} + \frac{\partial}{\partial s^2}\int_0^\eta h_{(1)}h_{(2)}\tau^{12}\right. \\
&\left. + \frac{\partial}{\partial s^3}\int_0^\eta h_{(1)}h_{(2)}\tau^{13}\right] + \frac{1}{\rho}\left(\int_0^\eta \frac{\tau^{11}}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^1} + 2\int_0^\eta \frac{\tau^{21}}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^2} - \int_0^\eta \tau^{22}\frac{h_{(2)}}{h^2_{(1)}}\frac{\partial h_{(2)}}{\partial s^1}\right) + \frac{\tau^1_b}{\rho}\ .
\end{aligned}$$

Again, we enforce an approximation of order $\epsilon_{\mathcal{G}}$ and recalling the expansions in eqs. (3.14) and (3.15), we obtain:

- for the left-hand-side

$$\frac{\partial \eta U^1_{(0)}}{\partial t} + \frac{1}{h_{(1)}h_{(2)}}\frac{\partial}{\partial s^1}\left(\eta\alpha_{11}(U^1_{(0)})^2 h_{(1)}h_{(2)}\right) + \frac{1}{h_{(1)}h_{(2)}}\frac{\partial}{\partial s^2}\left(\eta\alpha_{12}U^1_{(0)}U^2_{(0)}h_{(1)}h_{(2)}\right)$$
$$+\eta\alpha_{11}(U^1_{(0)})^2\frac{1}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^1} + 2\eta\alpha_{12}U^1_{(0)}U^2_{(0)}\frac{1}{h_{(1)}}\frac{\partial h_{(1)}}{\partial s^2} - \eta\alpha_{22}(U^2_{(0)})^2\frac{h_{(2)}}{h^2_{(1)}}\frac{\partial h_{(2)}}{\partial s^1} + \mathcal{O}(\epsilon_{\mathcal{G}}) \ ;$$

- for the right-hand-side

$$-\frac{\eta}{\rho h^2_{(1)}}\frac{\partial p}{\partial s^1} - \frac{\eta g}{h^2_{(1)}}\frac{\partial x^3}{\partial s^1} + \frac{1}{\rho h_{(1)}h_{(2)}}\left[\frac{\partial}{\partial s^1}\int_0^\eta \tau^{11}_{(0)}h_{(1)}h_{(2)} + \frac{\partial}{\partial s^2}\int_0^\eta \tau^{12}_{(0)}h_{(1)}h_{(2)}\right]$$
$$+\frac{1}{\rho h_{(1)}}\frac{\partial h_{(1)}}{\partial s^1}\int_0^\eta \tau^{11}_{(0)} + \frac{2}{\rho h_{(1)}}\frac{\partial h_{(1)}}{\partial s^2}\int_0^\eta \tau^{12}_{(0)} - \frac{h_{(2)}}{\rho h^2_{(1)}}\frac{\partial h_{(2)}}{\partial s^1}\int_0^\eta \tau^{22}_{(0)} + \frac{\tau^1_{b\,(0)}}{\rho} + \mathcal{O}(\epsilon_{\mathcal{G}}) \ .$$

Inserting the expression for the pressure given in eq. (3.21) yields:

$$-\frac{\eta}{h^2_{(1)}}\frac{\partial}{\partial s^1}\left(\eta g\frac{\partial x^3}{\partial s^3}\right) - \frac{\eta g}{h^2_{(1)}}\frac{\partial x^3}{\partial s^1} + \frac{1}{\rho}\left(\nabla_{\mathcal{G}}\cdot\mathbf{T}_{sw,(0)}\right)^1 + \frac{\tau^1_{b\,(0)}}{\rho} + \mathcal{O}(\epsilon_{\mathcal{G}}) \ .$$

The final divergence form of the model is obtained by applying the chain rule on the first term of the previous equation. This implies the assumption that the depth function η and $\frac{\partial x^3}{\partial s^3}$ are differentiable functions. We obtain:

$$\frac{\eta}{h^2_{(1)}}\frac{\partial}{\partial s^1}\left(\eta g\frac{\partial x^3}{\partial s^3}\right) = \frac{g}{h^2_{(1)}}\frac{\partial}{\partial s^1}\left(\frac{\eta^2}{2}\frac{\partial x^3}{\partial s^3}\right) + \frac{g\eta^2}{2h^2_{(1)}}\frac{\partial}{\partial s^1}\left(\frac{\partial x^3}{\partial s^3}\right) \ .$$

Recalling the definition of $\vec{q}$, the momentum equation can be re-written in compact form intrinsic to the bottom surface as:

$$\frac{\partial\vec{q}}{\partial t} + \nabla_{\mathcal{G}}\cdot\left(\frac{1}{\eta}(\vec{q}\otimes\vec{q})\circ\mathbb{A}_{sw} + \left(\frac{g\eta^2}{2}\frac{\partial x^3}{\partial s^3}\right)\mathcal{G}^{-1}_{sw}\right)$$
$$+\frac{g\eta^2}{2}\nabla_{\mathcal{G}}\left(\frac{\partial x^3}{\partial s^3}\right) + g\eta\,\nabla_{\mathcal{G}}(x^3) - \frac{1}{\rho}\nabla_{\mathcal{G}}\cdot\mathbf{T}_{sw} - \frac{\mathbf{f}_{\mathcal{B}}}{\rho} + \mathcal{O}(\epsilon_{\mathcal{G}}) = 0 \ .$$

Note that the higher order terms have the same form and are all proportional to η. Hence, if we add the further assumption of $\eta = \mathcal{O}(\epsilon_{\mathcal{G}})$ the final form of the momentum equation is an approximation of order $\mathcal{O}(\epsilon^2_{\mathcal{G}})$. Finally, note that the continuity equation eq. (3.11a) is exact and does not alter the order of approximation. □

3.2.1 Properties of the model

As already mentioned in the introduction, the ISWE model is similar to the model proposed by Bouchut and Westdickenberg [16]. The most important improvement of our new formulation eq. (3.19) is that only geometric terms appear in the source term and that the full divergence form is maintained for the conserved quantities. In addition to securing a more compact form of the equations, our formulation allows the development of a fully intrinsic discretization as will be seen in chapter 4. Here we prove some fundamental mathematical properties, namely that the proposed model is invariant under rotation, it admits an energy equation, and is well-balanced (preserves the steady state of lake-at-rest).

Proposition 3.2.2. *The ISWE system defined in eq.* (3.19) *is invariant under rotation, it admits a conserved energy in the absence of stresses, and is well-balanced.*

Proof. The first statement is proved by showing that all involved operators are rotation invariant. Consider first the gradient operator $\nabla_{\mathcal{G}}$. For any function $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$, and for any rotation operator $R \in \mathbb{R}^{d\times d}$, the following relation holds:

$$\nabla_{\mathcal{G}}\, \tilde{f}(x) = R^{\mathrm{T}}\, \nabla^{R}_{\mathcal{G}}\, f(\tilde{x}) \bigg|_{\tilde{x}=Rx} ,$$

where $\tilde{f}(x) = f(Rx)$, and $\nabla^{R}_{\mathcal{G}}$ is the gradient operator in the rotated coordinates. This follows immediately from the application of the chain rule formula. Note that in our case $d = 2$, since we are considering points lying on the bottom surface. Concerning the divergence operator, we consider first its application to a vector-valued function $u : \Omega \to \mathbb{R}^d$. We have:

$$\nabla_{\mathcal{G}} \cdot \tilde{u} = \nabla^{R}_{\mathcal{G}} \cdot u \bigg|_{\tilde{x}=Rx} ,$$

where $R \in \mathbb{R}^{d\times d}$ is a rotation matrix, $\tilde{u} = R^{\mathrm{T}}u$ is the rotated vector, and $\nabla^{R}_{\mathcal{G}}\cdot$ is the divergence in the rotated coordinate system. With the same notation, the following holds for the divergence of a tensor product $u \otimes u$:

$$\begin{aligned} \nabla^{R}_{\mathcal{G}} \cdot (\tilde{u} \otimes \tilde{u}) &= \tilde{u} \cdot \nabla^{R}_{\mathcal{G}}\, \tilde{u} + \tilde{u}\, \nabla^{R}_{\mathcal{G}} \cdot \tilde{u} = (R^T u \cdot R^T\, \nabla_{\mathcal{G}}) R^T u + R^T u\, \nabla_{\mathcal{G}} \cdot u = \\ &= R^T\, (u \cdot \nabla_{\mathcal{G}}\, u + u\, \nabla_{\mathcal{G}} \cdot u) = R^T\, \nabla_{\mathcal{G}} \cdot (u \otimes u)\ . \end{aligned}$$

The rotational invariance of eq. (3.19) follows by noting that the above relation can be directly extended also to a diagonal tensor.

The energy expression for the system in eq. (3.19) can be derived by setting $\mathbf{T}_{sw} = 0$ and $\mathbf{f}_{\mathcal{B}} = 0$ (zero stress) in eq. (3.19b). Scalar multiplication by the conservative velocity

$\vec{U}$ gives:

$$\underbrace{\left\langle \vec{U}, \frac{\partial \eta \vec{U}}{\partial t} \right\rangle_{\mathcal{G}}}_{[A]} + \underbrace{\left\langle \vec{U}, \nabla_{\mathcal{G}} \cdot (\eta \vec{U} \otimes \vec{U}) \right\rangle_{\mathcal{G}}}_{[B]} + \underbrace{\left\langle \vec{U}, \nabla_{\mathcal{G}} \left(\frac{1}{2} g \eta^2 \frac{\partial x^3}{\partial s^3} \right) \right\rangle_{\mathcal{G}}}_{[C]}$$

$$+ \frac{1}{2} g \eta^2 \underbrace{\left\langle \vec{U}, \nabla_{\mathcal{G}} \left(\frac{\partial x^3}{\partial s^3} \right) \right\rangle_{\mathcal{G}}}_{[D]} + g\eta \underbrace{\left\langle \vec{U}, \nabla_{\mathcal{G}} \, x^3 \right\rangle_{\mathcal{G}}}_{[E]} = 0 \,.$$

We proceed now by applying the chain rule to the different terms. We obtain for the term $[A]$:

$$[A] = g_{mj} U^m \partial_t \left(\eta U^j \right) = g_{mj} U^m U^j \partial_t \, \eta + \eta g_{mj} U^m \partial_t \, U^j = \left| \vec{U} \right|^2_{\mathcal{G}} \partial_t \, \eta + \eta \partial_t \left(\frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) =$$

$$= \left| \vec{U} \right|^2_{\mathcal{G}} \partial_t \, \eta + \partial_t \left(\eta \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) - \frac{\left| \vec{U} \right|^2_{\mathcal{G}}}{2} \partial_t \, \eta = \partial_t \left(\eta \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) + \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \partial_t \, \eta \,,$$

and for term $[B]$:

$$[B] = g_{mj} U^m \left[\frac{\partial}{\partial s^i} \left(\eta U^i U^j \right) + \Gamma^i_{ik} \eta U^k U^j + \Gamma^j_{ik} \eta U^i U^k \right] =$$

$$= \eta U^i \underbrace{g_{mj} U^m \frac{\partial U^i}{\partial s^i}}_{[*]} + \underbrace{\eta \left(g_{mj} U^m U^j \right) \frac{\partial U^i}{\partial s^i} + \eta g_{mj} U^m U^j \Gamma^i_{ik} U^k}_{\text{“}\eta \left|\vec{U}\right|^2_{\mathcal{G}} \nabla_{\mathcal{G}} \vec{U}\text{''}}$$

$$+ \Gamma^j_{ik} U^i U^k g_{mj} U^m \eta + \underbrace{g_{mj} U^m U^j \eta U^i \frac{\partial \eta}{\partial s^i}}_{\text{“}\vec{U} \cdot \vec{U} \otimes \vec{U} \, \nabla_{\mathcal{G}} \, \eta\text{''}} =$$

$$= \eta U^i \underbrace{\left[\frac{\partial}{\partial s^i} \left(\frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) - \tfrac{1}{2} \Gamma^i_{im} g_{kj} U^m U^j - \tfrac{1}{2} \Gamma^k_{ij} g_{mj} U^m U^j \right]}_{[*]} + \Gamma^j_{ik} U^i U^k g_{mj} U^m \eta$$

$$+ \eta \left| \vec{U} \right|^2_{\mathcal{G}} \nabla_{\mathcal{G}} \, \vec{U} + \vec{U} \cdot \vec{U} \otimes \vec{U} \, \nabla_{\mathcal{G}} \, \eta =$$

$$= \frac{\partial}{\partial s^i} \left(\eta U^i \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) - \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \frac{\partial}{\partial s^i} \left(\eta U^i \right) + \eta \left| \vec{U} \right|^2_{\mathcal{G}} \nabla_{\mathcal{G}} \, \vec{U} + \vec{U} \cdot \vec{U} \otimes \vec{U} \, \nabla_{\mathcal{G}} \, \eta =$$

$$= \frac{\partial}{\partial s^i} \left(\eta U^i \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \right) + \tfrac{1}{2} \eta g_{mj} U^m U^j \Gamma^i_{ik} U^k - \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \frac{\partial}{\partial s^i} \left(\eta U^i \right)$$

$$+ \left| \vec{U} \right|^2_{\mathcal{G}} \frac{\partial}{\partial s^i} \left(\eta U^i \right) + \tfrac{1}{2} \eta g_{mj} U^m U^j \Gamma^i_{ik} U^k - \left| \vec{U} \right|^2_{\mathcal{G}} U^i \frac{\partial \eta}{\partial s^i} + g_{mj} U^m U^j \eta U^i \frac{\partial \eta}{\partial s^i} =$$

$$= \nabla_{\mathcal{G}} \cdot \left(\eta \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \vec{U} \right) + \frac{\left|\vec{U}\right|^2_{\mathcal{G}}}{2} \nabla_{\mathcal{G}} \cdot \left(\eta \vec{U} \right) \,.$$

The analogous computations for the other terms are not reported here because they are similar. Finally, putting everything together and collecting common terms exploiting the relations in eqs. (3.4) to (3.6), the energy equation for the system becomes:

$$\frac{\partial}{\partial t}\left(\eta\frac{\left|\vec{U}\right|_{\mathcal{G}}^2}{2}+\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}+gx^3_{\mathcal{B}}\eta\right) \\ +\nabla_{\mathcal{G}}\cdot\left[\left(\eta\frac{\left|\vec{U}\right|_{\mathcal{G}}^2}{2}+\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}+gx^3_{\mathcal{B}}\eta+\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}\right)\vec{U}\right]=0\ ,$$

where $x^3_{\mathcal{B}}$ is the elevation of the bottom with respect to the the GCS, and

$$\mathcal{E}=\eta\frac{\left|\vec{U}\right|_{\mathcal{G}}^2}{2}+\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}+gx^3_{\mathcal{B}}\eta\ ,\qquad p=\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}$$

are the kinetic energy and the pressure term, respectively.

Next, we show that the system is well-balanced, in the sense that it preserves the steady-state of a lake-at-rest. Let the functions $\eta=\eta(s^1,s^2)$ and $\vec{q}=\vec{q}(s^1,s^2)$ be time independent. Then, the model results in a zero velocity field and thus the steady-state for a horizontal free fluid surface (lake-at-rest) is preserved. Explicitly, from the mass conservation equation we get $\vec{U}=0$, and in the momentum equation the remaining terms are:

$$\nabla_{\mathcal{G}}\left(\frac{1}{2}g\eta^2\frac{\partial x^3}{\partial s^3}\right)+\frac{1}{2}g\eta^2\,\nabla_{\mathcal{G}}\left(\frac{\partial x^3}{\partial s^3}\right)+g\eta\,\nabla_{\mathcal{G}}\,x^3=0\,.$$

Again, after the application of the chain rule, we obtain the following relation for the depth of the fluid:

$$\eta\frac{\partial x^3}{\partial s^3}+x^3_{\mathcal{B}}=\text{const}\ . \tag{3.22}$$

□

3.3 Balance law formulation of ISWE

In this final section of the chapter we want to write the ISWE system in a compact form, which highlights its formulation as balance law and is useful to the development of the numerical discretizations addressed in the following chapters. System (3.19) can be written in divergence form as the balance law:

$$\frac{\partial\mathbf{U}}{\partial t}+\operatorname{div}_{\mathcal{G}}\underline{\underline{F}}(\mathbf{s},\mathbf{U})+\mathbf{S}(\mathbf{s},\mathbf{U})=0\ . \tag{3.23}$$

Here the conservative variable is given by $\mathbf{U}=[\eta,\eta U^1,\eta U^2]^{\mathrm{T}}=[\eta,q^1,q^2]^{\mathrm{T}}$, where $\eta:\Gamma\times[0,t_f]\to\mathbb{R}$, and $\mathbf{q}=[q^1,q^2]$, $\mathbf{q}:\Gamma\times[0,t_f]\to\mathbb{R}^2$. From now on we make the assumption that $\alpha_{ij} = 1$ in $\mathbb{A}_{sw}$, as commonly done in practical applications. The flux function $\underline{\underline{F}}$ takes the form

$$\underline{\underline{F}}(\mathbf{s}, \mathbf{U}) = \begin{bmatrix} q^1 & q^2 \\ \dfrac{(q^1)^2}{\eta} + \dfrac{g\eta^2}{2h_{(1)}^2} \dfrac{\partial x^3}{\partial s^3} & \dfrac{q^1 q^2}{\eta} \\ \dfrac{q^1 q^2}{\eta} & \dfrac{(q^2)^2}{\eta} + \dfrac{g\eta^2}{2h_{(2)}^2} \dfrac{\partial x^3}{\partial s^3} \end{bmatrix} = \begin{bmatrix} \underline{F^\eta} \\ \\ \underline{\underline{F^{\mathbf{q}}}} \end{bmatrix} . \tag{3.24}$$

Note that the flux $\underline{\underline{F}}$ is a function of $\mathbf{s}$ because of the appearance of the components $h_{(i)}$ of the metric tensor $\mathcal{G}_{sw}$ and the presence of the bottom slope $\partial x^3/\partial s^3$. The symbol $\mathrm{div}_{\mathcal{G}}$ denotes the divergence operator applied to the flux function as divergence of a vector for the first row and divergence of a 2×2 tensor for the last two rows. We can define it as $\mathrm{div}_{\mathcal{G}} = [\nabla^\eta_{\mathcal{G}} \cdot, \nabla^{\mathbf{q}}_{\mathcal{G}} \cdot]^{\mathrm{T}}$. The source function $\mathbf{S}$ comprises the metric tensor coefficients, the bottom slope and its derivatives, the two-dimensional averaged stress tensor $\mathbf{T}_{sw}$, the bottom friction parameter τ_b, and the conserved variable η. We summarize this dependency by explicitly writing it out in $\mathbf{S}(\mathbf{s}, \eta)$. We have then:

$$\mathbf{S}(\mathbf{s}, \eta) = \begin{bmatrix} 0 \\ \dfrac{g\eta^2}{2h_{(1)}^2} \dfrac{\partial}{\partial s^1} \left(\dfrac{\partial x^3}{\partial s^3} \right) + \dfrac{g\eta}{h_{(1)}^2} \dfrac{\partial x^3}{\partial s^1} - \dfrac{1}{\rho} \left[\nabla_{\mathcal{G}} \cdot \mathbf{T}_{sw} \right]^{(1,\cdot)} - \dfrac{\tau_b^1}{\rho} \\ \dfrac{g\eta^2}{2h_{(2)}^2} \dfrac{\partial}{\partial s^2} \left(\dfrac{\partial x^3}{\partial s^3} \right) + \dfrac{g\eta}{h_{(2)}^2} \dfrac{\partial x^3}{\partial s^2} - \dfrac{1}{\rho} \left[\nabla_{\mathcal{G}} \cdot \mathbf{T}_{sw} \right]^{(2,\cdot)} - \dfrac{\tau_b^2}{\rho} \end{bmatrix} = \begin{bmatrix} S^\eta \\ \\ \mathbf{S}^{\mathbf{q}} \end{bmatrix} . \tag{3.25}$$

The regularity assumption on the bottom surface implies the uniform continuity of the flux and source functions with respect to $\mathbf{s}$.

4

Intrinsic Finite Volume Scheme

One of the standard approaches in dealing with shallow water flows on variable topography is to supplement the system of balance laws with an equation for the fixed geometry, and this leads to a system in non-conservative form [54, 35, 18, 19]. Here we are able to develop a first order upwind Godunov-type Finite Volume scheme defined intrinsically on the LCS coordinate system that solves directly eq. (3.23). We consider a low order discretization scheme because we want to highlight the importance of the correct geometrical formulation of the ISWE system rather than focus on their numerical discretization. Thus, we adopt a simple but robust solver, favoring resiliency over accuracy. The aim is to derive intrinsic definitions of the geometric differential operators.

We assume that the final system (3.19) is defined on a compact subset of the bottom surface, $\Gamma \subset \mathcal{S}_{\mathcal{B}}$, and that a well-defined curvilinear boundary, denoted by $\partial\Gamma = \partial\overline{\Gamma}$, exists. We start our work on $\mathcal{R}(\Gamma)$, a partition of the region Γ, where the divergence and integration by parts theorems are naturally defined. Following a standard development workflow for FV methods, we test eq. (3.23) with a piece-wise constant (in space and time) function and apply the divergence theorem to obtain the following set of equations valid for all regions $R_i \in \mathcal{R}(\Gamma)$ and for $t \in [t^k, t^{k+1}]$:

$$\mathbf{U}_i^{k+1} = \mathbf{U}_i^k - \frac{1}{\mathcal{A}_{R_i}} \sum_{j=1}^{N_{\sigma(i)}} \ell_{ij} \int_{t^k}^{t^{k+1}} \mathbf{F}_{ij}(\mathbf{U}) \ \mathrm{d}t - \int_{t^k}^{t^{k+1}} \mathbf{S}_i(\eta) \ \mathrm{d}t \ , \tag{4.1}$$

where we use the cell-averaged and edge-averaged quantities defined intrinsically in $\mathcal{R}(\Gamma)$ as:

$$\mathbf{U}_i = \frac{1}{\mathcal{A}_{R_i}} \int_{R_i} \mathbf{U} \ \mathrm{d}\mathbf{s} \ , \qquad \mathbf{F}_{ij} = \frac{1}{\ell_{ij}} \int_{\sigma_{ij}} \left\langle \underline{\underline{F}}, \nu_{ij} \right\rangle_{\mathcal{G}} \ \mathrm{d}\sigma \ , \qquad \mathbf{S}_i = \frac{1}{\mathcal{A}_{R_i}} \int_{R_i} \mathbf{S} \ \mathrm{d}\mathbf{s} \ . \tag{4.2}$$

Note that the quantities $\mathbf{F}_{ij}$ and $\mathbf{S}_i$ are depending only on the unknown $\mathbf{U}$ but not on the space variable $\mathbf{s}$, since they are integrated in space. Moreover, it is important to underline that no numerical approximations are done up to this point. Now we need to devise numerically computable approximations of the quantities appearing in eq. (4.1), namely in eq. (4.2). Thus, the following steps need to be appropriately defined: i) time stepping; ii) normal fluxes on edges; iii) quadrature rules; iv) Riemann problem. Moreover, we need to properly define the partition of the domain, namely a triangulation. The scope of the next sections is to give an exhaustive definition of all these steps, starting from the ones related to the approximation of the surface, then the time discretization and the definition of the numerical fluxes.

4.1 Surface triangulation

The derivation of a numerical scheme starts from the definition of the computational mesh.

Definition 4.1.1. *Let $\Gamma \subseteq \mathcal{S}$ be a regular region ($\Gamma = \mathcal{S}$ with $\mathcal{S}$ compact is allowed). A* (regular) triangulation *of Γ is a family $\mathcal{T}(\Gamma) = \{T_i\}_i$ of surface triangles such that:*

- $\Gamma = \cup_i T_i$*;*
- *the intersection of two different triangles is either empty, or consists of vertices, or is a whole side;*
- *every vertex of the boundary of R is a vertex of (at least) one triangle of the triangulation.*

Proposition 4.1.2. *Let $\Gamma \subseteq \mathcal{S}$ be a region on a regular surface $\mathcal{S}$, and $\mathcal{U}$ an open cover of Γ. Then, there exists a triangulation $\mathcal{T}(\Gamma)$ of Γ such that for all $T \in \mathcal{T}(\Gamma)$ there exists $U \in \mathcal{U}$ with $T \subset U$.*

Thus, triangulations exist and it is possible to choose arbitrarily small triangles. In particular, we can see triangulations as particular case of partitions. We work on the compact set Γ and we define the partition $\mathcal{R}(\Gamma) = \mathcal{T}(\Gamma)$, as a surface triangulation formed by the union of non-intersecting geodesic triangles (edges are geodesics). Obviously, we have that $\mathcal{T}(\Gamma) = \cup_{i=1}^{N_T} T_i = \overline{\Gamma}$ and $\sigma_{ij} = T_i \cap T_j$ is an internal geodesic edge. We will also use the approximate triangulation $\mathcal{T}_h(\Gamma)$ defined by the piece-wise linear surface identified by the union of 2-simplices in $\mathbb{R}^3$ (flat three-dimensional triangles) with vertices coinciding with the vertices of $\mathcal{T}(\Gamma)$ and edges corresponding to secants joining two vertices. We assume that this triangulation is closely inscribed in $\mathcal{T}(\Gamma)$ in the sense of Morvan [57] (the tangent spaces of $\mathcal{T}(\Gamma)$ and of $\mathcal{T}_h(\Gamma)$ are close in some sense). Quantities belonging to the approximated triangulation $\mathcal{T}_h(\Gamma)$ will be identified with the subscript h. Thus the symbol $\sigma_{h,ij}$ will identify the common (straight) edge between triangles $T_{h,i}$ and $T_{h,j}$. Note that the mesh parameter h denotes also the length of the longest triangle edge in $\mathcal{T}_h(\Gamma)$. We denote by r_T the radius of the circle inscribed in T_h and by h_T the longest side of T_h, and write $h = \max_T h_T$ and $r = \min_T r_T$. We assume that $\mathcal{T}_h(\Gamma)$ is *shape-regular*, i.e., there exists a strictly positive constant ρ independent of h such that:

$$\frac{r_T}{h_T} \geq \rho \qquad \forall T \in \mathcal{T}_h(\Gamma)\,.$$

Consequently, the surface triangulation $\mathcal{T}(\Gamma)$ is also shape-regular if the surface is regular. We will denote by $\mathcal{A}_T$ ($\mathcal{A}_i$) the area of cell T (T_i) in $\mathcal{T}(\Gamma)$, and by $\mathcal{A}_{T_h}$ ($\mathcal{A}_{h,i}$) the corresponding area in $\mathcal{T}_h(\Gamma)$. Analogously, we will denote with ℓ_σ (ℓ_{ij}) the length of the geodesic edge σ (σ_{ij}) in $\mathcal{T}(\Gamma)$, and ℓ_{σ_h} ($\ell_{h,ij}$) the corresponding length in $\mathcal{T}_h(\Gamma)$.

Let us explain more in details the relations between the surface triangulation $\mathcal{T}(\Gamma)$ and the approximated triangulation $\mathcal{T}_h(\Gamma)$. We recall from proposition 2.3.1 the notion of tubular neighborhood $\mathcal{N}_\Gamma$, and from eq. (2.6) the relation to project points to the surface Γ, and we give the following definition:

Definition 4.1.3. *A subset $W \subset \mathbb{R}^3$ is* closely near *to Γ if it lies in $\mathcal{N}_\Gamma$ and if the restriction of the orthogonal projection* pr *to W is one-to-one.*

In particular, given a point $\mathbf{q} \in T_h$, we indicate with $\mathrm{pr}(\mathbf{q})$ the point in $T \subset \mathcal{T}(\Gamma)$ of which $\mathbf{q}$ is the *orthogonal projection* along the surface normal direction $\mathbf{N}(\mathrm{pr}(\mathbf{q}))$. Thus, the orthogonal projection map given by

$$\mathrm{pr} : \mathcal{T}_h(\Gamma) \to \mathcal{T}(\Gamma) \tag{4.3}$$

is well defined for all the points of the approximated triangulation. From the regularity of the surface it is possible to prove that the projection map is continuous. Moreover, we

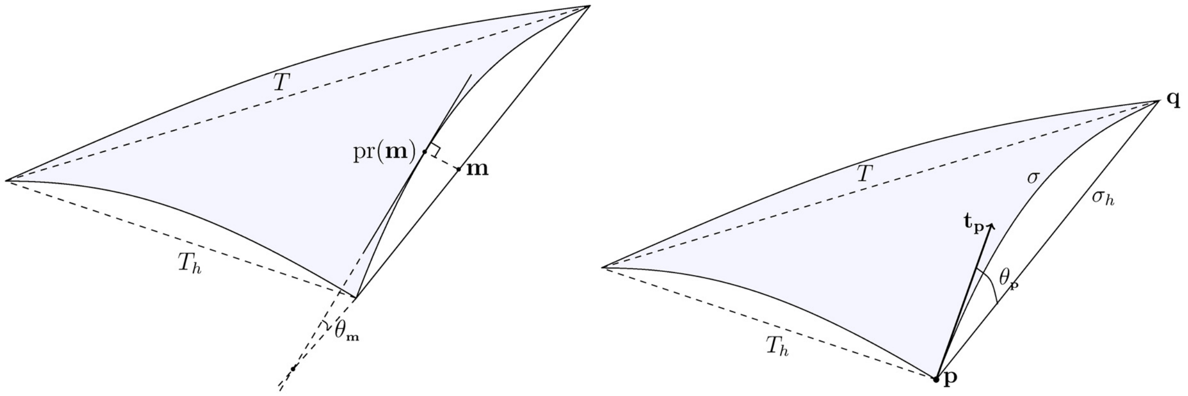

Figure 4.1: Representation of the orthogonal projection onto the surface of a point $\mathbf{m} \in T_h$ and deviation angle $\theta_{\mathbf{m}}$ (left panel). Tangent vector to the surface in $\mathbf{p}$ and deviation angle $\theta_{\mathbf{p}}$ (right panel).

say that $\mathcal{T}_h(\Gamma)$ is *closely inscribed* in $\mathcal{T}(\Gamma)$ if every $T_h \in \mathcal{T}_h(\Gamma)$ lies within a neighborhood $\mathcal{N}_{\mathrm{pr}(\mathbf{q})}$ of $\mathcal{T}(\Gamma)$ such that $\mathrm{pr}(\mathbf{q})$ is well-defined and one-to-one for all $\mathbf{q} \in T_h$.

Remark 4.1. Note that, if Γ is smooth, then the vector $\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}$ is orthogonal to the surface, in particular it is orthogonal to the tangent plane $T_{\mathrm{pr}(\mathbf{q})}\Gamma$ of Γ at the point $\mathrm{pr}(\mathbf{q})$.

In this setting, we can state the following proposition [36, lemma 4.1].

Proposition 4.1.4. *Given $\mathcal{T}(\Gamma)$, $\mathcal{T}_h(\Gamma)$ and the projection map* pr*, the following estimates hold:*

- *the distance between the approximate triangulation and the surface satisfies:*

$$\max_{\mathbf{q}\in\mathcal{T}_h(\Gamma)} \left|\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}\right| \leq Ch^2\,;$$

- *the ratio δ_h between the area measures* $\mathrm{d}\mathbf{s}$ *and* $\mathrm{d}\mathbf{x}$ *of the surface and its approximation, defined by* $\mathrm{d}\mathbf{s} = \delta_h\,\mathrm{d}\mathbf{x}$*, satisfies:*

$$\|1-\delta_h\|_{L^\infty(\mathcal{T}_h(\Gamma))} \leq Ch^2\,.$$

Now we will compare quantities of the two triangulations. In particular, the idea is to compare the tangent spaces of $\mathcal{T}_h(\Gamma)$ and $\mathcal{T}(\Gamma)$, crucial objects to find approximation errors and convergence of geometric quantities. Given $T \in \mathcal{T}(\Gamma)$ and $T_h \in \mathcal{T}_h(\Gamma)$ let $(\mathbf{q}, \mathrm{pr}(\mathbf{q})) \in T_h \times T$ be the pair of points related by the orthogonal projection $\mathbf{p} := \mathrm{pr}(\mathbf{q})$ with respect to $\mathbf{N}(\mathrm{pr}(\mathbf{q}))$. Denote by $\left\|\mathcal{W}_{\mathrm{pr}(\mathbf{q})}\right\|$ the norm of the Weingarten tensor, i.e., the maximum of the absolute values of the principal curvatures of Γ along $\mathbf{N}$ in $\mathrm{pr}(\mathbf{q})$. Then, we can define the deviation angle, the relative curvature, and the relative height of T_h with respect to T.

Definition 4.1.5. *Given a (regular) point* $\mathbf{q} \in T_h$*, let the angle* $\theta_{\mathbf{q}} \in [0, \frac{\pi}{2}]$ *be the angle between the tangent spaces* $T_{\mathbf{q}}T_h$ *and* $T_{\mathrm{pr}(\mathbf{q})}T$*. The* deviation angle θ_{max} *of* T_h *with respect to* T *is the maximum over all the points* $\mathbf{q} \in T_h$*, i.e.,* $\theta_{max} = \max_{\mathbf{q}\in T_h} \theta_{\mathbf{q}}$.

Remark 4.2. The geodesic edge σ is the projection by pr of σ_h. The deviation angle $\theta_{\mathbf{p}}$ in $\mathbf{p}$, shown in the right panel of fig. 4.1, is the angle formed by σ_h and $\mathbf{t}_{\mathbf{p}}$, the vector tangent to σ in $\mathbf{p}$.

For any point $\mathbf{q} \in T_h$, the relative curvature of T_h with respect to T in $\mathbf{q}$ is the function $\omega_\Gamma(\mathbf{q})$ formed by the product of the Euclidean distance $\left|\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}\right|$ between $\mathrm{pr}(\mathbf{q})$ and $\mathbf{q}$ times $\left\|\mathcal{W}_{\mathrm{pr}(\mathbf{q})}\right\|$, i.e., $\omega_\Gamma(\mathbf{q}) = \left|\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}\right| \left\|\mathcal{W}_{\mathrm{pr}(\mathbf{q})}\right\|$. Then, the relative curvature of a subset U of Γ is given by:

Definition 4.1.6. *Given* $U \subset \Gamma$*, closely near, its* relative curvature $\omega_\Gamma(U)$ *with respect to* Γ *is* $\omega_\Gamma(U) = \sup_{\mathbf{q}\in U} \omega_\Gamma(\mathbf{q})$.

Finally we define the relative height of T_h with respect to T as follows:

Definition 4.1.7. *The real number defined by*

$$\pi_\Gamma(T_h) = \sup_{\mathbf{q}\in T_h} h_T \left\|\mathrm{II}_{\mathrm{pr}(\mathbf{q})}\right\| , \tag{4.4}$$

where $\left\|\mathrm{II}_{\mathrm{pr}(\mathbf{q})}\right\|$ *denotes the usual norm of the second fundamental form of* Γ *at* $\mathrm{pr}(\mathbf{q})$*, is called the* relative height *of* T_h *with respect to* T.

Now we are in the position to describe some approximation results related to these geometric invariants [57]. We start with an estimate of the distance between triangles in $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$.

Proposition 4.1.8. *The Hausdorff distance* $\mathrm{dist}(T_h, T,)$ *between any cell* T_h *and its projection* T *satisfies:*

$$\mathrm{dist}(T_h, T) \leq h_T \leq 2r_T\,.$$

Moreover, $\omega_\Gamma(T_h) \leq \pi_\Gamma(T_h)$.

Next we move to the angular deviation between the surface and the approximate triangulation:

$$\sin\theta_{max} \leq \left(\frac{\sqrt{10}}{rig(T_h)(1-\omega_\Gamma(T_h))} + \frac{1}{1-\omega_\Gamma(T_h)}\right)\pi_\Gamma(T_h)\,,$$

where $rig(T_h)$ is the maximum among the cell vertices of the absolute value of the sinus of the vertex angle, and is called the *rightness* of T_h. We also need the following lemma that is part of the proof of the previous result and is a direct application of the mean value theorem:

Lemma 4.1.9. *Given two points* $\mathbf{p}, \mathbf{q} \in T$*, the angle* $\theta_{\mathbf{pq}} \in [0, \frac{\pi}{2}]$ *between the two unit vectors* $\mathbf{t_p}, \mathbf{t_q}$ *tangent to the surface at* $\mathbf{p}$ *and* $\mathbf{q}$*, respectively, satisfies:*

$$\sin(\theta_{\mathbf{pq}}) \leq |\mathrm{II}_T|\, \ell_{\mathbf{pq}}\,, \tag{4.5}$$

where $|\mathrm{II}_T|$ *denotes the supremum over* Γ *of the norm of the second fundamental form of* Γ*, and* $\ell_{\mathbf{pq}}$ *is the geodesic distance between* $\mathbf{p}$ *and* $\mathbf{q}$*.*

Using the above results, it is now easy to prove the following proposition, detailing the estimates that will be later used in the proofs of the convergence of our approximations of Γ.

Proposition 4.1.10. *The following relations hold:*

1. *the curvilinear length* ℓ_σ *of edge* σ *is related to the Euclidean length* ℓ_{σ_h} *via the inequalities:*
$$\ell_{\sigma_h} \leq \ell_\sigma \leq \frac{1}{1 - \omega_\Gamma(\sigma_h)} \ell_{\sigma_h}\,,$$
where $\omega_\Gamma(\sigma_h)$ *is the relative curvature with respect to* Γ*;*

2. *the difference between the unit vector* $\mathbf{v}_{\overrightarrow{\mathbf{pq}}}$ *aligned to the chord* σ_h *and the unit tangent vector* $\mathbf{t_p}$ *to the geodesic edge at* $\mathbf{p}$ *satisfies:*
$$\left|\mathbf{v}_{\overrightarrow{\mathbf{pq}}} - \mathbf{t_p}\right| \leq \frac{1}{2} |\mathrm{II}_T|\, \ell_\sigma\,,$$
where $|\mathrm{II}_T|$ *is the supremum over* Γ *of the norm of the second fundamental form of* Γ*;*

3. *the surface area of the cell* T *is related to the planar area of* T_h *by the relation:*
$$|\mathcal{A}_T - \mathcal{A}_{T_h}| \leq C_\Gamma \left(\theta^2_{\max} + \omega_\Gamma(T_h)\right),$$
where C_Γ *is a constant depending on* Γ*.*

Remark 4.3. For any $U \subseteq T_h$ we have that for all $\mathbf{q} \in U$ the length of the orthogonal projection $\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}$ is bounded by the square of the length of the longest triangle edge, i.e., $\left|\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}\right| \leq C\ell^2_{\sigma_h}$ [36]. From the definition of the relative curvature we find immediately $\omega_\Gamma(U) = \sup_{\mathbf{q}\in U} \left|\overrightarrow{\mathrm{pr}(\mathbf{q})\mathbf{q}}\right| \left\|\mathcal{W}_{\mathrm{pr}(\mathbf{q})}\right\| \leq C\ell^2_{\sigma_h}$, where here C is a generic constant.

Remark 4.4. Analogously, the deviation angle $\theta_{\max}$ of T_h with respect to T is bounded by the radius of the triangle circumcircle, leading to the bound $\theta_{\max} \leq C\ell_{\sigma_h}$ [57].

Remark 4.5. Note that the spectral norm of the second fundamental form of Γ at $\mathbf{p}$ is exactly the principal curvature of Γ at $\mathbf{p}$. Thus, $|\mathrm{II}_T|$ can be interpreted as the maximum principal curvature of Γ.

Intrinsic quadrature rules and sampling points

The aim is to exploit these relation in order to define numerical quadrature rules for our surface integrals consistent with the first order Godunov method we plan to use and with minimal sampling points. This will allow to solve a minimal number of Riemann problems per triangle edge. Linear consistency with a single sampling point is achieved by the midpoint formula, which can be written as:

$$\int_\sigma f(\mathbf{s})\,\mathrm{d}\sigma \approx \ell_\sigma f(\mathbf{m}_\sigma)\ , \tag{4.6}$$

$$\int_T f(\mathbf{s})\,\mathrm{d}\mathbf{s} \approx \mathcal{A}_T f(\mathbf{m}_T)\ , \tag{4.7}$$

where ℓ_σ is the length of the geodesic edge and $\mathcal{A}_T$ the area of the surface triangle, and $f(\mathbf{m}_\sigma)$ and $f(\mathbf{m}_T)$ denote the evaluation of the surface function at the edge and cell midpoints, respectively.

To address step iii) of the FV development mentioned at the beginning of this chapter, we need to establish the approximations of the necessary geometric quantities. An important feature we want to maintain in our numerical approach is the exclusive use of geometrically intrinsic quantities. Up to this point, our FV scheme is defined intrinsically on the LCS attached on the bottom surface. To continue our development within this setting, we assume that all the relevant intrinsic information, namely the tangent plane, is known (in exact or approximate form) at the vertices of the triangulation. Then the task is to develop intrinsic approximations of the geometric quantities starting from these data.

Approximation on edges. With reference to fig. 4.2, we directly approximate the LCS (or equivalently the tangent plane) at the edge midpoint $\mathbf{m}_\sigma$ using nodal information, without actually resorting to an explicit expression for $\mathbf{m}_\sigma$. We start by approximating the tangent plane $T_{\mathbf{m}_\sigma}\Gamma$ from the linear interpolation of the tangent vectors $\mathbf{t}_i(A)$ and $\mathbf{t}_i(B)$, $i = 1, 2$, with re-orthogonalization. The frame completing normal vector is naturally obtained using the normalized outer product of $\tilde{\mathbf{t}}_i$. The

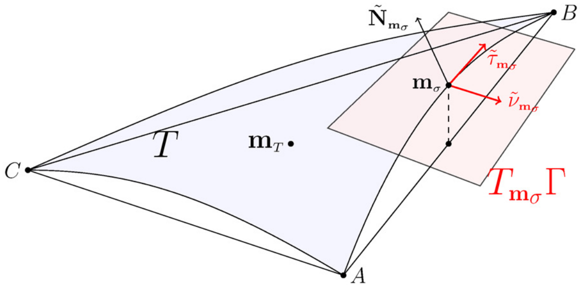

Figure 4.2: Curvilinear cell T and corresponding approximate gravity center $\mathbf{m}_T$. The curvilinear edge σ of cell T is the minimal geodesic curve connecting the two vertices. The midpoint of σ is denoted by $\mathbf{m}_\sigma$.

three LCS vectors are thus given by:

$$\tilde{\mathbf{t}}_1(\mathbf{m}_\sigma) = \frac{1}{2}\mathbf{t}_1(A) + \frac{1}{2}\mathbf{t}_1(B),$$

$$\mathbf{t}_2'(\mathbf{m}_\sigma) = \frac{1}{2}\mathbf{t}_2(A) + \frac{1}{2}\mathbf{t}_2(B), \qquad \tilde{\mathbf{t}}_2(\mathbf{m}_\sigma) = \mathbf{t}_2' - \frac{\left\langle \mathbf{t}_2', \tilde{\mathbf{t}}_1 \right\rangle}{\left\langle \tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_1 \right\rangle}\tilde{\mathbf{t}}_1,$$

$$\tilde{\mathbf{t}}_3(\mathbf{m}_\sigma) = \frac{\tilde{\mathbf{t}}_1 \wedge \tilde{\mathbf{t}}_2}{\left\|\tilde{\mathbf{t}}_1\right\| \left\|\tilde{\mathbf{t}}_2\right\|} .$$

The tangent plane at $\mathbf{m}_\sigma$ is then identified by $T_{\mathbf{m}_\sigma}\Gamma = \langle \tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_2 \rangle$, and the associated metric tensor is:

$$\tilde{\mathcal{G}}_{\mathbf{m}_\sigma} = \begin{bmatrix} \left\|\tilde{\mathbf{t}}_1\right\|^2 & 0 \\ 0 & \left\|\tilde{\mathbf{t}}_2\right\|^2 \end{bmatrix} .$$

Note that this approximation is fully intrinsic and does not require information on the geodesic edge between A and B.

The intrinsic definition of the Riemann problem requires the approximation of the normal and tangent directions to the geodesic edge at the midpoint. This is achieved as follows. The tangent plane is identified by its normal, given by:

$$\tilde{\mathbf{N}}_{\mathbf{m}_\sigma} = \tilde{\mathbf{t}}_3(\mathbf{m}_\sigma) .$$

The tangent vector is approximated by a second order accurate linear interpolation of the vectors tangent to the geodesic edge at the two vertices. Let $\sigma(\lambda)$ be a parametrization of the geodesic curve connecting the two nodes A and B and forming edge σ, with $\lambda \in [0,1]$ (see fig. 4.2). In our case, we approximate this parametrization using the following expression:

$$\underline{\sigma}(\lambda) = \begin{cases} s^1(\lambda) = \left(s^1_B - s^1_A\right)\lambda + s^1_A \\ s^2(\lambda) = \left(s^2_B - s^2_A\right)\lambda + s^2_A \\ s^3(\lambda) = \mathcal{B}\left(s^1(\lambda), s^2(\lambda)\right) \end{cases} , \tag{4.8}$$

where the third component of this parametric curve is the vertical projection of the chord onto the surface and not the real geodesic curve connecting A and B. Then, we compute the derivatives at the extremal points of this curve, obtaining the vectors tangent to the relevant edge at the triangle nodes, which in our case are:

$$\tau_A = \underline{\dot{\sigma}}(0), \qquad \tau_B = \underline{\dot{\sigma}}(1) \; .$$

The tangent vector at $\mathbf{m}_\sigma$ of the geodesic edge is then given by linear interpolation of τ_A and τ_B, orthogonalized with respect to the $\tilde{\mathbf{N}}_{\mathbf{m}_\sigma}$ to project it on $T_{\mathbf{m}_\sigma}\Gamma$:

$$\tau'_{\mathbf{m}_\sigma} = \frac{1}{2}\left(\tau_A + \tau_B\right) \; , \qquad \tilde{\tau}_{\mathbf{m}_\sigma} = \tau'_{\mathbf{m}_\sigma} - \frac{\left\langle \tau'_{\mathbf{m}_\sigma}, \tilde{\mathbf{N}}_{\mathbf{m}_\sigma} \right\rangle}{\left\langle \tilde{\mathbf{N}}_{\mathbf{m}_\sigma}, \tilde{\mathbf{N}}_{\mathbf{m}_\sigma} \right\rangle} \tilde{\mathbf{N}}_{\mathbf{m}_\sigma} \; .$$

Hence, the outer product of $\tilde{\tau}_{\mathbf{m}_\sigma}$ and $\tilde{\mathbf{N}}_{\mathbf{m}_\sigma}$ gives the approximation of the intrinsic normal to the geodesic edge, namely:

$$\tilde{\nu}_{\mathbf{m}_\sigma} = \tilde{\tau}_{\mathbf{m}_\sigma} \wedge \tilde{\mathbf{N}}_{\mathbf{m}_\sigma} \; . \tag{4.9}$$

To complete the definition we need to express $\tilde{\nu}_{\mathbf{m}_\sigma}$ and $\tilde{\tau}_{\mathbf{m}_\sigma}$ using the LCS bases. This provides the approximate vectors that define a proper rotation and projection in the direction orthogonal to the edge at the midpoint.

Approximation on cells. Analogous approximations need to be done in order to define the geometric information on the cells, namely the tangent plane at the gravity center of the cell $\mathbf{m}_T$. As before, we start from nodal data and construct the following vectors:

$$\begin{aligned}
\tilde{\mathbf{t}}_1(\mathbf{m}_T) &= \frac{1}{3}\mathbf{t}_1(A) + \frac{1}{3}\mathbf{t}_1(B) + \frac{1}{3}\mathbf{t}_1(C), \\
\mathbf{t}'_2(\mathbf{m}_T) &= \frac{1}{3}\mathbf{t}_2(A) + \frac{1}{3}\mathbf{t}_2(B) + \frac{1}{3}\mathbf{t}_2(C), \qquad \tilde{\mathbf{t}}_2(\mathbf{m}_T) = \mathbf{t}'_2 - \frac{\left\langle \mathbf{t}'_2, \tilde{\mathbf{t}}_1 \right\rangle}{\left\langle \tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_1 \right\rangle}\tilde{\mathbf{t}}_1, \\
\tilde{\mathbf{t}}_3(\mathbf{m}_T) &= \frac{\tilde{\mathbf{t}}_1 \wedge \tilde{\mathbf{t}}_2}{\left\|\tilde{\mathbf{t}}_1\right\| \left\|\tilde{\mathbf{t}}_2\right\|},
\end{aligned}$$

with the associated metric tensor

$$\tilde{\mathcal{G}}_{\mathbf{m}_T} = \begin{bmatrix} \left\|\tilde{\mathbf{t}}_1\right\|^2 & 0 \\ 0 & \left\|\tilde{\mathbf{t}}_2\right\|^2 \end{bmatrix} \; .$$

Approximation error estimates

In this paragraph we derive approximation error estimates for the quantities defined in the previous paragraphs. In particular, we need to control the accuracy of the quadrature rule and, as a consequence, of the divergence theorem so that consistency of the IFV is ensured. In addition, the given estimates will relate these errors with the geometrical properties of the bottom surface.

For a regular surface, the piece-wise linear interpolant $I_h f$ of a smooth function is characterized by quadratic error [36]:

$$\|f - I_h f\|_{L^2(T)} \leq C\ell_{\sigma_h}^2 \left\|D^2 f\right\|_{L^2(T)} .$$

In section 4.3.2 we will show the experimental convergence results confirming this fact in our setting. Since we assume that the flux and source functions are Lipschitz regular, their evaluation at the sampling points maintains the order of accuracy. Thus, to complete our analysis, we are left with the study of the approximation errors introduced in the quadrature rules by the definition of the tangent/normal reference frame in $T_{\mathbf{m}_\sigma}\Gamma$. In particular, we analize here the quadrature rule for an integral over the edge σ of a general Lipschitz regular vector function F. Here we denote with $\tilde{F}^\nu(\mathbf{m}_\sigma)$ the evaluation of the function F on the approximate $T_{\mathbf{m}_\sigma}\Gamma$ at the edge midpoint $\mathbf{m}_\sigma$ projected along the approximate normal direction $\tilde{\nu}_{\mathbf{m}_\sigma}$. Then, we have the following result:

Proposition 4.1.11. *Given a Lipschitz regular vector function F defined on the geodesic edge σ we have:*

$$\int_\sigma \langle F, \nu\rangle_{\mathcal{G}} \, \mathrm{d}\sigma = \ell_{\sigma_h} \tilde{F}^\nu(\mathbf{m}_\sigma) + \mathcal{O}\left(\ell_{\sigma_h}^2\right) .$$

Proof. Direct application of the mid-point rule yields [39]:

$$\int_\sigma \langle F, \nu\rangle_{\mathcal{G}} \, \mathrm{d}\sigma = \ell_\sigma \langle F, \nu\rangle_{\mathcal{G}} (\mathbf{m}_\sigma) + \mathcal{O}\left(\ell_\sigma^3\right) .$$

Next we address explicitly the different approximations made in the definition of $T_{\mathbf{m}_\sigma}\Gamma$. We first note that, by proposition 4.1.10, item 2, we have:

$$|\tau - \tilde{\tau}_{\mathbf{m}_\sigma}| \leq \left|\tau - \mathbf{v}_{\overrightarrow{AB}}\right| + \left|\mathbf{v}_{\overrightarrow{AB}} - \tilde{\tau}_{\mathbf{m}_\sigma}\right| \leq C\ell_\sigma ,$$

where C is a constant depending on the surface curvature. By construction, the approximate normal $\tilde{\nu}_{\mathbf{m}_\sigma}$ shares the same first order error estimate. Recalling remark 4.3, essentially proposition 4.1.10, item 1 states that $|\ell_\sigma - \ell_{\sigma_h}| \leq C\ell_{\sigma_h}^3$. Linear interpolation of the nodal quantities ensures that the approximate metric and tangent plane in $\mathbf{m}_\sigma$ are second order accurate, i.e.:

$$\mathcal{G}(\mathbf{m}_\sigma) = \tilde{\mathcal{G}}_{\mathbf{m}_\sigma} + \mathcal{O}\left(\ell_\sigma^2\right) \qquad \text{and} \qquad F(\mathbf{m}_\sigma) = \tilde{F}(\mathbf{m}_\sigma) + \mathcal{O}\left(\ell_\sigma^2\right).$$

Thus, expanding the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ and combining all the terms, we can write:

$$\begin{aligned}
&\ell_\sigma \langle F, \nu \rangle_{\mathcal{G}}(\mathbf{m}_\sigma) + \mathcal{O}\Big(\ell_\sigma^3\Big) = \\
&\qquad = \ell_\sigma \left(\tilde{F}(\mathbf{m}_\sigma) + \mathcal{O}\Big(\ell_\sigma^2\Big)\right)\left(\tilde{\mathcal{G}}_{\mathbf{m}_\sigma} + \mathcal{O}\Big(\ell_\sigma^2\Big)\right)(\tilde{\nu}_{\mathbf{m}_\sigma} + \mathcal{O}(\ell_\sigma)) + \mathcal{O}\Big(\ell_\sigma^3\Big) = \\
&\qquad = \ell_\sigma \left(\tilde{F}(\mathbf{m}_\sigma)\tilde{\mathcal{G}}_{\mathbf{m}_\sigma}\tilde{\nu}_{\mathbf{m}_\sigma} + \mathcal{O}(\ell_\sigma)\right) + \mathcal{O}\Big(\ell_\sigma^3\Big) = \\
&\qquad = \ell_{\sigma_h}\tilde{F}^\nu(\mathbf{m}_\sigma) + \mathcal{O}\Big(\ell_{\sigma_h}^2\Big) \ ,
\end{aligned}$$

which yields the desired result. □

Because of its importance in our IFV scheme, we report here the following corollary whose proof is an immediate consequence of the previous proposition.

Corollary 4.1.12. *Given a Lipschitz regular vector function F defined on the geodesic triangle T we have:*

$$\int_T \nabla_{\mathcal{G}} \cdot F \, \mathrm{d}\mathbf{s} = \sum_{j=1}^{N_\sigma} \ell_{h,j} \tilde{F}^\nu(\mathbf{m}_{\sigma_j}) + \mathcal{O}\Big(\ell_{\sigma_h}^2\Big) \ .$$

We end this section by noting that all these estimates can be directly related to the global mesh parameter h. This is classically defined as the maximum of the lengths of the triangulation edges, i.e., $h = \max_{\sigma_h \in \mathcal{T}_h(\Gamma)} \ell_{\sigma_h}$. Noting that, by remarks 4.3 and 4.4 and proposition 4.1.10, item 3, the area of T_h converges quadratically to the area of T, we can conclude that all our geometric approximations involved in the calculation of the right-hand-side of eq. (4.10) are consistent with the global accuracy of our IFV discretization.

4.2 The scheme in the geometric setting

The previous section dealt with the definition of the surface approximation, presenting theoretical results on convergence. This section progresses by presenting the complete FV scheme in our geometric setting. Step i) uses a first order explicit Euler time stepping scheme. To maintain a well-balanced scheme we use an adaptation of the approach proposed by Audusse et al. [8], Bouchut [14] and include the source terms in the flux. Then, the following FV equations are defined for each T_i:

$$\mathbf{U}_{h,i}^{k+1} = \mathbf{U}_{h,i}^k - \frac{\Delta t}{\mathcal{A}_{h,i}} \sum_{j=1}^{N_{\sigma(i)}} \ell_{h,ij} \left[\mathbf{F}_{h,ij}(\mathbf{U}_{h,i}^k, \mathbf{U}_{h,j}^k) + \mathbf{S}_{h,ij}(\mathbf{U}_{h,i}^k, \mathbf{U}_{h,j}^k)\right] \ , \tag{4.10}$$

where $\mathbf{F}_{h,ij}$ is the numerical approximation of the edge-averaged normal flux $\mathbf{F}_{ij}$ at σ_{ij}, and $\mathbf{S}_{h,ij}$ is calculated so that $\sum_{j=1}^{N_{\sigma(i)}} \ell_{h,ij}\mathbf{S}_{h,ij}$ is a consistent quadrature rule for the last integral in eq. (4.2) and maintains the discrete version of eq. (3.22).

4.2.1 The Riemann problem

As customary in upwind-Godunov methods, the numerical fluxes $\mathbf{F}_{h,ij}$ in eq. (4.10) are defined by evaluating the flux function $\mathbf{F}^{\nu}_{ij}$ from solutions of Riemann Problems (RPs) at quadrature points of the cell edges. Since our equations are invariant under linear transformations and, specifically, rotations, each RP can be formulated as a one-dimensional problem in the tangent plane $T_{\mathbf{m}_\sigma}\Gamma$, passing through the quadrature point $\mathbf{m}_\sigma$ in the direction of the local edge normal $\tilde{\nu}_{\mathbf{m}_\sigma}$.

Given the tangent plane $T_{\mathbf{m}_\sigma}\Gamma$ on the edge, with associated metric tensor $\tilde{\mathcal{G}}_{\mathbf{m}_\sigma}$, and denoting by $\tilde{\nu}_{\mathbf{m}_\sigma} = [\nu^1, \nu^2]^{\mathrm{T}}$ and $\tilde{\tau}_{\mathbf{m}_\sigma} = [\tau^1, \tau^2]^{\mathrm{T}}$ the normal and tangent vectors to the geodesic edge, the normal flux function can be written as:

$$\mathbf{F}^{\nu}(\mathbf{s}, \mathbf{U}) = \left\langle \underline{\underline{F}}, \tilde{\nu}_{\mathbf{m}_\sigma} \right\rangle_{\mathcal{G}} = \begin{bmatrix} q^1 h^2_{(1)} \nu^1 + q^2 h^2_{(2)} \nu^2 \\ \dfrac{(q^1)^2}{\eta} h^2_{(1)} \nu^1 + \dfrac{1}{2} g \eta^2 \dfrac{\partial x^3}{\partial s^3} \nu^1 + \dfrac{q^1 q^2}{\eta} h^2_{(2)} \nu^2 \\ \dfrac{q^1 q^2}{\eta} h^2_{(1)} \nu^1 + \dfrac{(q^2)^2}{\eta} h^2_{(2)} \nu^2 + \dfrac{1}{2} g \eta^2 \dfrac{\partial x^3}{\partial s^3} \nu^2 \end{bmatrix} . \tag{4.11}$$

Using an edge-based LCS aligned with the normal and tangent vectors, we can write:

$$\begin{aligned} q^N &= \langle \vec{q}, \tilde{\nu}_{\mathbf{m}_\sigma} \rangle_{\mathcal{G}} = q^1 h^2_{(1)} \nu^1 + q^2 h^2_{(2)} \nu^2 \\ q^T &= \langle \vec{q}, \tilde{\tau}_{\mathbf{m}_\sigma} \rangle_{\mathcal{G}} = q^1 h^2_{(1)} \tau^1 + q^2 h^2_{(2)} \tau^2 = -q^1 h^2_{(1)} \nu^2 + q^2 h^2_{(2)} \nu^1 \ , \end{aligned}$$

where we have used the relation $\tilde{\tau}_{\mathbf{m}_\sigma} = [\tau^1, \tau^2]^{\mathrm{T}} = [-\nu^2, \nu^1]^{\mathrm{T}}$. Finally, the normal flux function in the edge LCS takes on the expression:

$$\mathbf{F}^{\nu}(\mathbf{s}, \mathbf{U}) = \begin{bmatrix} q^N \\ \dfrac{q^1 q^N}{\eta} + \dfrac{1}{2} g \eta^2 \dfrac{\partial x^3}{\partial s^3} \nu^1 \\ \dfrac{q^2 q^N}{\eta} + \dfrac{1}{2} g \eta^2 \dfrac{\partial x^3}{\partial s^3} \nu^2 \ . \end{bmatrix}$$

Note that our flux function is non-autonomous as it depends on the space variable defined on the local coordinate system. In this case difficulties in the well-posedness of the RP may arise (see, e.g., Andreianov et al. [5, 6], Andreianov and Sbihi [4]). Thus, we write the Riemann problem as follows. Let $\mathbf{V} = \left[\eta, q^N, q^T\right]^{\mathrm{T}} = [\eta, \eta v^1, \eta v^2]^{\mathrm{T}}$ be the vector of conservative variables, where, q^N and q^T indicate the components of the flux (discharge) vector defined along the directions $\tilde{\nu}_{\mathbf{m}_\sigma}$ and $\tilde{\tau}_{\mathbf{m}_\sigma}$, respectively. Denote with

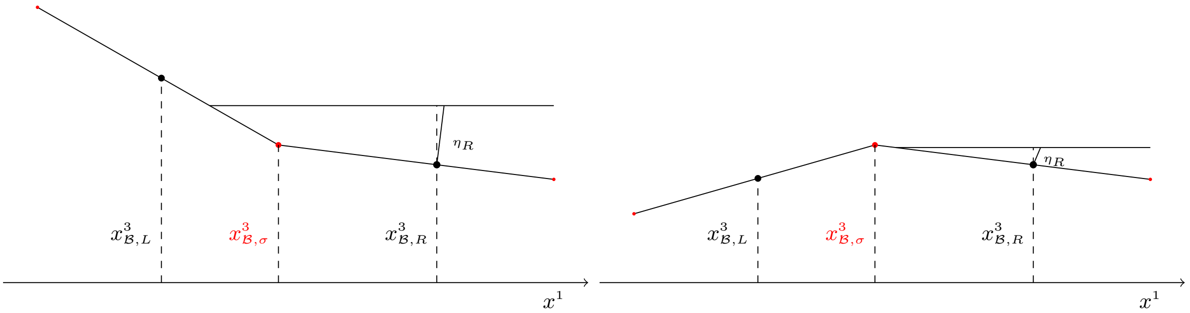

Figure 4.3: Examples of wet and dry regions around edge σ, over the triangulation $\mathcal{T}_h(\Gamma)$.

x the space variable along $\tilde{\nu}_{\mathbf{m}_\sigma}$. Then, the RP is formally given by:

$$\begin{aligned}
&\mathbf{V}_t + \mathbf{F}^\nu(x,\mathbf{V})_x = 0, \\
&\mathbf{F}^\nu : (x,\mathbf{V}) \in \Gamma \times T_{\mathbf{m}_\sigma}\Gamma \;\mapsto \begin{cases} \mathbf{F}_L(\mathbf{V}) & \text{if } x < 0 \\ \mathbf{F}_R(\mathbf{V}) & \text{if } x > 0 \end{cases}, \\
&\mathbf{V}(x,0) = \begin{cases} \mathbf{V}_L & \text{if } x < 0 \\ \mathbf{V}_R & \text{if } x > 0 \end{cases},
\end{aligned} \tag{4.12}$$

where $\mathbf{F}^\nu$ is the normal flux function as defined in eq. (4.11). To simplify the problem, we advocate the assumption of continuity of the flux function with respect to the first argument. This allows us to approximate the RP by assuming $\mathbf{F}_L(\mathbf{V}) = \mathbf{F}_R(\mathbf{V})$, thus recovering the standard form of the shallow-water RP. This approximation deserves further investigation, which is however outside the scope of this work and we leave it to future studies.

Consistently with the first order accuracy, we use cell values to define the left and right initial states $\mathbf{V}_L$ and $\mathbf{V}_R$. However, the physical quantities of depth and velocity need to be written in the correct reference system attached on the edge midpoint. For each cell, η represents the cell average of the water column (measured in the normal direction) at the cell midpoint $\mathbf{m}_T$. Since we consider a surface triangulation, the midpoint of the cell can have a different elevation with respect to the GCS than the midpoint of the edge. Thus, the depth value needs to be properly moved to the point $\mathbf{m}_\sigma$, where the RP is defined. For edge σ, we calculate the left and right edge elevations $\eta_{(L)}$ and $\eta_{(R)}$ from the cell elevations η_L and η_R of cells T_L and T_R as follows:

$$\eta_{(k)} = \max\left\{0, \frac{\eta_k \left.\dfrac{\partial x^3}{\partial s^3}\right|_k + x^3_{\mathcal{B},k} - x^3_{\mathcal{B},\sigma}}{\left.\dfrac{\partial x^3}{\partial s^3}\right|_\sigma}\right\} \qquad k = L, R\,,$$

where $\left.\frac{\partial x^3}{\partial s^3}\right|_k$ and $x^3_{\mathcal{B},k}$ are defined at the gravity center of the k-th cell. The zero bound is introduced to ensure non-negativity of water depth. We need to define the edge-quantities $\left.\frac{\partial x^3}{\partial s^3}\right|_\sigma$ and $x^3_{\mathcal{B},\sigma}$. Because of the regularity of the bottom surface, we can set $\left.\frac{\partial x^3}{\partial s^3}\right|_\sigma = \frac{\partial x^3}{\partial s^3}(\mathbf{m}_\sigma)$. Definition of the edge bottom elevation $x^3_{\mathcal{B},\sigma}$ must take into consideration the fact that the edge might be in a dry region (i.e., the edge midpoint is at a higher elevation with respect to the cell water elevation). Following [14, 8], we then set

$$x^3_{\mathcal{B},\sigma} = \max\{x^3_{\mathcal{B},L}, x^3_{\mathcal{B},R}, x^3_{\mathcal{B}}(\mathbf{m}_\sigma)\} \ ,$$

where $x^3_{\mathcal{B}}(\mathbf{m}_\sigma)$ is the approximated value at the edge center $\mathbf{m}_\sigma$. This ensures that dry regions with $\eta = 0$ are captured so that wetting-and-drying processes are handled seamlessly (fig. 4.3).

Concerning the velocity, the vector $\mathbf{q}$ must be "parallel" transported from the cell gravity center to the edge midpoint, so that all quantities in the RP belong to $T_{\mathbf{m}_\sigma}\Gamma$. This is achieved by a first order linearization of the ODE system defining parallel transport. Again, for edge σ we can write [63, 37]:

$$\begin{aligned} q^1{}_{(k)}(\mathbf{s}_{\mathbf{m}_\sigma}) &= q^1(\mathbf{s}_k) - (s^1_{\mathbf{m}_T} - s^1_k)\left(\Gamma^{1,(k)}_{11} q^1(\mathbf{s}_k) + \Gamma^{1,(k)}_{21} q^2(\mathbf{s}_k)\right) && \text{(4.13a)} \\ & && k = L, R \ . \\ q^2{}_{(k)}(\mathbf{s}_{\mathbf{m}_\sigma}) &= q^2(\mathbf{s}_k) - (s^2_{\mathbf{m}_T} - s^2_k)\left(\Gamma^{2,(k)}_{11} q^1(\mathbf{s}_k) + \Gamma^{2,(k)}_{21} q^2(\mathbf{s}_k)\right) && \text{(4.13b)} \end{aligned}$$

Note that this first order linearization maintains the accuracy of the overall scheme. However, its most important role is to perform the change of coordinate systems from the cell-attached to the edge-attached tangent planes.

The Riemann solver

We use two different types of Riemann solvers. The first is an extension to the curvilinear case of the exact Riemann solver described in Toro [70]. The second is the classical HLL and its variant HLLC [48, 71], which easily adapts to our equations.

Exact Riemann Solver. As typical, to unveil the wave structures of the problem, eq. (4.12) is written in non-conservative form, leading to:

$$\mathbf{V}_t + \mathbf{A}(\mathbf{V})\mathbf{V}_x = 0 \ ,$$

where $\mathbf{A}(\mathbf{V})$ is the Jacobian matrix of $\mathbf{F}^\nu$, with eigenvalues and eigenvectors denoted by $(\lambda_1, \lambda_2, \lambda_3)$, and $(\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \mathbf{r}^{(3)})$, respectively. The Jacobian takes on the form:

$$\mathbf{A}(\mathbf{V}) = \begin{bmatrix} 0 & 1 & 0 \\ \eta g \dfrac{\partial x^3}{\partial s^3} - \left(\dfrac{q^N}{\eta}\right)^2 & 2\dfrac{q^N}{\eta} & 0 \\ -\dfrac{q^N q^T}{\eta^2} & \dfrac{q^T}{\eta} & \dfrac{q^N}{\eta} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ a^2 \dfrac{\partial x^3}{\partial s^3} - (v^1)^2 & 2v^1 & 0 \\ -v^1 v^2 & v^2 & v^1 \end{bmatrix} ,$$

where $a = \sqrt{\eta g}$ indicates the wave celerity. Its eigenvalues are:

$$\lambda_1 = v^1 - a\sqrt{\frac{\partial x^3}{\partial s^3}} , \qquad \lambda_2 = v^1 , \qquad \lambda_3 = v^1 + a\sqrt{\frac{\partial x^3}{\partial s^3}} , \tag{4.14}$$

with corresponding eigenvectors given by:

$$\mathbf{r}^{(1)} = \begin{bmatrix} \dfrac{1}{v^2} \\ \left(v^1 - a\sqrt{\dfrac{\partial x^3}{\partial s^3}}\right) \dfrac{1}{v^2} \\ 1 \end{bmatrix} , \qquad \mathbf{r}^{(2)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} , \qquad \mathbf{r}^{(3)} = \begin{bmatrix} \dfrac{1}{v^2} \\ \left(v^1 + a\sqrt{\dfrac{\partial x^3}{\partial s^3}}\right) \dfrac{1}{v^2} \\ 1 \end{bmatrix} .$$

The wave structure is the same as the standard SW equations with the additional presence of the bottom slope term $\frac{\partial x^3}{\partial s^3}$. Thus, the solution described in Toro [70] applies with only slight modifications. We present here a brief summary of the possible wave patterns.

There are four possible wave patterns that may occur in the solution of the Riemann problem, and note that in each pattern there are three waves. The left and right eigenvalues corresponding to the one-dimensional shallow water equations are either shocks or rarefactions. See fig. 4.4 for all the possible cases:

(a) the left wave is a rarefaction wave and the right wave is a shock;

(b) the left wave is a shock wave and the right wave is a rarefaction;

(c) both the left and right waves are rarefaction;

(d) both the left and right waves are shock waves.

The middle wave arises from the presence of the second component in the momentum equation and it is always a shear wave. The three waves are associated with the three eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of the linearized problem. The waves separate four constant states denoted, from left to right, by $\mathbf{V}_L$, $\mathbf{V}_L^*$, $\mathbf{V}_R^*$ and $\mathbf{V}_R$. The region between the left and

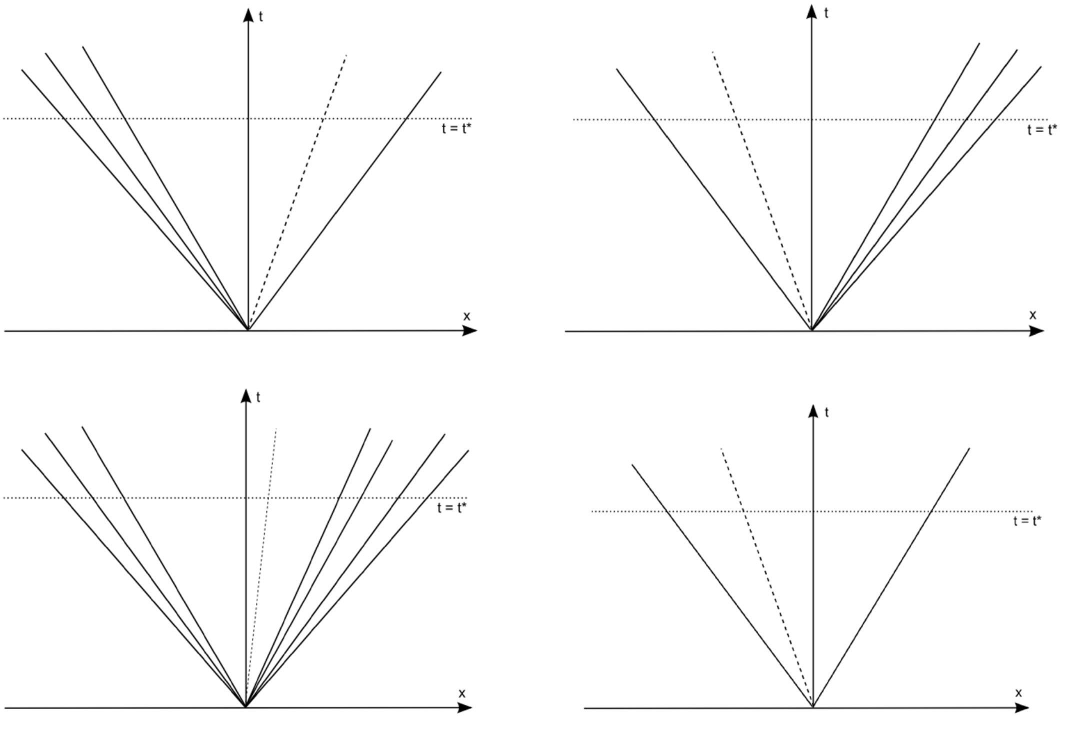

Figure 4.4: Wave patterns for the exact Riemann solver: cases (a) to (d) are shown from top left to bottom right panels.

right waves is called the *star region* and is subdivided in two sub-regions. We denote the constant states of the water depth and fluid velocity (along the normal direction) in the star region by η^* and v^* respectively.

The types of nonlinear left and right waves are determined by the following conditions:

$$\begin{aligned} &\eta^* > \eta_L && \text{the left wave is a shock wave,} \\ &\eta^* \leq \eta_L && \text{the left wave is a rarefaction wave,} \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} &\eta^* > \eta_R && \text{the right wave is a shock wave,} \\ &\eta^* \leq \eta_R && \text{the right wave is a rarefaction wave.} \end{aligned} \tag{4.16}$$

The basic step now is to derive a single algebraic nonlinear equation for the water depth η^* in the star region. To this aim, we will connect the state v^* in the star region to the data on the left and on the right considering two appropriate functions, namely $f_L(\eta, \eta_L)$ and $f_R(\eta, \eta_R)$. These functions govern the relations of flow quantities across the left and right waves respectively. In our case, these waves are either shock or rarefaction waves then, in deriving f_L and f_R we make use of either generalized Riemann invariants or Rankine-Hugoniot conditions, as appropriate.

- **Wet case solution:**

Proposition 4.2.1. *The solution η^* for the Riemann problem is given by the roots of the algebraic equation*

$$f(\eta) = f_L(\eta, \eta_L) + f_R(\eta, \eta_R) + (v_R - v_L) = 0, \tag{4.17}$$

where the functions f_L and f_R are

$$f_L = \begin{cases} 2\dfrac{\partial x^3}{\partial s^3}\left(\sqrt{g\eta} - \sqrt{g\eta_L}\right) & \text{if } \eta \leq \eta_L \text{ (rarefaction)} \\ (\eta - \eta_L)\sqrt{\dfrac{1}{2}g\dfrac{\partial x^3}{\partial s^3}\dfrac{\eta + \eta_L}{\eta\eta_L}} & \text{if } \eta > \eta_L \text{ (shock)} \end{cases},$$

$$f_R = \begin{cases} 2\dfrac{\partial x^3}{\partial s^3}\left(\sqrt{g\eta} - \sqrt{g\eta_R}\right) & \text{if } \eta \leq \eta_R \text{ (rarefaction)} \\ (\eta - \eta_R)\sqrt{\dfrac{1}{2}g\dfrac{\partial x^3}{\partial s^3}\dfrac{\eta + \eta_R}{\eta\eta_R}} & \text{if } \eta > \eta_R \text{ (shock)} \end{cases}.$$

The solution for the particle velocity v^* in the star region follows as

$$v^* = \frac{1}{2}(v_L + v_R) + \frac{1}{2}[f_R(\eta^*, \eta_R) - f_L(\eta^*, \eta_L)]\,. \tag{4.18}$$

Proof. There are four possible cases to consider, namely:

- left rarefaction wave

$$\begin{aligned} v^* &= v_L - f_L(\eta^*, \eta_L)\,, \\ f_L(\eta^*, \eta_L) &= 2\frac{\partial x^3}{\partial s^3}\left(\sqrt{g\eta_*} - \sqrt{g\eta_L}\right)\,; \end{aligned}$$

- right rarefaction wave

$$\begin{aligned} v^* &= v_R + f_R(\eta^*, \eta_R)\,, \\ f_R(\eta^*, \eta_R) &= 2\frac{\partial x^3}{\partial s^3}\left(\sqrt{g\eta^*} - \sqrt{g\eta_R}\right)\,; \end{aligned}$$

- left shock wave

$$\begin{aligned} v^* &= v_L - f_L(\eta^*, \eta_L)\,, \\ f_L(\eta^*, \eta_L) &= (\eta^* - \eta_L)\sqrt{\frac{1}{2}g\frac{\partial x^3}{\partial s^3}\frac{\eta^* + \eta_L}{\eta^*\eta_L}}\,; \end{aligned}$$

- right shock wave

$$\begin{aligned} v^* &= v_R + f_R(\eta^*, \eta_R)\,, \\ f_R(\eta^*, \eta_R) &= (\eta^* - \eta_R)\sqrt{\frac{1}{2}g\frac{\partial x^3}{\partial s^3}\frac{\eta^* + \eta_R}{\eta^*\eta_R}}\,. \end{aligned}$$

Elimination of v^* in all four cases gives

$$f_L(\eta^*, \eta_L) + f_R(\eta^*, \eta_R) + v_R - v_L = 0\,,$$

and thus η^* is the root of eq. (4.17). Assuming the root η^* is available, then v^* follows directly from any of the equations in the cases above, or from a mean value as

$$v^* = \frac{1}{2}(v_L + v_R) + \frac{1}{2}[f_R(\eta^*, \eta_R) - f_L(\eta^*, \eta_L)],$$

that is exactly eq. (4.18). Hence, the proposition is proved. □

The last step that we need to develop is how to solve the algebraic eq. (4.17) for the depth in the star region. There is no general closed-form solution, therefore, we solve it numerically using an iterative technique. We first re-write the algebraic nonlinear equation:

$$f(\eta) := f_L(\eta, \eta_L) + f_R(\eta, \eta_R) + \Delta v = 0\,, \qquad \text{with} \qquad \Delta v = v_R - v_L\,,$$

for the unknown η in the star region. Given the particular behavior of $f(\eta)$, a Newton-Raphson iteration scheme can be used. At the generic step ℓ, the scheme is:

$$\eta^{(\ell+1)} = \eta^{(\ell)} - \frac{f(\eta^{(\ell)})}{f'(\eta^{(\ell)})}, \tag{4.19}$$

for $\ell = 0, 1, \ldots$. To start the iteration we require a guess value $\eta^{(0)}$, that can be chosen using the two-rarefaction approximation, an expression that comes from the study of the depth function. We turn to Toro [70] for a precise study of the behavior of the depth function $f(\eta)$.

The iteration eq. (4.19) is stopped whenever the change in η is smaller that a tolerance value TOL, namely when

$$\Delta\eta = \frac{\left|\eta^{(\ell+1)} - \eta^{(\ell)}\right|}{(\eta^{(\ell+1)} + \eta^{(\ell)})/2} < TOL.$$

So far we have an algorithm for finding η^* and v^* in the star region. By comparing η^* with η_L and η_R (see eq. (4.15) and eq. (4.16)), we can determine when the left and right waves are shock or rarefaction waves and compute the exact solution of the one-dimensional Riemann problem for the variables η and v.

- **Dry case solution:**

  There are three possible cases as described in Toro [70] all for the rarefaction wave, because the a shock wave cannot be adjacent to a region of dry bed. We will consider the dry bed on the right side. The characteristic speed that coincides with the front is S_L^* and corresponds to the left eigenvalue $\lambda = v - a\sqrt{\frac{\partial x^3}{\partial s^3}}$.

  - Left rarefaction wave. Right dry bed. Consider a point of contact discontinuity, the characteristic of speed S_L^* that coincides with the front is

    $$S_L^* = \frac{dx}{dt} = v_c - \sqrt{g\eta_c \frac{\partial x^3}{\partial s^3}}\ .$$

    By connecting $\sqrt{g\eta_c}$ and v_c to the data on the left-hand-side via Riemann invariant $v - 2\sqrt{g\eta\frac{\partial x^3}{\partial s^3}} = const$, we obtain

    $$v_c + 2\sqrt{g\eta_c \frac{\partial x^3}{\partial s^3}} = v_L + 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}}\ .$$

    Along the discontinuity $\eta_c = 0$, so

    $$S_L^* = v_c = v_L + 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}}\ .$$

The limit speeds are:

a) on the left, the eigenvalue S_L

$$S_L = v_L - \sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}} \ ;$$

b) on the right, the speed S_L^*

$$S_L^* = v_L + 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}} \ ;$$

c) in the middle, it is necessary to compute the status and the solution

$$\begin{aligned} \sqrt{g\eta} &= \frac{1}{3}\left[2\sqrt{g\eta_L} + \frac{1}{\frac{\partial x^3}{\partial s^3}}(v_L - \frac{dx}{dt})\right] \ , \\ u &= \frac{1}{3}\left(v_L + 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}} + 2\frac{dx}{dt}\right) \ . \end{aligned}$$

– External rarefaction waves. Dry bed in the middle. This case is reached when $S_L^* < S_R^*$, so when

$$\begin{aligned} v_L - 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}} &\le v_R + 2\sqrt{g\eta_R \frac{\partial x^3}{\partial s^3}} \ , \\ \eta_{crit} = v_R - v_L + 2\sqrt{\frac{\partial x^3}{\partial s^3}}\,(\sqrt{g\eta_R} + \sqrt{g\eta_L}) &\ge 0 \ . \end{aligned}$$

Then, we apply this case of dry bed in the middle and we compute all the speeds:

$$\begin{aligned} S_{HL} &= v_L - \sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}}, & S_{SL} &= v_L + 2\sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}}, \\ S_{SR} &= v_R - 2\sqrt{g\eta_R \frac{\partial x^3}{\partial s^3}}, & S_{HR} &= v_R + \sqrt{g\eta_R \frac{\partial x^3}{\partial s^3}}. \end{aligned}$$

HLL/HLLC solvers. The HLL scheme estimates the smallest and largest wave speeds S_L, S_R in the solution of the Riemann problem starting from the left and right data $\mathbf{U}_L, \mathbf{U}_R$ defined in the neighboring cells to the edge and the corresponding fluxes $\mathbf{F}_L = \mathbf{F}^\nu(\mathbf{V}_L)$ and $\mathbf{F}_R = \mathbf{F}^\nu(\mathbf{V}_R)$. It is easy to calculate the extremal eigenvalues of the Jacobian as:

$$S_L = \min\left\{0, \min\left\{v_L - \sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}\Big|_L}, v_R - \sqrt{g\eta_R \frac{\partial x^3}{\partial s^3}\Big|_R}\right\}\right\} \ ,$$

$$S_R = \max\left\{0, \max\left\{v_L + \sqrt{g\eta_L \frac{\partial x^3}{\partial s^3}\Big|_L}, v_R + \sqrt{g\eta_R \frac{\partial x^3}{\partial s^3}\Big|_R}\right\}\right\} \ .$$

Considering the normal direction defined in eq. (4.12), the HLL flux on the edge is then computed as:

$$\mathbf{F}^{\nu}_{HLL} = \begin{cases} \mathbf{F}_L & \text{if } S_L \geq 0 \\ \dfrac{S_R\mathbf{F}_L - S_L\mathbf{F}_R + S_R S_L(\mathbf{U}_R - \mathbf{U}_L)}{S_R - S_L} & \text{if } S_L \leq 0 \leq S_R \\ \mathbf{F}_R & \text{if } S_R \leq 0 \end{cases} . \tag{4.20}$$

Note that, this approach ignores intermediate waves, such as shear waves and contact discontinuities. Consideration of these waves is achieved in the HLLC approach described in Toro et al. [71], whose extension to our setting is straightforward.

4.2.2 Well-balance

We look here at the well-balance property for the "lake-at-rest" condition. In this case, we have a steady-state condition with zero velocity and thus a time-independent horizontal free surface. Our intrinsic SW equation then becomes:

$$\begin{cases} \vec{q} = 0 \\ \nabla_{\mathcal{G}} \cdot \left(\tfrac{1}{2} g\eta^2 \dfrac{\partial x^3}{\partial s^3} \mathcal{G}^{-1}_{sw}\right) = -\tfrac{1}{2} g\eta^2 \, \nabla_{\mathcal{G}} \dfrac{\partial x^3}{\partial s^3} - g\eta \, \nabla_{\mathcal{G}} \, x^3_{\mathcal{B}} \end{cases} ,$$

which is equivalent to eq. (3.22). We would like to note that more accurate well-balance properties would be needed in case of higher order methods, but in this work we content ourselves with simple well-balance property in the form of the "lake-at-rest". We will consider as future studies the more complex situation of a fully well balanced scheme as done in [9, 11]. The divergence theorem can be used to express the cell-integral of the source term as a sum of integrals on the edges. In fact, from the steady momentum conservation equation we can write the following equality:

$$\begin{aligned} &\int_{T_i} \left(\frac{1}{2} g\eta^2 \, \nabla_{\mathcal{G}} \frac{\partial x^3}{\partial s^3} + g\eta \, \nabla_{\mathcal{G}} \, x^3_{\mathcal{B}}\right) \mathrm{d}\mathbf{s} = -\int_{T_i} \nabla_{\mathcal{G}} \cdot \left(\frac{1}{2} g\eta^2 \frac{\partial x^3}{\partial s^3} \mathcal{G}^{-1}_{sw}\right) \mathrm{d}\mathbf{s} = \\ &= -\sum_{j=1}^{3} \int_{\sigma_{ij}} \left\langle \frac{1}{2} g\eta^2 \frac{\partial x^3}{\partial s^3} \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} \mathrm{d}\sigma \approx -\sum_{j=1}^{3} \ell_{h,ij} \left\langle \frac{1}{2} g \left(\eta^*_{ij}\right)^2 \left.\frac{\partial x^3}{\partial s^3}\right|^*_{ij} \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} , \end{aligned} \tag{4.21}$$

where the starred quantities are evaluated at appropriate sampling points so that discrete well-balance is maintained. Then, the integral of the source $\mathbf{S}(\mathbf{s}, \eta)$ over T_i can be written as:

$$\int_{T_i} \mathbf{S}\,\mathrm{ds} \approx \begin{bmatrix} 0 \\ -\sum_{j=1}^{3} \ell_{h,ij} \left\langle \frac{1}{2} g \left(\eta_{ij}^*\right)^2 \left.\frac{\partial x^3}{\partial s^3}\right|_{ij}^{*} \mathcal{G}_{sw}^{-1}, \nu_{ij} \right\rangle_{\mathcal{G}} \end{bmatrix} =$$

$$= \sum_{j=1}^{3} \ell_{h,ij} \begin{bmatrix} 0 \\ -\left\langle \frac{1}{2} g \left(\eta_{ij}^*\right)^2 \left.\frac{\partial x^3}{\partial s^3}\right|_{ij}^{*} \mathcal{G}_{sw}^{-1}, \nu_{ij} \right\rangle_{\mathcal{G}} \end{bmatrix} = \sum_{j=1}^{3} \ell_{h,ij} \begin{bmatrix} 0 \\ \mathbf{S}_{ij} \end{bmatrix} .$$

The idea is to apply the latter relation as a quadrature rule for the source integral in the scheme also in the non steady-state case. This quadrature rule is, by definition, exact when the velocity is zero and we need to prove that its approximation error is of order $\mathcal{O}(h)$ when the velocity is non-zero. Moreover, it has to be consistent with the flux term $\mathbf{F}_{ij}$ computed via Riemann solvers.

We first define the sampling values η_{ij}^* and $\left.\frac{\partial x^3}{\partial s^3}\right|_{ij}^{*}$ used in $\mathbf{S}_{ij}$ that ensure well-balance and consistency, with the first condition satisfied exactly. In the case of a steady horizontal water table, discrete well-balance implies that, with respect to the GCS, the total water elevation at the cell-center must equal the total water elevation at the edge center:

$$\eta_i \left.\frac{\partial x^3}{\partial s^3}\right|_{i} + x^3_{\mathcal{B},i} = \eta_{ij} \left.\frac{\partial x^3}{\partial s^3}\right|_{ij} + x^3_{\mathcal{B},ij} \, , \tag{4.22}$$

which yields immediately:

$$\eta_{ij} = \frac{\eta_i \left.\frac{\partial x^3}{\partial s^3}\right|_{i} + x^3_{\mathcal{B},i} - x^3_{\mathcal{B},ij}}{\left.\frac{\partial x^3}{\partial s^3}\right|_{ij}} \, , \tag{4.23}$$

where we define the values at the edge as described in section 4.2.1. Then, we ensure non-negativity of water depth by defining $\eta_{ij}^* = \max\{0, \eta_{ij}\}$. We then have the following proposition.

Proposition 4.2.2. *The intrinsic finite volume scheme* (4.10) *i) preserves the non-negativity of η_i; ii) is well-balanced, i.e. preserves the steady-state of a lake-at-rest (eq.* (4.22)*); iii) is consistent with the continuous ISWE model (eq.* (3.19)*).*

Proof. Statement i) follows directly from the definition of η_{ij}^*, which ensures $0 \leq \eta_{ij}^* \leq \eta_i$ and $0 \leq \eta_{ji}^* \leq \eta_j$.

Property ii) of steady-state for lake-at-rest is maintained by the consistency of the flux $\mathbf{F}_{ij}$, valid by construction, and the definition of the source terms at the interface.

To prove Property iii) we need to establish the consistency of the scheme for the general case of a nonzero velocity, i.e., $\vec{q} \neq 0$. To this aim it is sufficient to prove that the left and right fluxes for edge σ_{ij}, $\mathcal{F}_L = \mathbf{F}_{ij} + \mathbf{S}_{ij}$ and $\mathcal{F}_R = \mathbf{F}_{ji} + \mathbf{S}_{ji}$, respectively, satisfy the following condition [14]:

$$\mathcal{F}_L + \mathcal{F}_R = \mathbf{f}(\eta)\, h + o(h) \ , \tag{4.24}$$

where $\mathbf{f}(\eta)$ indicates a general vector function of the depth only. Obviously, our numerical flux $\mathbf{F}_{ij}$ is consistent:

$$\mathbf{F}_{ij} = \left\langle \underline{\underline{F}}(\mathbf{U}^*), \nu_{ij} \right\rangle_{\mathcal{G}} = \left\langle \underline{\underline{F}}(\mathbf{U}^*), -\nu_{ji} \right\rangle_{\mathcal{G}} = -\mathbf{F}_{ji} \ .$$

It remains to show the consistency for the edge source terms, that are nonzero only for the two momentum conservation equations. Recall that, due to the assumption of regularity of the bottom surface $\mathcal{S}_{\mathcal{B}}$, Taylor expansions for functions living on Γ are well defined [see 33]. Moreover, for a general surface vector $\mathbf{v}$ written in physical coordinates and a scalar function f we can write $\langle \nabla_{\mathcal{G}} f, \mathbf{v} \rangle_{\mathcal{G}} = \langle \nabla f, \mathbf{v} \rangle$. Hence we have:

$$\begin{aligned} x^3_{\mathcal{B},ij} &= x^3_{\mathcal{B},i} + \nabla x^3_{\mathcal{B},i} \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + o(h) \ , \\ \left.\frac{\partial x^3}{\partial s^3}\right|_{ij} &= \left.\frac{\partial x^3}{\partial s^3}\right|_{i} + \nabla \left(\left.\frac{\partial x^3}{\partial s^3}\right|_{i} \right) \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + o(h) \ , \\ \eta_{ij} &= \eta_i + \nabla \eta_i \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + o(h) \ , \end{aligned}$$

where $\mathbf{s}_{ij}, \mathbf{s}_i$ are the LCS coordinates of $\mathbf{m}_{\sigma_{ij}}$ and $\mathbf{m}_{T_i}$, respectively, and we have used the fact that $|\mathbf{s}_{ij} - \mathbf{s}_i| = \mathcal{O}(h)$. Analogously, from eq. (4.23), we can write:

$$\eta^*_{ij} \left.\frac{\partial x^3}{\partial s^3}\right|^*_{ij} = \eta_i \left.\frac{\partial x^3}{\partial s^3}\right|_{i} + x^3_{\mathcal{B},i} - x^3_{\mathcal{B},ij} = \eta_i \left.\frac{\partial x^3}{\partial s^3}\right|_{i} + \nabla x^3_{\mathcal{B},i} \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + o(h) \ .$$

From corollary 4.1.12, applying the divergence theorem to the constant vectors $[1, 0]$ and $[0, 1]$ we obtain:

$$0 = \int_{\partial T_i} \nu \, \mathrm{d}\sigma = \sum_{j=1}^{3} \ell_{h,ij} \nu_{ij} + \mathcal{O}\left(h^2\right) \ .$$

Thus, we add to the edge-evaluated source a term that, when summed over all the cell edges, gives a $\mathcal{O}(h^2)$ contribution:

$$\mathbf{S}^{\vec{q}}_{ij} = \left\langle \frac{1}{2} g \left(\eta^*_{ij}\right)^2 \left.\frac{\partial x^3}{\partial s^3}\right|^*_{ij} \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} - \left\langle \frac{1}{2} g \eta^2_i \left.\frac{\partial x^3}{\partial s^3}\right|_{i} \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} + \mathcal{O}\left(h^2\right) \ .$$

Then we can write:

$$
\begin{aligned}
&\mathbf{S}^{\vec{q}}_{ij} + \mathbf{S}^{\vec{q}}_{ji} = \qquad\qquad (4.25)\\
&= \left\langle \frac{1}{2} g \left[\left(\eta^*_{ij}\right)^2 \frac{\partial x^3}{\partial s^3}\bigg|^*_{ij} - \eta_i^2 \frac{\partial x^3}{\partial s^3}\bigg|_i \right] \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} \\
&\qquad + \left\langle \frac{1}{2} g \left[\left(\eta^*_{ji}\right)^2 \frac{\partial x^3}{\partial s^3}\bigg|^*_{ji} - \eta_j^2 \frac{\partial x^3}{\partial s^3}\bigg|_j \right] \mathcal{G}^{-1}_{sw}, \nu_{ji} \right\rangle_{\mathcal{G}} + \mathcal{O}\left(h^2\right) = \\
&= \left\langle \frac{1}{2} g \left[\left(\eta^*_{ij}\right)^2 \frac{\partial x^3}{\partial s^3}\bigg|^*_{ij} - \eta_i^2 \frac{\partial x^3}{\partial s^3}\bigg|_i + \left(\eta^*_{ji}\right)^2 \frac{\partial x^3}{\partial s^3}\bigg|^*_{ji} - \eta_j^2 \frac{\partial x^3}{\partial s^3}\bigg|_j \right] \mathcal{G}^{-1}_{sw}, \nu_{ij} \right\rangle_{\mathcal{G}} + \mathcal{O}\left(h^2\right) \, .
\end{aligned}
$$

Substituting the Taylor expansions written above, the term within the square brackets becomes:

$$
\begin{aligned}
&\left[\eta^*_{ij} \left(\eta_i \frac{\partial x^3}{\partial s^3}\bigg|_i + \nabla x^3_{\mathcal{B},i} \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) \right) - \eta_i^2 \frac{\partial x^3}{\partial s^3}\bigg|_i \right] \\
&\qquad + \left[\eta^*_{ji} \left(\eta_j \frac{\partial x^3}{\partial s^3}\bigg|_j + \nabla x^3_{\mathcal{B},j} \cdot (\mathbf{s}_{ji} - \mathbf{s}_j) \right) - \eta_j^2 \frac{\partial x^3}{\partial s^3}\bigg|_j \right] + \mathcal{O}\left(h^2\right) = \\
&= \left(\eta_i + \nabla \eta_i \cdot (\mathbf{s}_{ij} - \mathbf{s}_i)\right) \left(\eta_i \frac{\partial x^3}{\partial s^3}\bigg|_i + \nabla x^3_{\mathcal{B},i} \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) \right) - \eta_i^2 \frac{\partial x^3}{\partial s^3}\bigg|_i \\
&\qquad + \left(\eta_j + \nabla \eta_j \cdot (\mathbf{s}_{ji} - \mathbf{s}_j)\right) \left(\eta_j \frac{\partial x^3}{\partial s^3}\bigg|_j + \nabla x^3_{\mathcal{B},j} \cdot (\mathbf{s}_{ji} - \mathbf{s}_j) \right) - \eta_j^2 \frac{\partial x^3}{\partial s^3}\bigg|_j + \mathcal{O}\left(h^2\right) = \\
&= \left(\eta_i \nabla x^3_{\mathcal{B},i} + \eta_i \frac{\partial x^3}{\partial s^3}\bigg|_i \nabla \eta_i \right) \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + \left(\eta_j \nabla x^3_{\mathcal{B},j} + \eta_j \frac{\partial x^3}{\partial s^3}\bigg|_j \nabla \eta_j \right) \cdot (\mathbf{s}_{ji} - \mathbf{s}_j) \\
&\qquad + \nabla \eta_i \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) \nabla x^3_{\mathcal{B},i} \cdot (\mathbf{s}_{ij} - \mathbf{s}_i) + \nabla \eta_j \cdot (\mathbf{s}_{ji} - \mathbf{s}_j) \nabla x^3_{\mathcal{B},j} \cdot (\mathbf{s}_{ji} - \mathbf{s}_j) + \mathcal{O}\left(h^2\right) \, ,
\end{aligned}
$$

which, once inserted in eq. (4.25), proves eq. (4.24). □

4.2.3 Boundary conditions

The implementation of boundary conditions in our FV is obtained by specifying the appropriate edge value to the Riemann solver. In practice, on a boundary edge σ we define an outer $\mathbf{U}^{out}_{\sigma}$ and an inner $\mathbf{U}^{in}_{\sigma}$ state that form, depending on the direction of the local edge normal $\tilde{\nu}_{\mathbf{m}_\sigma}$, the left and right states for the edge Riemann solver. The boundary value $\mathbf{U}^{out}_{\sigma}$ is assumed to be given directly with respect of the reference system in the tangent plane $T_{\mathbf{m}_\sigma}\Gamma$, while internal values $\mathbf{U}^{in}_{\sigma}$ are defined on edge σ by parallel transport (4.13). In practice, following Toro [70], we implemented two types of boundary conditions: transmissive boundaries and solid reflective boundaries. Transmissive boundaries are given by:

$$
\eta^{out} = \eta^{in}, \qquad \mathbf{q}^{out} = \mathbf{q}^{in} \, ,
$$

while, solid reflective boundaries are imposed by setting:

$$\eta^{out} = \eta^{in}, \qquad \mathbf{q}^{out} = -\mathbf{q}^{in} ,$$

where $\eta^{in}, \mathbf{q}^{in}$ are the state values in the internal cell. Other more complex types of boundary conditions, such those presented in [68], require adaptation to our setting and are not considered here.

4.3 Numerical results

Our numerical tests are designed to experimentally verify the applicability of the developed scheme to the simulation of the SWE on surfaces. Specifically, we want to test selected properties of the obtained numerical solution showing the robustness of the proposed approach. To this aim, we first show convergence of the discretization of the geometric quantities on smooth surfaces, then look at experimental convergence on full scale realistic tests.

4.3.1 Description of the test cases

We use a global parametrization $x^3 = \mathcal{B}(x^1, x^2)$ of the bottom surface $\mathcal{S}_{\mathcal{B}}$, with $\mathcal{B}$ a sufficiently smooth height function, whereby we start from a regular triangulation of a rectangular subset $U \subset \mathbb{R}^2$ and move the nodes vertically on Γ. In all cases, we assume that all relevant quantities of the bottom surface are known or can be approximated at the nodes of the triangulation, and use interpolation to define needed quantities at other points, as described in the previous section. We first consider simple one-dimensional domains, a rectangular flat domain, a sloping plane and one with simple one-dimensional curvatures. Then, we consider a centrally symmetric surface and finally examine a more realistic fully three-dimensional bottom topography taken from [37].

Flat case. In this case we consider a subset $U = [0, 10] \times [0, 1] \subset \mathbb{R}^2$ and a constant height function $\mathcal{B} \equiv 0$. We tested this case on a specific example of application for which an exact analytic solution for the SW model is available, in order to compare the exact solution with our numerical results.

Sloping plane. In this case we use the rectangular subset $U = [0, 10] \times [0, 1] \subset \mathbb{R}^2$ and the height function given by:

$$\mathcal{B}(x^1, x^2) = -\frac{1}{10} x^1 + 1 .$$

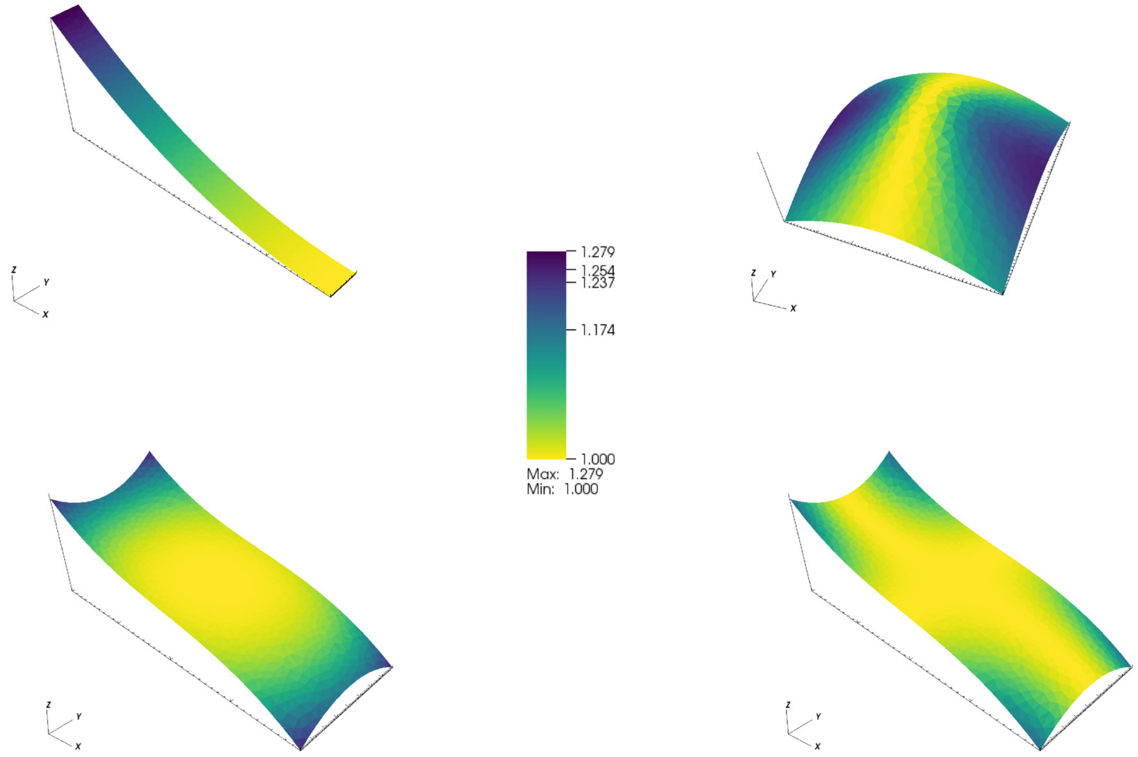

Figure 4.5: Spatial distribution of the metric coefficients $h_{(1)}$ for the parabola (top left), the hyperboloid-central-bump (top right), and of $h_{(i)}$, $i = 1, 2$, for the fully 3D surface (bottom).

Here the term $\frac{\partial x^3}{\partial s^3}$ is not equal to 1 and the curvilinear tangent vector in the s^1-direction is not unitary, yielding a non-trivial albeit constant metric tensor.

Parabola. Let U be the subset $[0, 10] \times [0, 1] \subset \mathbb{R}^2$ in a global reference frame. We use a height function that parameterizes a parabola and as given by:

$$\mathcal{B}(x^1, x^2) = \frac{1}{25}(x^1 - 10)^2 \; .$$

This case aims at verifying the effect of the curvatures in a simple one-dimensional flow, with metric coefficients that are different from one and vary along x^1 (fig. 4.5, top left).

Hyperboloid-central-bump. This test case is a complete three-dimensional benchmark. We consider a subset $U = [-3, 3] \times [-3, 3] \subset \mathbb{R}^2$ and a height function given

	$x^3_{\mathcal{B}}$				$\frac{\partial}{\partial s^1}\left(\frac{\partial x^3}{\partial s^3}\right)$			
h_ℓ	E_{L^∞}	eoc$_\ell$	E_{L^2}	eoc$_\ell$	E_{L^∞}	eoc$_\ell$	E_{L^2}	eoc$_\ell$
0.760	5.51E-03		1.79E-02		6.80E-05		1.67E-04	
0.381	1.38E-03	2.01	4.47E-03	2.01	1.70E-05	2.01	4.17E-05	2.01
0.191	3.45E-04	2.00	1.12E-03	2.00	4.26E-06	2.00	1.04E-05	2.00
0.095	8.63E-05	2.00	2.79E-04	2.00	1.07E-06	2.00	2.61E-06	2.00
0.048	2.17E-05	1.99	6.98E-05	2.00	2.66E-07	2.00	6.52E-07	2.00
0.024	5.57E-06	1.96	1.75E-05	2.00	6.65E-08	2.00	1.63E-07	2.00

Table 4.1: Parabola case: L^∞ and L^2 norms of the experimental errors on cells and order of convergence of the approximations to the bottom geometric quantities.

by:

$$\mathcal{B}(x^1, x^2) = -\frac{4}{5}\sqrt{(x^1)^2 + (x^2)^2 + 1}\,. \tag{4.26}$$

Figure 4.5, top right, shows the geometry of the bottom surface, named for simplicity hyperboloid-central-bump (HCB), and the spatial distribution of the metric coefficient in the s^1-direction, the s^2-distribution being analogous and thus not shown.

Fully 3D surface. The final test case is taken from [37] and is a fully three-dimensional bottom surface defined on the subset $U = [-10, 10] \times [-4, 4] \subset \mathbb{R}^2$ with height function expressed as:

$$\mathcal{B}(x^1, x^2) = -\frac{1}{500}(x^1)^3 - \frac{1}{100}x^1(x^2)^2\,.$$

In this case the bottom topography presents different curvature values in the two directions (fig. 4.5, bottom) and varying characteristics that influence the flow dynamics.

4.3.2 Convergence of the surface quantities

In this section we verify the accuracy of the proposed approximation to the needed surface quantities. Thus, we look at the L^∞- and L^2- norms of the difference between the approximated values and exact values at the gravity centers of cells and edges, respectively. Given the errors $E(h_\ell)$ and $E(h_{\ell+1})$ at grid levels ℓ and $\ell+1$, respectively, we calculate the experimental order of convergence eoc$_\ell$. In particular, we look at the approximation errors to some of the geometric information that enter the SW system as approximated in the FV approach, namely bottom elevation $x^3_{\mathcal{B}}$, metric coefficients $h_{(1)}$ and $h_{(2)}$, tangent vectors $\frac{\partial x^3}{\partial s^1}$ and $\frac{\partial x^3}{\partial s^3}$, and their derivative $\frac{\partial}{\partial s^1}\left(\frac{\partial x^3}{\partial s^3}\right)$. We look at these statistics for the bottom surfaces shown in fig. 4.5. The numerical results are

	$x^3_{\mathcal{B}}$				$\frac{\partial x^3}{\partial s^3}$			
h_ℓ	E_{L^∞}	eoc_ℓ	E_{L^2}	eoc_ℓ	E_{L^∞}	eoc_ℓ	E_{L^2}	eoc_ℓ
0.668	4.38E-02		9.10E-02		1.20E-02		2.43E-02	
0.334	1.17E-02	1.91	2.29E-02	1.99	3.72E-03	1.69	6.22E-03	1.96
0.167	3.05E-03	1.93	5.73E-03	2.00	9.71E-04	1.94	1.57E-03	1.99
0.083	7.81E-04	1.97	1.43E-03	2.00	2.45E-04	1.99	3.92E-04	2.00
0.042	1.97E-04	1.98	3.58E-04	2.00	6.10E-05	2.00	9.80E-05	2.00
0.021	4.96E-05	1.99	8.96E-05	2.00	1.53E-05	2.00	2.45E-05	2.00
	$\frac{\partial x^3}{\partial s^1}$				$\frac{\partial}{\partial s^1}\left(\frac{\partial x^3}{\partial s^3}\right)$			
0.668	2.68E-02		4.00E-02		6.07E-02		4.29E-02	
0.334	7.35E-03	1.86	1.02E-02	1.97	1.82E-02	1.74	1.08E-03	1.98
0.167	1.93E-03	1.93	2.57E-03	1.99	5.11E-03	1.83	2.78E-03	1.96
0.083	4.88E-04	1.98	6.44E-04	2.00	1.29E-03	1.98	6.99E-04	1.99
0.042	1.23E-04	1.99	1.61E-04	2.00	3.27E-04	1.98	1.75E-04	2.00
0.021	3.09E-05	1.99	4.03E-05	2.00	8.18E-05	2.00	4.38E-05	2.00

Table 4.2: HCB: L^∞ and L^2 norms of the experimental errors on cells and order of convergence of the approximations to the bottom geometric quantities.

	$h_{(1)}$				$h_{(2)}$			
h_ℓ	E_{L^∞}	eoc_ℓ	E_{L^2}	eoc_ℓ	E_{L^∞}	eoc_ℓ	E_{L^2}	eoc_ℓ
1.046	1.57E-03		5.96E-03		1.09E-03		3.63E-03	
0.530	4.07E-04	1.99	1.49E-03	2.05	2.99E-04	1.90	9.12E-04	2.04
0.268	1.04E-04	2.00	3.71E-04	2.02	7.90E-05	1.94	2.28E-04	2.02
0.134	2.61E-05	2.00	9.28E-05	2.01	2.03E-05	1.97	5.71E-05	2.01
0.067	6.57E-06	2.00	2.32E-05	2.01	5.13E-06	1.99	1.43E-05	2.01
0.034	1.65E-06	2.00	5.80E-06	2.00	1.29E-06	1.99	3.57E-06	2.00
	$\frac{\partial x^3}{\partial s^1}$				$\frac{\partial x^3}{\partial s^3}$			
1.046	4.26E-03		1.97E-02		9.23E-04		3.73E-03	
0.530	1.06E-03	2.05	4.92E-03	2.05	2.35E-04	2.02	9.29E-04	2.05
0.268	2.66E-04	2.02	1.23E-03	2.02	5.90E-05	2.02	2.32E-04	2.02
0.134	6.65E-05	2.01	3.07E-04	2.01	1.48E-05	2.01	5.80E-05	2.01
0.067	1.66E-05	2.01	7.68E-05	2.01	3.69E-06	2.01	1.45E-05	2.01
0.034	4.16E-06	2.00	1.92E-05	2.00	9.24E-07	2.00	3.62E-06	2.00

Table 4.3: Fully 3D surface: L^∞ and L^2 norms of the experimental errors on edges and order of convergence of the approximations to the bottom geometric quantities.

shown in table 4.1 for the parabola, in table 4.2 for the hyperboloid-central-bump, and in table 4.3 for the fully 3D surface, and invariably show second order of convergence both on cell centers and edge midpoints, in agreement with the theoretical results described in the previous section.

4.3.3 Simulations and experimental convergence

The design of the test cases needs to be aware of the assumption underlying our continuous formulation and the discretization approach. All the test cases simulate a gravity-driven fluid in a dam-break setting, without any stress tensor. The initial conditions are defined to initiate a dam-break phenomenon, with water depth in any case small enough to exclude the issue of the intersection of the local normals so that the coordinate transformation is always a diffeomorphism. Different initial conditions on water depth are considered depending on the shape of the bottom surface, while we always consider zero initial velocities. No-flow boundary conditions are imposed everywhere except at the outlet boundary, where a free outflow is enforced. The value of the time step Δt is calculated so that CFL $\approx$ 0.1. We choose a triangulation of the domain that is aligned with three cross sections uniformly distributed across the main flow directions that are used to evaluate stream-flows (i.e., discharge vs. time).

Analytic solutions to SWE are rare and each of them has a limited scope in terms of flow conditions. To first validate our scheme we are interested in solutions for a dam break problem on a flat bottom surface. Then, since no analytical solution is available for the case of variable bottom geometry if not for simple cases (i.e., planar free surface [32]), we investigate numerical stability and convergence by looking at errors calculated as differences with a fine-grid solution, assumed as reference solution. We look at L^1- and L^2- error norms and calculate the experimental order of convergence.

The well-balance property was tested on the sloping plane and the parabola test cases by verifying that a dam-break wave reached a horizontal steady-state that was stable in time. The simulations are not reported here.

Flat domain. Research on dam break started with Ritter (1892), who first studied the problem, deriving an analytic solution for the case of dam break over dry bed without friction, based on the method of characteristics. In the 1950's this solution was extended by Dressler and Whitham for dam breaks on a dry bed including the effect of friction (in particular, Chézy friction law). A few years later, Stoker generalized Ritter's solution for a wet bed downstream the dam to avoid wet/dry interactions. We consider here Stoker's solution, or dam break on a wet bed domain. See [32]. Let

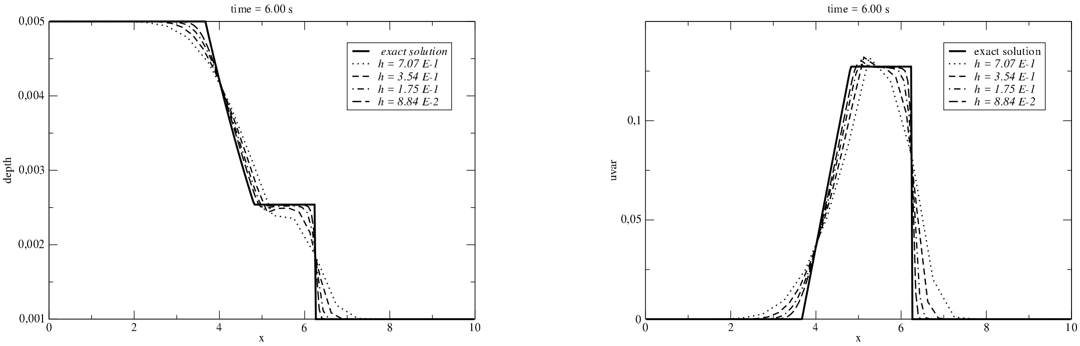

Figure 4.6: Dam-break wet bed problem: comparison between exact solution and numerical approximation on refined meshes, water depth (left panel) and velocity profile (right panel).

assume an ideal dam break on a wet domain, i.e. the dam break is instantaneous, over a flat bottom and without friction. The discretized domain is described in section 4.3.1, and we assume a regular triangulation of the domain. Essentially, this test case is a Riemann problem, whose initial conditions is defined as follows:

$$\eta(x^1, x^2, 0) = \begin{cases} 0.005 & \text{if } x^1 \leq 5 \\ 0.001 & \text{if } x^1 > 5 \end{cases} \qquad \text{and} \qquad v_i(x^1, x^2, 0) = 0.$$

Even if our solution comes out of a two-dimensional model, the problem in this case is one-dimensional, and so we can compare our approximated depth and velocity (in particular, the norm of our velocity vector) with the analytic results. Figure 4.6 shows the solution obtained for different spatial and temporal discretizations compared to the analytical solution at $t = 6$ s. The approximate solutions seem convergence to the exact one as $h \to 0$ with a monotone behavior for η, while small oscillations in the velocity magnitude are visible but tend to disappear as the mesh is refined. The large amount of numerical viscosity typically introduced by first Godunov methods is clearly noticeable.

Sloping plane. Initial conditions describe a dam located at x^1 =2.0 m with a water depth of 0.5 m upstream and 0.2 m downstream, with wall boundary conditions everywhere except at the outlet located at $x^1 = 10$ m. Figure 4.7 shows the numerically evaluated evolution of the gravity wave in terms of water depth η at $t = 0.0$ s, 0.50 s, 1.00 s and 1.50 s. The mesh used in these simulations is a Delaunay triangulation characterized by an average mesh parameter $h_\ell \approx 0.23$ m and a total of 588 surface

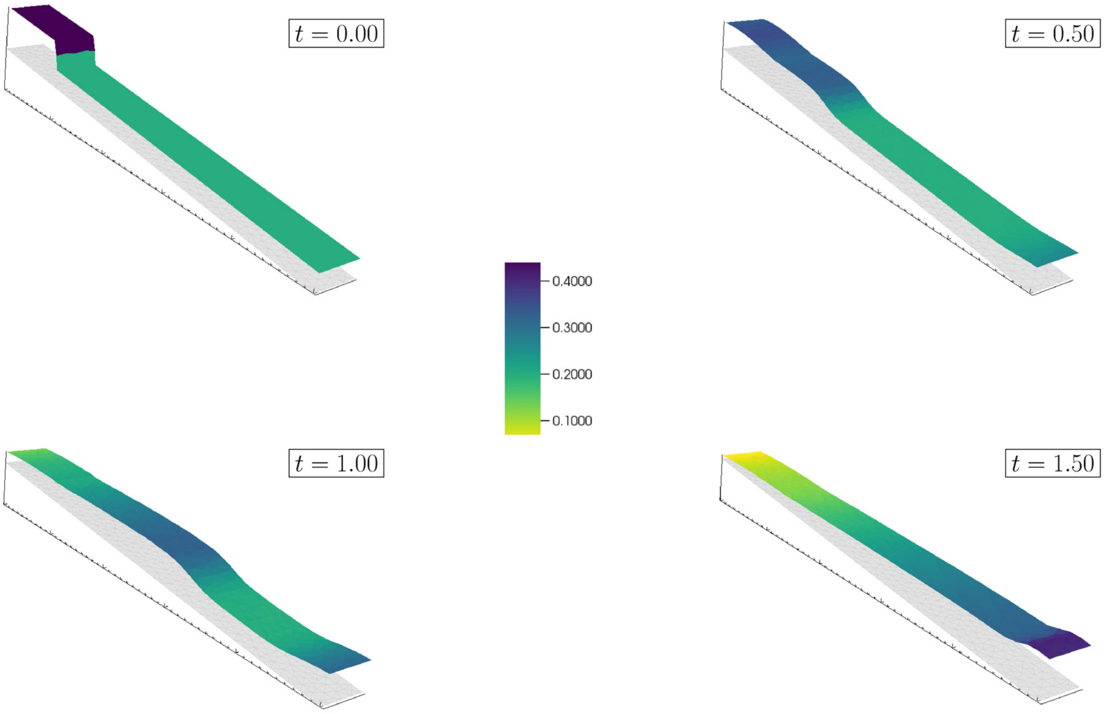

Figure 4.7: Sloping plane: evolution of the gravity wave shown both as color codes and water surface elevation, the latter with a magnification factor of 2.0.

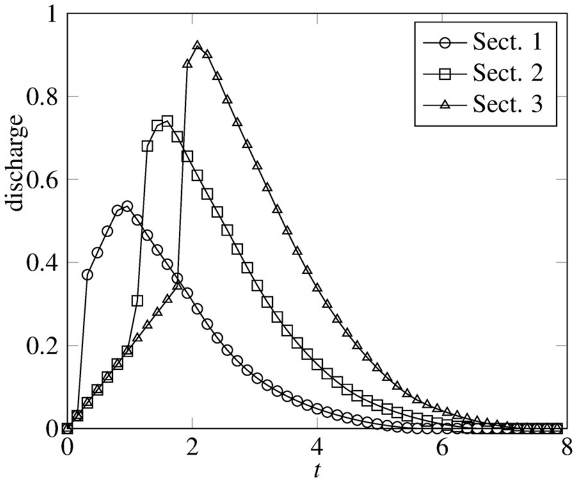

Figure 4.8: Sloping plane: stream-flow at the three cross sections.

	η				$\left\|\eta\vec{U}\right\|^2$			
h_ℓ	E_{L^1}	eoc$_\ell$	E_{L^2}	eoc$_\ell$	E_{L^1}	eoc$_\ell$	E_{L^2}	eoc$_\ell$
0.448	6.76E-02		4.78E-02		1.26E-01		9.23E-02	
0.224	4.30E-02	0.65	3.40E-02	0.49	7.97E-02	0.66	6.60E-02	0.48
0.112	2.57E-02	0.74	2.32E-02	0.55	4.69E-02	0.76	4.50E-02	0.55
0.056	1.38E-02	0.90	1.43E-02	0.70	2.48E-02	0.92	2.79E-02	0.69
0.028	5.58E-03	1.30	6.66E-03	1.10	9.91E-03	1.32	1.32E-02	1.08

Table 4.4: Sloping plane: experimental L^1- and L^2- error norms and order of convergence.

cells. Figure 4.8 shows the behavior of the discharge vs. time (stream-flow) at three sections located at x^1 =2.5 m, 5.0 m and 7.5 m. These results are compatible with the work presented in [37]. The trapezoidal rule is used to evaluate the area under the curves, showing that mass is accurately maintained throughout the simulation.

Experimental convergence is tested on a mesh sequence built by uniformly refining an initial Delaunay grid with average mesh parameter $h_0 \approx$ 0.448 m, for a total of 154 surface cells. The mesh sequence comprises 5 uniform refinements of the initial triangulation, with the finest grid having $h_5 \approx$ 0.014 m. We use the same initial conditions as before and we assume for simplicity no flow boundary conditions for all the boundary edges. Care is taken to ensure that the initial volumes are the same in the entire simulation sequence. Table 4.4 reports the convergence results at $t = 0.16$ s, a sufficiently early time to avoid reflections from the downstream boundary. The errors are calculated with respect to the reference solution obtained at $\ell = 5$. Both L^1 and L^2 relative error norms of the depth value and of the norm of the velocity magnitude are calculated (table 4.4) together with the experimental orders of convergence. As expected, first order convergence is achieved by both water depth and velocities.

Parabola. The parabola test case considers a simple one-dimensional flow where the effects of curvature in the model can be verified. We run our simulations on a grid with average mesh parameter $h = 0.26$ m, giving a total of 588 FV cells. We simulate the breaking of a dam located at x^1 =2.0 m, with initially zero velocity everywhere and water depth of 0.5 m upstream and 0.2 m downstream the dam. We use the HLL Riemann solver, but no differences are noted with respect to the HLLC RS. Figure 4.9 shows the calculated distribution of the water depth η at times $t = 0.00$ s, 0.50 s, 1.00 s and 1.50 s. The progress of the dam-break wave towards the outlet is characterized

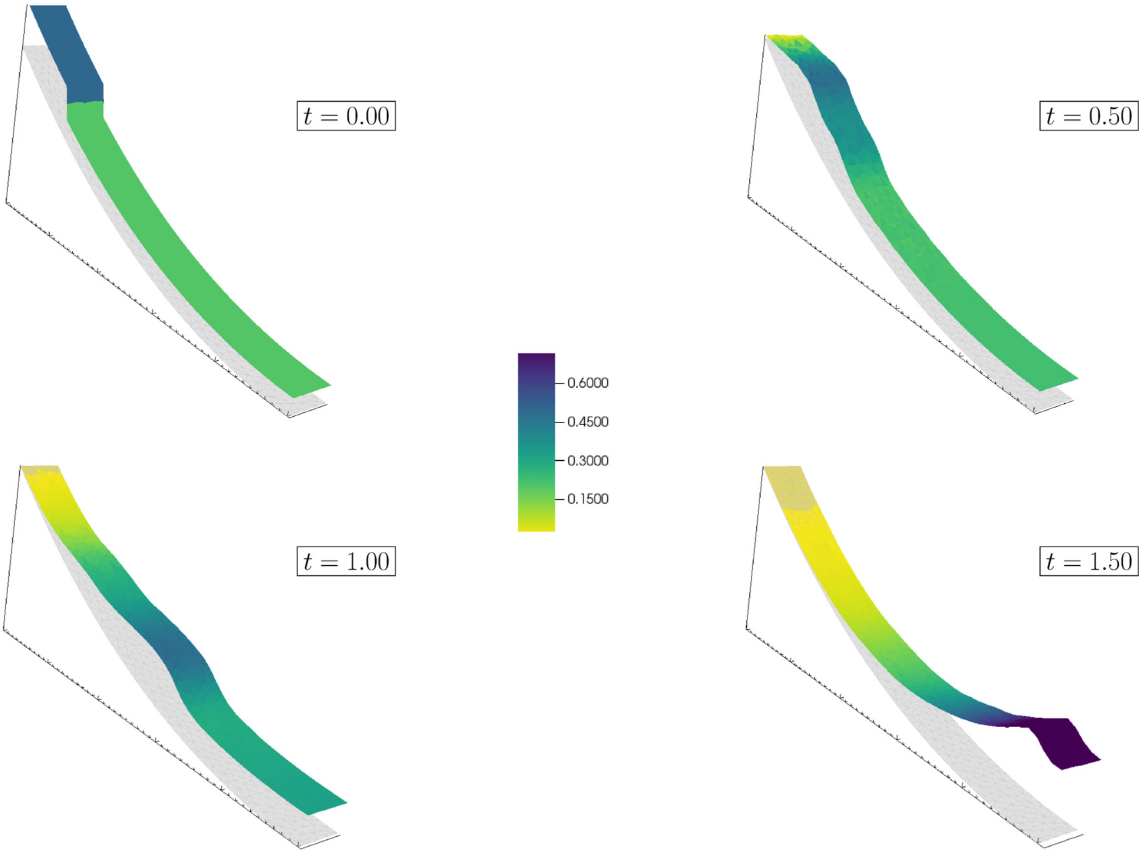

Figure 4.9: Parabola case: evolution of the gravity wave, shown both as color codes and depth elevation, the latter with a vertical magnification factor of 2.0.

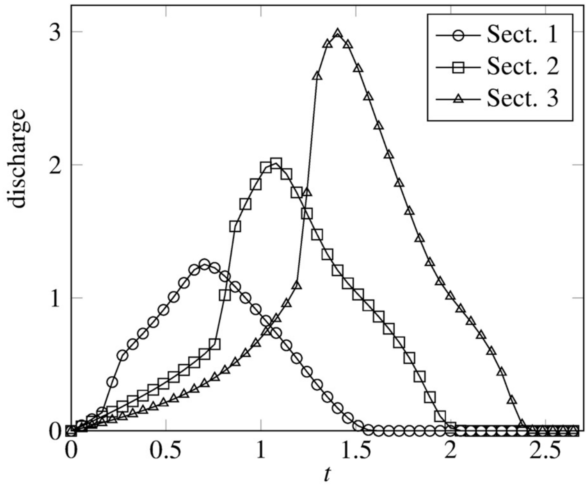

Figure 4.10: Parabola case: stream-flows at the three cross sections.

	η				$\left\|\eta\vec{U}\right\|^2$			
h_ℓ	E_{L^1}	eoc_ℓ	E_{L^2}	eoc_ℓ	E_{L^1}	eoc_ℓ	E_{L^2}	eoc_ℓ
0.494	4.14E-01		2.04E-01		1.71E-00		9.49E-01	
0.248	2.58E-01	0.69	1.42E-01	0.52	1.02E-00	0.76	6.50E-01	0.55
0.124	1.47E-01	0.81	9.29E-02	0.62	5.92E-01	0.78	4.36E-01	0.57
0.062	7.58E-02	0.96	5.50E-02	0.76	3.11E-01	0.93	2.68E-01	0.70
0.032	3.02E-02	1.32	2.51E-02	1.13	1.27E-01	1.30	1.26E-01	1.09

Table 4.5: Parabola case: experimental convergence rates of L^1 and L^2 cell-based error norms.

by a variable speed of propagation. The downwind shock initially smoothed by the numerical viscosity introduced by the 1st order solver is sharpened downstream by curvature effects, as the decreasing slope is decelerating the wave front. Also the upstream wave seems to sharpen, as evidenced by a shorter wave length at the end of the simulation. The results display some oscillations, in particular at the tail of the downstream wave, which remain always bounded and do not seem to interfere with the trailing wave. We attribute these oscillations to our treatment of the non-autonomous flux function, since they are not present in planar cases with a spatially constant metric. Figure 4.10 represents the time behavior of the simulated discharge at three channel cross sections located at $x^1 = 2.5$ m, 5.0 m and 7.5 m. Mass balance calculated a posteriori is exactly satisfied up to quadrature error.

We carried out a convergence test by using a mesh sequence starting from a coarse level characterized by $h_0 = 0.49$ m and 154 FV surface cells and composed by five mesh levels built by uniform refinement. Table 4.5 reports the L^1 and L^2-norms of the errors for the water depth and the velocity magnitude at time $t = 0.20$ s. We assume the numerical solution calculated on the finest grid ($\ell = 5$, average mesh parameter $h_5 \approx 0.016$) is the reference exact solution. First order convergence rates are attained by the L^1, L^2 error norms for both water depth and velocity.

Hyperboloid-central-bump. The following test case that presents an “almost” centrally symmetric domain is designed to verify the ability of the FV scheme to maintain symmetry on an unstructured grid. The square subset U defined in section 4.3.1 is discretized by a Delaunay triangulation with average mesh parameter $h = 0.34$ m, generating a total of 1238 FV surface cells. The nodal values are then raised using eq. (4.26). The initial conditions outline a central area of radius 0.5 m with upstream

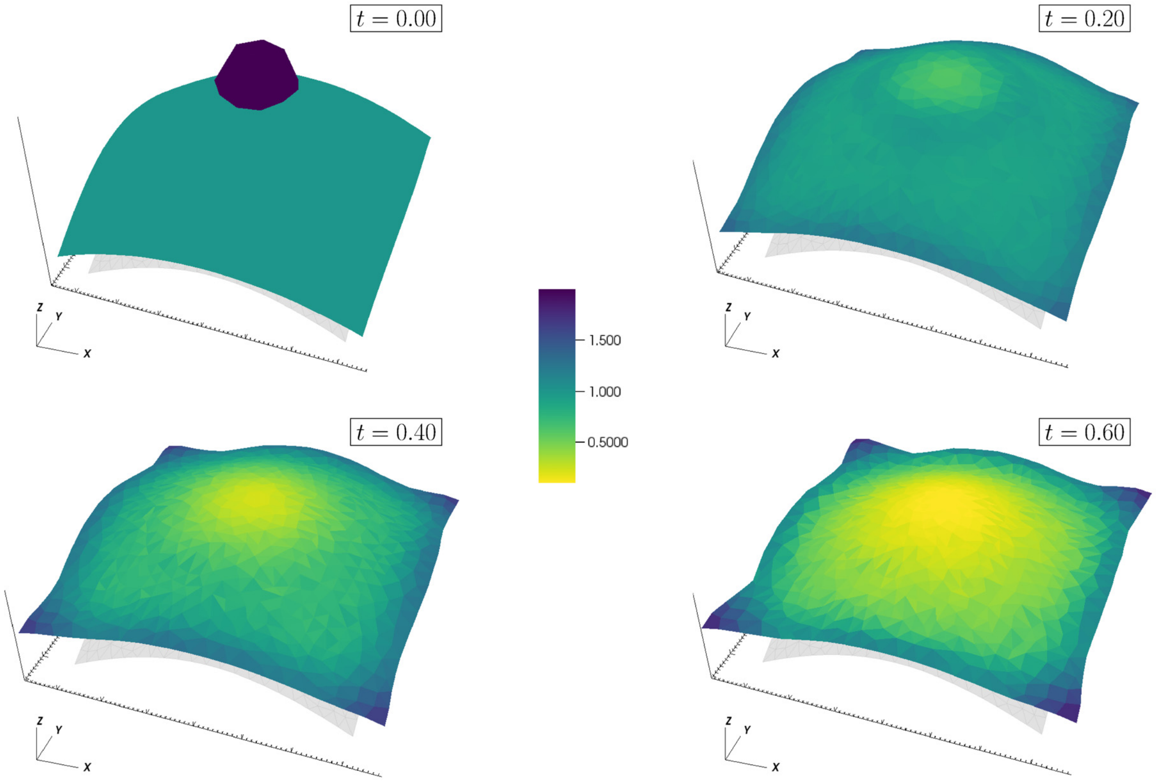

Figure 4.11: HCB: water depth at initial time ($t = 0.0$), and at $t = 0.20,\ 0.40,\ 0.60$.

water depth of 2.0 m and downstream water depth of 1.0 m, leading to an initially symmetric gravity wave. Outflow conditions are imposed on all boundaries. We use again the HLL solver.

Figure 4.11 shows the numerically evaluated evolution of the initial wave in terms of water depth η at times $t = 0.0$ s, 0.20 s, 0.40 s and 0.60 s. The initial wave moves downward with radial velocity vectors towards the outlet. The dynamics of the flow is such that the downstream portion of the initial dam-break wave accelerates faster than the upstream region because of the larger bottom slope. Some oscillations are created by the Riemann solver at the tail of the downstream wave, but these remain bounded and do not interfere with the trailing wave. Nonetheless, the numerical results shows a rather symmetric wave pattern, demonstrating the robustness of the chosen numerical approach. This is further evidenced in fig. 4.12, where the velocity vectors at $t = 0.20$ s are shown. The radial pattern of the direction of the fluid motion is maintained everywhere in the domain and at different times, showing again that the FV scheme well captures the essence of the gravity flow. The stream-flows at the three different sections, located at a radial distance from the center of 1.0 m, 1.75 m and

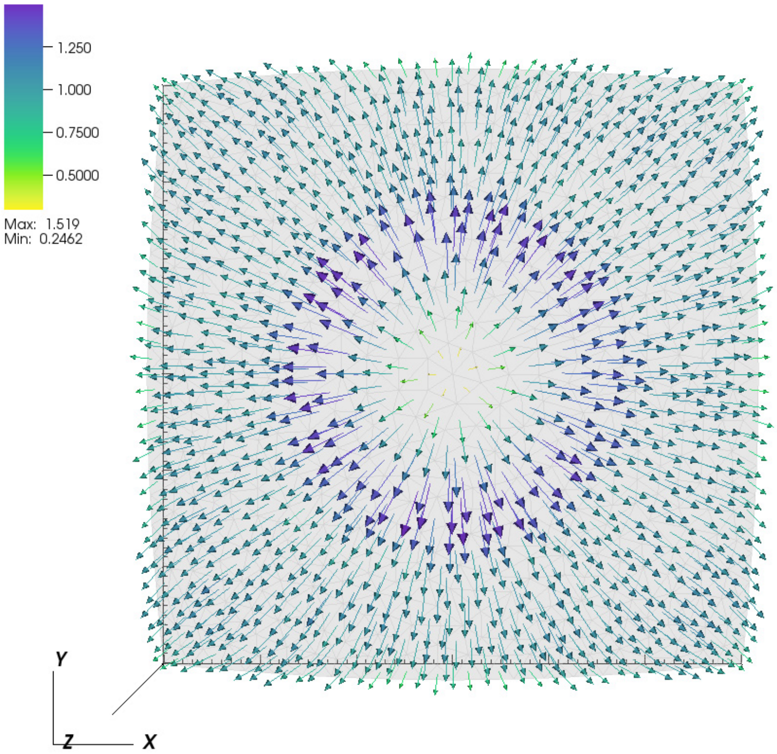

Figure 4.12: HCB: velocity vectors at $t = 0.20$.

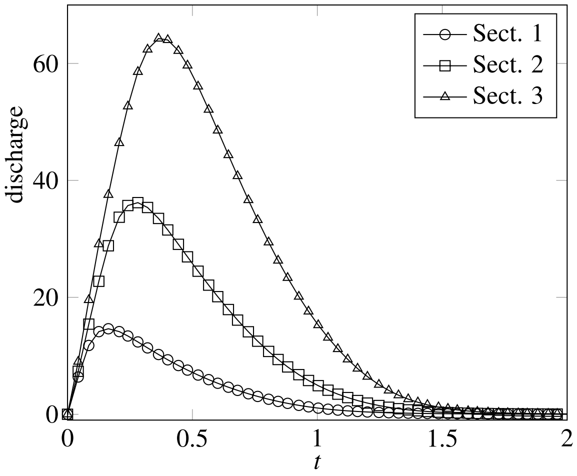

Figure 4.13: HCB: stream-flows at the three control sections.

	η				$\left\|\eta\vec{U}\right\|^2$			
h_ℓ	E_{L^1}	eoc$_\ell$	E_{L^2}	eoc$_\ell$	E_{L^1}	eoc$_\ell$	E_{L^2}	eoc$_\ell$
0.560	1.16E+00		3.05E-01		5.42E+00		1.32E+00	
0.281	7.20E-01	0.69	2.02E-01	0.60	2.98E+00	0.87	7.56E-01	0.81
0.140	4.03E-01	0.84	1.27E-01	0.67	1.49E+00	0.99	4.44E-01	0.77
0.070	2.13E-01	0.92	7.32E-02	0.79	8.61E-01	0.80	3.16E-01	0.49
0.035	9.54E-02	1.16	3.66E-02	1.00	3.22E-01	1.42	1.32E-01	1.26

Table 4.6: HCB: experimental convergence rates of L^1 and L^2 cell-based error norms.

2.5 m are shown in fig. 4.13.

For the convergence test we start from a coarse grid with average mesh parameter $h_0 = 0.56$ m and 424 FV surface cells. The mesh sequence is obtained by uniform refinements with a total of 6 mesh levels. The reference solution is obtained at level $\ell = 5$ (average mesh parameter $h = 0.018$). Table 4.6 presents L^1 and L^2 error norms and related experimental orders of convergence for the solutions at time $t = 0.08$ s. Consistently with the previous test cases, the expected first order convergence is achieved in both norms.

Fully 3D surface. The final test case considers a bottom surface with variable curvature. The discretization of U is obtained again with a Delaunay triangulation with average mesh parameter $h = 0.62$ m, in this case generating a total of 1656 FV surface cells. The initial conditions consider a uniform water depth of 2.0 m upstream of x^1 = -8.5 m, and 1.0 m downstream. We would like to note that the choice of initial conditions of a 2.0 m deep reservoir avoids the singularities of the coordinate transformation by ensuring that the water depth is sufficiently shallow, so that the local normals to the bottom surface do not intersect within the fluid layer. No flow conditions are imposed in all boundaries, with an outlet located at x^1 = 10 m. The approximate HLL Riemann solver is used.

Figure 4.14 describes the numerically evaluated water depth η at times $t = 0.0$ s, 0.80 s, 1.60 s and 2.40 s. The evolution of the profile is similar to what reported in Fent et al. [37]. Water accumulates within the central portion of the first bowl and then disperses towards the lateral boundaries in the concave region, until it finds the impermeable lateral walls. At the end water exits from the downstream edge, where the outlet is located. Figure 4.15 shows stream-flows at the three different sections located at x^1 = -5.0 m, 0.0 m and 5.0 m. Again, the results compare well with those

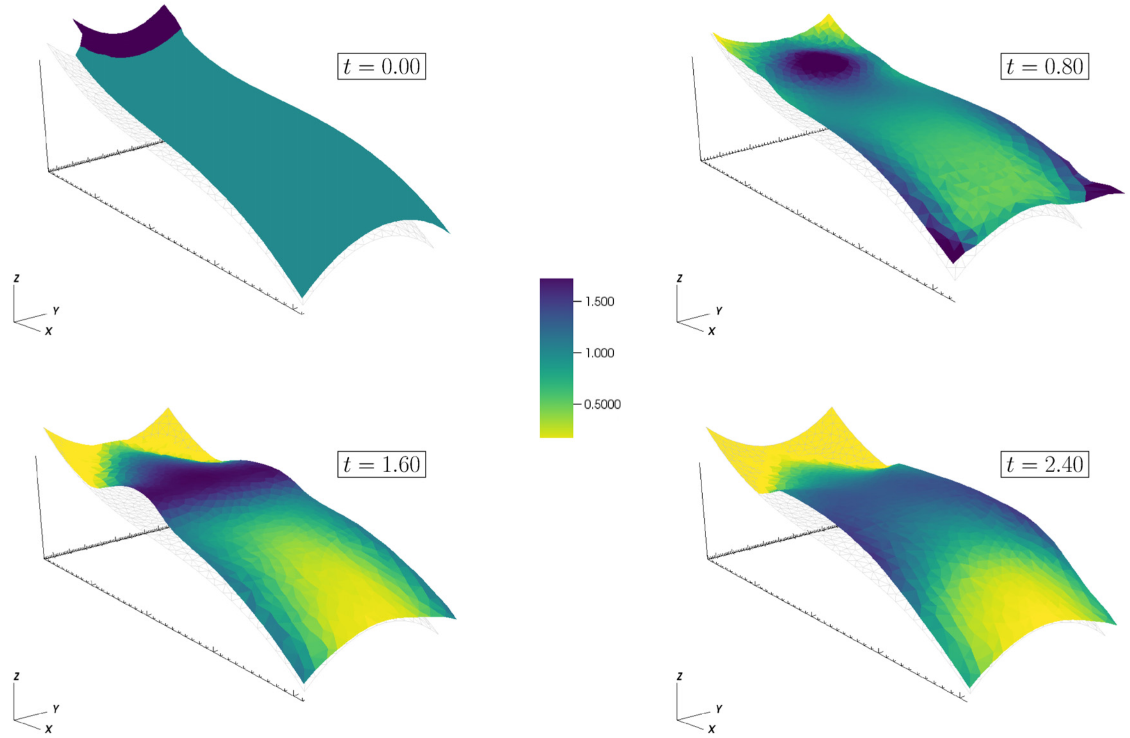

Figure 4.14: Fully 3D surface: water depth evolution of initial wave, shown both as color codes and depth elevation at initial time ($t = 0.0$) and at $t = 0.80,\ 1.60,\ 2.40$.

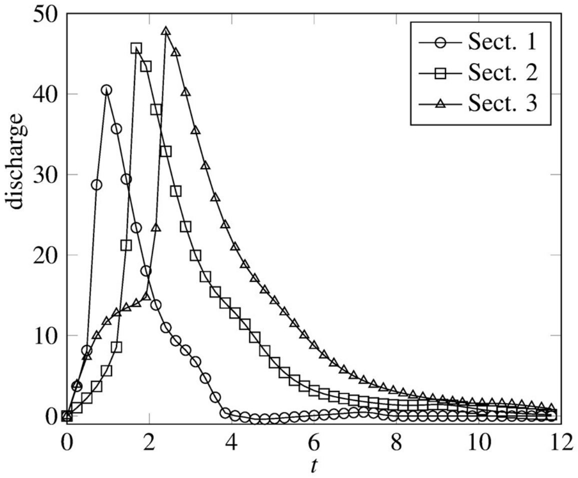

Figure 4.15: Fully 3D surface: stream-flows at the three control sections.

	η				$\left\|\eta\vec{U}\right\|^2$			
h_ℓ	E_{L^1}	eoc_ℓ	E_{L^2}	eoc_ℓ	E_{L^1}	eoc_ℓ	E_{L^2}	eoc_ℓ
1.113	6.95E+00		8.58E-01		3.14E+01		4.19E+00	
0.564	4.15E+00	0.76	5.48E-01	0.66	1.74E+01	0.86	2.54E+00	0.74
0.284	2.36E+00	0.82	3.47E-01	0.66	9.30E+00	0.92	1.53E+00	0.73
0.142	1.22E+00	0.96	2.03E-01	0.78	4.55E+00	1.04	8.77E-01	0.81
0.071	4.97E-01	1.29	9.32E-01	1.13	1.76E+00	1.37	3.99E-01	1.14

Table 4.7: Fully 3D surface: experimental convergence rates of L^1 and L^2 cell-based error norms.

obtained by Fent et al. [37].

The mesh sequence used to test convergence of the FV scheme starts with coarse grid with average mesh parameter $h_0 = 1.11$ m, for a total of a 480 FV surface cells, uniformly refined 5 times to yield 6 meshes. The finest, used for the reference solution, is characterized by h_5= 0.036 m and 491520 FV surface triangles. The initial conditions defined on the coarsest mesh $\ell = 0$ are projected on each mesh level using the local normals taking care that the initial water volume and the location of the dam are consistently the same across the entire mesh sequence.

Table 4.7 reports the L^1 and L^2-error norms of the depth and of the velocity magnitude at $t = 0.20$ s together with the experimental order of convergence. The same behavior as for the previous test is observed, with optimal rates being reached by the water depth and the velocity field.

5

Intrinsic Discontinuous Galerkin Scheme

The promising results obtained in the previous chapters lead us to consider higher order discretizations in the numerical resolution of eq. (3.23). An option would be to set up a second order finite volume scheme. This will be addressed as a future work, since the reconstruction of gradients at the interface of surface cells needs additional investigation. We decided to consider here the implementation of a discontinuous Galerkin scheme. First, the use of higher-order polynomials is naturally built into the DG method. Instead of computing these higher-order terms through some sort of ad hoc post-processing procedure, they are defined through the variational equation. Second, diffusive terms can be easily incorporated into the method, contrary to the FV method that provides only large-stencil mechanisms for dealing with second-order derivatives [see e.g. 12]. Furthermore, the DG method allows for the use of non-conforming grids, i.e., grids with non-matching faces, without the use of hanging nodes or mortar spaces. This feature could be very useful in dealing with complicated geometries and adaptive meshes.

We consider here the so-called Runge-Kutta Discontinuous Galerkin (RKDG) scheme, a combination of DG discretization in space with an explicit Runge-Kutta time-marching algorithm. The aim of this chapter is to set the scheme in the geometric framework. We first address the linear scalar transport equation and define the variational formulation (in its geometric form), the discrete function spaces and the quadrature rules. Then we extend this framework to the ISWE approach. We build our construction on the development of the FV scheme described in the previous chapter. Thus, for the ISWE system, we follow the work of Aizinger and Dawson [3] that generalizes and extends the Godunov FV method [21] in several ways, thus forming an ideal starting point. Due to the limitations of our approximation of the surface quantities, as presented in section 4.1, we are able to exploit only up to second order discretization schemes, and thus we will describe DG with linear basis functions.

5.1 The intrinsic scalar transport equation

The aim of this section is to present the discontinuous Galerkin method applied to a scalar transport equation defined on a surface, eq. (2.12). The starting hypotheses are the same as in the previous chapter. We assume that the surface PDE is defined on a compact subset of the bottom surface, $\Gamma \subset \mathcal{S}$, and that a well-defined curvilinear boundary, denoted by $\partial\Gamma = \partial\overline{\Gamma}$, exists. We work on the compact set Γ and we assume that the triangulations $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$ are given and have the properties presented in section 4.1. We assume Γ is fixed in time. Then, denoting with $u : \Gamma \times [0, t_f] \to \mathbb{R}$ the scalar conserved variable, and $\mathbf{w} : \Gamma \times [0, t_f] \to \mathbb{R}^2$ a given velocity field everywhere tangent to the surface, we recall the transport equation:

$$\frac{\partial u}{\partial t} + \nabla_{\mathcal{G}} \cdot (\mathbf{w}\, u) = 0\,. \tag{5.1}$$

The derivation of the DG scheme proceeds as follows. First, we discretize the equation in space using the discontinuous Galerkin approach. For each time $t \in [0, t_f]$, the approximated solution u_h is sought in the finite element space of discontinuous functions:

$$\mathcal{V}_h^{\Gamma} = \{ v \in L^{\infty}(\Gamma) \ : \ v|_T \in V(T) \quad \forall\, T \in \mathcal{T}(\Gamma) \}\,, \tag{5.2}$$

where $V(T)$ is the so-called local space. In this work we assume $V(T) = \mathcal{P}_1(T)$, the space of affine functions defined in T. We can test eq. (5.1) with $v_h \in \mathcal{V}_h^{\Gamma}$ and integrate in space. Recalling that $\Gamma = \cup_{i=1}^{N_T} T_i$, we can localize the integration over each single cell $T \in \mathcal{T}(\Gamma)$, i.e., the following equation must be valid for all $T \in \mathcal{T}(\Gamma)$:

$$\int_T \partial_t\, u\, v_h \,\mathrm{d}\mathbf{s} + \int_T \nabla_{\mathcal{G}} \cdot (\mathbf{w}u)\ v_h \,\mathrm{d}\mathbf{s} = 0\,.$$

Then, we substitute the exact solution u with its approximation u_h and apply the divergence theorem:

$$\int_T \partial_t\, u_h\, v_h \,\mathrm{d}\mathbf{s} + \int_{\partial T} \langle \mathbf{w}u_h, \nu \rangle_{\mathcal{G}}\ v_h \,\mathrm{d}\sigma - \int_T \langle \mathbf{w}u_h, \nabla_{\mathcal{G}}\, v_h \rangle_{\mathcal{G}} \,\mathrm{d}\mathbf{s} = 0\,,$$

where ν is the vector tangent to the surface and normal to the boundary of the cell, pointing outward of T. Note that, when we consider the union of all triangles, the trace of u_h is discontinuous on ∂T and the scalar product $\langle \mathbf{w} u_h, \nu \rangle_{\mathcal{G}}$ does not have a precise meaning. Thus, we replace the scalar product by a numerical flux function $\tilde{f}_{\partial T}(u)$, which takes into consideration the jumps of u_h. The value of the numerical flux $\tilde{f}_{\partial T}(u_h(\mathbf{p}))$ at the point $\mathbf{p} \in \partial T$ depends on the two values at either sides of ∂T of the solution in $\mathbf{p}$:

$$\tilde{f}_{\partial T}(u(\mathbf{p})) = \tilde{f}(u_L(\mathbf{p}), u_R(\mathbf{p})) .$$

This will be suitably discussed later.

We can now introduce the weak discrete formulation of eq. (5.1):

Problem 5.1. Find $u_h \in \mathcal{V}_h^{\Gamma}$ such that

$$\frac{d}{dt} \int_T u_h \, v_h \, \mathrm{d}\mathbf{s} = - \int_{\partial T} \tilde{f}_{\partial T}(u_h) \, v_h \, \mathrm{d}\sigma \, + \int_T \langle \mathbf{w} u_h, \nabla_{\mathcal{G}} \, v_h \rangle_{\mathcal{G}} \, \mathrm{d}\mathbf{s} \, ,$$

for any function $v_h \in \mathcal{V}_h^{\Gamma}$.

Note that the formulation of the problem is completely intrinsic. We would like to remain within this setting as much as possible. Approximation issues arise when we need to practically compute quantities. As proceeded for the FV method, we devise appropriate steps in order to set up the entire resolution algorithm. In practice, we need to define: i) basis functions for $\mathcal{V}_h^{\Gamma}$; ii) quadrature rules; iii) numerical fluxes on edges; iv) time discretization for the resulting ODE. We will give a full characterization of all these steps in the next sections.

5.1.1 Basis functions

We start this paragraph by underlying that the choice of basis and degrees of freedom does not affect the algorithm, which is completely determined by the choice of the function space $\mathcal{V}_h^{\Gamma}$, the numerical fluxes, the quadrature rules and the time discretization. However, a suitable choice of basis and degrees of freedom may simplify the implementation and computations, and guarantee stability of the algorithm.

We examine the case of first order affine functions on each cell, and we define the basis functions $\varphi_1, \varphi_2, \varphi_3$ spanning $\mathcal{V}_h^{\Gamma}$ by means of the classical interpolation property:

$$\varphi_j \in \mathcal{V}_h^{\Gamma} \; , \quad \varphi_j(\mathbf{p}_i) = \delta_{ij} \quad i, j = 1, 2, 3 \; ,$$

where $\mathbf{p}_i \in \Gamma$ are the vertices of the cell, for each cell of $\mathcal{T}(\Gamma)$, see fig. 5.1.

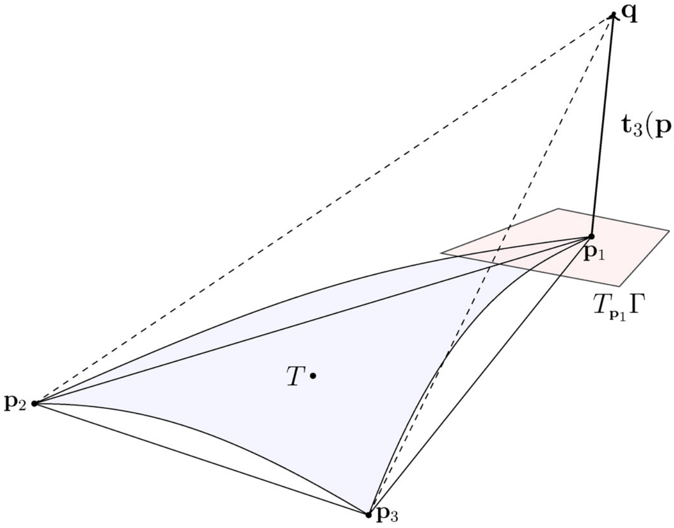

Figure 5.1: Geometrical representation of the basis function φ_1 of the cell T.

The definition of φ_j proceeds as follows. Given the global coordinates $\mathbf{x}(\mathbf{p})$ of $\mathbf{p} \in T$, we can define the affine function $\tilde{\varphi}_j(\mathbf{x}) = \tilde{a} + \tilde{b}x^1 + \tilde{c}x^2 + \tilde{d}x^3$ as a function in $\mathbb{R}^3$. Composition of $\tilde{\varphi}_j$ with the surface parametrization expresses the basis function in local coordinates:

$$\varphi_j(s^1, s^2) = \tilde{\varphi}_j \circ \phi(s^1, s^2) \; .$$

Since the integrals in problem 5.1 are calculated by means of Gaussian quadrature rules, we need to evaluate the basis functions at quadrature points $\mathbf{p}_j$ inside the cell T and on the edge σ, for the boundary integrals. To properly define the basis functions we make explicit use of the parametrization $\phi_{\mathbf{p}_j}$ at point $\mathbf{p}_j$. However, as will be seen in the practical example, only the knowledge of the tangent plane $T_{\mathbf{p}_j}\Gamma$ is actually necessary. Assuming $T \subset \phi_{\mathbf{p}_j}(U)$ for some open set $U \subset \mathbb{R}^2$, eq. (2.1) allows us to work on the linear approximation of our surface, i.e., $\phi_{\mathbf{p}_j}(T) = T_{\mathbf{p}_j}\Gamma + \mathcal{O}(h^2)$. As a consequence, the local basis functions $\varphi_j(s^1, s^2) \approx \tilde{\varphi}_j \circ \phi_{\mathbf{p}_j}(s^1, s^2)$ obtained by neglecting the higher order terms span the set of affine functions in T.

Example 5.1. We consider here a practical example of calculation of the basis function φ_1 related to the first node $\mathbf{p}_1$ of T. The coefficients of $\tilde{\varphi}_1$ are the solution of the following linear system:

$$\begin{bmatrix} 1 & x^1_{\mathbf{p}_1} & x^2_{\mathbf{p}_1} & x^3_{\mathbf{p}_1} \\ 1 & x^1_{\mathbf{p}_2} & x^2_{\mathbf{p}_2} & x^3_{\mathbf{p}_2} \\ 1 & x^1_{\mathbf{p}_3} & x^2_{\mathbf{p}_3} & x^3_{\mathbf{p}_3} \\ 1 & x^1_{\mathbf{q}} & x^2_{\mathbf{q}} & x^3_{\mathbf{q}} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} ,$$

where $\mathbf{q} = \mathbf{p}_1 + \mathbf{t}_3(\mathbf{p}_1)$, and $\mathbf{t}_3(\mathbf{p}_1)$ the unitary normal to the surface in $\mathbf{p}_1$ (see fig. 5.1). Then, given a point $\mathbf{p}_j \in T$ and the normal vector to the surface at that point, it is possible to compute the expression of $T_{\mathbf{p}_j}\Gamma$. Let us consider the same point $\mathbf{p}_1$, the normal vector to the surface $\mathbf{t}_3(\mathbf{p}_1) =: \mathbf{N} = (\mathbf{N}^1, \mathbf{N}^2, \mathbf{N}^3)$, then the tangent plane passing through $\mathbf{p}_1$ and orthogonal to $\mathbf{N}$ is described by:

$$\mathbf{N}^1(x^1 - x^1_{\mathbf{p}_1}) + \mathbf{N}^2(x^2 - x^2_{\mathbf{p}_1}) + \mathbf{N}^3(x^3 - x^3_{\mathbf{p}_1}) = 0\,.$$

Substituting the value of x^3 of the points of the plane into the expression of $\tilde{\varphi}_1$, we obtain:

$$\tilde{\varphi}_1(\mathbf{x}) = \tilde{a} + \tilde{b}x^1 + \tilde{c}x^2 + \tilde{d}\left(\frac{-\mathbf{N}^1(x^1 - x^1_{\mathbf{p}_1}) - \mathbf{N}^2(x^2 - x^2_{\mathbf{p}_1})}{\mathbf{N}^3} + x^3_{\mathbf{p}_1}\right),$$

and collecting terms:

$$a = \tilde{a} + \tilde{d}\left(\frac{\mathbf{N}^1 x^1_{\mathbf{p}_1} + \mathbf{N}^2 x^2_{\mathbf{p}_1}}{\mathbf{N}^3} + x^3_{\mathbf{p}_1}\right), \quad b = \tilde{b} - \tilde{d}\,\frac{\mathbf{N}^1}{\mathbf{N}^3}\,, \quad c = \tilde{c} - \tilde{d}\,\frac{\mathbf{N}^2}{\mathbf{N}^3}\,,$$

which are the new coefficients of φ_1 in the LCS.

We are now able to write every function v_h in the functional space $\mathcal{V}_h^\Gamma$ as:

$$v_h(\mathbf{s}) = \sum_{j=1}^{3} v_j \varphi_j(\mathbf{s}) \qquad \mathbf{s} \in T\,, \tag{5.3}$$

with v_j the j-th coefficient and by the above interpolation properties they coincide with $v_h(\mathbf{s}_j)$. In particular, we express our approximate solution u_h inside the cell T as follows:

$$u_h(\mathbf{s}, t) = \sum_{i=1}^{3} u_i(t)\varphi_i(\mathbf{s})\,,$$

where the degrees of freedom $u_i(t)$ are the values of the numerical solution at the nodes of the cell. Then, the weak formulation takes the form:

$$\sum_{i=1}^{3}\left(\frac{du_i}{dt}\int_T \varphi_i\varphi_k\,\mathrm{d}\mathbf{s}\right) = -\sum_{j=1}^{3}\int_{\sigma_j}\tilde{f}_{j,T}(u_h)\varphi_k\,\mathrm{d}\sigma + \sum_{i=1}^{3}\int_T u_i\varphi_i\,\langle\mathbf{w}, \nabla_{\mathcal{G}}\,\varphi_k\rangle_{\mathcal{G}}\,\mathrm{d}\mathbf{s}\,, \tag{5.4}$$

where now $\tilde{f}_{j,T}(u_h)$ denotes the numerical flux along the j-th edge of cell T. Moreover, for each cell $T \in \mathcal{T}(\Gamma)$ we can define, for $i, k = 1, 2, 3$,

$$\mathbf{M}_{ik} = \int_T \varphi_i\varphi_k\,\mathrm{d}\mathbf{s} \tag{5.5}$$

$$\mathbf{R}_k(\mathbf{U}_h) = -\sum_{j=1}^{3}\int_{\sigma_j}\tilde{f}_{j,T}(u_h)\varphi_k\,\mathrm{d}\sigma + \sum_{i=1}^{3}\int_T u_i\varphi_i\,\langle\mathbf{w}, \nabla_{\mathcal{G}}\,\varphi_k\rangle_{\mathcal{G}}\,\mathrm{d}\mathbf{s}\,, \tag{5.6}$$

respectively the local mass matrix and the vector of the right-hand-side, and write the following system of ODE:

$$\mathbf{M}\,\frac{d\mathbf{U}_h}{dt} = \mathbf{R}(\mathbf{U}_h)\,, \tag{5.7}$$

where we denote with $\mathbf{U}_h = \{u_i(t)\}$ the vector of the degrees of freedom.

The formulation then proceeds by the appropriate definition of the quadrature rules on the surface and the approximation of the parametrization at the quadrature points $\mathbf{p}_j$, or equivalently of the tangent plane $T_{\mathbf{p}_j}\Gamma$, needed to define the numerical fluxes along the edges. It is completed by the definition of a time discretization scheme.

5.1.2 Quadrature rules and Gaussian points

Theoretical analysis of the method [29] shows that, using $\mathcal{P}_1$ basis functions, optimal (second order) convergence is obtained if the quadrature rule for the edges of the elements is exact for polynomials of degree 3, and the quadrature rules for the interior of the elements is exact for polynomials of degree 2. We replace the space integrals by Gaussian quadrature rules:

$$\int_T f(\mathbf{s})\,\mathrm{d}\mathbf{s} \approx \mathcal{A}_T \sum_{m=1}^{3} \omega_m f(\mathbf{p}_m)\ ,$$

$$\int_\sigma f(\mathbf{s})\,\mathrm{d}\sigma \approx \ell_\sigma \sum_{l=1}^{2} \omega_l f(\mathbf{p}_l)\ .$$

Concerning the interior of the cell T, we need a rule based on three points and, in particular, we apply the three mid-point rule:

$$\int_T f(\mathbf{s})\,\mathrm{d}\mathbf{s} \approx \frac{\mathcal{A}_{T_h}}{3} \sum_{m=1}^{3} \tilde{f}(\mathbf{m}_{\sigma_m})\ ,$$

where $\mathbf{m}_{\sigma_m}$ are the mid-points of the edges, and $\tilde{f}(\mathbf{m}_{\sigma_m})$ denote the evaluation of the surface function at the edge midpoints. We refer to section 4.1 for the explicit definition of the tangent vectors and the relative metric tensor at the mid-point of edges, so we have already the complete surface information to perform computations. Hence, the following proposition is easily proved.

Proposition 5.1.1. *Given a Lipschitz regular function f defined on the surface triangle T we have:*

$$\int_T f(\mathbf{s})\,\mathrm{d}\mathbf{s} = \frac{\mathcal{A}_{T_h}}{3} \sum_{i=1}^{3} \tilde{f}(\mathbf{m}_{\sigma_i}) + \mathcal{O}\Big(h_T^2\Big)\ ,$$

where $\tilde{f}(\mathbf{m}_{\sigma_i})$ denote the approximate value of the surface function at the edge mid-points, in the approximated LCS.

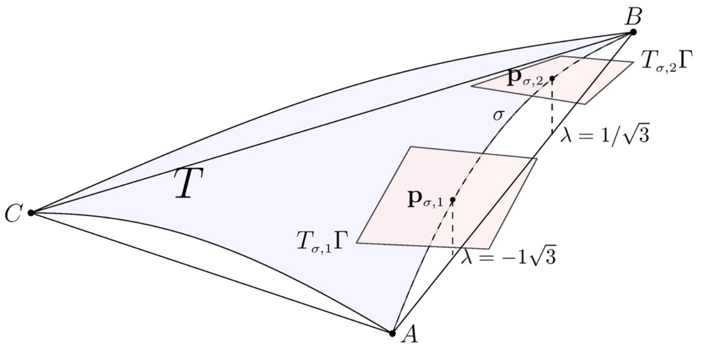

Figure 5.2: Tangent planes at the Gaussian points on the edge σ.

For the edge integrals we need two quadrature points and we apply the two point Gaussian quadrature rule:

$$\int_{-1}^{1} f(\mathbf{s})\,\mathrm{d}\mathbf{s} \approx f\left(-\tfrac{1}{\sqrt{3}}\right) + f\left(\tfrac{1}{\sqrt{3}}\right) ,$$

which is based on a parametrization of the edge $\sigma(\lambda)$ with $\lambda \in [-1, 1]$. In our case, we use an approximate parametrization of the form:

$$\underline{\sigma}(\lambda) = \begin{cases} s^1(\lambda) = \frac{\left(s^1_B - s^1_A\right)(1+\lambda)}{2} + s^1_A \\ s^2(\lambda) = \frac{\left(s^2_B - s^2_A\right)(1+\lambda)}{2} + s^2_A \\ s^3(\lambda) = \mathcal{B}\left(s^1(\lambda), s^2(\lambda)\right) \end{cases} ,$$

where again the third component of this parametric curve is the vertical projection of the chord onto the surface, like we did in eq. (4.8).

The computation proceeds following the ideas of section 4.1, by performing a direct approximation of the LCS (or equivalently of the tangent plane) at the quadrature points (fig. 5.2). However, in this case we need to apply a quadratic interpolation, and thus we need to use the tangent vectors $\mathbf{t}_i(A)$ and $\mathbf{t}_i(B)$ at the nodes, and $\mathbf{t}_i(\mathbf{m}_\sigma)$ at the mid-point of the edge. The needed information at the mid-point of the edges is assumed given data of the problem. In practice, we consider the vertical projection onto the bottom surface of the mid-point of the chord between the nodes A, B, and compute the information from the parametrization of the bottom surface. Note that the use of the vertical projection in place of the normal projection provides a second order accurate approximation of the geometry (see proposition 4.1.10 and the discussion contained in the corresponding section). As a consequence, our quadrature rule cannot enjoy better accuracy. Thus the edge quadrature, which in principle is exact up to polynomial of third order, is only second order accurate, and the following proposition is easily proved.

Proposition 5.1.2. *Given a Lipschitz regular function f defined on the geodesic edge σ we have:*

$$\int_\sigma f(\mathbf{s})\,\mathrm{d}\sigma = \frac{\ell_{\sigma_h}}{2}\sum_{i=1}^{2}\tilde{f}(\mathbf{p}_{\sigma,i}) + \mathcal{O}\left(\ell^2_{\sigma_h}\right) ,$$

where $\tilde{f}(\mathbf{p}_{\sigma,i})$ denote the approximate value of the surface function at the edge Gaussian points in the approximated LCS system.

Remark 5.1. Note that, in our particular case, $f(\mathbf{s}) = \langle \mathbf{w}u_h, \nu\rangle_{\mathcal{G}}$ is the numerical flux along the local normal to the edge. As seen in the previous chapter, we are introducing a number of second order approximations in the calculation of $\tilde{f}(\mathbf{s})$. Indeed, the first approximation is done in calculating the tangent plane at the central interpolation point, which is evaluated using the vertical instead of the normal projection. As a consequence, the metric forms computed at the two interpolated points is only second order accurate. Then, the edge normal vectors $\nu(\mathbf{p}_{\sigma,i})$ are approximated with a first order error. Following the proof of proposition 4.1.11, global second order accuracy is obtained, as stated in the proposition.

Explicitly, the procedure of computing geometric information over the edge is the following. The tangent vectors at the two Gaussian points $\mathbf{p}_{\sigma,i}$, $i = 1, 2$, are given by:

$$\tilde{\mathbf{t}}_1(\mathbf{p}_{\sigma,i}) = \mathbf{t}_1(\mathbf{m}_\sigma) + \frac{\mathbf{t}_1(B) - \mathbf{t}_1(A)}{2}\lambda_i + \left(\frac{\mathbf{t}_1(A) + \mathbf{t}_1(B)}{2} - \mathbf{t}_1(\mathbf{m}_\sigma)\right)\lambda_i^2 ,$$

$$\tilde{\mathbf{t}}_2(\mathbf{p}_{\sigma,i}) = \mathbf{t}'_2 - \frac{\left\langle \mathbf{t}'_2, \tilde{\mathbf{t}}_1\right\rangle}{\left\langle \tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_1\right\rangle}\tilde{\mathbf{t}}_1 ,$$

$$\text{with}\quad \mathbf{t}'_2(\mathbf{p}_{\sigma,i}) = \mathbf{t}_2(\mathbf{m}_\sigma) + \frac{\mathbf{t}_2(B) - \mathbf{t}_2(A)}{2}\lambda_i + \left(\frac{\mathbf{t}_2(A) + \mathbf{t}_2(B)}{2} - \mathbf{t}_2(\mathbf{m}_\sigma)\right)\lambda_i^2 ,$$

$$\tilde{\mathbf{t}}_3(\mathbf{p}_{\sigma,i}) = \frac{\tilde{\mathbf{t}}_1 \wedge \tilde{\mathbf{t}}_2}{\left\|\tilde{\mathbf{t}}_1\right\|\left\|\tilde{\mathbf{t}}_2\right\|} ,$$

where $i = 1, 2$ and $\lambda_1 = -1/\sqrt{3}, \lambda_2 = 1/\sqrt{3}$. The final vectors are orthogonalized by Gram-Schmidt.

The computation of the numerical fluxes in the DG weak formulation of the problem requires the approximation of the normal/tangent reference frame to the geodesic edge at the quadrature points. This is achieved again by considering quadratic interpolation. For $i = 1, 2$, the tangent plane is identified by its normal and given by $\tilde{\mathbf{N}}_{\sigma,i} = \tilde{\mathbf{t}}_3(\mathbf{p}_{\sigma,i})$. Then, we compute the derivatives of the approximate parametrization $\underline{\sigma}$ at the extremal points and at the mid-point of this curve, obtaining the following vectors tangent to the edge:

$$\tau_A = \dot{\underline{\sigma}}(0), \qquad \tau_B = \dot{\underline{\sigma}}(1), \qquad \tau_{\mathbf{m}_\sigma} = \dot{\underline{\sigma}}(1/2) .$$

The tangent vectors to the geodesic edge at quadrature points are the evaluated using the following quadratic interpolation:

$$\begin{aligned}
\tau'_{\sigma,i} &= \tau_{\mathbf{m}_\sigma} + \frac{\tau_B - \tau_A}{2}\lambda_i + \left(\frac{\tau_A + \tau_B}{2} - \tau_{\mathbf{m}_\sigma}\right)\lambda_i^2 \ , \\
\tilde{\tau}_{\sigma,i} &= \tau'_{\sigma,i} - \frac{\left\langle \tau'_{\sigma,i}, \tilde{\mathbf{N}}_{\sigma,i} \right\rangle}{\left\langle \tilde{\mathbf{N}}_{\sigma,i}, \tilde{\mathbf{N}}_{\sigma,i} \right\rangle}\tilde{\mathbf{N}}_{\sigma,i} \ ,
\end{aligned}$$

and the approximation of the intrinsic normal to the geodesic edge is obtained by applying the outer product to $\tilde{\tau}_{\sigma,i}$ and $\tilde{\mathbf{N}}_{\sigma,i}$:

$$\tilde{\nu}_{\sigma,i} = \tilde{\tau}_{\sigma,i} \wedge \tilde{\mathbf{N}}_{\sigma,i} \ . \tag{5.8}$$

Finally, the vectors $\tilde{\nu}_{\sigma,i}$ and $\tilde{\tau}_{\sigma,i}$ have to be expressed using the LCS bases at the quadrature points of interest.

5.1.3 Numerical flux

One of the main idea behind RKDG methods is that they are constructed in such a way that they are high-order accurate schemes that reduce to a monotone Godunov-like scheme when a piece-wise constant approximation is used. As a consequence, the numerical flux needs to satisfy some conditions: the function $\tilde{f}$ is defined to be locally Lipschitz, it is consistent with the flux, i.e., $\tilde{f}(u,u) = \langle \mathbf{w}u, \nu\rangle_{\mathcal{G}}$, and it is calculated via an exact or approximate Riemann solver. Moreover, it has to be conservative, that is:

$$\tilde{f}_{\sigma,T}(u_L^T(\mathbf{p}), u_R^T(\mathbf{p})) + \tilde{f}_{\sigma,T'}(u_L^{T'}(\mathbf{p}), u_R^{T'}(\mathbf{p})) = 0 \, , \qquad T \cap T' = \sigma \, .$$

As we did in the FV scheme, we take into consideration the Godunov approach, i.e., the numerical flux $\tilde{f}_{j,T}$ in eq. (5.4), or eq. (5.6), is defined by evaluating the flux function from solutions of Riemann problems at quadrature points of the cell edges. We know from proposition 3.2.2 that the intrinsic transport equation is invariant under rotations. Hence each RP can be formulated as a one-dimensional problem in the tangent plane $T_{\mathbf{p}_j}\Gamma$, at quadrature point $\mathbf{p}_j$, in the direction of the local edge normal $\nu_{\sigma,j}$. The Godunov flux can be written as:

$$\tilde{f}^G(u_L^\nu, u_R^\nu) = \begin{cases} \min_{u_L^\nu \le u \le u_R^\nu} \langle \mathbf{w}u, \nu\rangle_{\mathcal{G}} \ , & \text{if } u_L^\nu \le u_R^\nu \\ \max_{u_L^\nu \ge u \ge u_R^\nu} \langle \mathbf{w}u, \nu\rangle_{\mathcal{G}} \ , & \text{if } u_L^\nu > u_R^\nu \end{cases} .$$

Other definitions of numerical flux are allowed, e.g., the local Lax-Friedrichs flux, the Lax-Friedrichs flux or the Roe flux, with different characteristics (e.g., amount of artificial viscosity) but similar performances. These have the advantage over the

Godunov flux that they do not need to solve a Riemann problem, which, as seen in the previous chapter, may be problematic in the case of non autonomous fluxes. However, we stick this method in order to unveil all the difficulties related to our geometric framework.

5.1.4 Time discretization

As suggested by several authors and to maintain the accuracy achieved by the spatial discretization, the last step iv) is addressed by discretizing in time the ODE eq. (5.7), with a second-order accurate Runge-Kutta method. Let $\{t^n\}_{n=0}^N$ be a partition of the time interval $[0, t_f]$, and denote by $\Delta t^n = t^{n+1} - t^n$ the n-th time step for $n = 0, \ldots, N-1$, a s-step time-marching algorithm reads as follows:

- Set $u_h^0 = P_h(u_0)$;
- For $n = 1, \ldots, N-1$ compute u_h^{n+1} from u_h^n as follows:
    1. $u_h^{(0)} = u_h^n$;
    2. for $k = 1, \ldots, s$ compute the intermediate functions:
    $$u_h^{(k)} = \sum_{l=0}^{k-1} \alpha_{kl}\, u_h^{(l)} + \beta_{kl}\, \Delta t^n\, \mathbf{R}\left(u_h^{(l)}\right) \; ;$$
    3. set $u_h^{n+1} = u_h^{(s)}$.

Note that the method is very easy to code and only a single routine computing the right-hand-side is needed. Different Runge-Kutta methods depend on the choice of the parameters α and β. Since we are considering a second order discretization in space we are interested in a 2-step time discretization. The Runge-Kutta update for the approximate solution then simplifies in:

$$u_h^{n+1} = u_h^n + \frac{\Delta t^n}{2}\, \left[\mathbf{R}(u_h^n) + \mathbf{R}\left(u_h^n + \Delta t^n\, \mathbf{R}(u_h^n)\right)\right] \, ,$$

or

$$u_h^{n+1} = u_h^n + \Delta t^n\, \mathbf{R}\left(u_h^n + \tfrac{\Delta t^n}{2}\, \mathbf{R}(u_h^n)\right) \, ,$$

called the *Heun* and *Runge* formula, respectively. The former can be interpreted as an explicit version of the Crank-Nicolson method, where we have resolved the implicitness deriving by the use of the trapezoidal rule by estimating the right-hand-side $\mathbf{R}(u_h^{n+1})$ with one step of Forward Euler. Analogously, Runge's scheme can be thought of as the use of the mid-point quadrature rule where $\mathbf{R}(u_h^{n+1/2})$ is again evaluated using Forward Euler with a step-size $\Delta t/2$.

5.2 The ISWE formulation

The aim of this section is to apply the Runge-Kutta Discontinuous Galerkin method to the final system ISWE (3.19). The extension to a multidimensional system comes out directly, and the description of the main steps has been already addressed in the previous section. Here we discuss only the main differences.

We rewrite here for convenience the final system written as balance law, namely eq. (3.23):

$$\frac{\partial \mathbf{U}}{\partial t} + \operatorname{div}_{\mathcal{G}} \underline{\underline{F}}(\mathbf{s}, \mathbf{U}) + \mathbf{S}(\mathbf{s}, \mathbf{U}) = 0 \,, \tag{5.9}$$

defined on a compact subset of the bottom surface, $\Gamma \subset \mathcal{S}_{\mathcal{B}}$, with a well-defined curvilinear boundary $\partial\Gamma$. We recall that $\mathbf{U} = [\eta, q^1, q^2]^{\mathrm{T}}$ is the three-components conservative variable. We assume that the triangulations $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$ are given. For each time $t \in [0, t_f]$, we test the system 5.9 with the vector-valued test function $v_h \in \mathcal{V}_h^{\Gamma}$ and integrate in space, over each single cell $T \in \mathcal{T}(\Gamma)$:

$$\int_T \partial_t \mathbf{U} \, v_h \, \mathrm{d}\mathbf{s} + \int_T \operatorname{div}_{\mathcal{G}} \underline{\underline{F}}(\mathbf{s}, \mathbf{U}) \, v_h \, \mathrm{d}\mathbf{s} + \int_T \mathbf{S}(\mathbf{s}, \mathbf{U}) \, v_h \, \mathrm{d}\mathbf{s} = 0 \,.$$

Then, application of the divergence theorem yields:

$$\int_T \partial_t \mathbf{U} \, v_h \, \mathrm{d}\mathbf{s} + \int_{\partial T} \mathbf{F}_*^{\nu}(\mathbf{s}, \mathbf{U}) \, v_h \, \mathrm{d}\sigma - \int_T \left\langle \underline{\underline{F}}(\mathbf{s}, \mathbf{U}), \nabla_{\mathcal{G}} \, v_h \right\rangle_{\mathcal{G}} \mathrm{d}\mathbf{s} + \int_T \mathbf{S}(\mathbf{s}, \mathbf{U}) \, v_h \, \mathrm{d}\mathbf{s} = 0 \,,$$

where $\mathbf{F}_*^{\nu}(\mathbf{s}, \mathbf{U})$ denotes the numerical flux at the boundary of the cell. We substitute the solution $\mathbf{U}$ with the approximate solution $\mathbf{U}_h$, and the weak formulation of eq. (5.9) reads:

Problem 5.2. Find $\mathbf{U}_h \in \mathcal{V}_h^{\Gamma}$ such that

$$\frac{d}{dt} \int_T \mathbf{U}_h \, v_h \, \mathrm{d}\mathbf{s} = - \int_{\partial T} \mathbf{F}_*^{\nu}(\mathbf{s}, \mathbf{U}_h) \, v_h \, \mathrm{d}\sigma + \int_T \left\langle \underline{\underline{F}}(\mathbf{s}, \mathbf{U}_h), \nabla_{\mathcal{G}} \, v_h \right\rangle_{\mathcal{G}} \mathrm{d}\mathbf{s} - \int_T \mathbf{S}(\mathbf{s}, \mathbf{U}_h) \, v_h \, \mathrm{d}\mathbf{s} \,,$$

for any function $v_h \in \mathcal{V}_h^{\Gamma}$.

The function space $\mathcal{V}_h^{\Gamma}$ and the basis functions are obvious extensions of the definitions given in section 5.1.1. Indeed, we consider first order affine functions over each cell, and denote with $\varphi_1, \varphi_2, \varphi_3$ the basis functions spanning $\mathcal{V}_h^{\Gamma}$. For each cell $T \in \mathcal{T}(\Gamma)$ we can represent the approximate conservative variable $\mathbf{U}_h$ by means of the basis functions:

$$\mathbf{U}_h(\mathbf{s}, t) = \sum_i^3 \mathbf{U}_{h,i}(t) \varphi_i(\mathbf{s}) \,,$$

where $\mathbf{U}_{h,i}(t)$ are the values of the numerical solution at the nodes at time t. Thus, we can write the weak formulation as:

$$\sum_i \left(\frac{d\mathbf{U}_{h,i}}{dt} \int_T \varphi_i \varphi_k \, \mathrm{d}\mathbf{s} \right) = - \int_{\partial T} \mathbf{F}^{\nu}_{*}(\mathbf{s}, \mathbf{U}_h) \, \varphi_k \, \mathrm{d}\sigma \\ + \int_T \left\langle \underline{\underline{F}}(\mathbf{s}, \mathbf{U}_h), \nabla_{\mathcal{G}} \, \varphi_k \right\rangle_{\mathcal{G}} \mathrm{d}\mathbf{s} - \int_T \mathbf{S}(\mathbf{s}, \mathbf{U}_h) \varphi_k \, \mathrm{d}\mathbf{s} \, .$$

Moreover, for each $i, k = 1, 2, 3$, we have that

$$\begin{aligned} \mathbf{M}_{ik} &= \int_T \varphi_i \varphi_k \, \mathrm{d}\mathbf{s} \, , \\ \mathbf{R}_k &= - \int_{\partial T} \mathbf{F}^{\nu}_{*}(\mathbf{s}, \mathbf{U}_h) \, \varphi_k \, \mathrm{d}\sigma + \int_T \left\langle \underline{\underline{F}}(\mathbf{s}, \mathbf{U}_h), \nabla_{\mathcal{G}} \, \varphi_k \right\rangle_{\mathcal{G}} \mathrm{d}\mathbf{s} - \int_T \mathbf{S}(\mathbf{s}, \mathbf{U}_h) \varphi_k \, \mathrm{d}\mathbf{s} \, , \end{aligned}$$

are respectively the local mass matrix and the vector of the right-hand-side for each of the components of $\mathbf{U}_h$. Then, the ODE system becomes:

$$\mathbf{M} \, \frac{d\mathbf{U}_h}{dt} = \mathbf{R} \, ,$$

for each cell $T \in \mathcal{T}(\Gamma)$.

The definition of the scheme proceeds as described in the previous sections. One of the main difference is in the choice of the numerical fluxes on the edges. Due to the complexity of the equations, in this case we rely on the Riemann solvers described in section 4.2.1, and we need to solve one Riemann problem for each quadrature point on the edge. The quadrature rules we consider are the same as for the scalar transport case.

5.3 Numerical results

The numerical tests have the purpose to experimentally verify the applicability of the developed scheme, both to the simulation of the scalar transport equation and to the SWE on surfaces. Specifically we want to validate the code on a simple case, with experimental verification of the numerical convergence, and then run simulations over more complex surfaces. We consider the same bottom surfaces as in section 4.3: a flat domain, a sloping plane and a parabola case to test the scheme in one-dimensional cases, the hyperboloid-central-bump (HCB) and the fully 3D surface for the more complex three-dimensional cases. A complete description of the surfaces is presented in section 4.3.1, and the spatial distribution of the metric coefficients is reported in fig. 4.5.

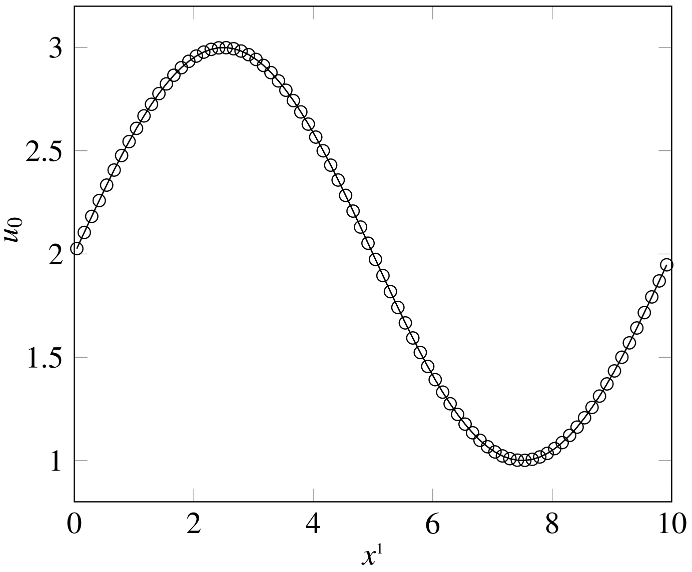

Figure 5.3: Flat domain: initial conditions of a one-dimensional manufactured solution.

5.3.1 Scalar transport equation

We start by solving problem 5.1. We run a simulation considering a manufactured smooth solution over the flat domain to test the convergence of the scheme. Non-smooth solutions are considered over the parabola and the fully 3D surface, taking into account both the cases with a constant and variable advection velocity vectors. The Heun formula is applied for the time discretization in all the test cases.

Flat domain. In this case we verify the convergence of the scheme over the domain $U = [0, 10] \times [0, 1] \subset \mathbb{R}^2$ with constant height function $x^3 \equiv 0$. We consider the smooth solution:

$$u(\mathbf{x}, t) = \sin\left(\frac{\pi(x^1 + 10) - t}{5}\right) + 2\,,$$

and impose the initial condition $u_0 = u(\mathbf{x}, 0)$ (see fig. 5.3). We run the simulation in the time interval $[0, t_f]$, with $t_f = 6.0$ s.

Table 5.1 shows the L^2-norm of the error between the exact and numerical solutions at final time over a mesh sequence. The mesh sequence in built by uniform refinements of a structured grid with a coarse mesh parameter of $h_0 = 0.707$ m, with a total of 40 triangles. The sequence comprises 4 meshes, with the finest grid having $h_3 = 0.088$ m. The error shows second order convergence as expected.

Parabola. Here we consider a simple one-dimensional case where the effects of curvature can be verified. We run our simulations on a grid with average mesh parameter

h_ℓ	E_{L^2}	eoc_ℓ
0.707	6.15E-03	
0.354	1.50E-03	2.04
0.177	3.70E-04	2.02
0.088	9.19E-05	2.01

Table 5.1: Flat domain: L^2 norm of the error between the exact and numerical solution (at $t = 6.0$ s) on refined meshes and order of convergence.

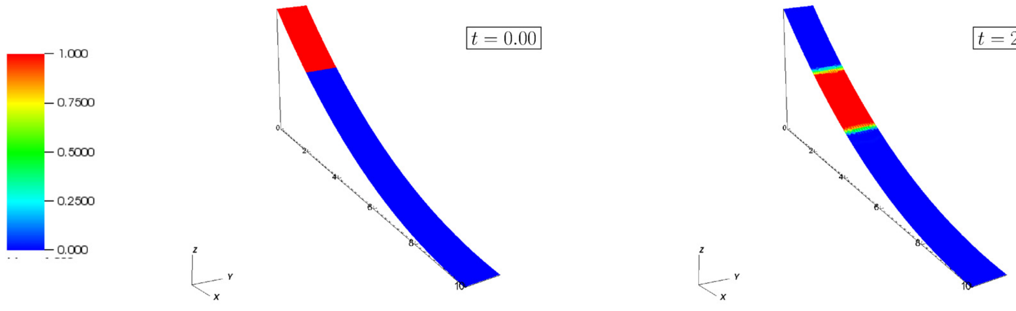

Figure 5.4: Parabola case: transport equation over a parabola. Initial condition (left panel) and numerical solution at time $t = 2.0$ s.

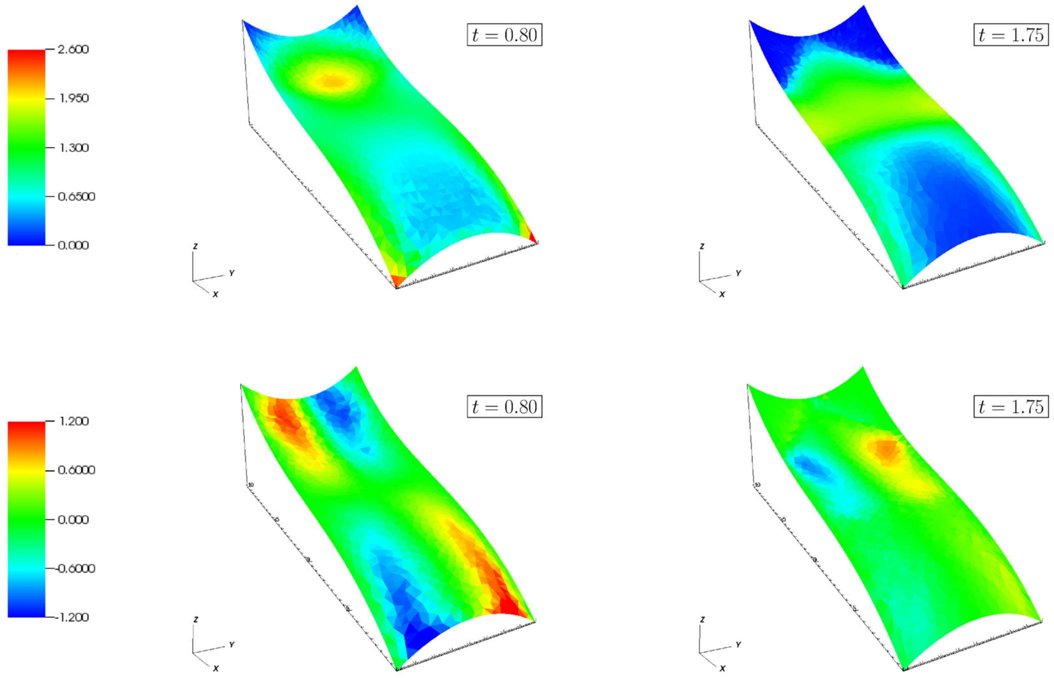

Figure 5.5: Fully 3D surface: shallow water conservative velocity $\vec{q}$ calculated by FV-ISWE. Values of the s^1-component (top line) and s^2-component (bottom line) at times $t =$ 0.80 s and 1.75 s.

$h = 0.12$ m, giving a total of 2464 cells. We simulate the transport of a piece-wise constant initial concentration (fig. 5.4, left panel) over the surface, with a constant surface velocity field given by $\mathbf{w} = (1/h_{(1)}, 0)$. The results show, as expected, a steep front moving downwards maintaining its shape and with minimal numerical viscosity localized within one or two triangle layers of the upstream and downstream fronts. Note that no oscillations are visible even though no limiting (flux or concentration) was used, showing the robustness of the DG scheme.

Fully 3D surface. Then we consider the more complex surface, addressed already in the previous chapter as fully three-dimensional surface, defined starting from the subset $U = [-10, 10] \times [-4, 4] \subset \mathbb{R}^2$ and using the height function:

$$x^3(x^1, x^2) = -\frac{1}{500}(x^1)^3 - \frac{1}{100}x^1(x^2)^2 \,. \tag{5.10}$$

The mesh used in this test case was obtained from a Delaunay triangulation with mesh parameter is $h = 0.62$ m for a total of 1656 surface triangles. The advection velocity vector is non-constant and varies in space and in time, and is given by the

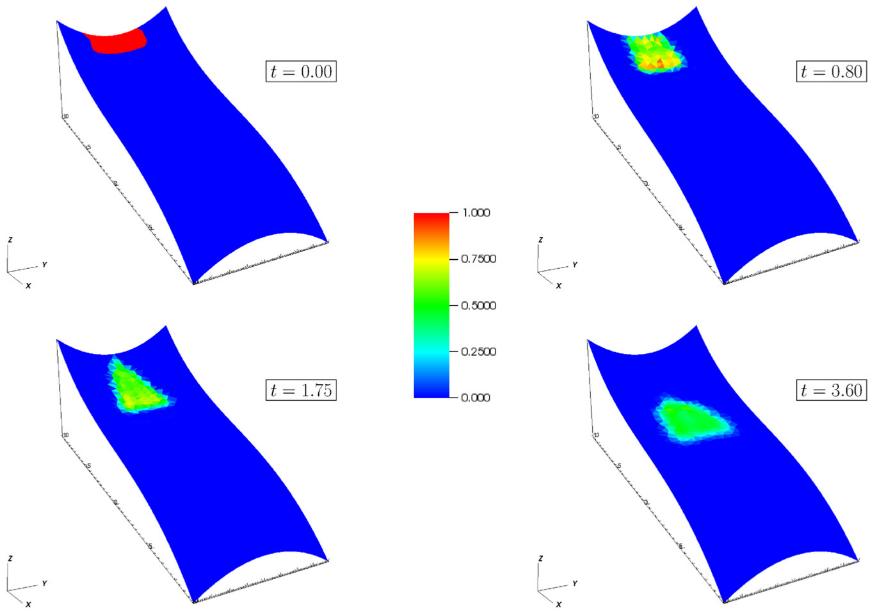

Figure 5.6: Fully 3D surface: transport equation on a mesh with $h = 0.62$ m. The concentration in reported at different times, $t = 0.00, 0.80, 1.75, 3.60$ s.

finite volume solution of the Intrinsic Shallow Water Equations defined in section 4.3. The simulation for the ISWE is set to be a dam-break wet bed problem, with initial condition that considers uniform water depth of 2.0 m upstream of $x^1 = -8.5$ m, and 1.0 m downstream. No flow conditions are imposed in all boundaries, with an outlet located at $x^1 = 10$ m. We run the simulation up to 3.6 s and store at each time step the value of the time and the solution.

Then, we define an initial condition for a scalar concentration u, as represented in fig. 5.6, top left. The advection velocities $\mathbf{w}$ are specific discharge $\vec{q}$ calculated by the ISWE simulation, while the time steps are evaluated to maintain a value of CFL approximately equal to 0.1. As a consequence, we need to proceed by linear interpolation to assign the correct values of the velocities at the actual times of the simulation. The numerical solution is shown in fig. 5.6. The results show the initial front moving down by gravity, concentrating initially in the upper bowl of the bottom surface and spreading wider when reaching the central mound. The behavior is qualitatively plausible as it follows the distribution of the specific discharge field $\vec{q}$ calculated by the ISWE and shown in fig. 5.5.

	η		q^1		q^2	
h_ℓ	E_{L^2}	eoc_ℓ	E_{L^2}	eoc_ℓ	E_{L^2}	eoc_ℓ
1800.0	2.30E-03		4.54E-03		4.54E-03	
900.0	5.79E-04	1.99	1.15E-03	1.98	1.15E-03	1.98
450.0	1.45E-04	2.00	2.89E-04	1.99	2.89E-04	1.99
225.0	3.62E-05	2.00	7.22E-05	2.00	7.22E-05	2.00

Table 5.2: Manufactured solution: L^2 norm of the error between the exact and numerical solution on refined meshes and order of convergence.

5.3.2 Shallow water system

Now we move to the resolution of the ISWE system. We start presenting a test case over a flat domain that considers a manufactured smooth solution, while all the other test cases simulate a gravity-driven fluid in a dam-break setting, without any stress tensor. Again, the Heun formula is applied for the time discretization for all the tests.

A manufactured solution

We consider here an example with the exact solution presented in [73]. We define a triangulation of the domain $U = [x_a^1, x_b^1] \times [x_a^2, x_b^2] \subset \mathbb{R}^2$ and a constant height function $x^3 = 0.0$. The manufactured solution has the following form:

$$\begin{aligned}
\eta &= 2\,\eta_0 \frac{\cos\left(\omega\left(x^1 - x_a^1\right)\right)\cos\left(\omega\left(x^2 - x_a^2\right)\right)}{\cos\left(\omega\left(x_b^1 - x_a^1\right)\right)\cos\left(\omega\left(x_b^2 - x_a^2\right)\right)} \cos\left(\omega\left(t+\tau\right)\right) + 2.0\,,\\
q^1 &= q_0^1 \frac{\sin\left(\omega\left(x^1 - x_a^1\right)\right)\cos\left(\omega\left(x^2 - x_a^2\right)\right)}{\cos\left(\omega\left(x_b^1 - x_a^1\right)\right)\cos\left(\omega\left(x_b^2 - x_a^2\right)\right)} \sin\left(\omega\left(t+\tau\right)\right)\,,\\
q^2 &= q_0^2 \frac{\cos\left(\omega\left(x^1 - x_a^1\right)\right)\sin\left(\omega\left(x^2 - x_a^2\right)\right)}{\cos\left(\omega\left(x_b^1 - x_a^1\right)\right)\cos\left(\omega\left(x_b^2 - x_a^2\right)\right)} \sin\left(\omega\left(t+\tau\right)\right)\,,
\end{aligned}$$

where ω is the angular frequency that we set to $2\pi/43200$, τ is the phase shift and is set to 0. The domain is set to exactly fit one period, i.e. $x_b^1 - x_a^1$ and $x_b^2 - x_a^2$ are set to 43200, and $q_0^1 = q_0^2 = \eta_0 = 0.25$. Note that, the initial depth is translated in the vertical direction by a value 2.0 in order to have always a positive depth value.

The convergence is tested over a mesh sequence that comprises 4 mesh levels ($\ell = 0, \dots, 3$). The coarse mesh is characterized by a mesh parameter h_0 =1800.0 m and 1250 cells, then by uniform refinements the finest mesh contains 80000 triangles, with a mesh parameter h_3 =225.0 m. The simulation runs in the interval $[0, t_f]$ with $t_f = 10000$ s, and the value of the time steps are calculated so that CFL $\approx$ 0.1. Table 5.2 shows the

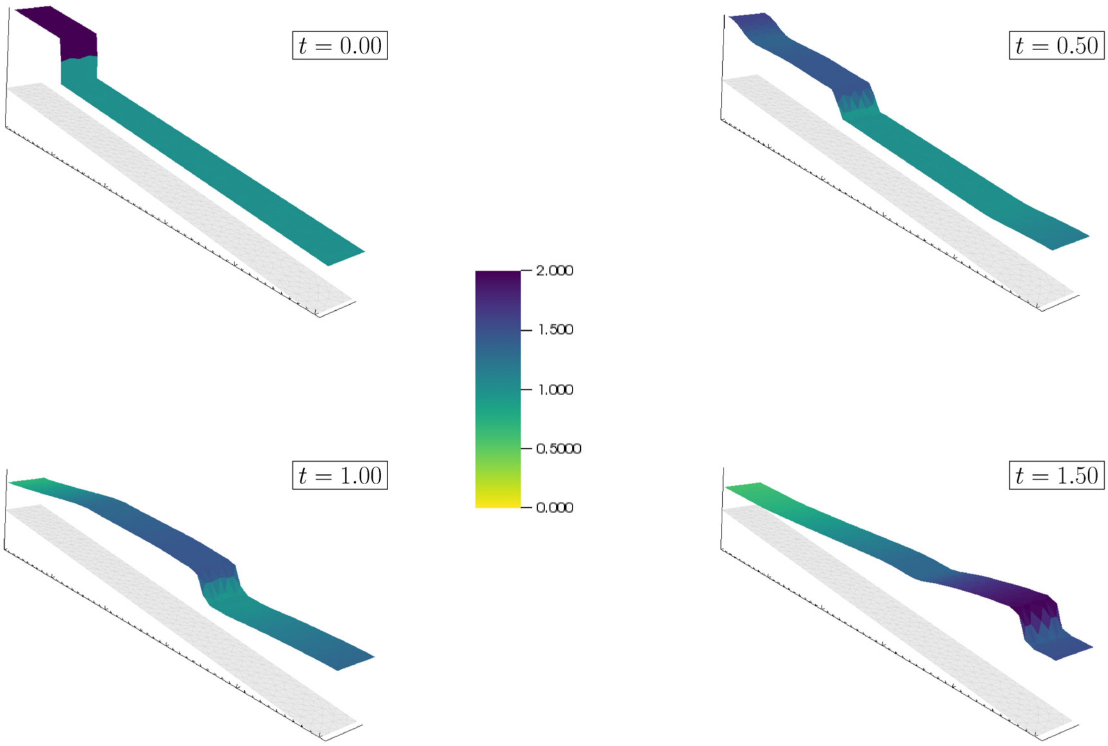

Figure 5.7: Sloping plane: evolution of the gravity wave shown both as color codes and water surface elevation.

L^2-norm of the error between the exact and numerical solutions at the final time over the mesh sequence, and a second order of convergence is reached as expected.

Dam-break simulations

In these test cases we consider all the four different bottom surfaces of the previous chapter (see section 4.3). The initial conditions are defined to initiate a dam-break phenomenon, with water depth in any case small enough to exclude the issue of the intersection of the local normals, as detailed in the FV chapter. Different initial conditions on water depth are considered depending on the shape of the bottom surface, while we always consider zero initial velocities. No-flow boundary conditions are imposed everywhere. The value of the time step Δt is calculated so that CFL $\approx$ 0.1. The HLL Riemann solver is applied.

Sloping plane. Initial conditions over the sloping plane domain describe a dam located at x^1 =2.0 m with a water depth of 2.0 m upstream and 1.0 m downstream, with wall boundary conditions everywhere. Figure 5.7 shows the numerically evaluated evolution of the gravity wave in terms of water depth η at $t = 0.0$ s, 0.50 s, 1.00 s and 1.50 s. The mesh used in these simulations is a Delaunay triangulation characterized by an average mesh parameter $h_\ell \approx 0.23$ m and a total of 588 surface cells. We can note that, compared to the FV simulations, smaller numerical viscosity is present and the wave front is sharper. Note that no oscillations are visible although no limiters were used. This is expected since the metric form is constant over the domain. As a consequence, the fluxes are autonomous and pose no difficulties in the definition of the Riemann problem and its solution. A small wave trailing the main shock is visible especially at $t = 1.50$ s. This phenomenon is actually observed in experiments and known to be difficult to be modeled [64, 65]. In our case we believe that it is the higher order approximation and the stability and robustness of the developed DG approach that allows the simulation of these trailing waves. The first order approximation property together with high numerical viscosity of the Godunov FV scheme hinders the appearance of these phenomena in the numerical solution.

Parabola. The simulation over the parabola considers a grid with average mesh parameter $h = 0.26$ m, giving a total of 588 cells. We simulate the breaking of a dam located at x^1 =2.0 m, with initially zero velocity everywhere and water depth of 2.0 m upstream and 1.0 m downstream the dam. Figure 5.8 shows the calculated distribution of the water depth η at times $t = 0.00$ s, 0.30 s, 0.60 s and 0.90 s.

The distribution of the depth at different times shows a compact downhill shock wave reproduced with very little smearing, thus providing further confirmation to the observation reported in the previous test case that second order accuracy of DG drastically improves the resolution of steep fronts. Very small oscillations are barely detectable at the foot of the shock (see fig. 5.9, right panel), attributable to the non-autonomous fluxes. This is more visible upstream of the front, in particular at the latest time (see fig. 5.9, left panel). In this region the numerical solution shows a number of trailing waves that dissipate moving upward. This is the same phenomenon discussed in the previous test case, but here it is more developed because of the presence of steeper slopes in the bottom surface. We can conclude that the use of high-order numerical approximations is fundamental to fully capture the physical behavior of gravity flows over non-flat terrains.

Hyperboloid-central-bump. This test case considers the hyperboloid-central-bump case, defined starting from a subset $U = [-3, 3] \times [-3, 3] \subset \mathbb{R}^2$ and a height function

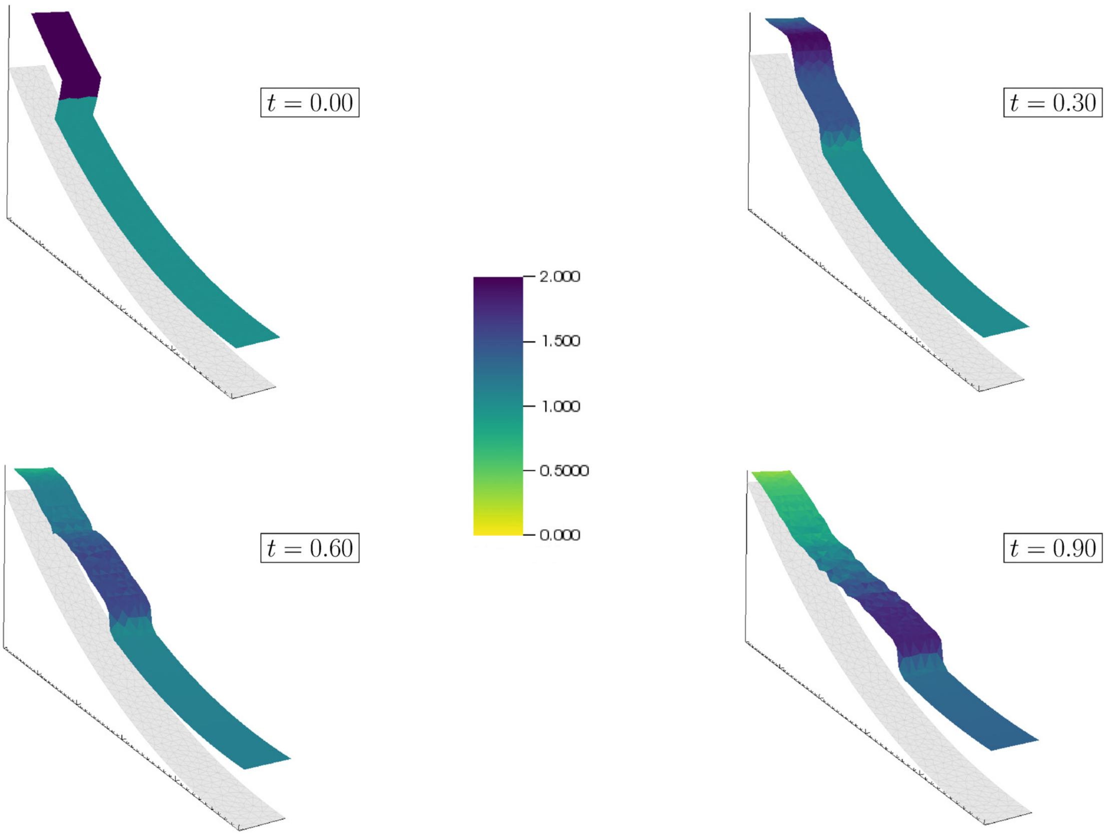

Figure 5.8: Parabola case: evolution of the gravity wave, shown both as color codes and depth elevation.

Figure 5.9: Parabola case: zoom at $t = 0.90$ s.

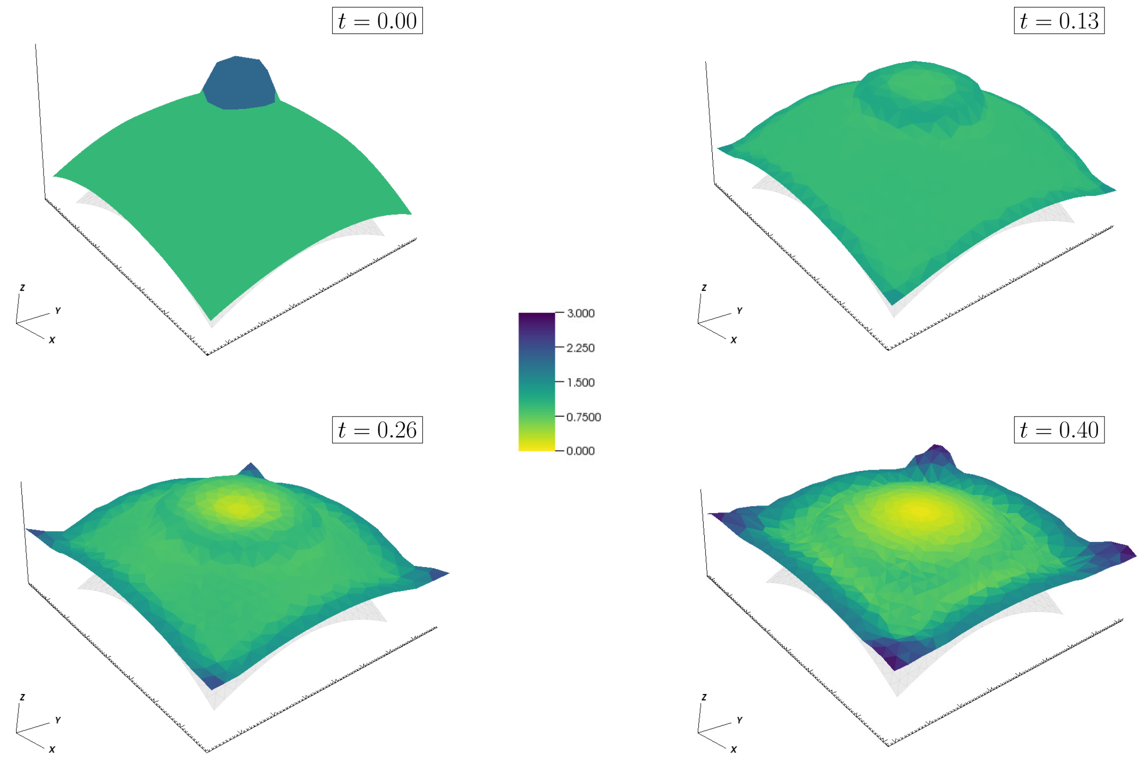

Figure 5.10: HCB: evolution of the gravity wave. Water depth at initial time ($t = 0.0$), and at $t = 0.13,\ 0.26,\ 0.40$ s.

given by:

$$x^3(x^1, x^2) = -\frac{4}{5}\sqrt{(x^1)^2 + (x^2)^2 + 1}\,. \tag{5.11}$$

The mesh is generated by a Delaunay triangulation of U with average mesh parameter $h = 0.34$ m, generating a total of 1238 surface cells. The nodal values are then raised using eq. (5.11). The initial conditions outline a central area of radius 0.5 m with upstream water depth of 2.0 m and downstream water depth of 1.0 m, leading to an initially symmetric gravity wave. Wall conditions are imposed on all boundaries.

Figure 5.10 shows the numerically evaluated evolution of the initial wave in terms of water depth η at times $t = 0.0$ s, 0.13 s, 0.26 s and 0.40 s. The initial wave moves downward with accurately reproduced radial velocity vectors (fig. 5.11).

Fully 3D surface. The final test case considers the fully 3D bottom surface with variable curvature defined on the subset $U = [-10, 10] \times [-4, 4] \subset \mathbb{R}^2$ with height function expressed as:

$$x^3(x^1, x^2) = -\frac{1}{500}(x^1)^3 - \frac{1}{100}x^1(x^2)^2\,.$$

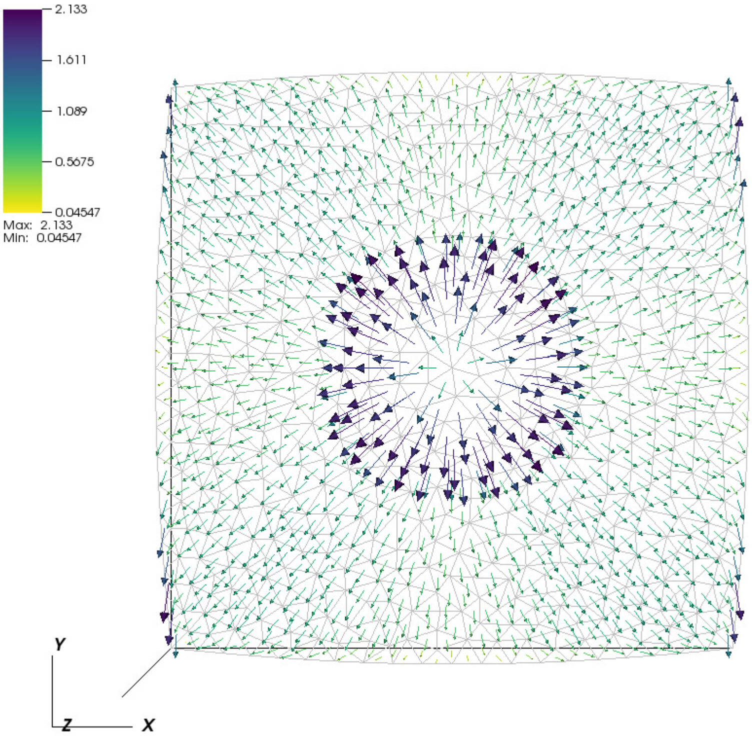

Figure 5.11: HCB: velocity vectors at $t = 0.13$ s.

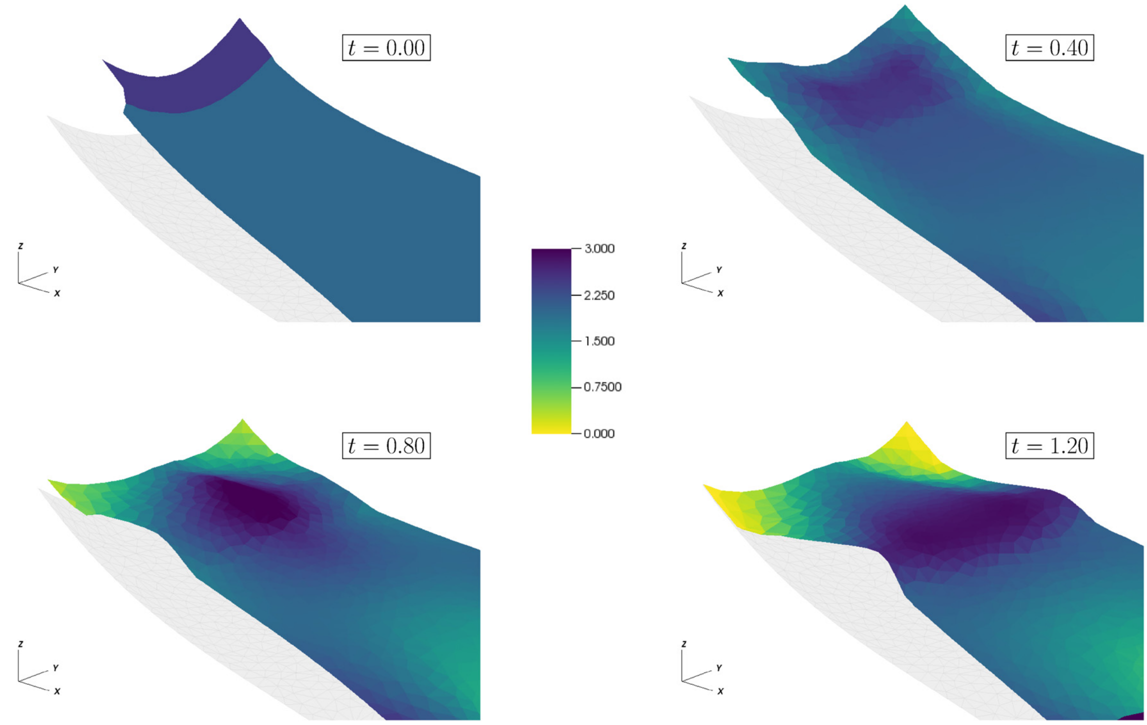

Figure 5.12: Fully 3D surface: water depth evolution of initial wave, shown both as color codes and depth elevation at initial time ($t = 0.0$) and at $t = 0.40,\ 0.80,\ 1.20$.

The discretization of U is obtained again with a Delaunay triangulation with average mesh parameter $h = 0.62$ m, in this case generating a total of 1656 surface cells. The initial conditions consider a uniform water depth of 2.5 m upstream of x^1 = -8.5 m, and 2.0 m downstream. Figure 5.12 shows the numerically evaluated water depth η at times $t = 0.0$ s, 0.40 s, 0.80 s and 1.20 s, displaying a behavior qualitatively similar to the one reported in the FV case.

6

Intrinsic Finite Element Method

Given the encouraging results obtained in the extension of DG, namely the definition of affine basis functions and consistent surface quadrature rules developed in the previous chapters, it is natural to extend our intrinsic approach to finite element methods. Thus, in this chapter we move from hyperbolic to elliptic and parabolic equations on surfaces, within the framework of finite element methods. Among first contributions in the study of PDEs on surfaces and numerical methods to solve them we cite Dziuk and Elliott [36], which consider a piece-wise polygonal approximation of the surface and introduce a finite element space defined on this triangular surface mesh, the so-called surface finite element methods (SFEM). Another contribution considered in this chapter is the work of Olshanskii et al. [60] that provides a new method of discretize a surface PDE based on an outer mesh. The test cases proposed in this chapter are all taken form these two references.

Here we want to focus on the study of the intrinsic formulation for FEM. As shown in the previous chapter, we have already a framework for the definition of a conforming function space considering the geometry of the surface. The aim of this chapter is to apply the same idea to the continuous Galerkin discretization. Our proposed approach seems ideally adaptable to this application. We are interested in developing continuous P1 Galerkin FEM for advection-diffusion-reaction equations on surfaces (see eq. (2.13)).

We will develop an innovative intrinsic FE scheme and compare it with the embedded technique described by [36]. While the geometric data requirements are similar between the two approaches, the advantage of the intrinsic approach is that it is automatically tailored to the surface (that we recall needs to be regular for both approaches) and can use different charts for different regions. Moreover, it maintains the symmetries of the PDEs, and is hopefully more robust as it does not need to project vector quantities back to the surface. Similarly to chapter 5, we start the development from a scalar equation, in this case a scalar advection-diffusion-reaction equation. In what follows we will describe both the approach of Dziuk and Elliott [36], which we term *embedded*, and our intrinsic approach so that a critical comparison can be made.

6.1 Surface advection-diffusion-reaction equation

Consider a compact surface $\Gamma \subset \mathbb{R}^3$ over which we would like to solve an advection-diffusion-reaction equation of the form:

$$\frac{\partial u}{\partial t} - \epsilon\, \Delta u + \langle \mathbf{w}, \nabla\, u \rangle + c\, u = f \qquad \text{on} \quad \Gamma\,, \tag{6.1}$$

where Γ is assumed to be fixed in time, the solution $u : \Gamma \times [0, t_f] \to \mathbb{R}$ is a scalar function defined on the surface and $\mathbf{w} : \Gamma \times [0, t_f] \to \mathbb{R}^3$ is a given divergence-free velocity field everywhere tangential to the surface. The coefficient ϵ is the diffusion coefficient, $\epsilon \in (0, 1]$, and we assume the function $c : \Gamma \to \mathbb{R}$ non-negative and $f \in L^2(\Gamma)$ to ensure well-posedness of the elliptic terms of eq. (6.1). The differential operators Δ and ∇, the Laplacian and gradient, respectively, and the scalar product $\langle \cdot, \cdot \rangle$ need to be properly defined on the surface to follow the geometric setting of the problem. In the following sections we are going to develop two different approaches for the numerical solution of eq. (6.1) with different specifications for these operators, we will use the above symbols for the generic version, and add specific subscripts when adapted to the different approaches. If the surface has a boundary (i.e., $\partial\Gamma \neq \{\emptyset\}$) we assume zero Neumann boundary conditions.

The classical variational formulation for eq. (2.13) on Γ is given by:

Problem 6.1. Find $u \in H^1(\Gamma)$ such that

$$m(\partial_t\, u, v) + a(u, v) + b(u, v) + c(u, v) = F(u) \qquad \forall\, v \in \mathcal{V}^\Gamma\,,$$

where the bilinear forms $m(\cdot, \cdot), a(\cdot, \cdot), b(\cdot, \cdot)$ and $c(\cdot, \cdot)$, are called the mass, stiffness, transport, and reaction forms, respectively. They are defined as:

$$\begin{aligned} m(\partial_t u, v) &= \int_\Gamma \partial_t uv \, \mathrm{d}\mathbf{s}\,, & a(u,v) &= \epsilon \int_\Gamma \langle \nabla u, \nabla v \rangle \, \mathrm{d}\mathbf{s}\,, \\ b(u,v) &= \int_\Gamma \langle \mathbf{w}, \nabla u \rangle \, v \, \mathrm{d}\mathbf{s}\,, & c(u,v) &= \int_\Gamma c \, uv \, \mathrm{d}\mathbf{s}\,, \end{aligned}$$

together with the linear (load) form $F(\cdot)$:

$$F(v) = \int_\Gamma f v \, \mathrm{d}\mathbf{s}\,.$$

The test space $\mathcal{V}^\Gamma$ is given by:

$$\mathcal{V}^\Gamma = \begin{cases} \{v \in H^1(\Gamma) \mid \int_\Gamma v \, \mathrm{d}\mathbf{s} = 0\} & \text{if } \ c = 0 \\ H^1(\Gamma) & \text{if } \ c > 0 \end{cases},$$

where $H^1(\Gamma)$ is the classical Hilbert space of functions defined on Γ that are square integrable and with all the first derivatives that are square integrable.

Note that this is the formal definition for the variational problem. We will define the problem rigorously after the specification of the operators for both the approaches we are using.

Given the shape-regular triangulations $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$ of Γ, the classical FEM formulation for Problem 6.1 is given by:

Problem 6.2. Find $u_h \in \mathcal{V}_h^\Gamma$ such that

$$m(\partial_t u_h, v) + a(u_h, v) + b(u_h, v) + c(u_h, v) = F(u_h) \qquad \forall \, v \in \mathcal{V}_h^\Gamma\,,$$

where $\mathcal{V}_h^\Gamma \subset H^1(\Gamma)$ is one of the classical conforming finite dimensional FEM spaces.

More specifically, we will be interested in piece-wise linear Galerkin methods and we will indicate with $\mathcal{P}_1(\mathcal{T}(\Gamma))$ and $\mathcal{P}_1(\mathcal{T}_h(\Gamma))$ the relative discrete functional spaces of affine functions on $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$, respectively.

6.1.1 Embedded approach

In this section we describe what we call the *embedded* FEM method as proposed by Dziuk and Elliott [36]. We term this approach as *embedded* as opposed to our intrinsic approach. In this setting, a function $f : \Gamma \to \mathbb{R}$ is actually defined on the points of Γ seen as points in $\mathbb{R}^3$. We define also a companion differentiable function $\bar{f}$ as the smooth extension of f in the set of points belonging to the tubular neighborhood of Γ such that $\bar{f}|_\Gamma = f$.

Differential operators

We can define the tangential gradient of f to the surface Γ as follows:

Definition 6.1.1 ((Embedded) Tangential Gradient). *The* tangential gradient *of f at a point $\mathbf{p} \in \Gamma$ is given by*

$$\nabla_\Gamma f(\mathbf{x}) = \nabla \bar{f}(\mathbf{x}) - \left\langle \nabla \bar{f}(\mathbf{x}), \mathbf{N}(\mathbf{x}) \right\rangle \mathbf{N}(\mathbf{x}) = \mathbb{P}(\mathbf{x}) \nabla \bar{f}(\mathbf{x}), \tag{6.2}$$

where $\mathbf{x} = \mathbf{x}(\mathbf{p}) = (x^1_\mathbf{p}, x^2_\mathbf{p}, x^3_\mathbf{p})$ denotes the canonical Cartesian coordinates in $\mathbb{R}^3$, the projection tensor $\mathbb{P}$ is $\mathbb{P}(\mathbf{x})_{ij} = \delta_{ij} - \mathbf{N}^i(\mathbf{x})\mathbf{N}^j(\mathbf{x})$ $(i, j = 1, 2, 3)$. The symbol ∇ denotes the gradient in $\mathbb{R}^3$, $\mathbf{N}(\mathbf{x})$ is the unit normal to the surface Γ at $\mathbf{p}$, and the bilinear form $\langle \cdot, \cdot \rangle$ is the standard scalar product in $\mathbb{R}^3$.

Note that the unit normal $\mathbf{N}(\mathbf{x})$ is always well defined for all $\mathbf{p} \in \Gamma$ because we assume Γ is a regular surface. Since $\langle \nabla_\Gamma f(\mathbf{x}), \mathbf{N}(\mathbf{x}) \rangle = 0$, the gradient $\nabla_\Gamma f(\mathbf{x})$ belongs to $T_\mathbf{p}\Gamma$, the tangent plane to Γ at $\mathbf{p}$. We will use the notation

$$\nabla_\Gamma f(\mathbf{x}) = [\underline{D}_1 f(\mathbf{x}), \underline{D}_2 f(\mathbf{x}), \underline{D}_3 f(\mathbf{x})] \ .$$

for the components of the tangential gradient.

Definition 6.1.2 ((Embedded) Laplace-Beltrami operator). *The* Laplace-Beltrami operator *applied to a twice differentiable function $f \in C^2(\Gamma)$ is given by:*

$$\Delta_\Gamma f = \nabla_\Gamma \cdot \nabla_\Gamma f = \sum_{i=1}^{3} \underline{D}_i \underline{D}_i f \ . \tag{6.3}$$

Definition 6.1.3. *For $\Gamma \in C^2$ the matrix*

$$\mathcal{H}_{ij} = \underline{D}_i \mathbf{N}^j \quad (i, j = 1, \dots, 3) \ , \tag{6.4}$$

is a symmetric matrix having a zero eigenvalue in the normal direction, i.e. $\mathcal{H}\mathbf{N} = 0$. The other two eigenvalues are the principal curvatures *of Γ. The quantity*

$$H(\mathbf{x}) = \operatorname{tr}(\mathcal{H}(\mathbf{x})) \tag{6.5}$$

is the mean curvature *of Γ at the point $\mathbf{x} \in \Gamma$.*

For an embedded surface we can express the theorem of integration by parts in terms of the projected differential operators. Thus, given a compact surface $\Gamma \in \mathbb{R}^3$ possibly with smooth boundary $\partial\Gamma$, integration by parts holds, as the following theorem states [36].

Theorem 6.1.4. *Assume that $\Gamma \subset \mathbb{R}^3$ is a regular surface with smooth boundary $\partial\Gamma$ and $u \in C^1(\bar{\Gamma})$. Then,*

$$\int_\Gamma \nabla_\Gamma u \, \mathrm{d}\mathbf{x} = \int_\Gamma u H \mathbf{N} \, \mathrm{d}\mathbf{x} + \int_{\partial\Gamma} u\mu \, \mathrm{d}\sigma \, ,$$

where μ denotes the co-normal vector, i.e., the vector field normal to $\partial\Gamma$ and tangent to Γ.

This theorem directly implies the following version of Green's Lemma.

Theorem 6.1.5 (Green's formula)**.** *Let $\Gamma \subset \mathbb{R}^3$ be a regular surface with smooth boundary $\partial\Gamma$ and $u \in C^2(\bar{\Gamma})$ be a continuously differentiable function over $\bar{\Gamma}$. Then*

$$\int_\Gamma \langle \nabla_\Gamma u, \nabla_\Gamma v \rangle \, \mathrm{d}\mathbf{x} = -\int_\Gamma \Delta_\Gamma u \, v \, \mathrm{d}\mathbf{x} + \int_{\partial\Gamma} \langle \nabla_\Gamma u, \mu \rangle \, v \, \mathrm{d}\sigma \, , \tag{6.6}$$

where μ is the co-normal vector.

Surface finite element method

We can now proceed with the development of the embedded Surface Finite Element Method (SFEM) as described in [36]. We can define the piece-wise affine FEM space as follows:

$$\mathcal{V}_h^\Gamma = \{ v \in C^0(\mathcal{T}_h(\Gamma)) : \; v|_{T_h} \in \mathcal{P}_1(T_h) \quad \forall \, T_h \in \mathcal{T}_h(\Gamma) \} \, ,$$

where $\mathcal{P}_1(T_h)$ is the space of affine functions defined in $T_h \subset \mathbb{R}^3$. The space $\mathcal{V}_h^\Gamma$ is spanned by the nodal basis functions $\varphi_1, \ldots, \varphi_{N^{dof}}$ defined by:

$$\varphi_j \in \mathcal{V}_h^\Gamma \, , \quad \varphi_j(\mathbf{x}_i) = \delta_{ij} \quad i, j = 1, \ldots, N^{dof} \, ,$$

where $\mathbf{x}_i \in \Gamma \cap \Gamma_h$ are the vertices of the triangulation. Then, every function v in the functional space $\mathcal{V}_h^\Gamma$ can be written as

$$v(\mathbf{x}) = \sum_{j=1}^{N^{dof}} v_j \varphi_j(\mathbf{x}) \qquad \mathbf{x} \in \Gamma_h \, ,$$

with v_j the nodal coefficients. The projection tensor is naturally restricted to the approximate triangulation as:

$$\mathbb{P}_h = \mathbb{I} - \mathbf{N}_h \mathbf{N}_h^{\mathrm{T}} \, ,$$

and the differential operator ∇_{Γ_h} is changed accordingly. Moreover, the advective velocity field is also projected onto Γ_h as $\mathbf{w}_h = \mathbb{P}_h \mathbf{w}$. The surface FEM for Problem 6.1 can now be written as follows:

Problem 6.3. Find $u_h \in \mathcal{V}_h^\Gamma$ such that

$$m_h(\partial_t u_h, v) + a_h(u_h, v) + b_h(u_h, v) + c_h(u_h, v) = F_h(v) \quad \forall v \in \mathcal{V}_h^\Gamma ,$$

where the linear and bilinear forms are mesh-dependent and are given by:

$$m_h(\partial_t u_h, v) = \int_{\Gamma_h} \partial_t u_h \, v \, \mathrm{d}\mathbf{x} , \qquad a_h(u_h, v) = \epsilon \int_{\Gamma_h} \left\langle \nabla_{\Gamma_h} u_h, \nabla_{\Gamma_h} v \right\rangle \mathrm{d}\mathbf{x} ,$$

$$b_h(u_h, v) = \int_{\Gamma_h} \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} u_h \right\rangle v \, \mathrm{d}\mathbf{x} , \qquad c_h(u_h, v) = \int_{\Gamma_h} c \, u_h \, v \, \mathrm{d}\mathbf{x} ,$$

and

$$F_h(v) = \int_{\Gamma_h} f_h \, v \, \mathrm{d}\mathbf{x} .$$

The implementation of the method proceeds similarly to the standard case with the caution that a logically two-dimensional triangulation is used with nodes defined in $\mathbb{R}^3$. We refer to Dziuk and Elliott [36] for the analysis of the described method. We want to mention only the fact that optimal $\mathcal{O}(h^2)$ convergence is achieved by this embedded SFEM in the $L^2(\Gamma)$ norm if $u \in H^2(\Gamma)$ (i.e., u has square integrable second derivatives).

We can represent the numerical solution as linear combination of the basis functions, i.e., $u_h(\mathbf{x}, t) = \sum_j u_j(t) \varphi_j(\mathbf{x})$. For every $i = 1, \dots, N^{dof}$ we obtain:

$$\sum_{j=1}^{N^{dof}} \partial_t u_j m_h(\varphi_j, \varphi_i) + \sum_{j=1}^{N^{dof}} u_j \Big(a_h(\varphi_j, \varphi_i) + b_h(\varphi_j, \varphi_i) + c_h(\varphi_j, \varphi_i) \Big) = F_h(\varphi_i) ,$$

and in particular, in matrix form we can write:

$$\mathbf{M} \, \partial_t \mathbf{u} + (\mathbf{A} + \mathbf{B}(t) + \mathbf{C}) \, \mathbf{u} = \mathbf{b} , \tag{6.7}$$

where $\mathbf{u} = \{u_i\}$ is the vector of the solution coefficients, and $\mathbf{b} = \{F_h(\varphi_i)\}$ is the right-hand side vector. The above matrices take on the following expressions:

$$\mathbf{M}_{ij} = \int_{\Gamma_h} \varphi_j \varphi_i \, \mathrm{d}\mathbf{x} , \qquad \mathbf{A}_{ij} = \epsilon \int_{\Gamma_h} \left\langle \nabla_{\Gamma_h} \varphi_j, \nabla_{\Gamma_h} \varphi_i \right\rangle \mathrm{d}\mathbf{x} ,$$

$$\mathbf{B}_{ij}(t) = \int_{\Gamma_h} \left\langle \mathbf{w}_h(t), \nabla_{\Gamma_h} \varphi_j \right\rangle \varphi_i \, \mathrm{d}\mathbf{x} , \qquad \mathbf{C}_{ij} = \int_{\Gamma_h} c \, \varphi_j \varphi_i \, \mathrm{d}\mathbf{x} .$$

Note that we still have to define a discretization scheme for the time derivative and a quadrature rule for the matrix elements. Concerning the time, we consider an Implicit Euler scheme. Let $\{t^n\}_{n=0}^N$ be a partition of $[0, t_f]$, and $\Delta t^n = t^{n+1} - t^n$ the n-th time step, for $n = 0, \dots, N-1$, the scheme reads:

$$\Big(\frac{1}{\Delta t^n} \mathbf{M} + \mathbf{A} + \mathbf{B}^{n+1} + \mathbf{C} \Big) \mathbf{u}^{n+1} = \mathbf{b} + \frac{1}{\Delta t^n} \mathbf{M} \, \mathbf{u}^n ,$$

where the superscripts indicate the time-step evaluation. Then, we compute the integrals in space applying the surface mid-point rule, which is a consistent rule with the same accuracy of the linear SFEM. For the solution of the linear system, because of its non-symmetric property, we use the ILU-preconditioned BiCGSTAB solver [72].

6.1.2 An intrinsic SFEM approach

In this section we present our intrinsic approach to SFEM. The idea is to consider a local curvilinear coordinate system (LCS) anchored to the surface and write the equations with respect to this local reference system, as described in chapter 2. The resulting intrinsic differential operators contain all the geometric information arising from the surface where the LCS is established (see proposition 2.3.2). All the involved variables and quantities are intrinsic to the surface and no projection is needed for their definition. The intrinsic SFEM variational formulation presented in problem 6.3 can be then written in the LCS as:

Problem 6.4. Find $u_h \in \mathcal{V}_h^\Gamma(\mathcal{T}(\Gamma))$ such that

$$m(\partial_t u_h, v) + a(u_h, v) + b(u_h, v) + c(u_h, v) = F(v) \quad \forall\, v \in \mathcal{V}_h^\Gamma(\mathcal{T}(\Gamma)),$$

where the linear and bilinear forms are given by:

$$\begin{aligned} m(\partial_t u_h, v) &= \int_\Gamma \partial_t u_h\, v \,\mathrm{d}\mathbf{s}, & a(u_h, v) &= \epsilon \int_\Gamma \langle \nabla_{\mathcal{G}} u_h, \nabla_{\mathcal{G}} v\rangle_{\mathcal{G}} \,\mathrm{d}\mathbf{s}, \\ b(u_h, v) &= \int_\Gamma \langle \mathbf{w}, \nabla_{\mathcal{G}} u_h\rangle\, v \,\mathrm{d}\mathbf{s}, & c(u_h, v) &= \int_\Gamma c\, u_h\, v \,\mathrm{d}\mathbf{s}, \end{aligned}$$

and

$$F(v) = \int_\Gamma f_h\, v \,\mathrm{d}\mathbf{s}.$$

Note that in this intrinsic formulation the original advective velocity $\mathbf{w}$ is used directly once expressed in the proper reference frame, contrary to the embedded approach where the velocity $\mathbf{w}_h$ is projected onto $\mathcal{T}_h(\Gamma)$.

Up to the definition of the test space $\mathcal{V}_h^\Gamma(\mathcal{T}(\Gamma))$, no numerical approximations are done until this point, since all the operators and integrals are defined on $\mathcal{T}(\Gamma)$ whose interior coincides with the surface Γ. We would like to remain within this setting as much as possible. Approximation issues arise when we need to practically compute quantities. To this aim, we assume that all the relevant geometric information related to the surface are known (in exact or approximate but consistent form) at the vertices of the triangulation. Then, we need to define the functional space $\mathcal{V}_h^\Gamma$ and appropriate quadrature rules for the surface integrals, and proceed with the approximation of the relevant quantities on cells.

$\mathcal{P}_1$-basis functions. We start from the definition of the basis functions generating the finite-dimensional FEM space $\mathcal{V}_h^\Gamma$. Again, we use the lowest order conforming approach, so that the FEM space is given by:

$$\mathcal{V}_h^\Gamma = \{v \in C^0(\mathcal{T}(\Gamma)) \,:\, v|_T \in \mathcal{P}_1(T) \quad \forall\ T \in \mathcal{T}(\Gamma)\}\,, \tag{6.8}$$

the space of continuous functions defined on the surface whose restriction on each $T \in \mathcal{T}(\Gamma)$ is an affine function. We define the nodal basis functions $\varphi_1, \ldots, \varphi_{N^{dof}}$ spanning $\mathcal{V}_h^\Gamma$ by means of the classical interpolation property:

$$\varphi_j \in \mathcal{V}_h^\Gamma\ , \quad \varphi_j(\mathbf{p}_i) = \delta_{ij} \quad i,j = 1, \ldots, N^{dof}\ ,$$

where $\mathbf{p}_i \in \Gamma$ are the nodes of the triangulation. Note that these functions are defined on the curvilinear (geodesic) triangles T, and not, as in the embedded case, on the approximate (flat) triangles T_h.

The practical definition of φ_j works on an element-by-element basis and proceeds similarly to section 5.1.1. In the following we adapt the previous definition to the conforming FEM case. Given the global coordinates $\mathbf{x}(\mathbf{p})$ of $\mathbf{p} \in T$, we can define the affine function $\tilde{\varphi}_j{}^T(\mathbf{x}) = \tilde{a} + \tilde{b}x^1 + \tilde{c}x^2 + \tilde{d}x^3$ as a function in $\mathbb{R}^3$. Composition of $\tilde{\varphi}_j{}^T$ with the surface parametrization expresses the basis function in local coordinates:

$$\varphi_j^T(s^1, s^2) = \tilde{\varphi}_j{}^T \circ \phi(s^1, s^2)\ .$$

Since the integrals in problem 6.4 are calculated by means of Gaussian quadrature rules, we need to evaluate of the nodal basis functions at quadrature points $\mathbf{p}_j$ inside the cell T. Noting that, when using $\mathcal{P}_1$ basis functions optimal convergence is ensured by the mid-point rule, the evaluation of the integrand at the cell mid-point $\mathbf{m}_T$ is required. As done in the DG setting, assuming $T \subset \phi_{\mathbf{m}_T}(U)$ for some open set $U \subset \mathbb{R}^2$, eq. (2.1) allows us to work on the linear approximation of our surface, i.e., $\phi_{\mathbf{m}_T}(T) = T_{\mathbf{m}_T}\Gamma + \mathcal{O}(h^2)$. As a consequence, the local basis functions

$$\varphi_j^T(s^1, s^2) \approx \tilde{\varphi}_j{}^T \circ \phi_{\mathbf{m}_T}(s^1, s^2)$$

obtained by neglecting the higher order terms span the set of affine functions in T. The global basis function φ_j is then formed as usual by gluing together the elemental components. We are now able to write every function v in the functional space $\mathcal{V}_h^\Gamma$ as:

$$v(\mathbf{s}) = \sum_{j=1}^{N^{dof}} v_j \varphi_j(\mathbf{s}) \qquad \mathbf{s} \in \Gamma\,, \tag{6.9}$$

with v_j the nodal coefficients, and proceed as in the standard finite element method obtaining the linear system as in eq. (6.7).

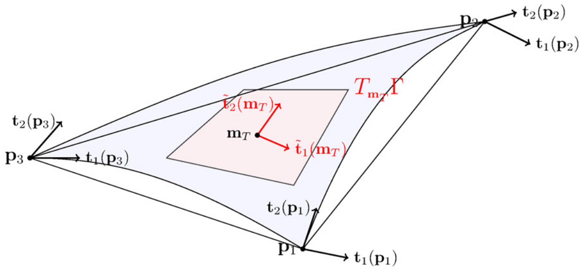

Figure 6.1: Curvilinear cell T and corresponding approximate tangent plane $T_{\mathbf{m}_T}\Gamma$.

The proposed intrinsic formulation is completed by the appropriate definition of the mid-point rule on the surface and approximation of the parametrization in $\mathbf{m}_T$, or equivalently of the tangent plane $T_{\mathbf{m}_T}\Gamma$. In order to numerically approximate the surface integral we can consider the direct extension to the surface of the trapezoidal and the mid-point quadrature rules for triangles [39]:

$$\int_T f \,\mathrm{d}S \approx \frac{1}{3}\sum_{j=1}^{3} f(\mathbf{p}_j)\mathcal{A}_{T_h}\,, \tag{6.10}$$

$$\int_T f \,\mathrm{d}S \approx f(\mathbf{m}_T)\mathcal{A}_{T_h}\,, \tag{6.11}$$

where $f(\mathbf{p}_j)$ and $f(\mathbf{m}_T)$ are the evaluation of the function f at the cell nodes and mid-points, respectively. With the results presented in section 4.1 it is possible to prove that these rules are exact on polynomials of degree one and thus are consistent with the accuracy of the FEM scheme. As standard in finite element methods, we apply the trapezoidal rule for node-based functions, e.g., when lumping the mass matrix, and the mid-point rule for element-based functions, as for example in the stiffness and convection matrices when the diffusion coefficients and the velocity vectors are given element-wise.

The last step is to establish the approximations of the necessary geometric quantities at the quadrature points. We assume to have the complete surface information at the triangle nodes, so that trapezoidal rule can be directly applied, while application of the mid-point rule requires approximation of the geometric quantities at the cell mid-points. We use the interpolation procedure presented in section 4.1, which yields the necessary second order accuracy in h. Given a surface triangle $T \in \mathcal{T}(\Gamma)$ with vertices A, B, C, for the gravity center of the cell $\mathbf{m}_T \in T$ (see fig. 6.1) we construct the following vectors:

$$\begin{aligned}
&\tilde{\mathbf{t}}_1(\mathbf{m}_T) = \frac{1}{3}\mathbf{t}_1(A) + \frac{1}{3}\mathbf{t}_1(B) + \frac{1}{3}\mathbf{t}_1(C), \\
&\mathbf{t}_2'(\mathbf{m}_T) = \frac{1}{3}\mathbf{t}_2(A) + \frac{1}{3}\mathbf{t}_2(B) + \frac{1}{3}\mathbf{t}_2(C)\,, \qquad \tilde{\mathbf{t}}_2(\mathbf{m}_T) = \mathbf{t}_2' - \frac{\left\langle \mathbf{t}_2', \tilde{\mathbf{t}}_1 \right\rangle}{\left\langle \tilde{\mathbf{t}}_1, \tilde{\mathbf{t}}_1 \right\rangle}\tilde{\mathbf{t}}_1\,,
\end{aligned}$$

and the associated metric tensor

$$\tilde{\mathcal{G}}_{\mathbf{m}_T} = \begin{bmatrix} \left\|\tilde{\mathbf{t}}_1\right\|^2 & 0 \\ 0 & \left\|\tilde{\mathbf{t}}_2\right\|^2 \end{bmatrix}.$$

6.1.3 Stabilization

For advection-diffusion-reaction equation the surface finite elements method is unstable when the advective term is dominant with respect to the diffusive one, as the classical plain FEM. More precisely, spurious oscillations appear when the mesh Peclet number is higher than one. We define the cell *Peclet number* as

$$\mathrm{Pe}_T = \frac{h_T \left\|\mathbf{w}\right\|_\infty}{2\epsilon}\,, \tag{6.12}$$

where h_T is the diameter of the cell T. To overcome the problem of spurious oscillations the stabilized finite elements method adds some stabilization terms to the weak formulation. The stabilized finite element method for both the embedded and intrinsic approaches can be written as:

Problem 6.5. Find $u_h \in \mathcal{V}_h^\Gamma$ such that:

$$\begin{aligned}
m_h(\partial_t\, u_h, v) + a_h(u_h, v) + b_h(u_h, v)& \\
+ c_h(u_h, v) + \mathrm{s}_h(u_h, v) &= F_h(v) + \mathrm{s}_{F_h}(v) \quad \forall\, v \in \mathcal{V}_h^\Gamma\,,
\end{aligned}$$

where $\mathrm{s}_h(u_h, v), \mathrm{s}_{F_h}(v)$ are the stabilization terms.

Note that the actual bilinear forms used in the intrinsic FEM contain the h subscript to indicate that numerical quadrature is used in the evaluation of the integrals.

We specify the surface versions of the stabilized FEM based on the standard SUPG approach. In the embedded approach the stabilization terms are defined projecting the relevant operators onto the approximate triangulation $\mathcal{T}_h(\Gamma)$, yielding:

$$\begin{aligned}
\mathrm{s}_h(u_h, v) &= \sum_{T_h \in \mathcal{T}_h(\Gamma)} \delta_{T_h} \int_{T_h} \left(\partial_t\, u_h - \epsilon\Delta_{\Gamma_h} u_h + \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} u_h \right\rangle + c\, u_h\right) \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} v \right\rangle \mathrm{d}\mathbf{x}\,, \\
\mathrm{s}_{F_h}(v) &= \sum_{T_h \in \mathcal{T}_h(\Gamma)} \delta_{T_h} \int_{T_h} f_h \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} v \right\rangle \mathrm{d}\mathbf{x}\,.
\end{aligned}$$

The stabilization parameter δ_{T_h} depends on $T_h \in \mathcal{T}_h(\Gamma)$, and can be written as [30]:

$$\delta_{T_h} = \left(\frac{4\epsilon}{h_T^2} + \frac{2\,\|\mathbf{w}_h\|_{T,\infty}}{h_T} + \tilde{c}_{T_h} \right)^{-1} , \tag{6.13}$$

where $\tilde{c}_{T_h}$ takes into account the coefficient of the mass matrix and the reaction term, i.e., $\tilde{c}_{T_h} = \frac{1}{\Delta t^n} + c_{T_h}$.

The intrinsic formulation reads:

$$\begin{aligned} \mathrm{s}_h(u_h, v) &= \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T \Big(\partial_t\, u_h - \epsilon \Delta_{\mathcal{G}} u_h + \langle \mathbf{w}, \nabla_{\mathcal{G}}\, u_h \rangle_{\mathcal{G}} + c\, u_h\Big) \langle \mathbf{w}, \nabla_{\mathcal{G}}\, v\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \\ \mathrm{s}_{F_h}(v) &= \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T f\, \langle \mathbf{w}, \nabla_{\mathcal{G}}\, v\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \end{aligned}$$

with δ_T taking on the form:

$$\delta_T = \left(\frac{4\epsilon}{h_T^2} + \frac{2\,\|\mathbf{w}\|_{T,\infty}}{h_T} + \tilde{c}_T \right)^{-1} , \tag{6.14}$$

where $\tilde{c}_T = \frac{1}{\Delta t^n} + c_T$. We note that many alternative definitions of the stabilization parameter δ_T have been explored in the literature. In particular, the formulation proposed by Bazilevs et al. [10] has shown promise for highly dynamic problems in complex geometries. The adaptation of this approach to our intrinsic scheme is given by:

$$\delta_{T,\mathcal{G}} = \left(4\tilde{c}_T^2 + \|\mathbf{w}\|_{\mathcal{G}} + 9\epsilon^2 \sum_i g_{ii} \right)^{-1/2} ,$$

where the surface metric tensor replaces the Jacobian of the change of coordinate between the reference and physical triangle. In all the test problems considered in this work we observed similar behavior across all the definitions of δ_T and report only the results using eqs. (6.13) and (6.14).

The stabilized problem at time t can be re-written in matrix form as:

$$\tilde{\mathbf{M}}(t)\, \partial_t\, \mathbf{u} + \Big(\tilde{\mathbf{A}}(t) + \tilde{\mathbf{B}}(t) + \tilde{\mathbf{C}}(t)\Big)\, \mathbf{u} = \tilde{\mathbf{b}}(t)\,,$$

where now every term depends on time through the stabilization term. We denote with a tilde the matrices that contain the relative stabilization term, i.e.:

$$\begin{aligned} \tilde{\mathbf{M}}_{ij}(t) &= \int_\Gamma \varphi_j \varphi_i \,\mathbf{ds} + \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T \varphi_j\, \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_i\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \\ \tilde{\mathbf{A}}_{ij}(t) &= \epsilon \int_\Gamma \langle \nabla_{\mathcal{G}}\, \varphi_j, \nabla_{\mathcal{G}}\, \varphi_i\rangle_{\mathcal{G}} \,\mathbf{ds} + \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T \left(-\epsilon\, \Delta_{\mathcal{G}} \varphi_j\right) \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_i\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \\ \tilde{\mathbf{B}}_{ij}(t) &= \int_\Gamma \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_j\rangle_{\mathcal{G}}\, \varphi_i \,\mathbf{ds} + \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_j\rangle_{\mathcal{G}}\, \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_i\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \\ \tilde{\mathbf{C}}_{ij}(t) &= \int_\Gamma c\, \varphi_j \varphi_i \,\mathbf{ds} + \sum_{T\in\mathcal{T}(\Gamma)} \delta_T \int_T c\, \varphi_j\, \langle \mathbf{w}(t), \nabla_{\mathcal{G}}\, \varphi_i\rangle_{\mathcal{G}} \,\mathbf{ds}\,, \end{aligned}$$

and analogously for the vector of the right-hand-side:

$$\tilde{\mathbf{b}}_i(t) = \int_\Gamma f_h \, \varphi_i \, \mathrm{d}\mathbf{s} + \sum_{T \in \mathcal{T}(\Gamma)} \delta_T \int_T f_h \, \langle \mathbf{w}(t), \nabla_\mathcal{G} \, \varphi_i \rangle_\mathcal{G} \, \mathrm{d}\mathbf{s} \, .$$

Note that, in the $\mathcal{P}_1$ case, $\tilde{\mathbf{A}} = \mathbf{A}$ because the Laplacian of an affine function in zero. The system is then discretized in time in the interval $[0, t_f]$ by means of an implicit Euler scheme, which takes the form:

$$\left(\frac{1}{\Delta t^n} \, \tilde{\mathbf{M}}^{n+1} + \tilde{\mathbf{A}}^{n+1} + \tilde{\mathbf{B}}^{n+1} + \tilde{\mathbf{C}}^{n+1} \right) \mathbf{u}^{n+1} = \tilde{\mathbf{b}}^{n+1} + \frac{1}{\Delta t^n} \, \tilde{\mathbf{M}}^{n+1} \, \mathbf{u}^n \, ,$$

where $\Delta t^n = t^{n+1} - t^n$, as the usual notation.

6.2 Numerical results

Our numerical tests are designed to experimentally verify the applicability of the developed schemes to the solution of advection-diffusion-reaction equation on surfaces, eq. (6.1). Specifically, we want to test the experimental convergence of the schemes on known solutions, from simple ones to more complex expressions.

First, we validate the codes on literature examples. In these test cases we consider first a simple flat unit square domain. Next we look at the example proposed in [36] and work on the surface defined as $\Gamma = \{\mathbf{x} \in \mathbb{R}^3 \, | \, \mathcal{H}(\mathbf{x}) = 0\}$ with the level set function:

$$\mathcal{H}(\mathbf{x}) = \frac{(x^1)^2}{4} + (x^2)^2 + \frac{4(x^3)^2}{\left(1 + \frac{1}{2}\sin(\pi x^1)\right)^2} - 1 \, . \tag{6.15}$$

Then, we run simulations on the following three surfaces. We consider the same hyperboloid (hyperboloid-central-bump (HCB)) used in the FV and DG test cases (see section 4.3) and described on the subset $U = [-3, 3] \times [-3, 3] \subset \mathbb{R}^2$ by the parametrization given by the height function:

$$x^3(x^1, x^2) = -\frac{4}{5}\sqrt{(x^1)^2 + (x^2)^2 + 1} \, . \tag{6.16}$$

Subsequently, we consider the unit sphere. For simplicity we take only a portion of the sphere so that an explicit parametrization can be used with one single chart. However, no real difficulties would arise if an atlas formed by more than one chart was used. Thus the boundary of the surface corresponds to $\partial U = \{(x^1, x^2) \, | \, (x^1)^2 + (x^2)^2 = 0.75\} \subset \mathbb{R}^2$. Finally we use the more complex surface proposed in [37], and already used in the previous chapters to addresses a fully three-dimensional case. This surface is defined starting from the subset $U = [-10, 10] \times [-4, 4] \subset \mathbb{R}^2$ and using the height function:

$$x^3(x^1, x^2) = -\frac{1}{500}(x^1)^3 - \frac{1}{100}x^1(x^2)^2 \, . \tag{6.17}$$

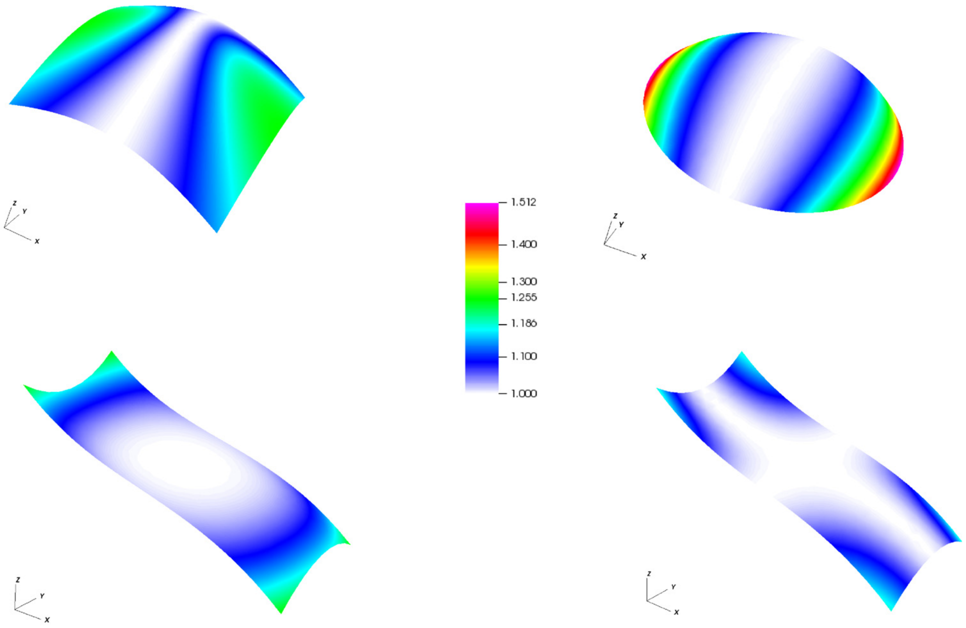

Figure 6.2: Surfaces and spatial distribution of the metric coefficients, $h_{(1)}$ for the hyperboloid-central-bump (top left) and the sphere (top right) and of $h_{(i)}$, $i = 1, 2$, for the Fully 3D surface (bottom), respectively.

	$\epsilon = 1$				$\epsilon = 10^{-4}$			
h_ℓ	$E_{L^2}(u)$	eoc$_\ell$	$E_{L^2}(\nabla_{\mathcal{G}}\, u)$	eoc$_\ell$	$E_{L^2}(u)$	eoc$_\ell$	$E_{L^2}(\nabla_{\mathcal{G}}\, u)$	eoc$_\ell$
0.400	7.10E-04		1.62E-01		3.36E-03		1.63E-01	
0.200	1.79E-04	1.99	8.10E-01	1.00	1.14E-03	1.55	8.13E-02	1.00
0.100	4.44E-05	2.01	4.05E-02	1.00	3.10E-04	1.88	4.06E-02	1.00
0.050	1.11E-05	2.00	2.02E-02	1.00	7.82E-05	1.99	2.03E-02	1.00
0.025	2.77E-06	2.00	1.01E-02	1.00	1.96E-05	2.00	1.01E-02	1.00
0.125	6.92E-07	2.00	5.06E-03	1.00	4.89E-06	2.00	5.06E-03	1.00

Table 6.1: Flat domain, intrinsic approach: L^2 error norms for u and $\nabla_{\mathcal{G}}\, u$ and corresponding experimental convergence rates.

Figure 6.2 shows the geometry of the three different surfaces described by the parametrization, and the spatial distribution of the metric coefficients. In all cases, we assume that the relevant surface quantities are known or can be approximated at the nodes of the triangulation, and use the interpolation described above to define needed quantities at other points.

Flat domain

We start considering stationary problems with a reaction coefficient constant in space. The equation simplifies as follow:

$$-\,\epsilon\,\Delta u + \langle \mathbf{w}, \nabla\, u\rangle + c\, u = f \qquad \text{on} \quad \Gamma\,. \tag{6.18}$$

Here the equation contains the general symbols for gradient and divergence that will be specify later in the specific examples based on the approach we apply.

We consider a structured triangulation of a flat domain defined by a subset $U = [0,1]\times[0,1] \subset \mathbb{R}^2$ and a constant height function $x^3 \equiv 0$. We apply both the embedded and intrinsic approach. Note that, on this flat test case both approaches should collapse to the standard $\mathcal{P}_1$ Galerkin method. The convergence tests are carried out by using a mesh sequence built by refining an initial structured mesh with $h_0 = 0.4$ m and a total of 16 nodes. The mesh sequence is composed by 7 mesh levels built by uniform refinement, with the finest level characterized by $h_6 \approx 0.125$ m and 9409 nodes. We consider the stationary problem eq. (6.18), with constant velocity field $\mathbf{w} = (1,1)$, for the intrinsic case ($\mathbf{w} = (1,1,0) = \mathbf{w}_h$ for the embedded case), and reaction coefficient $c = 1$. We work with the exact solutions $u(\mathbf{x}) = x^1$ (to test exactness on linears) with diffusion coefficient $\epsilon = 10^{-4}$, and $u(\mathbf{x}) = x^1x^2$ with diffusion coefficients $\epsilon = 1, 10^{-4}$.

	$\epsilon = 1$			
h_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\Gamma_h} u)$	eoc_ℓ
0.428	2.02E-01		8.34E-01	
0.225	4.14E-02	2.47	2.85E-01	1.67
0.115	1.03E-02	2.07	1.29E-01	1.17
0.058	2.57E-03	2.02	6.29E-02	1.05
0.029	6.41E-04	2.01	3.13E-02	1.01

Table 6.2: Literature example, embedded approach: L^2 error norms for u and $\nabla_{\Gamma_h} u$ and corresponding experimental convergence rates.

In this case the mesh Peclet number $\mathrm{Pe} = \max_{T \in \mathcal{T}(\Gamma)} \mathrm{Pe}_T$ varies in the range 0.01-0.20 with a diffusion coefficient equal to 1, while Pe is in the range 10^1-10^4 in the second case.

The results show that the equation is solved exactly in the case of the linear solution $u(\mathbf{x}) = x^1$. Table 6.1 shows the L^2 errors and experimental orders of convergence for the solution and its gradient for the exact solution $u(\mathbf{x}) = x^1 x^2$. As expected, the results show second order convergence rates for the solution and first order for the gradient. Similar results are obtained applying the embedded approach, whose convergence rates are not reported here.

Comparison with a literature example

In this test case we consider an example presented in [36]. We are interested in this example to verify the implementation of the embedded code, and we applied only the embedded approach due to the complexity of the surface requiring more that one chart and of the corresponding mesh generation procedure. The surface is described in eq. (6.15). The example considers the Laplace-Beltrami equation on the surface. We want to verify the convergence for the exact solution $u = x^1 x^2$. To this aim, the forcing function is computed as $f(\mathbf{x}) = -\Delta_{\Gamma_h} u(\mathbf{x})$. We can calculate the forcing function in terms of the mean curvature of the surface as follows:

$$f(\mathbf{x}) = 2\mathbf{N}^1(\mathbf{x})\mathbf{N}^2(\mathbf{x}) + H(\mathbf{x}) \left(x^2 \mathbf{N}^1(\mathbf{x}) + x^1 \mathbf{N}^2(\mathbf{x}) \right) ,$$

where $\mathbf{N} = \frac{\nabla \mathcal{H}}{|\nabla \mathcal{H}|}$ is the surface normal.

In practice, we impose $\epsilon = 1$, $\mathbf{w} = \mathbf{0}$ and $c = 0$. For the convergence test we start from a coarse grid with average mesh parameter $h_0 = 0.428$ m and 162 surface nodes. The mesh sequence contains 5 levels, the finest characterized by a mesh parameter

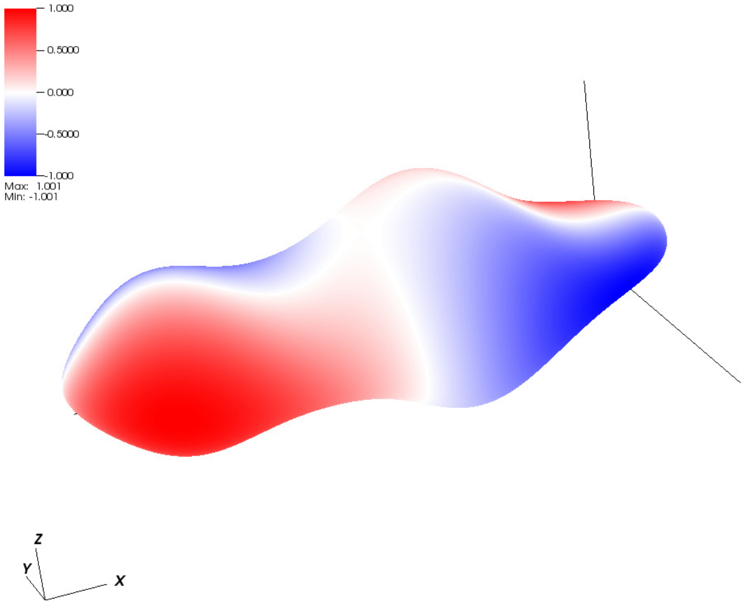

Figure 6.3: Literature example, embedded approach: numerical solution over the finest mesh ($h_4 = 0.029$ m) for the $u(\mathbf{x}) = x^1 x^2$ solution.

$h_4 = 0.029$ m and 40962 surface nodes. Figure 6.3 shows the numerical solution over the finest mesh. Table 6.2 shows the L^2 errors and experimental orders of convergence for the solution and its gradient. Consistently with the example presented in [36] we notice second order convergence rates for the solution and first order for the gradient as expected from theoretical results.

Hyperboloid-central-bump

We consider again the stationary problem eq. (6.18), with reaction coefficient $c = 1$. Two diffusion coefficients are tested $\epsilon = 1, 10^{-4}$, with the lowest value requiring stabilization. We work with the exact solution $u(\mathbf{x}) = x^1 x^2$ and compute the resulting forcing function f.

Embedded approach. In the case of the embedded approach, the surface is given by $\Gamma = \{\mathbf{x} \in \mathbb{R}^3 \mid \mathcal{H}(\mathbf{x}) = 0\}$, where the level function is:

$$\mathcal{H}(\mathbf{x}) = -\frac{4}{5}\sqrt{(x^1)^2 + (x^2)^2 + 1} - x^3(x^1, x^2)\,.$$

	$\epsilon = 1$				$\epsilon = 10^{-4}$			
h_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\Gamma_h} u)$	eoc_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\Gamma_h} u)$	eoc_ℓ
0.560	6.14E-02		8.86E-01		2.43E-01		1.31E+00	
0.281	1.37E-02	2.17	4.44E-01	1.00	7.87E-02	1.63	6.94E-01	0.92
0.140	3.23E-03	2.09	2.22E-01	1.00	2.38E-02	1.73	3.63E-01	0.94
0.070	7.90E-04	2.03	1.11E-01	1.00	6.91E-03	1.79	1.95E-01	0.90
0.035	1.96E-04	2.01	5.56E-02	1.00	1.92E-03	1.85	9.91E-02	0.98
0.018	4.90E-05	2.00	2.78E-02	1.00	4.87E-04	1.98	4.30E-02	1.21

Table 6.3: HCB, embedded approach: L^2 error norms for u and $\nabla_{\Gamma_h} u$ and corresponding experimental convergence rates.

The right-hand side forcing function is:

$$f(\mathbf{x}) = -\epsilon\, \Delta_{\Gamma_h} u + \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} u \right\rangle + u\,,$$

where the operators are defined as in section 6.1.1. The velocity field is constant and given by $\mathbf{w} = (1, 1)$, yielding a projected velocity $\mathbf{w}_h = \mathbb{P}_h \mathbf{w}$. Experimental convergence of the scheme is tested on a mesh sequence built by refining an initial Delaunay triangulation of U with mesh parameter $h_0 \approx 0.560$ m for a total of 424 surface nodes. The nodes of this planar triangulation are moved vertically according to eq. (6.16). The mesh sequence comprises 5 uniform refinements of the initial triangulation, with the finest level $\ell = 5$ characterized by $h_5 \approx 0.018$ m. We look at the L^2-norm of the difference between the approximated and exact values of the solution u and its gradient $\nabla_{\Gamma_h} u$, calculated by means of the trapezoidal and midpoint rules described in eqs. (6.10) and (6.11). Table 6.3 reports the convergence results for the cases $\epsilon = 1$ and $\epsilon = 10^{-4}$. Second order of convergence for the solution and first order convergence for its gradient are invariably obtained.

Intrinsic approach. For the intrinsic approach the right-hand side forcing function is computed by:

$$f(\mathbf{x}) = -\epsilon\, \Delta_{\mathcal{G}} u + \left\langle \mathbf{w}, \nabla_{\mathcal{G}} u \right\rangle_{\mathcal{G}} + u\,,$$

with the differential operators, defined in proposition 2.3.2, which contain the surface geometric information. The velocity field is defined by $\mathbf{w} = (1/h_{(1)}, 1/h_{(2)})$. The convergence test is carried out on the same mesh sequence used in the embedded approach, starting from a coarsest level characterized by $h_0 = 0.560$ m and 424 surface nodes, and consisting of 5 refinement levels. Table 6.4 reports the L^2-norms of the errors

	$\epsilon = 1$				$\epsilon = 10^{-4}$			
h_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\mathcal{G}}\, u)$	eoc_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\mathcal{G}}\, u)$	eoc_ℓ
0.560	4.68E-02		8.72E-01		1.30E-01		1.00E+00	
0.281	1.07E-02	2.13	4.38E-01	1.00	4.43E-02	1.56	5.12E-01	0.97
0.140	2.56E-03	2.07	2.19E-01	1.00	1.27E-02	1.80	2.50E-01	1.03
0.070	6.27E-04	2.03	1.10E-01	1.00	3.40E-03	1.91	1.24E-01	1.01
0.035	1.56E-04	2.01	5.49E-02	1.00	9.14E-04	1.90	6.27E-02	0.99
0.018	3.89E-05	2.00	2.75E-02	1.00	2.39E-04	1.93	3.08E-02	1.03

Table 6.4: HCB, intrinsic approach: L^2 error norms for u and $\nabla_{\mathcal{G}}\, u$ and corresponding experimental convergence rates.

	$\epsilon = 10^{-6}$			
h_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\Gamma_h}\, u)$	eoc_ℓ
0.329	2.13E-02		1.90E-01	
0.176	3.80E-03	2.76	5.72E-02	1.93
0.091	1.27E-03	1.65	1.55E-02	1.96
0.046	3.30E-04	2.00	6.78E-03	1.22
0.023	8.89E-05	1.92	3.94E-03	0.79
0.012	2.36E-05	1.92	2.38E-03	0.74

Table 6.5: Sphere, embedded approach: L^2 error norms for u and $\nabla_{\Gamma_h}\, u$ and corresponding experimental convergence rates.

for the solution and its gradient. The expected second and first order of convergence for the solution and its gradient is obtained.

Sphere

Here we solve eq. (6.18) on the unit sphere, considering a variant of the example proposed in Olshanskii et al. [60]. The diffusion parameter is set to be $\epsilon = 10^{-6}$, the reaction coefficient is $c = 1$, while the advection velocity vector is chosen to be tangent to the sphere and will be specified in the following paragraphs. We consider the exact solution given by:

$$u(\mathbf{x}) = \frac{x^1}{\pi} \arctan\left(\frac{x^2}{\sqrt{\epsilon}}\right) .$$

	$\epsilon = 10^{-6}$			
h_ℓ	$E_{L^2}(u)$	eoc_ℓ	$E_{L^2}(\nabla_{\mathcal{G}}\, u)$	eoc_ℓ
0.329	1.94E-02		1.96E-01	
0.176	3.41E-03	2.78	5.54E-02	2.02
0.090	1.09E-03	1.71	1.53E-02	1.94
0.046	2.87E-04	1.98	6.88E-03	1.18
0.023	7.93E-05	1.88	4.12E-03	0.75
0.012	2.14E-05	1.90	2.50E-03	0.72

Table 6.6: Sphere, intrinsic approach: L^2 error norms for u and $\nabla_{\mathcal{G}}\, u$ and corresponding experimental convergence rates.

Embedded approach. The sphere is defined by a level set function as $\Gamma = \{\mathbf{x} \in \mathbb{R}^3 \mid (x^1)^2 + (x^2)^2 + (x^3)^2 - 1 = 0\}$. Here, the tangential velocity field is expressed in the three-dimensional $\mathbf{x}$ components and is given by:

$$\mathbf{w} = \left(-x^3\sqrt{1-(x^2)^2}, 0, x^1\sqrt{1-(x^2)^2}\right) .$$

It is easy to verify that this vector field is tangent to the sphere. As before, the forcing function is computed by using the values of the exact solution to yield:

$$f(\mathbf{x}) = -\epsilon\, \Delta_{\Gamma_h} u + \left\langle \mathbf{w}_h, \nabla_{\Gamma_h} u \right\rangle + u\,.$$

The circular subset $U = \{(x^1, x^2) \,|\, (x^1)^2 + (x^2)^2 \leq 0.75\}$ is discretized by a Delaunay triangulation and then the nodal values are raised using the height function $x^3(x^1, x^2) = \sqrt{1-(x^1)^2-(x^2)^2}$. For the convergence test we start from a coarse grid with average mesh parameter $h_0 = 0.329$ m and 39 surface nodes. The mesh sequence is obtained by uniform refinements for a total of 6 mesh levels (the finest mesh level $\ell = 5$ has an average mesh parameter $h_5 = 0.012$). Due to the boundary layer we compute the error in the domain $D = \{\mathbf{x} \in \Gamma \mid |x^2| \geq 0.3\}$. Table 6.5 shows the L^2 errors and experimental orders of convergence for the solution and its gradient. Consistently with the previous test cases, we notice second order convergence rates for the solution and first order for the gradient. Convergence rates slightly smaller that one are attributable to the complexity of the solution in the pole of the sphere, as shown in fig. 6.4

Intrinsic approach. The mesh sequence used to test convergence is the same as in the embedded approach. In this case, we define the advection velocity vector using

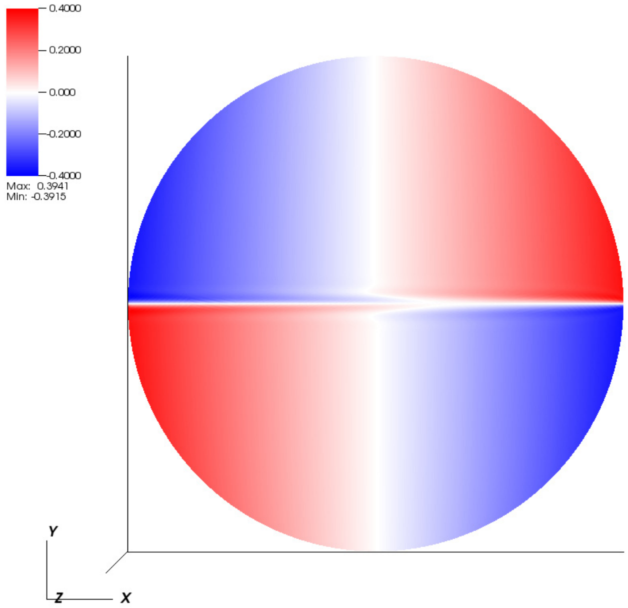

Figure 6.4: Sphere, intrinsic approach: numerical solution over the finest mesh ($h_5 = 0.012$ m).

local coordinates and, in the **s** components, we can write:

$$\mathbf{w} = \left(-\frac{1-(s^2)^2}{h_{(1)}}, 0 \right) .$$

Figure 6.4 shows the numerical solution over the finest grid characterized by h_5=0.012 m and 30977 surface nodes. Table 6.6 reports the L^2-error norms computed in $D = \{\mathbf{x} \in \Gamma \mid |x^2| \geq 0.3\}$. Again, second and first order convergence for the solution and its gradient, respectively, are observed, with the same slight loss of convergence in the solution gradient already seen in the embedded approach.

Fully 3D surface (intrinsic approach)

For the final test case we solve the complete non-stationary problem eq. (6.1) with the intrinsic approach. The mesh used in this test case was obtained from a Delaunay triangulation of $U = [-10, 10] \times [-4, 4]$ and elevated using the height function of eq. (6.17). The mesh parameter is $h = 0.62$ m for a total of 880 surface nodes. We consider a diffusion coefficient $\epsilon = 10^{-2}$, and a zero reaction term. The advection velocity vector varies in space and in time and is given by the finite volume solution of the Intrinsic Shallow Water Equations defined in section 4.3. The simulation for the

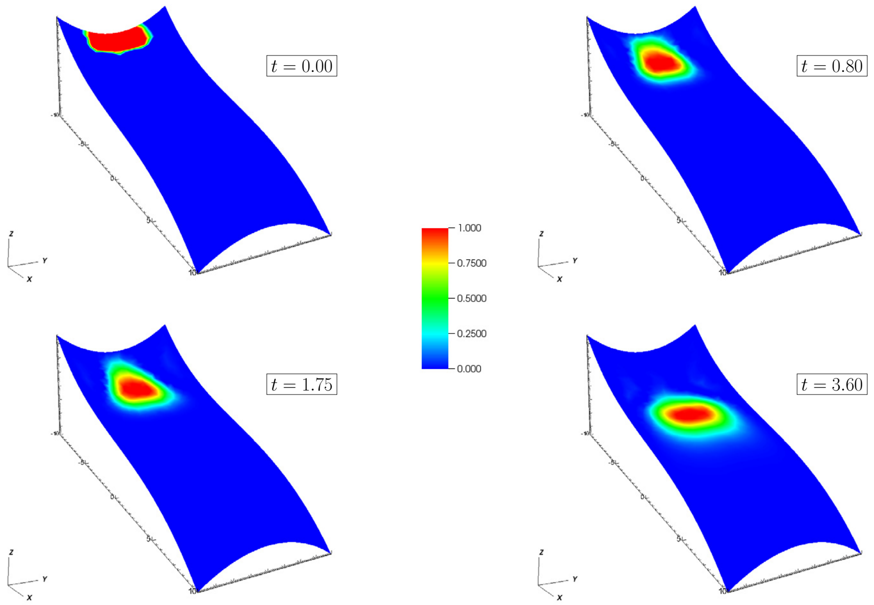

Figure 6.5: Fully 3D surface: advection-diffusion equation solved with FEM code on a mesh with $h =$ 0.62 m. The concentration in reported at different times, $t =$ 0.00, 0.80, 1.75, 3.60 s.

ISWE is set to be a dam-break wet bed problem, with initial condition that considers uniform water depth of 2.0 m upstream of $x^1 =$ -8.5 m, and 1.0 m downstream. No flow conditions are imposed in all boundaries, except at $x^1 = 10$ m where the outlet is located. We run the simulation up to 2 s and store all the values of the velocities. The time steps are calculated in order to satisfies a CFL condition, with value CFL $\approx$ 0.1.

The initial condition for the concentration u is represented in fig. 6.5, top left. The advection velocity $\mathbf{w}$ is read in from the FV solution and the time steps are defined as the time difference between successive input velocity fields (velocities are not interpolated in time). Because of the relatively large velocity magnitude and small diffusion, the mesh Peclet numbers vary in the range 10^1-10^4, thus requiring stabilization. The simulated concentration is reported in fig. 6.5 at three different times. Qualitatively the solution shows the expected behavior. The initial plume moves following the velocity field, concentrating at the beginning towards the upper depression and starting to spread wider when reaching the central mound. This simulation shows that SUPG stabilization with parameter eq. (6.14) is very effective in maintaining monotonicity, although possibly overcompressive.

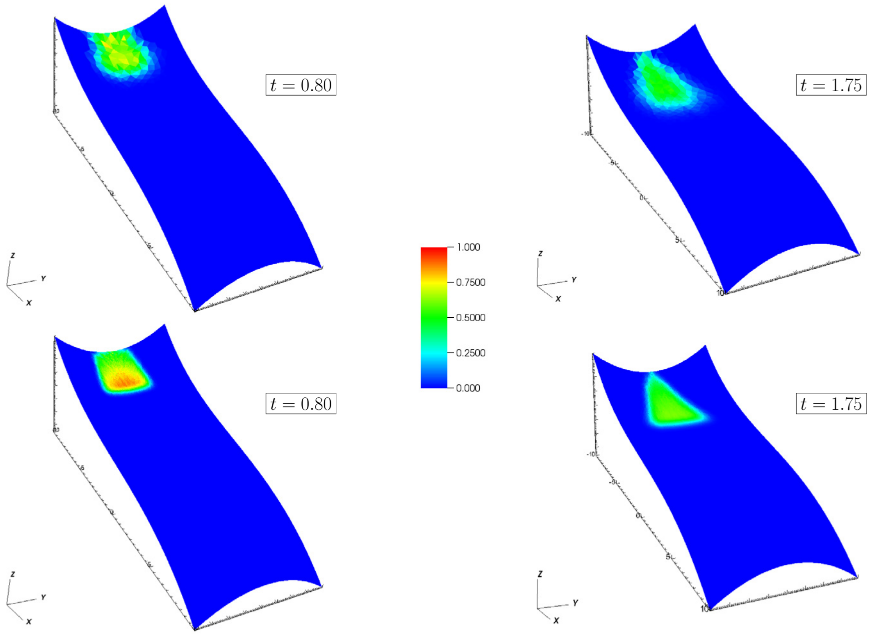

Figure 6.6: Fully 3D surface: transport equation solved with FV code on two different levels of refinement of a mesh (a coarse mesh with $h = 0.62$ m, top panels, and a refined mesh with mesh parameter $h = 0.142$, bottom panels) at times $t = 0.80,\ 1.75$ s.

Comparison with Intrinsic Finite Volume. To complete the treatment of this test case, we go back to our IFV scheme applied to the surface scalar transport equation. We take in input the advection velocities from the shallow water problem, and we consider a simple upwind method for the flux on the edges. We run the simulation over two different refinements of the mesh: a coarse mesh (the same as the FEM simulation) with $h =$ 0.62 m and 1656 cells; a fine mesh, obtained by two uniform refinement of the coarse one, with $h = 0.142$ and 34496 surface triangles. As in section 5.3.1, in this case the time steps need to adapt to the time interval of the input velocity, thus imposing a variable in time CFL number to the simulation. We proceed with linear interpolation to assign the correct values of the velocities at the actual simulation times, making sure to never exceed CFL≈ 0.5. The resulting values of the CFL number at each time steps are far from optimal and large numerical viscosity is expected.

Figure 6.6, top panels, show the simulation over the coarse mesh at times $t = 0.80, 1.75$ s, while fig. 6.6, bottom panels, contain the simulation over the fine mesh at the same times $t = 0.80, 1.75$ s. We can see that the behavior of the solution is qualitatively similar to the solution obtained by intrinsic FEM. The most striking difference is that the plume is more spread spatially as a consequence of the large numerical diffusion typically introduced by a first order Godunov scheme. This is enhanced by the fact that times step sizes are dictated by input data and are not optimal in minimizing numerical diffusion.

6.3 Future developments

Starting from the promising results of the intrinsic $\mathcal{P}_1$-FEM approach developed in this thesis, it is natural to take into consideration the method of edge-based entropy-viscosity (EV) recently proposed by [44], and extend it to PDEs on surfaces. In our developed intrinsic approach this extension becomes straight forward, showing again the advantages of our approach.

We work on the scalar conservation law written as:

$$\partial_t u + \nabla_{\mathcal{G}} \cdot \mathbf{F}(u) = 0 \quad \text{in } \Gamma \,, \tag{6.19}$$

subject to the initial condition $u(\mathbf{s}, 0) = u_0$ and appropriate boundary conditions. The EV algorithm is based on an explicit $\mathcal{P}_1$ finite element approximation. We consider the computational meshes $\mathcal{T}(\Gamma)$ and $\mathcal{T}_h(\Gamma)$, and we define on the surface triangulation the finite element space $\mathcal{V}_h^{\Gamma}$ as in the previous section, see eq. (6.8). For $i = 1, \ldots, N^{dof}$, we denote with φ_i the i-th basis functions of the space, with S_i its support and $\mathcal{I}(S_i)$ is the set of indices of basis functions that have a non-zero value in S_i.

Entropy-viscosity method. The hyperbolic behavior of an equation makes the computations challenging and high-order methods are known to produce spurious oscillations in shocks, as previously seen. There exists a large class of methods to solve this problem in an efficient way, but the majority of them relies on limiters to avoid spurious oscillations, which are non-trivial to develop. We start by presenting the original entropy-viscosity approach formulated in [45], which avoids the use of limiters and non-oscillatory reconstructions by adding a degenerate nonlinear dissipation to the numerical discretization of the equation or the system at hand. The viscosity coefficient is based on the local size of an entropy production. Scalar conservation equations have many entropy pairs and most physical systems have at least one entropy function satisfying an auxiliary entropy inequality. The entropy satisfies a conservation equation only in the regions where the solution is smooth and satisfies an inequality in shocks; this inequality then becomes a selection principle for the physically relevant solution. The amount of violation of the entropy equation is called entropy production. By making the numerical diffusion to be proportional to the entropy production, we add a large numerical dissipation in the shock regions and almost no dissipation in the regions where the solution remains smooth. This simple idea is mesh and approximation independent and can be applied to equations or physical systems that are supplemented with an entropy inequality.

Consider eq. (6.19). Using a Galerkin approximation, the idea is to construct a stabilization term based on a nonlinear entropy viscosity formulation as:

$$\int_\Gamma \partial_t u_h \, v \, \mathrm{d}\mathbf{s} + \int_\Gamma \nabla_{\mathcal{G}} \cdot \mathbf{F}(u_h) v \, \mathrm{d}\mathbf{s} + \sum_{T \in \mathcal{T}} \int_T \nu_T \left\langle \nabla_{\mathcal{G}} u_h, \nabla_{\mathcal{G}} v \right\rangle_{\mathcal{G}} \mathrm{d}\mathbf{s} = 0 \, .$$

Here ν_T is a nonlinear artificial viscosity coefficient for the cell T and it is defined by:

$$\nu_T = \min\{\nu_T^L, \nu_T^E\} \, ,$$

where

$$\nu_T^L = C_{\max} l_T \left\| \mathbf{J}_{\mathbf{F}} \right\|_{L^\infty(T)} \qquad \text{and} \qquad \nu_T^E = \frac{C_E l_T^2 \left\| R_h(E(u_h)) \right\|_{L^\infty(T)}}{E(u_h) - \bar{E}(u_h)} \, ,$$

with $C_{\max}$ and C_E are user-defined constants, l_T is the length of the smallest edge of T, E is the entropy that is combined with an entropy flux $\mathbf{G}$ to satisfy an entropy inequality:

$$R(u_h) = \partial_t E(u_h) + \nabla_{\mathcal{G}} \cdot \mathbf{G}(u_h) \le 0 \, ,$$

and $\bar{E}$ is the globally-averaged value of the normalized entropy viscosity, i.e.,

$$\bar{E} = \frac{1}{\mathcal{A}_\Gamma} \int_\Gamma E(u_h) \, \mathrm{d}\mathbf{s} \, .$$

We denote by $R(u_h)$ the entropy residual and by R_h its time approximation, which for example can be computed using first-order differences in time:

$$\begin{aligned} R_h(E(u_h)) &= \frac{E(u_h^n) - E(u_h^{n-1})}{\Delta t^n} + \nabla_{\mathcal{G}} \cdot \mathbf{F}(u_h^n) \, , \\ \text{or} \quad &= \frac{E(u_h^n) - E(u_h^{n-1})}{\Delta t^n} + \left\langle \mathbf{J}_{\mathbf{F}}(u_h^n), \nabla_{\mathcal{G}} E(u_h^n) \right\rangle_{\mathcal{G}} \, . \end{aligned}$$

Examples of entropy-entropy flux pairs for the linear transport equation (i.e., $\mathbf{F}(u_h) = \mathbf{w} \, u_h$) include $E(u_h) = \frac{1}{2} u_h^2$, $E(u_h) = |u_h|^p$, $p \ge 1$ or $E(u_h) = \log\left(|u_h(1 - u_h)| + 10^{-14}\right)$, and $\mathbf{G} = \mathbf{w} E$.

Second order maximum principle preserving scheme. Subsequently, the approach in [41] updates this original EV scheme and as a result offers a fairly general second-order scheme for nonlinear scalar problems that satisfies a maximum principle. The setting is very similar to the original one, and look at H^1 conforming spaces based on polynomial approximations. It does require a function space where the lumped mass matrix is positive definite. The main ingredients of the method are: i) the use of the graph Laplacian in the viscous regularization term; ii) the use of the Flux-Corrected-Transport (FCT) paradigm to perform limiting between the low and high-order approximations; iii) the use of a mass-lumping correction term. The basic weak form is:

$$\int_\Gamma \partial_t u_h \, v \, \mathrm{d}\mathbf{s} + \int_\Gamma \nabla_{\mathcal{G}} \cdot \mathbf{F}(u_h) v \, \mathrm{d}\mathbf{s} + \mathrm{s}_h(u_h, v) = 0 \,,$$

where again the third term is a viscous stabilization term. Here, the bilinear form $\mathrm{s}_h(\cdot,\cdot)$ can be written as:

$$\mathrm{s}_h(u_h, v) = \sum_{T\in\mathcal{T}(\Gamma)} \sum_{i,j\in\mathcal{I}(T)} \nu_T \, u_j \, v_i \; \mathrm{s}_h(\varphi_j, \varphi_i) \,,$$

with

$$\mathrm{s}_h(\varphi_j, \varphi_i) = \begin{cases} -\frac{1}{2}\mathcal{A}_T & i \neq j \,, i, j \in \mathcal{I}(T) \\ \mathcal{A}_T & i = j \,, i, j \in \mathcal{I}(T) \\ 0 & \text{otherwise} \end{cases} \,,$$

a definition that reminds of a graph Laplacian. The local artificial coefficient ν_T can be defined by:

$$\nu_T = \max_{i\neq j\in\mathcal{I}(T)} \frac{\left| \displaystyle\int_{S_{ij}} \langle \mathbf{J}_{\mathbf{F}}(u_h), \nabla_{\mathcal{G}} \varphi_j \rangle_{\mathcal{G}} \, \varphi_i \, \mathrm{d}\mathbf{s} \right|}{- \displaystyle\sum_{\substack{K\in\mathcal{T}(\Gamma) \\ K\subset S_{ij}}} \mathrm{s}_h(\varphi_j, \varphi_i)} \,, \tag{6.20}$$

where $S_{ij} = S_i \cap S_j$ is the intersection of the supports of the i, j-th basis functions. In this case, the solution update can be written as:

$$u_i^{n+1} = u_i^n - \Delta t \, m_i^{-1} \sum_{T\in S_i} \left(\nu_T^n \, \mathrm{s}_h(u_h^n, \varphi_i) + \int_T \nabla_{\mathcal{G}} \cdot \mathbf{F}(u_h^n) \varphi_i \, \mathrm{d}\mathbf{s} \right) \,,$$

where $m_i := \int_{S_i} \varphi_i \, \mathrm{d}\mathbf{s}$ is the i-th coefficient of the lumped mass matrix.

Remark 6.1. In [41] the authors show that the low-order scheme satisfies a local discrete maximum principle. Let β be a suitable upper bound on the wave speed, i.e.,

$$\beta = \sup_{v\in\{u^*_{min}, u^*_{max}\}} \|\mathbf{J}_{\mathbf{F}}(v)\| \,,$$

then the low-order order scheme satisfies a local discrete maximum principle under the CFL condition:

$$\frac{\beta \Delta t^n}{l} \leq \frac{1}{\lambda(1 + \rho^{-1})} \,.$$

Here, λ is a constant based on the integrals of the basis functions, l is the global minimum mesh size, i.e., $l = \min_T l_T$, and $\rho = 1/2$ in the case of a triangulation.

The above technique is then extended to make it higher-order using the notion of entropy-viscosity. We call *low-order viscosity* ν_T^L the coefficient defined in eq. (6.20). Then, we define an *high-order viscosity* term based on the entropy residual. Explicitly,

$$\nu_T^H = \min\left(\nu_T^L, \frac{C_T R_T + C_{\mathbf{J}} \max_{\sigma\in\partial T} \mathbf{J}_\sigma}{\left\|E(u_h) - \bar{E}(u_h)\right\|_{L^\infty(T)}}\right), \tag{6.21}$$

where $\mathbf{J}_\sigma$ is the edge-based stabilization term, which at time t^n is given by:

$$\mathbf{J}_\sigma^n = \left\|\langle \mathbf{J}_{\mathbf{F}}(u_h^n), \mu\, \partial_\nu E(u_h^n)\rangle_{\mathcal{G}}\right\|_{L^\infty(\sigma)},$$

and μ the local normal vector to the edge. The key improvement over earlier versions of the method is that the high-order viscosity coefficient does not require a notion of mesh-size. However, the constants C_T and $C_{\mathbf{J}}$ are still to be defined, thus the method is not yet parameter free. The applications investigated in [41] work satisfactorily with $C_T = 1$ and $C_{\mathbf{J}} = 4$.

Moreover, due to the high dispersion induced in working with a lumped mass matrix, it is convenient to work with a consistent mass matrix even though this strategy makes the method loose the maximum principle property. To partially remedy this issue, [42] proposes a conservative treatment of the mass matrix. Specifically, they define $\mathcal{B} = (\mathbf{M}_L - \mathbf{M}_C)\,\mathbf{M}_L^{-1}$, where $\mathbf{M}_L$ and $\mathbf{M}_C$ are the lumped and consistent mass matrix, respectively. Then, the consistent mass inverse can be formally approximated as:

$$\mathbf{M}_C^{-1} = \mathbf{M}_L^{-1}\left(\mathbb{I} + \mathcal{B} + \mathcal{B}^2 + \dots\right)$$

Finally, the high-order solution can be approximated formally by:

$$\mathbf{U}_H^{n+1} = \mathbf{U}^n - \Delta t^n\, \mathbf{M}_L^{-1}\,(\mathbb{I} + \mathcal{B})\,\mathbf{F}_H\,,$$

where $\mathbf{U}^n = \{u_i^n\}$ is the vector of the degrees of freedom at time t^n.

Edge-based viscosities. More recently, the entropy-viscosity technique moved to using edge-based viscosities. Guermond and Popov [44] lay out a second-order invariant domain preserving approximation based on the edge-based formulation. The approach differs from [41] and suggests that limiting viscosities attain better solutions than directly limiting the solution. Authors also identify some methods that are maximum principle preserving but do not converge to the correct entropy solution when the flux is non-convex. In particular, Galerkin FEM with just FCT may converge to entropy violating weak solutions if the flux is nonlinear.

The scheme uses the observation that $\mathbf{F}(u_h^n) = \sum_{j=1}^{N^{dof}} \mathbf{F}(u_j^n)\varphi_j$ if $\mathbf{F}$ is linear, to consider a formally second-order accurate approximation of the form:

$$\int_\Gamma \nabla_{\mathcal{G}} \cdot \mathbf{F}(u_h^n)\varphi_i \,\mathrm{d}\mathbf{s} \approx \sum_{j\in\mathcal{I}(S_i)} \mathbf{F}(u_j^n) \int_\Gamma (\nabla_{\mathcal{G}}\, \varphi_j)\varphi_i \,\mathrm{d}\mathbf{s}\,,$$

where we denote by $u_h^n = \sum_i u_i^n \varphi_i$ the approximate solution at time t^n written with respect to the nodal basis functions φ_i, for $i = 1, \dots, N^{dof}$. The generic formulation of the edge-based updates with lumped mass matrix can be written as:

$$m_i \,\frac{u_i^{n+1} - u_i^n}{\Delta t^n} + \sum_{j\in\mathcal{I}(S_i)} \left(\left\langle \mathbf{F}(u_j^n), \mathbf{c}_{ij} \right\rangle_{\mathcal{G}} - d_{ij}^n u_j^n \right) = 0\,, \tag{6.22}$$

where

$$m_i = \int_\Gamma \varphi_i \,\mathrm{d}\mathbf{s}\,, \qquad \mathbf{c}_{ij} = \int_\Gamma \varphi_i\, \nabla_{\mathcal{G}}\, \varphi_j \,\mathrm{d}\mathbf{s}\,,$$

and the coefficient d_{ij}^n is an artificial viscosity for the pair (i, j) that requires:

$$d_{ij}^n \geq 0 \ \text{ if } i \neq j\,, \quad d_{ij}^n = d_{ji}^n\,, \quad d_{ii}^n := \sum_{i\neq j\in\mathcal{I}(S_i)} -d_{ij}^n\,.$$

Thus, $\sum_{j\in\mathcal{I}(S_i)} d_{ij}^n = 0$, which implies that the method is conservative. In fact, summing (6.22) over all the degrees of freedom gives:

$$\begin{aligned}
\int_\Gamma u_h^{n+1} \,\mathrm{d}\mathbf{s} &= \int_\Gamma u_h^n \,\mathrm{d}\mathbf{s} - \Delta t^n \sum_{i=1}^{N^{dof}} \sum_{j\in\mathcal{I}(S_i)} \left[\left\langle \mathbf{F}(u_j^n)), \mathbf{c}_{ij} \right\rangle_{\mathcal{G}} - d_{ij}^n u_j^n \right] = \\
&= \int_\Gamma u_h^n \,\mathrm{d}\mathbf{s} - \Delta t^n \sum_{i=1}^{N^{dof}} \sum_{j\in\mathcal{I}(S_i)} \left\langle \mathbf{F}(u_j^n), \mathbf{c}_{ij} \right\rangle_{\mathcal{G}} = \\
&\approx \int_\Gamma u_h^n \,\mathrm{d}\mathbf{s} - \Delta t^n \int_\Gamma \nabla_{\mathcal{G}} \cdot \left(\sum_{j\in\mathcal{I}(S_i)} \mathbf{F}(u_j^n)\varphi_j \right) \mathrm{d}\mathbf{s}\,.
\end{aligned}$$

Moreover, this implies also that eq. (6.22) can be re-written as:

$$m_i \,\frac{u_i^{n+1} - u_i^n}{\Delta t^n} + \sum_{j\in\mathcal{I}(S_i)} \left[\left\langle \mathbf{F}(u_j^n) - \mathbf{F}(u_i^n), \mathbf{c}_{ij} \right\rangle_{\mathcal{G}} - d_{ij}^n (u_j^n + u_i^n) \right] = 0\,, \tag{6.23}$$

or equivalently:

$$u_i^{n+1} = u_i^n \left(1 - \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\Delta t^n d_{ij}^n}{m_i} \right) + \sum_{i\neq j\in\mathcal{I}(S_i)} \frac{2\Delta t^n d_{ij}^n}{m_i}\, \bar{u}_{ij}^{n+1}\,, \tag{6.24}$$

where the auxiliary variable $\bar{u}_{ij}$ is defined at time t^{n+1} by

$$\bar{u}_{ij}^{n+1} = \frac{1}{2}(u_j^n + u_i^n) - \left\langle \mathbf{F}(u_j^n) - \mathbf{F}(u_i^n), \frac{\mathbf{c}_{ij}}{2d_{ij}^n} \right\rangle_{\mathcal{G}}\,. \tag{6.25}$$

The following observation motivates the choice of the artificial viscosity coefficient d_{ij}^n. In fact, if we set $\nu_{ij} := \mathbf{c}_{ij} / \left|\mathbf{c}_{ij}\right|$ and $\tau := \left|\mathbf{c}_{ij}\right| / 2d_{ij}^n$, eq. (6.25) takes the form:

$$\bar{u}_{ij}^{n+1} = \frac{1}{2}(u_j^n + u_i^n) - \tau \left\langle \mathbf{F}(u_j^n) - \mathbf{F}(u_i^n), \nu_{ij} \right\rangle_{\mathcal{G}} ,$$

which can be related to the solution of a Riemann problem. In particular, τ denote a "fake time" value, and the CFL condition motivates the definition for d_{ij}^n:

$$d_{ij}^n = \max\left\{ \lambda_{max}(\nu_{ij}, u_i^n, u_j^n) \left|\mathbf{c}_{ij}\right|, \lambda_{max}(\nu_{ji}, u_j^n, u_i^n) \left|\mathbf{c}_{ji}\right| \right\} , \qquad (6.26)$$

where $\lambda_{max}(\nu_{ij}, u_i^n, u_j^n)$ is the maximum speed of propagation at the interface.

Concluding remarks. This method has not been implemented concretely yet. It is our intention to do it in the near future. However, as seen from the above developments, its extension to surfaces is straight forward in our intrinsic approach, thus showing the advantage of our method with respect to previously used surface based schemes. Not only is the extension of the method immediate, but also all the theoretical properties are maintained and extended directly to the surface case. The only requirements in the development of sufficiently accurate quadrature rules.

7

Conclusions

We have presented a novel formulation of the intrinsic shallow water equations with variable topography. The SW equations are obtained by integrating the Navier-Stokes equations along local normals defined on a local reference system anchored on the bottom surface. The resulting reduced model is written in a form intrinsic to the bottom geometry. The formulation is a second order approximation of the NS equation, is rotational invariant, maintains the lake-at-rest solution, and admits a conserved energy in case of no stresses.

The main advantage of the proposed intrinsic balance system of hyperbolic equations is that source terms contain only information related to bottom slope and curvatures, and not to the velocity field. The geometrically intrinsic description of the equations, and thus the existence of an intrinsic diverge theorem, allows a direct derivation of a well-balanced Godunov finite volume discretization defined on a bottom triangulation. The geometric information of the bed surface is assumed to be available only at the triangulation nodes. Careful interpolation together with a discretized version of parallel transport is used to define approximate tangent planes on the triangle edges and barycenters. As typical of the Godunov approach, flux evaluation proceeds by approximately solving a Riemann problem on the edges. Appropriately modified midpoint quadrature rules are used to evaluate integrals over curvilinear edges and triangles.

The first order FV method is then extended to the Discontinuous Galerkin framework. The function space is set in the intrinsic formulation and quadrature rules on surfaces of appropriate accuracy are developed. We implemented $\mathcal{P}_1$ basis functions together with a second order time-discretization technique to obtain a second order scheme in space and time for the numerical solution of both the scalar transport equation and the ISWE model on surfaces.

The FV and DG schemes are tested on several realistic examples showing that optimal order of convergence is obtained for smooth solutions. Mass conservation properties are verified by looking at streamflows across several control sections, and results are shown in the FV setting. The results show that the approaches are accurate and robust and can be effectively used to solve hyperbolic systems on general bottom topographies. Further studies concerning well-balance properties need to be done for the DG formulation. However, the main issue that needs to be addressed in the future is how to deal with the presence of non-autonomous fluxes and thus the definition of the Riemann problems.

Finally, the $\mathcal{P}_1$ Finite Element Method has been adapted to the intrinsic approach for the solution of advection-diffusion-reaction equations on surfaces. Numerical test cases to verify the applicability of the scheme are reported. The intrinsic formulation is compared to another surface finite element code already presented in the literature, and test cases considering exact solutions are used to show the convergence rates of the scheme.

Several improvements of out intrinsic formulation merit attention in the near future. Errors involved in the evaluation of the geometrical quantities on the surface need to be assessed in the case of surfaces defined starting from measured data, such as remotely sensed digital elevation maps. Moreover, the modeling of a movable bed, to take into account erosion and sedimentation, has to be properly addressed within the context of complex terrains to obtain robust and reliable predictions of natural phenomena as hyper-concentrated and debris flows.

Bibliography

[1] M. Abate and F. Tovena. *Geometria Differenziale.* Springer-Verlag Italia, Milano, Italy, 2011.

[2] M. Abate and F. Tovena. *Curves and Surfaces.* Springer-Verlag Italia, Milano, Italy, 2012.

[3] V. Aizinger and C. Dawson. A discontinuous Galerkin method for two-dimensional flow and transport in shallow water. *Adv. Water Resour.*, 25(1):67–84, Jan. 2002.

[4] B. Andreianov and K. Sbihi. Well-posedness of general boundary-value problems for scalar conservation laws. *Trans. AMS*, 367(6):pp. 3763–3806, June 2015. doi: 10.1090/S0002-9947-2015-05988-1.

[5] B. Andreianov, K. H. Karlsen, and N. H. Risebro. On vanishing viscosity approximation of conservation laws with discontinuous flux. *Netw. Heterog. Media*, 5(3): pp.617–633, 2010. doi: 10.3934/nhm.2010.5.617.

[6] B. Andreianov, K. H. Karlsen, and N. H. Risebro. A theory of l^1-dissipative solvers for scalar conservation laws with discontinuous flux. *Arch. Rational Mech. Anal.*, 201(1):pp.27–86, Jan. 2011. doi: 10.1007/s00205-010-0389-4.

[7] C. Aricò and C. Lo Re. A non-hydrostatic pressure distribution solver for the nonlinear shallow water equations over irregular topography. *Adv. Water Resour.*, 98:47–69, 2016.

[8] E. Audusse, F. Bouchut, M. Bristeau, R. Klein, and B. Perthame. A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. *SIAM J. Sci. Comput.*, 25(6):2050–2065, 2004. doi: 10.1137/S1064827503431090.

[9] E. Audusse, C. Chalons, and P. Ung. A simple well-balanced and positive numerical scheme for the shallow-water system. *Comm. Math. Sci.*, 13(5):1317–1332, 2015.

[10] Y. Bazilevs, V. Calo, T. Tezduyar, and T. J. R. Hughes. Yzβ discontinuity-capturing for advection-dominated processes with application to arterial drug delivery. *Int. J. Numer. Methods Fluids*, 54:593–608, 06 2007. doi: 10.1002/fld.1484.

[11] C. Berthon and C. Chalons. A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations. *Math. Comp.*, 85(299): 1281–1307, May 2016.

[12] E. Bertolazzi and G. Manzini. A cell-centered second-order accurate finite volume method for convection–diffusion problems on unstructured meshes. *ESAIM - Math. Model. Num. Anal.*, 2004.

[13] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, 1975.

[14] F. Bouchut. *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*, volume 2/2004. Birkhäuser Verlag, Basel, Switzerland, 01 2004.

[15] F. Bouchut and S. Boyaval. A new model for shallow viscoelastic fluids. *Math. Mod. Meth. Appl. Sci.*, 23(8):1479–1526, 2013.

[16] F. Bouchut and M. Westdickenberg. Gravity driven shallow water models for arbitrary topography. *Comm. Math. Sci.*, 2(3):359–389, Sept. 2004.

[17] F. Bouchut, A. Mangeney-Castelnau, B. Perthame, and J.-P. Vilotte. A new model of Saint Venant and Savage–Hutter type for gravity driven shallow water flows. *C. R. Math. Acad. Sci. Paris*, 336:531–536, Jan. 2003.

[18] A. Canestrelli, A. Siviglia, M. Dumbser, and E. F. Toro. Well-balanced high-order centered schemes for non-conservative hyperbolic systems. Applications to shallow water equations with fixed and mobile bed. *Adv. Water Resour.*, 32(6):834–844, June 2009.

[19] A. Canestrelli, M. Dumbser, A. Siviglia, and E. F. Toro. Well-balanced high-order centered schemes on unstructured meshes for shallow water equations with fixed and mobile bed. *Adv. Water Resour.*, 33(3):291–303, Mar. 2010.

[20] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab. Optimal a priori error estimates for the hp-version of the local discontinuous galerkin method for convection-diffusion problems. *Math. Comp.*, 71(238):455–478, Apr. 2002. doi: 10.1090/S0025-5718-01-01317-5.

[21] S. Chippada, C. Dawson, M. Martinez, and M. Wheeler. A Godunov-type finite volume method for the system of shallow water equations. *Comp. Methods App. Mech. Engrg.*, 151(1):105–129, 1998.

[22] V. T. Chow. *Open-channel hydraulics.* McGraw-Hill, New York, 1959.

[23] B. Cockburn and C. Dawson. Some extensions of the local discontinuous Galerkin method for convection-diffusion equations in multidimensions. In *The Proceedings of the Conference on the Mathematics of Finite Elements and Applications: MAFELAP X*, pages 225–238, 01 2000.

[24] B. Cockburn and C. Shu. The local discontinuous galerkin method for time-dependent convection-diffusion systems. *SIAM J. Num. Anal.*, 35(6):2440–2463, 1998. doi: 10.1137/S0036142997316712.

[25] B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Math. Comp.*, 52(186):411–435, 1989.

[26] B. Cockburn and C.-W. Shu. The Runge-Kutta local projection P^1 -discontinuous-Galerkin finite element method for scalar conservation laws. *ESAIM-Math. Model. Num.*, 25(3):337–361, 1991.

[27] B. Cockburn and C.-W. Shu. The Runge-Kutta discontinuous Galerkin method for conservation laws V: Multidimensional systems. *J. Comp. Phys.*, 141(2):199 – 224, 1998.

[28] B. Cockburn, S.-Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-dimensional systems. *J. Comp. Phys.*, 84(1):90 – 113, 1989.

[29] B. Cockburn, S. Hou, and C.-W. Shu. The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. the multidimensional case. *Math. Comp.*, 54:545–581, 1990.

[30] R. Codina. Comparison of some finite element methods for solving the diffusion-convection-reaction equation. *Comp. Methods App. Mech. Engrg.*, 156(1-4):185–210, 1998. doi: 10.1016/S0045-7825(97)00206-5.

[31] A. Decoene, L. Bonaventura, E. Miglio, and F. Saleri. Asymptotic derivation of the section-averaged shallow water equations for natural river hydraulics. *Math. Mod. Meth. Appl. Sci.*, 19(03):387–417, 2009.

[32] O. Delestre, C. Lucas, P.-A. Ksinant, F. Darboux, C. Laguerre, T. N. T. Vo, F. James, and S. Cordier. SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies. *Int. J. Numer. Methods Fluids*, 72(3):269–300, 2013.

[33] M. P. Do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice-Hall, Englewood Cliffs, New Jersey, 1976.

[34] A. S. Donahue, Y. Zhang, A. B. Kennedy, J. J. Westerink, N. Panda, and C. N. Dawson. A Boussinesq-scaled, pressure-Poisson water wave model. *Ocean Model.*, 86(C):36–57, Feb. 2015.

[35] M. Dumbser, M. Castro, C. Parés, and E. F. Toro. ADER schemes on unstructured meshes for nonconservative hyperbolic systems: Applications to geophysical flows. *Comput. Fluids*, 38(9):1731–1748, Oct. 2009.

[36] G. Dziuk and C. M. Elliott. Finite element methods for surfaces PDEs. *Acta Num.*, 22:289–396, 2013.

[37] I. Fent, M. Putti, C. Gregoretti, and S. Lanzoni. Modeling shallow water flows on general terrains. *Adv. Water Resour.*, 121:316–332, 12 2017. doi: 10.1016/j.advwatres.2017.12.017.

[38] E. D. Fernández-Nieto, F. Bouchut, D. Bresch, M. J. Castro Díaz, and A. Mangeney-Castelnau. A new Savage–Hutter type model for submarine avalanches and generated tsunami. *J. Comp. Phys.*, 227(16):7720–7754, Aug. 2008.

[39] K. Georg and J. Tausch. Some error estimates for the numerical approximation of surface integrals. *Math. Comp.*, 62(206):755–763, Apr. 1994.

[40] J. M. N. T. Gray, M. Wieland, and K. Hutter. Gravity-driven free surface flow of granular avalanches over complex basal topography. *Phil. Trans. R. Soc. A*, 455 (1985):1841–1874, May 1999.

[41] J. L. Guermond and M. Nazarov. A maximum-principle preserving c^0 finite element method for scalar conservation equations. *Comp. Methods App. Mech. Engrg.*, 271:198–213, 2014.

[42] J. L. Guermond and R. Pasquetti. A correction technique for the dispersive effects of mass lumping for transport problems. *Comp. Methods App. Mech. Engrg.*, 253: 186 – 198, 2013. doi: https://doi.org/10.1016/j.cma.2012.08.011.

[43] J. L. Guermond and B. Popov. Viscous regularization of the euler equations and entropy principles. *SIAM J. Appl. Math.*, 74(2):284–305, 2014.

[44] J. L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Num. Anal.*, 54(4):2466–2489, 2016.

[45] J. L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *J. Comp. Phys.*, 230:4248–4267, 2011.

[46] J. L. Guermond, M. Nazarov, B. Popov, and Y. Yang. A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. *SIAM J. Num. Anal.*, 52(4):2163–2182, 2014.

[47] J. L. Guermond, M. Quezada de Luna, B. Popov, C. Kees, and M. Farthing. Well-balanced second-order finite element approximation of the shallow water equations with friction. *SIAM J. Sci. Comput.*, in review 2017.

[48] A. Harten, P. D. Lax, and B. v. Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Review*, 25(1):35–61, 1983.

[49] R. L. Higdon. Numerical modelling of ocean circulation. *Acta Num.*, 15:385, May 2006.

[50] J. R. Holton. *An introduction to dynamic meteorology.* Burlington, MA: Elsevier Academic Press, 2004.

[51] R. M. Iverson and D. L. George. A depth-averaged debris-flow model that includes the effects of evolving dilatancy. I. Physical basis. *Proc. R. Soc. London*, 470(2170): 20130819–20130819, July 2014.

[52] P. P. Jansen, L. van Bengedom, J. van den Berg, M. de Vries, and A. Zanen. *Principles of river engineering: the non-tidal alluvial river.* Pittman, London, San Francisco, 1979.

[53] S. Lanzoni, A. Siviglia, A. Frascati, and G. Seminara. Long waves in erodible channels and morphodynamic influence. *Water Resour. Res.*, 42:W06D17, 2006.

[54] P. G. LeFloch, M. D. Thanh, et al. The Riemann problem for the shallow water equations with discontinuous topography. *Comm. Math. Sci.*, 5(4):865–885, 2007.

[55] X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comp. Phys.*, 115(1):200 – 212, 1994. doi: https://doi.org/10.1006/jcph.1994.1187.

[56] L. Moretti, K. Allstadt, A. Mangeney-Castelnau, Y. Capdeville, E. Stutzmann, and F. Bouchut. Numerical modeling of the Mount Meager landslide constrained by its force history derived from seismic data. *J. Geophys. Res.*, 120(4):2579–2599, 2015.

[57] J. M. Morvan. *Generalized Curvatures*, volume 2 of *Geometry and Computing.* Springer Science & Business Media, Berlin, Heidelberg, May 2008.

[58] P. Nevalainen, I. Jambor, J. Pohjankukka, J. Heikkonen, and T. Pahikkala. Triangular curvature approximation of surfaces - filtering the spurious mode. In *6th International Conference on Pattern Recognition Applications and Methods*, pages 684–692. SCITEPRESS - Science and Technology Publications, Mar. 2017.

[59] S. Nigam and V. Agrawal. Review. Curvature approximation on triangular meshes. *Int J Eng Sci Innov Technol(IJESIT)*, 2013.

[60] M. A. Olshanskii, A. Reusken, and X. Xu. A stabilized finite element method for advection-diffusion equations on surfaces. *IMA J. Num. Anal.*, 34(2):732–758, Apr. 2014.

[61] J. Pedlosky. *Geophysical fluid dynamics.* Springer Verlag, New York, 1979.

[62] M. Pelanti, F. Bouchut, and A. Mangeney-Castelnau. A Roe-type scheme for two-phase shallow granular flows over variable topography. *ESAIM-Math. Model. Num.*, 42(5):851–885, July 2008.

[63] J. A. Rossmanith, D. S. Bale, and R. J. LeVeque. A wave propagation algorithm for hyperbolic systems on curved manifolds. *J. Comp. Phys.*, 199(2):631–662, Sept. 2004.

[64] C. Ruyer-Quil and P. Manneville. Modeling film flows down inclined planes. *Eur. Phys. J. B*, 6(2):277–292, 1998.

[65] C. Ruyer-Quil and P. Manneville. Improved modeling of flows down inclined planes. *Eur. Phys. J. B*, 15(2):357–369, 2000.

[66] S. B. Savage and K. Hutter. The motion of a finite mass of granular material down a rough incline. *J. Fluid Mech.*, 199(-1):177–215, Feb. 1989.

[67] S. B. Savage and K. Hutter. The dynamics of avalanches of granular materials from initiation to runout. Part I: Analysis. *Acta Mech.*, 86(1-4):201–223, 1991.

[68] T. Song, A. Main, G. Scovazzi, and M. Ricchiuto. The shifted boundary method for hyperbolic systems: Embedded domain computations of linear waves and shallow water flows. *J. Comp. Phys.*, 369:45–79, Sept. 2018.

[69] H. Tennekes and J. L. Lumley. *A First Course in Turbulence.* MIT Press, 1972.

[70] E. Toro. *Shock-capturing methods for free-surface shallow flows.* John Wiley, 2001.

[71] E. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the Harten-Lax-van Leer Riemann solver. *Shock Waves*, 4:25–34, 1994.

[72] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 1992. doi: 10.1137/0913035.

[73] D. Wirasaet, E. Kubatko, C. Michoski, S. Tanaka, J. Westerink, and C. Dawson. Discontinuous galerkin methods with nodal and hybrid modal/nodal triangular, quadrilateral, and polygonal elements for nonlinear shallow water flow. *Comp. Methods App. Mech. Engrg.*, 270:113 – 149, 2014. doi: https://doi.org/10.1016/j.cma.2013.11.006.

[74] Y. Zhang, A. B. Kennedy, N. Panda, C. N. Dawson, and J. J. Westerink. Boussinesq–Green–Naghdi rotational water wave theory. *Coast. Eng.*, 73(C):13–27, Mar. 2013.

[75] G. Zolezzi and G. Seminara. Downstream and upstream influence in river meandering. Part 1. General theory and application to overdeepening. *J. Fluid Mech.*, 438:183–211, 2001.