Path-consistent schemes for interacting boundary layers

a discontinuous Galerkin approach

MASTER OF SCIENCE THESIS

Jan Hindrik Seubers

Monday 2nd June, 2014

Supervision:
ECN: Dr. Hüseyin Özdemir
TU Delft: Dr. Marc Gerritsma

Faculty of Aerospace Engineering, Delft University of Technology

Abstract

This study is aimed at modelling of the unsteady interacting boundary layer around wind turbine rotor blades. It is shown that nonconservative mechanisms are present in this system, affecting the definition and approximation of solutions to the problem. The objective is to overcome the inconsistency of classical (dissipative) schemes, and find a suitable approximation method. A path-consistent discontinuous Galerkin method is proposed to correctly model the dynamics of this system.

In the future, the current implementation is to be joined with the existing part of the RotorFlow simulation code. This will enable the incorporation of unsteady aerodynamic phenomena in the early design stage of new and large wind turbine rotors, which is currently unfeasible due to the high computational cost.

Keywords

Integral boundary layer, Nonconservative, Viscous-inviscid interaction

Contents

Nonconservative DG notations

d	Number of space dimensions	–
n	Number of components in the solution	–
p	Degree of (polynomial) representation	–
b_j	Basis function	–
$\mathbf{f}$, $\mathbf{h}$	Flux tensor (and numerical version)	$[\Omega]$*
$\mathbf{g}$, $\tilde{\mathbf{g}}_{\pm}$	Balance tensor (and numerical version)	–*
K_k	Finite element	m^d
$\boldsymbol{l}_k$	Left eigenvector	$[\Omega]$
$\boldsymbol{n}$	Outward unit normal vector	m
$\boldsymbol{r}_k$	Right eigenvector	$[\Omega]$
S, S_n	Surface, face	m^{d-1}
$\boldsymbol{s}$	Source function	$[\Omega]/s$
$\boldsymbol{u}$	Vector of unknowns	$[\Omega]$
$\boldsymbol{v}(\boldsymbol{u})$	Entropy variables	$[\Omega]$
$\boldsymbol{v}(\sigma)$	Parametrization of discontinuity	$[\Omega]$
β	Viscosity scaling parameter	–
Γ	Evaluation path for nonconservative product	$[\Omega]$
ϵ	Shock strength parameter	m/s
$\boldsymbol{\phi}$	Test function	–
λ_k	Eigenvalue of hyperbolic system	m/s
ξ	Transformed spatial coordinate	m
σ	Wave propagation speed	m/s
$\boldsymbol{\psi}$	Parametrization of evaluation path Γ	–
Ω	Domain of (conserved) quantities	$[\Omega]$
$\mathcal{L}(a;b)$	Linear operator from a to b	$[b]/[a]$
$\mathcal{B}$	Boundary residual	$[\Omega]m^d$
$\mathcal{E}$	Element residual	$[\Omega]m^d$
$\mathcal{F}$	Face residual	$[\Omega]m^d$
$\mathcal{J}$	Jacobian of the residual	mixed
$\mathcal{Q}_h$	Collection of entities of zero measure	mixed
$\mathcal{R}$	Total residual	$[\Omega]m^d$
$\mathcal{S}_h$	Collection of faces	m^{d-1}
$\mathcal{T}_h$	Collection of elements (tesselation)	m^d
$\mathcal{V}$	Spatial domain	m^d
$\mathbf{C}$	Connectivity matrix	–
$\mathbf{D}$	Incidence matrix	–
$\mathbf{F}$	Element flux matrix	–
$\mathbf{J}$	Jacobian of the element transformation	s*
$\mathbf{M}$	Element mass matrix	–
$\mathbf{P}$	Trace weighting matrix	–
$\mathbf{T}$	Trace matrix	–

*) multiply unit of spatial component by m/s

Boundary layer notations

B	Boundary layer budget factor	–
H	Boundary layer shape factor	–
H^*	Kinetic energy shape factor	–
H_ψ	Streamline displacement factor	–
p^*	Pressure thickness	m
p	Pressure	kg/ms^2
q	Fluid velocity magnitude	m/s
Re	Reynolds number	–
S_{me}	Mechanical energy dissipation	kgm^2/s^3
u	Wall-parallel fluid velocity	m/s
Γ_e	Upstream circulation	m^2/s
δ_ψ	Streamline-to-surface distance	m
δ^*	Displacement thickness	m
δ^k	Kinetic energy thickness	m
ε	Mechanical energy thickness	m
θ	Momentum thickness	m
ρ	Fluid density	kg/m^3
τ	Stress	kg/ms^2
μ	Dynamic viscosity	kg/ms
ν	Kinematic viscosity	m^2/s

Asymptotic integral quantities
(all with dimension of distance):

$$\begin{aligned}
\delta^* &\overset{\text{def}}{=} \int_0^{\delta_\psi} \left(\frac{u_e - u}{q_e}\right) \mathrm{d}\tilde{z}, \\
\theta &\overset{\text{def}}{=} \int_0^{\delta_\psi} \frac{u}{u_e}\left(1 - \frac{u}{U_e}\right) \mathrm{d}z, \\
p^* &\overset{\text{def}}{=} \int_0^{\delta_\psi} \left(1 - 2\frac{p_0 - p}{\rho q_e^2}\right) \mathrm{d}z + \frac{2\delta_\psi}{\rho q_e^2}\frac{\partial \Gamma_e}{\partial t} \\
\varepsilon &\overset{\text{def}}{=} \int_0^{\delta_\psi} \left(1 - \frac{\boldsymbol{u}\cdot\boldsymbol{u}}{q_e^2}\right) \mathrm{d}\tilde{z}, \\
\delta_k &\overset{\text{def}}{=} \int_0^{\delta_\psi} \frac{u}{q_e}\left(1 - \frac{\boldsymbol{u}\cdot\boldsymbol{u}}{q_e^2}\right) \mathrm{d}z.
\end{aligned}$$

Conserved integral quantities
(mass, momentum and energy):

$$\delta_\psi \overset{\text{def}}{=} \text{‘Orthogonal distance from the wall to a streamline or stream surface’}$$

$$\begin{aligned}
\dot{m} &\overset{\text{def}}{=} \int_0^{\delta_\psi} \rho u \mathrm{d}z = \rho(u_e\delta_\psi - q_e\delta^*), \\
\mathcal{E}_\psi &\overset{\text{def}}{=} \int_0^{\delta_\psi} \rho\boldsymbol{u}\cdot\boldsymbol{u}\mathrm{d}\tilde{z} = \rho q_e^2(\delta_\psi - \varepsilon)
\end{aligned}$$

Nondimensional factors:

$$\begin{aligned}
B &\overset{\text{def}}{=} \frac{u_e\delta^* - q_e\theta}{2u_e\delta^* - q_e\varepsilon} \\
H &\overset{\text{def}}{=} \delta^*/\theta\ , \\
H^* &\overset{\text{def}}{=} \delta_k/\theta \\
H_\psi &\overset{\text{def}}{=} \delta^*/\delta_\psi\ .
\end{aligned}$$

$$\begin{aligned}
\mathcal{D} &\overset{\text{def}}{=} \frac{2\mathrm{Re}_\theta}{\rho q_e^3}\int_0^{\delta_\psi} S_{\mathrm{me}}\mathrm{d}\tilde{z}, \\
C_f &\overset{\text{def}}{=} \frac{2\tau_w}{\rho u_e^2}.
\end{aligned}$$

Introduction

Due to economical, political and ecological reasons, renewable energy sources will probably gain an increased share in the energy market in the next decades [23]. Wind power generation, being one of the important technologies in this field, therefore needs to be deployed and expanded in an efficient way. The challenges of designing larger rotors and dealing with proximity of multiple turbines, require more accurate prediction of blade loading and wake generation. Unsteady aerodynamics play an important role in both problems. This includes the changing inflow conditions due to incident wakes, the effects on instantaneous blade loading conditions, and time history effects from the wake (such as dynamic stall). If during the design phase, accurate modeling for these phenomena is available, technological improvements can be achieved that will affect performance, control, production and maintenance during the entire turbine life cycle.

As part of an effort to make unsteady aerodynamic modelling available in the design phase, an interacting boundary layer- potential flow solver was proposed [39]. The resulting unsteady coupling terms however affect the stability and characteristics of the boundary layer system by introducing a nonconservative mechanism. The focus of this thesis lies with modelling and simulation of such nonconservative systems, because the classical solutions and approximations for conservation laws do not apply. Instead, the solutions become path-dependent, and it is argued that the conservation property for a numerical scheme must be replaced by a path-consistency property.

The contents of this thesis are organised as follows. It starts with a short historical overview of interacting boundary layer methods in section 1. In the next section, the important aspects of conservation laws (section 2) serve as an introduction to the generalised theory of nonconservative systems (section 3). A proposed formulation for path-consistent schemes in a discontinuous Galerkin framework is given in section 4, including a full quadrature-free implementation of the resulting scheme. The application to the unsteady interacting boundary layer (IBL) system, including the physical modelling is analysed in section 6. Finally the conclusions and recommendations are given in sections 7 and 8.

The main contributions can be found in the following sections of this thesis.

On nonconservative systems

- Clear definitions of conservation laws (2.1) and nonconservative systems (3.1).
- How to recognise and maintain the conservation property, regardless of mathematical formulation (quasilinear or conservation form) (3.3).
- Derivation of error bounds for paths of symmetric systems with commuting viscosity (3.4).
- Interpretation of the Mass, Momentum and Energy Conserving (MaMEC) scheme as an entropy-stabilized scheme (3.7).

Discontinuous Galerkin:

- Incorporate paths for test functions into path integral (4.1).
- Path-consistency in combination with classical Riemann solvers leads to a Petrov-Galerkin type scheme.
- Focus on space partitioning instead of an isolated reference element (4.2)
- Incorporated time-nonconservative products in formulation (3.1).

Boundary layer and interaction method:

- Streamline based control volume approach, including terms for non-classical (separating) boundary layers (6.1, 6.2, 6.3).
- Quasi-simultaneous interaction with off-surface velocity influence (6.4).
- Fully unsteady formulation without viscous-inviscid iteration.

Implementation:

- Projection between sinusoid/polynomial/rational spaces (A.1 and A.3).
- Construction of arbitrary-dimensional quadrature-free elements with Legendre basis.
- Implementation of an object-oriented discontinuous Galerkin (DG) solver in Fortran for generic nonconservative hyperbolic systems (NCHS).

1 Historical overview

Current design tools for rotor aerodynamics are mainly based on the Blade Element-Momentum (BEM) approach. These methods, based on local blade section aerodynamics combined with an azimuthal averaged momentum balance, provide quick and efficient estimates. However the range of applicability is not large, due to the assumptions at the base of the approach: it is limited to steady state, yaw aligned uniform flow, quasi-2d blade aerodynamics, spanwise independent induction, without rotor cone angle. To overcome these limits, many empirical corrections have been applied. These include the empirically and theoretically well established tip and root loss corrections, as well as corrections that conflict with the basic assumptions (skewed wake correction, dynamic stall models). Such empirical corrections may improve the model for some situations, but limited predictive confidence can be gained. This especially applies to predicting many relevant phenomena that could influence the design (o.a. load oscillations, dynamic stall, wake interference).

Other approaches, such as inviscid panel methods [22], nonlinear lifting line methods [21] and viscous-inviscid splitting methods are based on a less severe set of assumptions and can predict these physical phenomena intrinsically (without corrections). Hence they can be applied to a wider range of rotor operating conditions. The most accurate of these approaches is the splitting of the flow domain into viscous and inviscid regions. Since the introduction of the boundary layer by Prandtl in 1904, this approach has been used in many variants, since both regions allow a range of models to be selected, which can be combined using a range of interaction methods (see e.g. [10]). Historically, the two regions were solved separately and glued together (weak or no interaction). This approach works for attached flow, but leads to singularities associated with flow separation, both for steady (Goldstein [26]) and unsteady flow (Van Dommelen/Shen [16]). The latter result was interpreted by Matsushita et al. [33] as the result of discontinuities in the boundary layer (analogous to shocks in compressible flow). They used a dissipative finite difference method to capture separating flow including the singularity.

Later approaches take into account the two-way coupling of the viscous and inviscid regions during computation (strong interaction), thereby successfully extending the validity to (mildly) separated flow. The interaction of the two systems is essentially stabilizing: the breakdown due to singularity does not occur until more strongly separated flow regimes are entered, as can be seen in the analysis of Coenen[13]. These approaches saw major development in the 1980s, with the quasi-simultaneous [47] and fully simultaneous [17] approach. The XFOIL code [18] is a well known example, for 2d steady flow around an airfoil. For rotor aerodynamics, this code has been extended into RFOIL [43], which includes corrections for rotational effects on the boundary layer. This improves the prediction of stall delay, thereby overcoming some limitations of the quasi-2d assumption. Other fully 3d methods also exist, developed by Nishida [36], by Coenen [13] and very recently by Drela [19]. See the overview in the table below. Few truly unsteady applications are noted: a dynamic stall simulation by Cebeci [11] and a flutter computation by Zhang and Liu [50], both of which show significant effects due to the viscous part. The recent IBL3 method includes an unsteady formulation which seems very promising, but no validation results have emerged yet. However the shock-like behaviour of the characteristics at separation does re-emerge and is mentioned explicitly in this paper.

Author	year	dim.	viscous model	inviscid model	interaction
v. Dommelen/Shen	1980 [16]	2d	unsteady Lagrangian	prescribed	none
Veldman	1981 [47]	2d	steady Prandtl	thin airfoil	quasi-simultaneous
Matsushita et al.	1985 [33]	2d	unsteady integral	prescribed	none
Drela/Giles (ISES)	1985 [17]	2d	steady integral	steady Euler	simultaneous
Drela (XFOIL)	1989 [18]	2d	steady integral	steady panel	simultaneous
Cebeci	1993 [11]	2d	unsteady Prandtl	unsteady panel	quasi-simultaneous
v. Rooij (RFOIL)	1996 [43]	2d/3d	steady integral	steady panel	simultaneous
Nishida	1996 [36]	3d	steady integral	steady full potential	simultaneous
Coenen	2001 [13]	3d	steady integral	steady panel	quasi-simultaneous
Zhang/Liu	2004 [50]	2d	steady integral	unsteady Euler	semi-inverse
Bijleveld	2013 [4]	3d	quasi-steady integral	steady panel	quasi-simultaneous
Drela (IBL3)	2013 [19]	3d	unsteady integral defect	steady panel	simultaneous

Unsteady effects are expected to become important near separation since the timescale of the boundary layer increases: the characteristics form envelopes around separated regions and information propagates more slowly. High pitch rates, aeroelastic behaviour and wake interference further decrease the timescale at which unsteady effects play a role. Another reason to search for unsteady solutions is that steady solutions may simply not exist for separated flow (e.g. the Von Karman vortex street behind a cylinder). The nonconservative effects are expected in all (quasi-)simultaneous integral formulations, affecting the unsteady behaviour and the location of equilibria. In effect this causes the resulting solutions to depend not only upon the physics of the problem but on the numerical strategy as well, see section 3. To prevent this undesirable condition, specialized nonconservative numerical schemes should be applied.

To limit the scope of this investigation, the modelling approach used was based on the unsteady integral boundary layer, combined with an existing unsteady panel method [22] through quasi-simultaneous interaction. The considered methods will be applied in the framework of a discontinuous Galerkin scheme, which has been used in previous investigations[5].

2 Theory of conservative systems

In order to develop a better understanding of the nature of the interacting boundary layer system, the causes and consequences of the nonconservative terms will be made clear using the theory. Part of the basic mathematical machinery for the treatment of conservation laws, does not apply to nonconservative systems. This makes both the analytical and numerical treatment of nonconservative systems more involved; in addition it has received less attention historically[12]. However, there are common points in the treatment of conservative and nonconservative systems, including the theory of characteristics. Therefore this chapter deals with some basics of conservation laws, presented in a way that will be useful in the interpretation of nonconservative systems. The chapter consists roughly of three parts: first a formal description of conservation, secondly the theory of characteristics and finally a relation between entropy and integral curves.

2.1 Formulation and interpretation

A set of conservation laws express the rate of change of certain quantities inside a domain in terms of interactions at its boundary only. The quantities are called *conserved variables* and the interactions are called *fluxes*. Mathematically this can be described in integral form as

$$\overset{\text{time-continuous formulation}}{\frac{\partial}{\partial t}\int_{\mathcal{V}} \boldsymbol{u}\,\mathrm{d}\mathcal{V} + \oint_{\partial\mathcal{V}} \mathbf{f}(\boldsymbol{u})\cdot\boldsymbol{n}\,\mathrm{d}S = 0,} \qquad \overset{\text{space-time formulation}}{\oint_{\partial\underline{\mathcal{V}}} \underline{\mathbf{f}}(\boldsymbol{u})\cdot\underline{\boldsymbol{n}}\,\mathrm{d}\underline{S} = 0.} \tag{1}$$

We define the spatial domain $\mathcal{V}$ as an open subset of $\mathbb{R}^d$ and the set of states Ω as an open subset of $\mathbb{R}^n$. The dimension of space d and the dimension of the state n are independent of each other. To visualize this in the context of this work, $\mathcal{V}$ would be a map of the curved rotor blade surface and Ω would be the momentum and energy density (or another parametrization of the local state). The solution $\boldsymbol{u}$ is a field (taking values in Ω) that represents the density of conserved quantities over the volume $\mathcal{V}$. When integrated over space, with respect to the metric $\mathrm{d}\mathcal{V}$ and an arbitrary conjugate vector field $\boldsymbol{\phi} \in \Omega^*$, it yields the amount of the conserved quantity within the volume detected by the test function $\boldsymbol{\phi}$. When $\boldsymbol{\phi}$ is not specified in the integral, it is assumed globally constant. The conservation law (2.1) then consists of n equations, one for each quantity in Ω. The flux density $\mathbf{f}$ is a field (taking values in $\mathcal{L}(\mathrm{T}_x\mathcal{V};\Omega)$), representing the flow of conserved quantities through a surface $\partial\mathcal{V}$. The notation $\mathcal{L}(a;b)$ denotes a linear operator with domain a and codomain b, and $\mathrm{T}_x\mathcal{V}$ denotes the tangent space through a given point x, on which the surface normal vectors $\boldsymbol{n}$ are defined[1]. Essentially this defines the flux density as a tensor, because the flow is supposed to depend linearly on the direction in space in a coordinate-independent way. When $\mathbf{f}$ is integrated over a surface and over a time interval (with respect to the local surface normal $\boldsymbol{n}$, metric $\mathrm{d}\mathcal{S}\mathrm{d}t$ and the arbitrary conjugate vector field $\boldsymbol{\phi} \in \Omega^*$), one obtains the total amount of the conserved quantity flowing through the surface within the time interval, as detected by $\boldsymbol{\phi}$.

In the space-time formulation no distinction is made between the space and time domain, and to-

[1] Introducing exterior product spaces could replace the use of normal vectors here, but the author considers this unneccessary for the understanding of the subject

gether they are denoted $\underline{\mathcal{V}}$ (an open subset of $\mathbb{R}^{d+1}$). The space-time flux density $\underline{\mathbf{f}}$ is now the only field occuring in equation (2.1), taking values in $\mathcal{L}(\mathrm{T}_x\underline{\mathcal{V}};\Omega)$. The extra time component to the flux density is identical to the previously defined field of conserved variables. As a consequence of this interpretation, integrals over volumes and surfaces need no longer be separated as in the time-continuous formulation. Only integrals over space-time 'surfaces' remain in equation .

Two functional relations are required to complete the model: the state as a function of position $\boldsymbol{u}(\boldsymbol{x})$ and the flux density as a function of the state $\mathbf{f}(\boldsymbol{u})$ (see figure 1). Together with appropriate initial and boundary conditions this describes the conservation of n quantities in d dimensions. The boundary conditions can be given as an imposed value $\boldsymbol{u}_b$ or an imposed boundary normal flux $\boldsymbol{f}_b$:

$$\overset{\text{time-continuous formulation}}{\begin{aligned} \boldsymbol{u} &= \boldsymbol{u}_0(\boldsymbol{x}) && \text{on } \{t=0\} \\ \boldsymbol{u} &= \boldsymbol{u}_b(t) && \text{on } (\partial\mathcal{V})_{\text{dirichlet}} \\ \mathbf{f}\cdot\boldsymbol{n} &= \boldsymbol{f}_b && \text{on } (\partial\mathcal{V})_{\text{neumann}} \end{aligned}}\,, \qquad \overset{\text{space-time formulation}}{\begin{aligned} \boldsymbol{u} &= \boldsymbol{u}_b && \text{on } (\partial\underline{\mathcal{V}})_{\text{dirichlet}} \\ \underline{\mathbf{f}}\cdot\underline{\boldsymbol{n}} &= \boldsymbol{f}_b && \text{on } (\partial\underline{\mathcal{V}})_{\text{neumann}} \end{aligned}}\,. \tag{2}$$

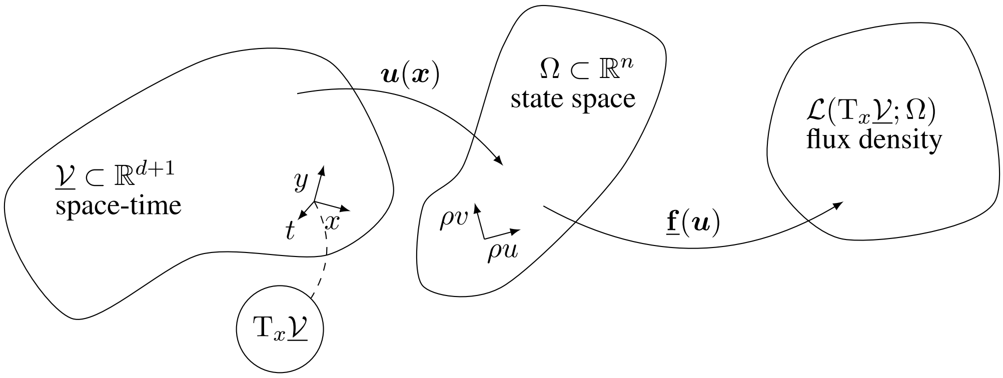
$\boldsymbol{u}(\boldsymbol{x})$
$\Omega \subset \mathbb{R}^n$
state space
$\mathcal{L}(\mathrm{T}_x\underline{\mathcal{V}};\Omega)$
flux density
$\underline{\mathcal{V}} \subset \mathbb{R}^{d+1}$
space-time
y
t
x
ρv
ρu
$\underline{\mathbf{f}}(\boldsymbol{u})$
$\mathrm{T}_x\underline{\mathcal{V}}$

Figure 1: *Distinction between the spaces in the mathematical problem formulation. Space-time $\underline{\mathcal{V}}$ is a metric space, the 'state' Ω and 'direction' $\mathrm{T}_x\underline{\mathcal{V}}$ are vector spaces, and the flux density is a linear map between them.*

This set of conservation laws can be equivalently expressed as a system of partial differential equations, if the solution is assumed to be sufficiently smooth. This can be done by using Green's identity to convert the boundary integral to divergence form, and requiring that conservation holds for an arbitrary domain. Then, the resulting differential equation (3) will hold pointwise.

$$\overset{\text{time-continuous formulation}}{\frac{\partial\boldsymbol{u}}{\partial t} + \operatorname{div}\mathbf{f}(\boldsymbol{u}) \quad = 0} \qquad \overset{\text{space-time formulation}}{\operatorname{div}\underline{\mathbf{f}}(\boldsymbol{u}) \quad = 0} \tag{3}$$

This differential conservative formulation of the original law is not unique: equivalent conservative formulations of (3) exist, expressed in other variables $\boldsymbol{v}(\boldsymbol{u})$ that will yield the same continuous solution. This even holds for quasi-linear formulations that are not in conservation form. Although these equivalent formulations may hold point-wise, their discretization will in general differ from that of the original form. Looking further, the integral form allows a broader range of solutions, not limited to continuous smooth functions. In fact, nonlinear conservation laws can develop these discontinuous solutions within finite time, starting from a smooth initial condition [25]. In such cases, the derivatives in (3) become locally undefined. Therefore, it is necessary to define exactly what functions are solutions to the conservation law. These solutions are called weak solutions of (2.1), (2) if they are locally bounded and satisfy the following equation (see [25]) for all continuously differentiable and compactly supported test functions $\boldsymbol{\phi} : \underline{\mathcal{V}} \to \mathbb{R}^n$:

$$\int_{\underline{\mathcal{V}}} \underline{\mathbf{f}}(\boldsymbol{u}) : \operatorname{grad}\boldsymbol{\phi}\,\mathrm{d}(\underline{\mathcal{V}}) - \oint_{\partial\underline{\mathcal{V}}} \underline{\mathbf{f}} \cdot \underline{\boldsymbol{n}} \cdot \boldsymbol{\phi}\,\mathrm{d}\underline{S} = 0. \tag{4}$$

A time-continuous version of this weak solution definition does not exist, since moving spatial discontinuities imply temporal discontinuities as well. The nonuniqueness of conservative or non-conservative formulations becomes important here, as the weak solutions to distinct formulations do not coincide when discontinuities are present [46]. A weak solution to a certain formulation may still not be unique, and some weak solutions may be unphysical. The weak solutions that, in addition to equation (4) satisfy an *entropy condition* (which will be formulated later) will be called entropy weak solutions of the initial value problem given by (2.1) and (2).

2.2 Source terms

The requirement that conservation laws should determine the state of certain quantities completely by the boundary interactions is rather strict. Yet due to this requirement it is possible to predict the large-scale behaviour of a system without further details on the small-scale behaviour inside a given region. When source terms are added resulting in the formulation (5), the requirement is violated and this desirable property could disappear.

$$\overset{\text{time-continuous formulation}}{\frac{\partial}{\partial t}\int_{\mathcal{V}} \boldsymbol{u}\,\mathrm{d}\mathcal{V} + \oint_{\partial\mathcal{V}} \mathbf{f}(\boldsymbol{u}) \cdot \boldsymbol{n}\,\mathrm{d}S = \int_{\mathcal{V}} \boldsymbol{s}(\boldsymbol{u})\,\mathrm{d}\mathcal{V}}, \qquad \overset{\text{space-time formulation}}{\oint_{\partial\underline{\mathcal{V}}} \underline{\mathbf{f}}(\boldsymbol{u}) \cdot \underline{\boldsymbol{n}}\,\mathrm{d}\underline{S} = \int_{\underline{\mathcal{V}}} \boldsymbol{s}(\boldsymbol{u})\,\mathrm{d}\underline{\mathcal{V}}}. \tag{5}$$

For example, consider the heat equation with some heat sources placed inside the domain. In every region that does not include a source, the amount of heat is determined completely by the heat flux through the region's boundary. For the region with sources this does not hold. Especially if the heat sources depend on the solution, the local behaviour will influence the outcome in terms of heat conservation, since it is unknown in advance how much will be added. In any of the following cases this influence may be approximated

- The source is independent of the solution.
- The source depends only on the local value of the solution.
- The source depends only on the local gradient of the solution.

For the last two cases, the length and times scales at which the system can be approximated as a conservation law depend on the (finite) sensitivity of the source term to the solution. Taking this into account, it is very well possible to admit sources while still modelling such processes as conservation laws. For systems with stiff sources this will have severe consequences for the discretization. But more importantly, source terms that violate all three of the above conditions may cause truly nonconservative behaviour that can not be predicted by conservative schemes.

2.3 Characteristic form

If the system of conservation laws is hyperbolic, it is possible to apply a real transformation to the system that will reduce it to a system of ordinary differential equations. This characteristic form describes the evolution of the system in terms of moving waves. The theory of characteristics is presented here in space-time formulation, which is slightly unusual but will be advantageous because it also applies to the general nonconservative case. The characteristics can be obtained from the one-dimensional quasi-linear form (6) of equation (5). Here the n-by-n matrices A^t and A^x are the flux Jacobians corresponding to the flux components $\boldsymbol{f}^x$ and $\boldsymbol{f}^t$, the superscripts distinguishing time and space directions.

$$\begin{aligned} \frac{\partial \boldsymbol{f}^t(\boldsymbol{u})}{\partial t} + \frac{\partial \boldsymbol{f}^x(\boldsymbol{u})}{\partial x} &= \boldsymbol{s}(\boldsymbol{u}), \\ A^t \frac{\partial \boldsymbol{u}}{\partial t} + A^x \frac{\partial \boldsymbol{u}}{\partial x} &= \boldsymbol{s}(\boldsymbol{u}). \end{aligned} \tag{6}$$

where

$$A^t_{ij} = \frac{\partial f^t_i(\boldsymbol{u})}{\partial u_j}, \qquad\qquad A^x_{ij} = \frac{\partial f^x_i(\boldsymbol{u})}{\partial u_j}.$$

Introduce the change of coordinates[2] $\xi = x - \sigma t$ and consider the homogeneous system (without source term),

$$\begin{aligned} A^t \frac{\partial \boldsymbol{u}}{\partial t} + A^x \frac{\partial \boldsymbol{u}}{\partial x} &= A^t \frac{\partial \boldsymbol{u}}{\partial \xi}\frac{\mathrm{d}\xi}{\mathrm{d}t} + A^x \frac{\partial \boldsymbol{u}}{\partial \xi}\frac{\mathrm{d}\xi}{\mathrm{d}x}, \\ \left(A^x - \sigma A^t\right) \frac{\partial \boldsymbol{u}}{\partial \xi} &= 0. \end{aligned} \tag{7}$$

This is a generalized eigenproblem[3] for $\partial \boldsymbol{u}/\partial \xi$. We call the associated left and right eigenvectors $\boldsymbol{l}_k$ and $\boldsymbol{r}_k$ respectively, corresponding to the eigenvalues λ_k, where $k \in \{1 \ldots n\}$. For a nonlinear system, the flux Jacobians (and therefore the eigenvalues and eigenvectors) depend on the solution. For each k, the right eigenvectors can be seen as a vector field, since each state in Ω is associated with a unique eigenvector in Ω. Hence it is not a vector field in the spatial domain, but in the solution domain, and it is called a characteristic field. Integrating this field from a given initial point defines a characteristic curve. To obtain a visual impression of this concept, such curves have been plotted in figure 2 for the interacting boundary layer system. The nonlinearity of the characteristic fields can be clearly observed.

[2] This coordinate system is relative to a characteristic or shock surface when σ is equal to the characteristic or shock speed, hence it is called the shock frame.

[3] Generalized eigenvalue problems are explained in Appendix B.

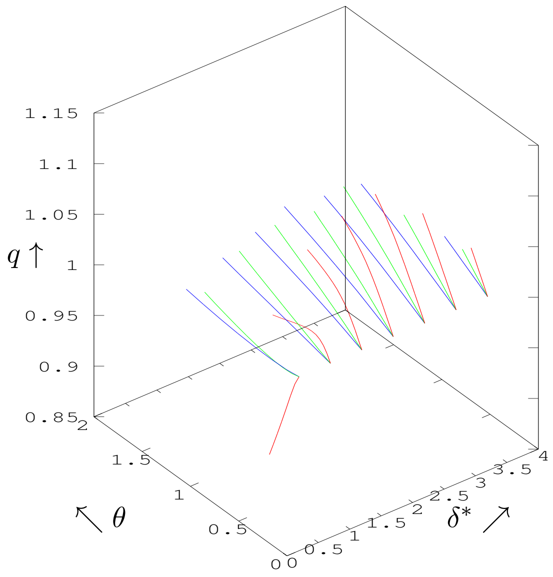

Figure 2: *Characteristic curves for the interacting IBL equations*

Each characteristic field corresponds to infinitesimal changes in the solution being propagated with a certain wave velocity. The fields can be classified as either *genuinely nonlinear* or *linearly degenerate* by considering whether the eigenvalue changes over its own characteristic curve.

$$D\lambda_k(\boldsymbol{u}) \cdot \boldsymbol{r}_k = 0 \qquad \text{if the field is linearly degenerate,} \tag{8}$$

$$D\lambda_k(\boldsymbol{u}) \cdot \boldsymbol{r}_k \neq 0 \qquad \text{if the field is genuinely nonlinear.} \tag{9}$$

The differential operator D denotes the Fréchet derivative, which can be seen as the gradient of the function with respect to its argument. In case of the genuinely nonlinear field, the convention is to normalize the right eigenvectors such that $D\lambda_k(\boldsymbol{u}) \cdot \boldsymbol{r}_k = 1$.

To show that the characteristics indeed propagate as waves, a derivation from [25], adapted to the space-time formulation is presented here. The wave solutions for the characteristic variables are found by pre-multiplying the quasilinear form (6) with the conjugated left eigenvectors $\boldsymbol{l}_k{}^{\mathrm{T}}$,

$$\begin{aligned} \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} \left(A^t \frac{\partial \boldsymbol{u}}{\partial t} + A^x \frac{\partial \boldsymbol{u}}{\partial x} \right) &= \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} \boldsymbol{s}(\boldsymbol{u}), \\ \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} A^t \left(\frac{\partial \boldsymbol{u}}{\partial t} + \lambda_k(\boldsymbol{u}) \frac{\partial \boldsymbol{u}}{\partial x} \right) &= \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} \boldsymbol{s}(\boldsymbol{u}), \qquad \forall k \in \{1 \ldots n\}. \end{aligned} \tag{10}$$

If these equations are transformed according to $\{\xi = x - \sigma t, \eta = t + x/\sigma\}$, the system in characteristic coordinates becomes

$$\boldsymbol{l}_k{}^{\mathrm{T}}(\boldsymbol{u}) A^t \left[(\lambda_k(\boldsymbol{u}) - \sigma) \frac{\partial \boldsymbol{u}}{\partial \xi} + (1 + \frac{\lambda_k(\boldsymbol{u})}{\sigma}) \frac{\partial \boldsymbol{u}}{\partial \eta} \right] = \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} \boldsymbol{s}(\boldsymbol{u}). \tag{11}$$

If σ is chosen equal to one of the eigenvalues λ_k, one obtains an independent ordinary differential equation. The combination of all n such equations is the characteristic or diagonalized system (12) (here expressed in the original coordinates (x, t)). The solution curves to $\mathrm{d}x/\mathrm{d}t = \lambda_k$ are the characteristic lines (in more dimensions: surfaces) of wave propagation.

$$\left.\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \lambda_k(\boldsymbol{u}) \\ \boldsymbol{l}_k{}^{\mathrm{T}}(\boldsymbol{u}) A^t \frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} &= \boldsymbol{l}_k(\boldsymbol{u})^{\mathrm{T}} \boldsymbol{s}(\boldsymbol{u}) \end{aligned}\right\}, \quad \forall k \in \{1 \ldots n\}. \tag{12}$$

2.4 Riemann problem

A basic building block in the analysis and numerical approximation of hyperbolic systems is the Riemann problem. It is the one-dimensional initial value problem for (4). The initial value consists of a discontinuous interface at $(x = 0, t = 0)$ between two constant states that extend left and right in space to infinity. A solution to this problem describes the behaviour for $t > 0$ of equations (4, 13), and can be visualized in terms of different wave regions emanating from the origin (see figure 3).

$$\boldsymbol{u}(x, t = 0) = \begin{cases} \boldsymbol{u}_L, & x < 0 \\ \boldsymbol{u}_R, & x \geq 0 \end{cases} \tag{13}$$

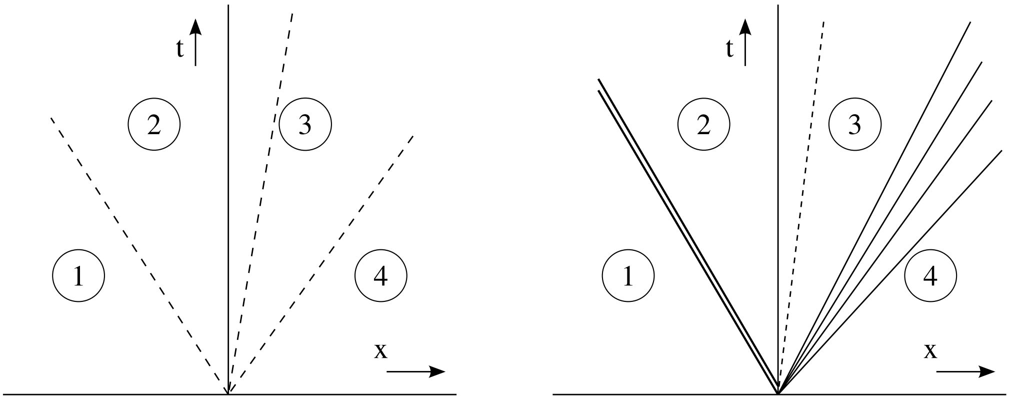

Figure 3: *Example of a solution to Riemann problems with 3 waves: linear (left) and nonlinear (right)*

Depending on the nature of the flux function $\boldsymbol{f}(\boldsymbol{u})$ different types of wave patterns exist: contact discontinuities (parallel wavefronts), expansion waves (diverging wavefronts) or shock waves (coinciding wavefronts). For linearly degenerate fields, all waves are of the contact type and carry a finite jump. Only genuinely nonlinear fields can generate shocks and expansions. Because the wave speed differs on both sides of a shock or expansion, the method of characteristics (12) fails to provide a unique solution.

2.5 Entropy conditions and integral curves

Genuinely nonlinear waves require an additional constraint, due to the failure of the method of characteristics at discontinuities. In general, this gives rise to multiple solutions and the physically relevant one has to be selected by an additional entropy condition, which can take several forms.

- Analytic formulation using an entropy pair
- Geometric formulation using the Lax entropy condition
- Physical formulation using a viscous limiting process

These three interpretations will be explained in this section.

In the **analytic interpretation**, an entropy pair $(E, \boldsymbol{F})$ is defined as a smooth convex function (the entropy) $E : \Omega \to \mathbb{R}$ and an entropy flux $\boldsymbol{F} : \Omega \to \mathbb{R}^d$ that satisfy [25, eq. 3.2]:

$$DE(\boldsymbol{u}) \cdot A^{x_j}(\boldsymbol{u}) = \boldsymbol{e}_j \cdot D\boldsymbol{F}(\boldsymbol{u}), \quad \forall j \in \{1 \dots d\} \tag{14}$$

where A^{x_j} is the flux Jacobian, $\boldsymbol{e}_j$ the unit vector in spatial direction j and D denotes the Fréchet derivative. Once an entropy pair for a system is found, the entropy condition is given by

$$\frac{\partial}{\partial t} \int_{\mathcal{V}} E(\boldsymbol{u}) \mathrm{d}\mathcal{V} + \oint_{\partial \mathcal{V}} \boldsymbol{F}(\boldsymbol{u}) \cdot \boldsymbol{n} \mathrm{d}S \leq 0. \tag{15}$$

For smooth solutions the above inequality becomes an equality, which follows directly from the definition (14): using the notation of entropy variables [6, 25] $\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{u}) = DE(\boldsymbol{u})$, premultiplying the original conservation law (2.1) pointwise by $\boldsymbol{v}^{\mathrm{T}}$ gives

$$\int_{\mathcal{V}} \boldsymbol{v}^{\mathrm{T}} \frac{\partial \boldsymbol{u}}{\partial t} \mathrm{d}\mathcal{V} + \int_{\mathcal{V}} \boldsymbol{v}^{\mathrm{T}} D\mathbf{f}(\boldsymbol{u}) : \nabla \boldsymbol{u} \mathrm{d}\mathcal{V} = 0. \tag{16}$$

which, by definition of the entropy makes the condition (15) an equality. This entropy condition provides a selection criterion that determines the unique admissible weak solution to systems which include a genuinely nonlinear field. Consider a family of Riemann problems with a fixed left state $\boldsymbol{u}_L$ connected through a simple k-wave to a variable right state $\boldsymbol{u}_R(\epsilon)$. The variable ϵ denotes the difference with the left state in the k-wave velocity, such that the initial conditions for the curve $\boldsymbol{u}_R(\epsilon)$ become

$$\begin{aligned} \boldsymbol{u}_R(\epsilon = 0) &= \boldsymbol{u}_L \\ D\boldsymbol{u}_R(\epsilon = 0) &= \boldsymbol{r}_k \end{aligned} \tag{17}$$

for a genuinely nonlinear field $\boldsymbol{r}_k$ (different families are distinguished based on the associated field). For each $\epsilon > 0$ (the expansion case), there is a unique [4] continuous, self-similar solution [25]: $\boldsymbol{u}(x,t) = \boldsymbol{v}(x/t)$. The solution for $\boldsymbol{v}(x/t)$ as a simple k-wave can be found by substituting into equation (7). It turns out this solution is exactly an integral curve of the k-th

[4] at least in the scalar one dimensional case

characteristic field. Using a parametrization $\sigma = x/t = \lambda_k(\boldsymbol{u}_L) + \epsilon$, this curve is defined as the solution to

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\sigma} &= \boldsymbol{r}_k(\boldsymbol{v}(\sigma)) \\ \boldsymbol{v}(\lambda_k(\boldsymbol{u}_L)) &= \boldsymbol{u}_L \end{cases} \tag{18}$$

The solution is the *Poisson* curve, and it is valid for $\lambda_k(\boldsymbol{u}_R) > \sigma > \lambda_k(\boldsymbol{u}_L)$. Because it is a characteristic curve, all k-Riemann invariants are constant over the expansion wave. Furthermore the solution is continuous for $t > 0$, thus overall entropy is conserved.

On the other hand, for the case $\epsilon < 0$ the entropy condition admits a single discontinuous wave solution to the conservation law. This solution is determined by evaluating the integral conservation law over the limiting volume containing the discontinuity. The result is the Rankine-Hugoniot relation

$$\overset{\text{time-continuous formulation}}{\big(\mathbf{f}(\boldsymbol{u}_R) - \mathbf{f}(\boldsymbol{u}_L)\big) \cdot \boldsymbol{n} = \sigma(\boldsymbol{u}_R - \boldsymbol{u}_L)}, \quad \overset{\text{space-time formulation}}{\big(\underline{\mathbf{f}}(\boldsymbol{u}_R) - \underline{\mathbf{f}}(\boldsymbol{u}_L)\big) \cdot \underline{\boldsymbol{n}}(\sigma) = 0.} \tag{19}$$

Differentiating this relation with respect to the shock speed σ, the admissible right shock states form another type of integral curve.

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\sigma} &= \big[A^x(\boldsymbol{v}(\sigma)) - \sigma\mathbf{I}\big]^{-1}(\boldsymbol{v}(\sigma) - \boldsymbol{u}_L) \\ \boldsymbol{v}(\lambda_k(\boldsymbol{u}_L)) &= \boldsymbol{u}_L \end{cases} \tag{20}$$

To connect this *Hugoniot* or shock curve to the existing curve, the parameter ϵ is related to the shock speed as $\sigma = \lambda_k(\boldsymbol{u}_L) + \frac{1}{2}\epsilon$, which satisfies the initial condition (17). Consequently, the Hugoniot curve is valid for $\lambda_k(\boldsymbol{u}_L) > \sigma > \lambda_k(\boldsymbol{u}_R)$. This shock curve is different from the characteristic curve, hence the k-Riemann invariants change over the shock (although proportional to ϵ^2). The entropy also changes, and integrating the inequality (15) over the shock gives a Rankine-Hugoniot type inequality for the entropy

$$\sigma \, [\![E(\boldsymbol{u})]\!] > [\![\boldsymbol{F}(\boldsymbol{u})]\!] \cdot \boldsymbol{n}.$$

The definition of the distinct solution curves resulting from the entropy condition (15) provides a **geometric interpretation** for the entropy condition: characteristics always point into shocks and outward from expansion fans. This is clear from the eigenvalue bounds specified for the arcs of each curve. This is summarized by the Lax entropy condition [25]:

$$\begin{aligned} \lambda_k(\boldsymbol{u}_L) > \sigma > \lambda_k(\boldsymbol{u}_R) \quad &\rightarrow \quad \text{Hugoniot curve} \\ \lambda_k(\boldsymbol{u}_L) = \sigma = \lambda_k(\boldsymbol{u}_R) \quad &\rightarrow \quad \text{Characteristic curve} \\ \lambda_k(\boldsymbol{u}_L) < \sigma < \lambda_k(\boldsymbol{u}_R) \quad &\rightarrow \quad \text{Poisson curve} \end{aligned}$$

The linearly degenerate case (middle) is added here for completeness. This alternative entropy condition is equivalent to (15) for strictly convex entropies, at least if ϵ is small enough.

The **physical interpretation** of the entropy condition is obtained by considering solutions $\boldsymbol{u}_\beta$ to the modified conservation law (21) in the limit of vanishing viscosity $\beta \downarrow 0$.

$$\frac{\partial}{\partial t}\int_{\mathcal{V}} \boldsymbol{u}_\beta \mathrm{d}\mathcal{V} + \oint_{\partial\mathcal{V}} (\mathbf{f}(\boldsymbol{u}_\beta) - \beta\boldsymbol{\nu}\nabla\boldsymbol{u}_\beta)\cdot\boldsymbol{n}\mathrm{d}S = 0. \tag{21}$$

For any symmetric positive definite viscosity tensor $\boldsymbol{\nu}$ this limit makes sense, and in that case the viscous solutions $\boldsymbol{u}_\beta$ converge to the entropy solution of the conservation law (2.1) independent of chosen viscosity $\boldsymbol{\nu}$. For this important result, see [12, 25]. The main difference between entropy solutions defined by the viscous limit and those defined by the analytic or geometric equivalents, is that the behaviour of the solution inside a shock is well-defined in the viscous limit case. In other words, the path taken by the solution between $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ (known as the viscous shock profile) is fixed by the choice of viscosity. Therefore $\boldsymbol{u}_\beta$ is a proper distribution whereas the entropy solution $\boldsymbol{u}$ is only defined 'almost everywhere'. Note that the viscous shock profiles are not the same as the Hugoniot curveb(20), which is independent of viscosity.

3 Theory of nonconservative systems

This chapter covers the extension of the theory of conservation laws to nonconservative systems. Essentially, nonconservative systems form a superset of conservation laws. Therefore it is important to make a distinction between the nonconservative systems that can be expressed in conservation form (2.1) and those that can not (genuinely nonconservative systems). As an example, the quasilinear forms of the Euler or Navier-Stokes equations are nonconservative, but they remain conservation laws. The interacting boundary layer system, the system of shallow water equations with topography [42] and multiphase pipe flows [51] are genuinely nonconservative. The scope is limited to the type of nonconservative systems occurring in the above cases, i.e. quasilinear and hyperbolic.

Firstly, a generic formulation for nonconservative systems of this type is presented, and it is shown how several genuinely nonconservative processes lead to this description. Secondly, a definition as suggested in literature for the weak solution of nonconservative systems is provided; the definition for conservation laws does not apply and additional information in the form of a family of paths is needed. The following sections address the implications of the choice of paths on convergence and the correct capture of waves or stationary solutions. Finally, some notes are made on the relation between the path-based solution and the entropy conditions as applied to conservation laws.

3.1 Formulation and interpretation

Many physical systems are conservative in nature: if a model is complete enough to describe conservation of mass, momentum and energy, any closed system is conservative. The previous sentence gives several reasons for a system to become nonconservative:

- Splitting a system into parts without tracking the interfaces
- Considering only part of the total mass/momentum/energy of a system
- Reducing the dimension of a system
- Imposing interaction with an external field

In the first case, exchanges between the subsystems can obviously be formulated as a conservative flux only if the location of the boundary between subsystems is known. Indeed, the flux is the only mechanism for exchange in a conservative system, and it needs a surface or interface to act on. Otherwise, the system as a whole may be conservative, but the subsystems are genuinely nonconservative (a situation occuring in certain models of multiphase flow [20, 31, 51]). In second case there is no interface to consider, but energy or momentum are exchanged between different modes (e.g. resolved and unresolved scales in turbulence modelling). The third situation applies to the integral boundary layer, where the outer streamline boundary becomes part of the interior domain of the depth-averaged formulation (see figure 4). This interface is therefore ‘lost’, and interaction terms at that boundary become nonconservative. The same occurs in the shallow water equations, when including a non-trivial bottom topography. Again, the nonconservative terms occur trough depth-averaging the interaction at the bottom. Steady state problems (without timestepping) involve dimension reduction as well, but due to the periodic boundary condition in time, any nonconservative terms at the time boundary cancel. The last situation concerning interaction with an external field occurs in many situations (e.g. gravity). Such interactions may be either conservative or non-conservative, depending on the model. The mathematical distinction between these two cases is postponed to the end of this section, after introducing the necessary mathematical description.

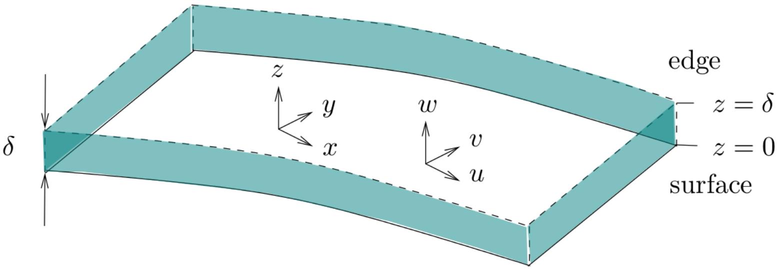

Figure 4: *Volume cell of the boundary layer. The boundaries of the reduced dimension cell are shown in blue. The surface at $z = 0$ and the outer streamline at $z = \delta$ become part of the domain interior.*

The generic formulation of a quasilinear nonconservative system is

$$\overset{\text{time-continuous formulation}}{\frac{\partial}{\partial t}\int_{\mathcal{V}} \boldsymbol{u}\mathrm{d}\mathcal{V} + \int_{\mathcal{V}} \mathbf{g}(\boldsymbol{u}) : \nabla\boldsymbol{u}\mathrm{d}\mathcal{V} = \int_{\mathcal{V}} \boldsymbol{s}(\boldsymbol{u})\mathrm{d}\mathcal{V},} \qquad \overset{\text{space-time formulation}}{\int_{\underline{\mathcal{V}}} \underline{\mathbf{g}}(\boldsymbol{u}) : \underline{\nabla}\boldsymbol{u}\mathrm{d}\underline{\mathcal{V}} = \int_{\underline{\mathcal{V}}} \boldsymbol{s}(\boldsymbol{u})\mathrm{d}\underline{\mathcal{V}}} \tag{22}$$

where ‘:’ denotes a double inner product between the state gradient tensor $\nabla\boldsymbol{u} \in \mathcal{L}(\mathrm{T}_x\underline{\mathcal{V}}; \Omega)$ and the *balance* tensor $\underline{\mathbf{g}} \in \mathcal{L}(\mathrm{T}_x\underline{\mathcal{V}}, \Omega; \Omega)$. This *nonconservative* product takes values in Ω, is comparable to the flux density in a conservative system (compare figures 5 and 1). However, two important distinctions with the flux density must be noted. Firstly, the nonconservative product is integrated over space(-time) volumes instead of surfaces. Secondly, it causes both transport and production of the state quantities, whereas the flux density purely causes transport. The source term does not contain any derivatives and is L^2 bounded. As stated here, the space-time formulation allows a nonconservative product with the time derivative. Although such a term can be equally well incorporated in the time-continuous formulation, this term is not found anywhere in the studied literature. Even while the space-time formulation of [42] naturally includes this term, the authors seem to neglect it, as the nonconservative product is only integrated over the space-like faces. These terms are not neglected in the present formulation.

There exist several ways to check if the nonconservative formulation (22) is a conservation law[5]. This is important, since in that case the much further developed conservative numerical schemes can be used. The following three conditions are equivalent for a given system,

- The balance tensor $\mathbf{g}$ is the Jacobian of a conservative flux $\mathbf{f}(\boldsymbol{u})$.
- The path integral (23) of $\mathbf{g}(\boldsymbol{u})\mathrm{d}\boldsymbol{u}$ yields zero for any closed path Γ in Ω.

$$\oint_\Gamma \boldsymbol{\phi}_0 \left(\mathbf{g}(\boldsymbol{u}) \cdot \boldsymbol{n}_0\right)\mathrm{d}\boldsymbol{u} = 0 \qquad \text{for all tests } \boldsymbol{\phi}_0 \in \Omega^* \text{ and all directions } \boldsymbol{n}_0 \in \mathrm{T}_x\mathcal{V} \quad (23)$$

- The system is a conservation law.

Hence to determine whether the system is a conservation law or not, it is sufficient to prove or disprove any of the first two statements. In literature, only the first condition is generally mentioned. This one is however not very constructive if $\mathbf{f}$ is not known, and certainly its negation is hard to demonstrate for a nonconservative system. The second condition is introduced here to show more easily that a given system is nonconservative (for an example, see section 3.3). It essentially claims that the nonconservative product is path-independent. If so, it can be reduced to conservative form; otherwise it is genuinely nonconservative. As a consequence, all nonconservative products in scalar partial differential equations can be reduced to a conservative form. Note that the test is based on a path in solution space Ω; it is irrelevant how far the states are separated in the spatial domain $\mathcal{V}$. Hence this test applies to smooth solutions as well as shocks.

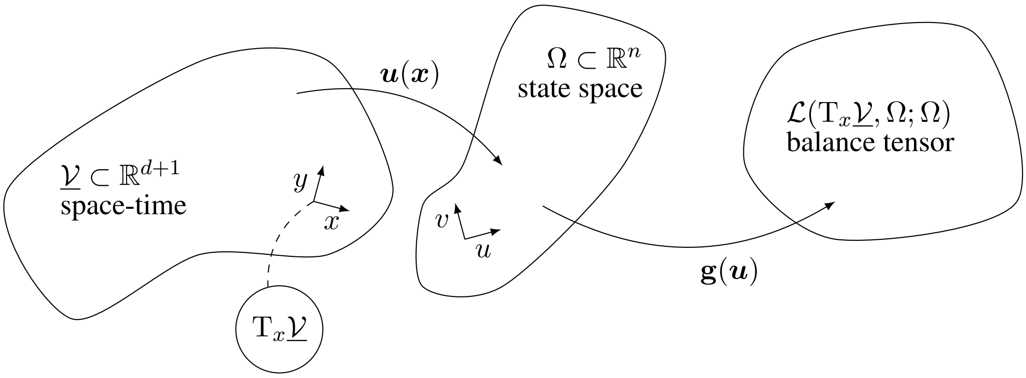
$\underline{\mathcal{V}} \subset \mathbb{R}^{d+1}$
space-time
y
x
$\mathrm{T}_x\underline{\mathcal{V}}$
$\boldsymbol{u}(\boldsymbol{x})$
$\Omega \subset \mathbb{R}^n$
state space
v
u
$\mathbf{g}(\boldsymbol{u})$
$\mathcal{L}(\mathrm{T}_x\underline{\mathcal{V}}, \Omega; \Omega)$
balance tensor

Figure 5: *Distinction between the spaces in the nonconservative problem formulation. Space-time $\underline{\mathcal{V}}$ is a metric space, the 'state' Ω and 'direction' $\mathrm{T}_x\underline{\mathcal{V}}$ are vector spaces, and the balance tensor is a multilinear map between them.*

Finally, the question returns whether a system interacting with an external field should be considered conservative or nonconservative. Direct source terms aside, a first order interaction takes the form of the gradient of a field p acting on the system as $\mathbf{a}\nabla p$:

$$\frac{\partial}{\partial t}\int_\mathcal{V} \boldsymbol{u}\mathrm{d}\mathcal{V} + \int_\mathcal{V} \mathbf{g}(\boldsymbol{u}, p) : \nabla\boldsymbol{u}\mathrm{d}\mathcal{V} = \int_\mathcal{V} \mathbf{a}(\boldsymbol{u}, p) \cdot \nabla p\mathrm{d}\mathcal{V}$$

[5] regarding the source term, see section 2.2

Since the source term in the generic formulation (22) is not allowed to contain derivatives, this must be reformulated into a system of $n+1$ equations for $(\boldsymbol{u}, p)$:

$$\frac{\partial}{\partial t}\int_{\mathcal{V}}\begin{pmatrix}\boldsymbol{u}\\ p\end{pmatrix}\mathrm{d}\mathcal{V}+\int_{\mathcal{V}}\begin{pmatrix}\mathbf{g}(\boldsymbol{u},p) & -\mathbf{a}(\boldsymbol{u},p)\\ \mathbf{0} & \mathbf{0}\end{pmatrix}:\nabla\begin{pmatrix}\boldsymbol{u}\\ p\end{pmatrix}\mathrm{d}\mathcal{V}=0 \tag{24}$$

If $\mathbf{g} = D_{\boldsymbol{u}}\mathbf{f}(\boldsymbol{u}, p)$ and $\mathbf{a} = D_p\mathbf{f}(\boldsymbol{u}, p)$, condition (23) holds and (24) is a conservation law. The evolution of the field p is necessarily included in this definition (here, $\partial p/\partial t = 0$). If the combined balance tensor $[\mathbf{g}, \text{-}\mathbf{a}]$ and the combined state $(\boldsymbol{u}, p)$ do not satisfy relation (23), the interaction with the external field is nonconservative.

3.2 Definition of weak solutions

As in the conservative case, discontinuities in the solution may occur in finite time. Therefore weak solutions to the nonconservative system should be defined. Take the sum of equations (2.1) and (22), apply the test function $\boldsymbol{\phi} \in \Omega^*$ pointwise and integrate the conservative term by parts. One obtains the equivalent of equation (4), the weak form for general nonconservative systems,

$$\begin{aligned}&\oint_{\partial\underline{\mathcal{V}}}\underline{\mathbf{f}}\cdot\underline{\boldsymbol{n}}\cdot\boldsymbol{\phi}\mathrm{d}\underline{\mathcal{S}}-\int_{\underline{\mathcal{V}}}\underline{\mathbf{f}}(\boldsymbol{u}):\underline{\nabla}\boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}}\\ &\qquad+\int_{\underline{\mathcal{V}}}\underline{\mathbf{g}}(\boldsymbol{u}):\underline{\nabla}\boldsymbol{u}\cdot\boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}}=\int_{\underline{\mathcal{V}}}\boldsymbol{s}(\boldsymbol{u})\cdot\boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}},\qquad\forall\boldsymbol{\phi}\in(H^1(\underline{\mathcal{V}}))^n.\end{aligned} \tag{25}$$

$$\text{with}\quad H^1:\left\{\phi:\mathcal{V}\to\mathbb{R}:\phi\in L^2\text{ and }\nabla\phi\in(L^2)^d\right\}$$

When looking for discontinuous solutions, instead of the Sobolev space H^1 the *broken* Sobolev space H^1_h (61) is used. In that case, one term in this equation is not well-defined: the nonconservative product $\underline{\mathbf{g}}(\boldsymbol{u}) : \underline{\nabla}\boldsymbol{u}$. Indeed, at discontinuities the gradient behaves as a delta distribution and the tensor $\underline{\mathbf{g}}$ makes a jump, rendering the contribution of their product to the integral undetermined. Several definitions for the evaluation of these type of products have been developed, that provide meaning to the above formulation. Colombeau [14], Cauret and Le Roux [9] provide a definition based on their theory of generalized functions, allowing a self-consistent formulation for the definition of weak solutions and the numerical approximation thereof. A later definition based on limiting paths is posed in the theory of Dal Maso, LeFloch and Murat [15] (known as DLM theory). They separate the domain into regions of continuity $\mathcal{C}$ and regions of discontinuity $\mathcal{S}$. The products are well-defined in continuous regions, and the definition is extended by considering the discontinuities as the limit of a continuous path connecting both sides. This leads to the DLM definition of the nonconservative product:

$$\int_{\mathcal{V}}\left[\mathbf{g}(\boldsymbol{u}):\nabla\boldsymbol{u}\right]_{\psi}\mathrm{d}\mathcal{V}\overset{\text{def}}{=}\int_{\mathcal{C}}\mathbf{g}(\boldsymbol{u}):\nabla\boldsymbol{u}\mathrm{d}\mathcal{V}+\int_{\mathcal{S}}\int_0^1\mathbf{g}(\boldsymbol{\psi}(\boldsymbol{u}_L,\boldsymbol{u}_R,s))\cdot\frac{\partial}{\partial s}\boldsymbol{\psi}(\boldsymbol{u}_L,\boldsymbol{u}_R,s)\mathrm{d}s\cdot\boldsymbol{n}\mathrm{d}\mathcal{S}, \tag{26}$$

where $\boldsymbol{\psi} : \Omega \times \Omega \times [0,1] \to \Omega$ is a path connecting the 'left' and 'right' side of the discontinuity satisfying the following properties [42]:

$$\boldsymbol{\psi}(\boldsymbol{u}_L, \boldsymbol{u}_R, 0) = \boldsymbol{u}_L \tag{27}$$

$$\boldsymbol{\psi}(\boldsymbol{u}_L, \boldsymbol{u}_R, 1) = \boldsymbol{u}_R \tag{28}$$

$$\boldsymbol{\psi}(\boldsymbol{u}_L, \boldsymbol{u}_L, \xi) = \boldsymbol{u}_L. \tag{29}$$

Furthermore, the path ψ should be Lipschitz continuous:

$$\|\boldsymbol{\psi}(\boldsymbol{u}_L, \boldsymbol{u}_R, a) - \boldsymbol{\psi}(\boldsymbol{u}_L, \boldsymbol{u}_R, b)\| \leq K|a-b|\|\boldsymbol{u}_L - \boldsymbol{u}_R\| \quad \text{for all } a,b \text{ in } [0,1] \text{ and all } \boldsymbol{u}_L,\boldsymbol{u}_R \text{ in } \Omega. \tag{30}$$

A third method to define the weak solutions is to consider the viscous limit of the nonconservative system, which leads to the integration of the nonconservative product over the viscous profiles. This will be further detailed in section 3.4

3.3 Example (non)conservative system

To illustrate some of the difficulties with the manipulation of nonconservative systems in practice, a simple system for elastic wave propagation (obtained from [9]) is analysed here. Consider the 2x2 system

$$\begin{aligned} u_t + uu_x &= \sigma_x \\ \sigma_t + u\sigma_x &= k^2 u_x \end{aligned} \quad \text{or in matrix form} \quad \frac{\partial}{\partial t}\underbrace{\begin{bmatrix} u \\ \sigma \end{bmatrix}}_{\boldsymbol{v}} + \underbrace{\begin{bmatrix} u & -1 \\ -k^2 & u \end{bmatrix}}_{G^x}\frac{\partial}{\partial x}\begin{bmatrix} u \\ \sigma \end{bmatrix} = 0. \tag{31}$$

The nonconservative property of this system is easily checked:

$$\oint G^x \mathrm{d}\boldsymbol{v} = \oint \begin{bmatrix} 0 & -1 \\ -k^2 & 0 \end{bmatrix} \mathrm{d}\boldsymbol{v} + \oint \begin{bmatrix} u\mathrm{d}u \\ u\mathrm{d}\sigma \end{bmatrix} = \oint \begin{bmatrix} \mathrm{d}(u^2/2) \\ u\mathrm{d}\sigma \end{bmatrix} = \oint \begin{bmatrix} 0 \\ u\mathrm{d}\sigma \end{bmatrix}.$$

For the path Γ depicted in figure 6, this integral is obviously nonzero (in fact it is equal to the enclosed area). However, the authors of [9] proceed to write down an 'equivalent' conservative

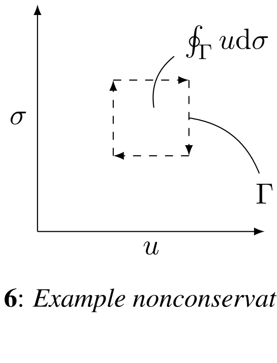

Figure 6: *Example nonconservative path*

system by the substitution $q = \sigma + \frac{1}{2}u^2$,

$$\frac{\partial}{\partial t}\underbrace{\begin{bmatrix} u \\ q \end{bmatrix}}_{\boldsymbol{f}^t} + \frac{\partial}{\partial x}\underbrace{\begin{bmatrix} u^2 - q \\ u^3/3 - k^2 u \end{bmatrix}}_{\boldsymbol{f}^x} = 0, \tag{32}$$

which they disfavour because its solutions become undefined for large shocks. It is remarkable that a genuinely nonconservative system allows a conservative formulation (even though the formulations are in fact only equivalent for continuous solutions). From here on, we further analyse these two systems using the theory from section 3, in order to pinpoint the crucial step that causes the difference between both formulations. First it is shown that the systems are (in a continuous sense) equivalent by writing the quasilinear form of (32),

$$\begin{aligned} \frac{\partial \boldsymbol{f}^t}{\partial t} &= \frac{\partial \boldsymbol{f}^t}{\partial u}\frac{\partial u}{\partial t} + \frac{\partial \boldsymbol{f}^t}{\partial \sigma}\frac{\partial \sigma}{\partial t} = A^t \frac{\partial \boldsymbol{v}}{\partial t}, \\ \frac{\partial \boldsymbol{f}^x}{\partial x} &= \frac{\partial \boldsymbol{f}^x}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial \boldsymbol{f}^x}{\partial \sigma}\frac{\partial \sigma}{\partial x} = A^x \frac{\partial \boldsymbol{v}}{\partial x}, \end{aligned}$$

$$\underbrace{\begin{bmatrix} 1 & 0 \\ u & 1 \end{bmatrix}}_{A^t} \frac{\partial}{\partial t}\begin{bmatrix} u \\ \sigma \end{bmatrix} + \underbrace{\begin{bmatrix} u & -1 \\ u^2 - k^2 & 0 \end{bmatrix}}_{A^x} \frac{\partial}{\partial x}\begin{bmatrix} u \\ \sigma \end{bmatrix} = 0. \tag{33}$$

Obviously, equation (33) is a conservation law, because A^t and A^x are Jacobians (the path property is easily checked). Since A^t is lower triangular with nonzero diagonal for all $\boldsymbol{v}$, it is invertible and equation (33) can be premultiplied by its inverse.

$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ \sigma \end{bmatrix} + \underbrace{\begin{bmatrix} 1 & 0 \\ -u & 1 \end{bmatrix}\begin{bmatrix} u & -1 \\ u^2 - k^2 & 0 \end{bmatrix}}_{G^x = [A^t]^{-1}[A^x]} \frac{\partial}{\partial x}\begin{bmatrix} u \\ \sigma \end{bmatrix} = 0. \tag{34}$$

But the result (34) is exactly equal to the original nonconservative system (31): just by the multiplication, equation (33) has become nonconservative. Since one system can not be both conservative and nonconservative at the same time, this is in fact a different system, which happens to have the same residual for continuous (approximate) solutions. As one can see, the transformation between the conserved variables $\boldsymbol{f}^t$ and the primitive variables $\boldsymbol{v}$ does *not* affect the system, but it is the multiplication that requires further investigation. To see the effects clearly, compare the coefficient tensors of equations (33) and (34):

$$\mathbf{a} = A^t \otimes \frac{\partial \cdot}{\partial t} + A^x \otimes \frac{\partial \cdot}{\partial x} \qquad \oint \mathbf{a} \cdot \mathrm{d}\boldsymbol{v} = \mathbf{0}, \tag{35}$$

$$\mathbf{g} = I \otimes \frac{\partial \cdot}{\partial t} + G^x \otimes \frac{\partial \cdot}{\partial x} \qquad \oint \mathbf{g} \cdot \mathrm{d}\boldsymbol{v} \neq \mathbf{0}. \tag{36}$$

The coefficient tensor has undergone a pre-multiplication by $[A^t]^{-1}$, making it nonconservative. Note that post-multiplication by an invertible matrix M would not change the system, since that is equivalent to a change of variables $\boldsymbol{w}(\boldsymbol{v})$ with $D\boldsymbol{w}(\boldsymbol{v}) = M$.

$$\mathbf{g}M\mathrm{d}\boldsymbol{v} = \mathbf{g}\mathrm{d}\boldsymbol{w}.$$

There are two special cases in which pre-multiplication preserves the conservation property of the system:

- The system is pre-multiplied by a matrix M that commutes with the coefficient tensor $\mathbf{G}$.
- The system is pre-multiplied by the transpose of the entropy variables $\boldsymbol{v}^{\mathrm{T}}$.

The first case corresponds to post-multiplication (i.e. change of variables) and the second case results in the entropy condition (the conservation property becomes an inequality). Thus it is very important to be careful when producing a mathematical description of a physical system, since seemingly innocent mathematical recipes can influence the conservation properties of the result. The derivations of the (boundary layer) mechanical energy equation is an example where these pre-multiplication recipes are applied (see e.g. White [49], Matsushita [34], Ozdemir [38], van 't Hof [29]). Sometimes this can be justified by other means (see section 6.3). Another example of a 'pre-multiplication' which is ubiquitous in numerical modelling, is the application of test functions $\boldsymbol{\phi} \in \Omega^*$. Many forms of numerical stabilization (both in FV and FE approaches) can be expressed in terms of a non-constant or even solution-dependent test function (see section 4.1). In light of the previous example, the test functions thus essentially modify the system being solved to an 'equivalent' system (compare the FD approach). It is also apparent when comparing (35) to (36) that this equivalence is broken for nonconservative systems (due to path-dependence), an important observation in anticipation of the convergence issues to be discussed in section 3.6.

3.4 Dependence on viscous profiles

The viscous profiles are defined (see [25]) as the vanishing viscosity solution $\boldsymbol{v}(\xi^*, \sigma)$ to the viscous system (21) for a shock between a left state $\boldsymbol{u}_L$ and a right state $\boldsymbol{u}_R$. The spatial variable $\xi^* = (x - \sigma t)/\beta$ scales with the viscous strength β and makes the formulation independent of the viscous length scale.

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\xi^*} & = \left[A^x(\boldsymbol{v}) - \sigma\mathbf{I}\right]^{-1} \dfrac{\partial}{\partial \xi^*}\left(B^x(\boldsymbol{v})\dfrac{\partial \boldsymbol{v}}{\partial \xi^*}\right), \\ \boldsymbol{v}(-\infty, \sigma) & \to \boldsymbol{u}_L, \qquad \dfrac{\partial \boldsymbol{v}}{\partial \xi^*}(-\infty, \sigma) \to 0, \\ \boldsymbol{v}(+\infty, \sigma) & \to \boldsymbol{u}_R, \qquad \dfrac{\partial \boldsymbol{v}}{\partial \xi^*}(+\infty, \sigma) \to 0. \end{cases} \tag{37}$$

This shock profile is valid for $\lambda_k(\boldsymbol{u}_R) < \sigma < \lambda_k(\boldsymbol{u}_L)$ (hence the trivial solution $\boldsymbol{u}_L = \boldsymbol{u}_R$ is excluded). Finding the viscous profiles, as defined by this singular boundary value problem on an infinite domain is not straightforward. At least their unique solution near a given left state can be found by reparametrizing the viscous profiles in terms of the departure from that state $\alpha = \|\boldsymbol{u}_L - \boldsymbol{v}\|$, see figure 7. When α is strictly increasing along the curve, this results in

$$\begin{cases} \big[A^x(\boldsymbol{v}) - \sigma\mathbf{I}\big]\dfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\alpha} & = \dfrac{\partial}{\partial\alpha}\left(B^x(\boldsymbol{v})\dfrac{\partial\boldsymbol{v}}{\partial\alpha}\dfrac{\mathrm{d}\alpha}{\mathrm{d}\xi^*}\right), \\ \boldsymbol{v}(0,\sigma) & = \boldsymbol{u}_L, \\ \boldsymbol{v}(\alpha_R,\sigma) & = \boldsymbol{u}_R \quad \text{with } \alpha_R = \|\boldsymbol{u}_R - \boldsymbol{u}_L\|. \end{cases} \tag{38}$$

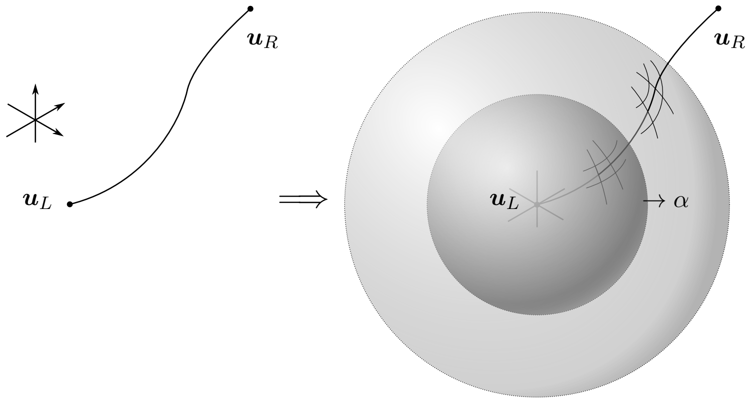

Figure 7: *Parametrization of viscous curve in the solution space Ω. The parameter α is associated with the radius of a sphere around $\boldsymbol{u}_L$. If the viscous curve is not tangential to any of these spheres, the intersection of the curve with the sphere provides a unique parametrization in terms of α.*

3.4.1 First order expansion

It is assumed that the advection and diffusion matrices $A^x(\boldsymbol{u})$ and $B^x(\boldsymbol{u})$ are continuously differentiable,

$$\begin{aligned} A(\boldsymbol{v}) &= A(\boldsymbol{v}(0)) + \alpha DA(\boldsymbol{v}(0))\{\boldsymbol{v}_\alpha(0)\} + \mathcal{O}(\alpha^2), \\ B(\boldsymbol{v}) &= B(\boldsymbol{v}(0)) + \alpha DB(\boldsymbol{v}(0))\{\boldsymbol{v}_\alpha(0)\} + \mathcal{O}(\alpha^2). \end{aligned}$$

The operator D again denotes differentiation with respect to the argument. The directional derivative of A in direction $\boldsymbol{v}$ is indicated by $DA\{\boldsymbol{v}\}$ (which is nothing more than an inner product). Then equation (38) can be expanded around the left state up to order $\mathcal{O}(\alpha^2)$. Using one-sided finite difference for the right hand side,

$$\big[A(\boldsymbol{u}_L) + \alpha DA(\boldsymbol{u}_L)\{\boldsymbol{v}_\alpha\} + \mathcal{O}(\alpha^2) - \sigma\mathbf{I}\big]\frac{\partial\boldsymbol{v}}{\partial\alpha} = \frac{1}{\alpha}\frac{\mathrm{d}\alpha}{\mathrm{d}\xi^*}\big[B(\boldsymbol{u}_L) + \alpha DB(\boldsymbol{u}_L)\{\boldsymbol{v}_\alpha\} + \mathcal{O}(\alpha^2)\big]\frac{\partial\boldsymbol{v}}{\partial\alpha}. \tag{39}$$

For a given σ and the limit of α to zero, this system becomes a generalized[6] eigenvalue problem for the left boundary condition

$$\big[A(\boldsymbol{u}_L) - \mu(0)B(\boldsymbol{u}_L) - \sigma\mathbf{I}\big]\boldsymbol{s}(0) = 0, \tag{40}$$

$$\text{with} \qquad \mu_k(0) = \lim_{\alpha\to 0}\frac{1}{\alpha}\frac{\partial\alpha}{\partial\xi^*}, \qquad\qquad \boldsymbol{s}_k(0) = \lim_{\alpha\to 0}\left(\frac{\partial\boldsymbol{v}}{\partial\alpha}\right).$$

The eigenvalue μ represents the decay of the profile at infinity, and the eigenvector $\boldsymbol{s}$ corresponds to the starting direction of the profile in Ω. Note that the decay should be exponential in ξ^* in order to produce a finite μ (and thus a solution). By changing the diffusion matrix B, arbitrary viscous profile directions can be obtained which can differ considerably from the Hugoniot and Poisson curves. Once the unique solution for (40) is found, the viscous profiles for each k-shock are well-defined and are found by integration of (38). Furthermore if B commutes with A (i.e. they have the same eigenvectors), then $\boldsymbol{s}_k = -\boldsymbol{r}_k/\|\boldsymbol{r}_k\|$ and $\mu_k = -\epsilon/2\nu_k$, where ν_k is the viscosity for the k-th field and ϵ the shock strength. This implies that the viscous profiles start off in the same direction as the Hugoniot and Poisson curves:

$$\boldsymbol{v}(\alpha,\sigma) = \boldsymbol{u}_L - \alpha\boldsymbol{r}_k + \mathcal{O}(\alpha^2) \quad \text{if } B \text{ commutes with } A. \tag{41}$$

3.4.2 Second order expansion

Because the viscous system is singular at the left boundary up to second order in α for a constant σ, it is possible to find the second derivative of the profile at the left end by differentiating (39). First, expressing this equation in terms of $Q(\alpha) = A(\boldsymbol{v}(\alpha)) - \mu_k(\alpha)B(\boldsymbol{v}(\alpha))$,

$$\big[Q(0) + \alpha DQ(0) + \alpha D\mu(0)B(\boldsymbol{u}_L) - \sigma\mathbf{I}\big]\boldsymbol{s}(\alpha) = \mathcal{O}(\alpha^2), \tag{42}$$

$$\text{with} \qquad \mu_k(\alpha) = \frac{1}{\alpha}\frac{\partial\alpha}{\partial\xi^*}, \qquad\qquad \boldsymbol{s}_k(\alpha) = \frac{\partial\boldsymbol{v}}{\partial\alpha},$$

then taking the derivative with respect to α results in

$$\big[\sigma\mathbf{I} - Q(0)\big]D\boldsymbol{s}(\alpha) = \big[DQ(0) + D\mu(0)B(\boldsymbol{u}_L)\big]\boldsymbol{s}(\alpha) + \mathcal{O}(\alpha).$$

Taking the limit for $\alpha \to 0$ and premultiplying by the pseudo-inverse of $[\sigma\mathbf{I} - Q]$, using the fact that $D\boldsymbol{s}(0)$ is orthogonal to the associated null space (i.e. orthogonal to $\boldsymbol{s}(0)$),

$$D\boldsymbol{s}(0) = \big[\sigma\mathbf{I} - Q(0)\big]^{\dagger}\big[DQ(0) + D\mu(0)B(\boldsymbol{u}_L)\big]\boldsymbol{s}(0).$$

[6] Generalized eigenvalue problems are explained in Appendix B.

The derivative of $\boldsymbol{r}_k$ (in the same direction $\boldsymbol{s}(0)$) on the other hand satisfies [32]

$$D\boldsymbol{r}(\boldsymbol{u}_L)\{\boldsymbol{s}(0)\} = \big[\lambda(\boldsymbol{u}_L)\mathbf{I} - A(\boldsymbol{u}_L)\big]^{\dagger}\big[(DA(\boldsymbol{u}_L) - D\lambda(\boldsymbol{u}_L)\mathbf{I})\{\boldsymbol{s}(0)\}\big]\boldsymbol{r}(\boldsymbol{u}_L).$$

These expressions are very similar, let us therefore consider how much they differ. We assume again that B commutes with A. Due to the normalization of $\boldsymbol{r}_k$ and $\boldsymbol{s}_k$ the eigenvalue derivative becomes $D\lambda\{\boldsymbol{s}_k\} = 1/\|\boldsymbol{r}_k\|$, which can be isolated in a term independent of ϵ. The remaining difference within the first brackets is $\frac{\epsilon}{2\nu}(\nu\mathbf{I} - B(\boldsymbol{u}_L))$ and within the second brackets $\frac{\epsilon}{2\nu}DB(\boldsymbol{u}_L)\{\boldsymbol{s}(0)\}$. Using the continuity of the pseudoinverse [45] for the given perturbation,

$$\begin{aligned} D\boldsymbol{r}(\boldsymbol{u}_L)\{\boldsymbol{s}(0)\} &= \big[\sigma\mathbf{I} - Q(0)\big]^{\dagger}\big[DQ(0) + D\mu(0)B(\boldsymbol{u}_L) - \mathbf{I}\big]\boldsymbol{s}(0) + \mathcal{O}(\epsilon) \\ &= D\boldsymbol{s}(0) + \mathcal{O}(\epsilon) - \big[\lambda(\boldsymbol{u}_L)\mathbf{I} - A(\boldsymbol{u}_L)\big]^{\dagger}\frac{\boldsymbol{r}(\boldsymbol{u}_L)}{\|\boldsymbol{r}(\boldsymbol{u}_L)\|}. \end{aligned}$$

The last term vanishes if A is symmetric. Therefore the second order expansion of the viscous profile will agree with the Hugoniot and Poisson curves if and only if A is symmetric. The approximation around the left state then satisfies for either ($\sigma = \lambda_k + \epsilon/2$) or ($\alpha = -\epsilon/\|\boldsymbol{r}_k\|$)

$$\boldsymbol{v}\,(\alpha,\sigma) = \boldsymbol{v}(\epsilon) = \boldsymbol{u}_L + \epsilon\boldsymbol{r}_k + \frac{1}{2}\epsilon^2 D\boldsymbol{r}_k\cdot\boldsymbol{r}_k + \mathcal{O}(\epsilon^3). \tag{43}$$

3.5 Path-dependent effects

Upon choosing a family of paths, the definition (26) uniquely defines the weak solutions to the nonconservative system (25). Therefore different paths for shocks, expansions and contact discontinuities are needed. The nonconservative versions of the Hugoniot (shock) and Poisson (expansion) curves depend upon these paths. For conservation laws, due to property (23) the DLM definition becomes independent of the chosen paths. Whereas the paths for expansions and contact discontinuities are equal to the characteristic curves (due to self-similarity, see section 2.3), shock curves and paths are not equal in general (just like in conservation laws). For limiting viscosity solutions, the paths can be chosen as the viscous profiles. Since the generalised Rankine-Hugoniot condition for nonconservative systems is path-dependent

$$\sigma(\boldsymbol{u}_R - \boldsymbol{u}_L) = \int_0^1 \mathbf{g}(\boldsymbol{\psi}(\boldsymbol{u}_L,\boldsymbol{u}_R,s))\cdot\frac{\partial}{\partial\xi}\boldsymbol{\psi}(\boldsymbol{u}_L,\boldsymbol{u}_R,s)\mathrm{d}s, \tag{44}$$

the shock curves based on (44) also depend on the path, and therefore on the viscosity [12]:

$$\left\{\begin{aligned} \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}\sigma} &= \big[A^x(\boldsymbol{v}(\sigma)) - \sigma\mathbf{I}\big]^{-1}\big(\boldsymbol{v}(\sigma) - \boldsymbol{u}_L + \boldsymbol{e}(\psi)\big), \\ \boldsymbol{v}(\lambda_k(\boldsymbol{u}_L)) &= \boldsymbol{u}_L, \end{aligned}\right. \tag{45}$$

where the path dependent effect $\boldsymbol{e}(\psi)$ is given by

$$\boldsymbol{e}(\psi) = \int_0^1 D\mathbf{g}(s)\{\boldsymbol{\psi}_\sigma\} \cdot \boldsymbol{\psi}_s - D\mathbf{g}(s)\{\boldsymbol{\psi}_s\} \cdot \boldsymbol{\psi}_\sigma \mathrm{d}s. \tag{46}$$

The important result is that entropy solutions for nonconservative systems (as opposed to conservation laws) depend on the limiting viscosity. As noted by [12], this effect is zero when $\mathbf{g}$ is conservative, or when the shock paths and shock curves are equal. In that case, the generalized Hugoniot curve (45) reduces to the classical Hugoniot curve (20). Alouges and Merlet [1] propose to use the classical Hugoniot curve as an approximation to the nonconservative shock curve (effectively neglecting the effect (46)). They show for one-dimensional systems that when the viscosity and advection operators commute, classical shock curves approximate the true viscous shock curves up to third order in the jump magnitude, justifying this approximation for weak shocks. Using an entirely different approach, this follows as well from the results from section 3.4, by substituting (41) into equation (46). Doing the same for (43), the estimate is improved to fourth order for symmetric systems (an additional result hinted by Alouges and Merlet as well). On the other hand, for general (noncommuting) viscosities the classical shock curve agrees only up to second order. In that case, a simpler estimate of the same order is obtained using linear paths (see figure 8).

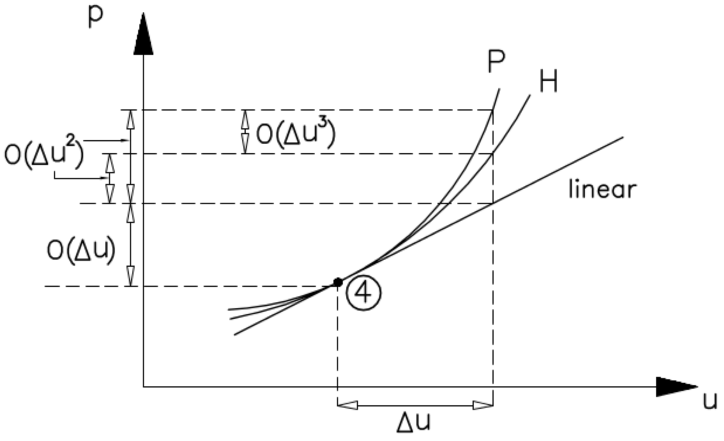

Figure 8: *Example of classical Hugoniot and Poisson curves [3]*

3.6 Convergence

When constructing numerical schemes for nonconservative systems, several sources of approximation error exist:

- Integration paths that differ from the exact paths[7]
- Stabilizing terms depending on the grid size (numerical diffusion)
- Solution components not captured by the finite approximation space

[7] Exact paths are the paths from which the weak solution is defined, i.e. poisson curves and viscous shock profiles

So one might expect (from the Lax equivalence theorem [24]) that a stable numerical scheme based on the exact paths converges to the proper weak solution as the grid is refined. This is not the case however for discontinuous solutions, as shown both theoretically and numerically in [6, 12]. This fact is due to the presence of numerical diffusion which is not included in the viscous profiles. There are several approaches to reduce the convergence problems

- Limiting the influence of diffusion by choosing good path approximations

- Limiting the influence of the path approximation by using the correct diffusion

- Selecting the paths and/or diffusion such that the scheme preserves steady states

Chalmers and Lorin [12] take the first approach and try to reduce the convergence error by using reversible Alouges-Merlet shock curves in combination with commuting numerical viscosity. This results in a bound for the convergence error by the third order of the shock strength. In contrast, in a recent paper [6] Fjordholm et al. take the second approach, constructing a numerical viscosity corresponding (up to the order of approximation) to the limiting physical viscosity, while using simple linear paths. This also results in stable, convergent results for small amplitude shocks, though an error bound is not provided. In addition, they show by example that the viscosity dependence of the weak solutions are stronger in some systems than in others[8], but no *a priori* estimate for this dependence has been found. The third approach of so-called well-balanced schemes attempts to eliminate the convergence errors completely, but only for equilibrium solutions (steady states). A scheme is said to be well balanced if it preserves these solutions. As shown by Parés et al. [40], this corresponds to the exact representation of contact discontinuities. This is a relatively simpler task, but it doesn't solve the convergence error for shocks (even steady ones), since their approximation is not considered by this approach. For schemes based on the complete Riemann problem (such as the Godunov or Roe schemes), well-balancing can be obtained by using a solver or Roe average matrix based upon the corresponding exact paths of the linearly degenerate field. Even with linear paths, the Roe scheme converges with order $\mathcal{O}(\Delta x^2)$ to continuous equilibrium solutions. This is not the case for schemes based on reduced or incomplete Riemann problems (Lax-Friedrichs,HLL(C)) as noted by Castro et al. [8]. In order to obtain well-balancing they introduce modifications to the schemes, removing the numerical viscosity over the linearly degenerate field. The reason that well-balanced schemes are so succesful for several nonconservative systems like the shallow water equations (with topography), is that for these systems the nonconservative product is associated to a linearly degenerate field. In [7] it is shown that the convergence error for this type of systems vanishes. However, it is not known yet whether the interacting IBL equations fall into this category.

In summary, the effects of the numerical viscosity are currently less predictable than that of the paths. No criteria have been found to evaluate the effects of (numerical) viscosity a priori. The effect of the (approximate) path on the other hand can be estimated. Therefore it seems more important to pay attention to the numerical viscosity of the scheme than to the approximate paths.

[8] a system of coupled Burgers equations appeared more susceptible than a two-layer shallow water system

3.7 Relation to entropy

As shown in previous sections, nonconservative systems require apart from the set of differential equations some additional information to guarantee uniqueness for discontinuous solutions. This is well known for conservation laws, where various entropy conditions provide this information (see section 2.5). Essentially there is no possibility to apply an entropy condition to a nonconservative system *after* choosing the paths, since the paths completely determine the production or destruction of entropy. So either the path must be chosen taking into account the entropy behaviour, or the entropy behaviour follows from the chosen path (plus the numerical scheme). So the paths replace the entropy condition, but they contain more information. In fact they determine the production or destruction of any quantity, since from the failure of the conservative property (23), the difference between two paths results in a source term

$$\int_{\boldsymbol{u}_L}^{\boldsymbol{u}_R} \mathbf{g}(\boldsymbol{\psi}_1)\mathrm{d}\boldsymbol{\psi}_1 - \int_{\boldsymbol{u}_L}^{\boldsymbol{u}_R} \mathbf{g}(\boldsymbol{\psi}_2)\mathrm{d}\boldsymbol{\psi}_2 = S(\boldsymbol{u}_L, \boldsymbol{u}_R) \neq \mathbf{0}. \tag{47}$$

The entropy equality (16) however can be enforced for a numerical flux by requiring

$$\oint \boldsymbol{v}^{\mathrm{T}}(\boldsymbol{\psi})\tilde{\mathbf{g}}_{\pm}(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi} = \mathbf{0}$$

This is exactly what Fjordholm et al. [6] have done by constructing a numerical scheme with a numerical diffusion that exactly cancels the source term in (47) for a given entropy $\boldsymbol{v}$. This has the advantage that the entropy behaviour becomes path-independent. On the other hand, this leads to ① shocks without entropy production and ② limited control over the production of other quantities. Since property ① is generally incompatible with (positive definite) limiting viscosities, their ESPC[9] scheme adds diffusion at shocks to produce entropy. This very ingenious scheme seems promising, as it guarantees stable solutions even for simple (linear) paths. The accuracy will still depend on property ② and the correct entropy dissipation.

Another interesting numerical scheme for the compressible Euler and shallow water equations is the recent MaMEC[10] scheme by Bas van 't Hof and Arthur Veldman [29]. It can be interpreted as an entropy-stabilized scheme (although the authors do not mention entropy). The construction is entirely different due to interpolation on staggered grids, but the derivation can be interpreted as a (mathematical) entropy transformation with

$$\boldsymbol{v} = \begin{bmatrix} \Phi + \phi - \frac{1}{2}|\boldsymbol{q}|^2 \\ \boldsymbol{q} \end{bmatrix}$$

$\boldsymbol{q}$	velocity vector
ϕ	potential of a conservative force
Φ	potential of energy transport

applied to the conservation laws with balance terms of the form (24). The result is used to enforce the discrete energy conservation. They show that a necessary prerequisite for this transformation is that the mass and momentum advection operators can be linearly combined into an antisym-

[9] ESPC: entropy-stable path-consistent

[10] MaMEC: Mass Momentum and Energy Conserving

metric form (both in continuous and discrete sense), which is necessary to obtain an additional divergence equation like (16). This requirement can be applied as well in the nonconservative case: it holds for the IBL equations. The requirement is not sufficient however, since the viscous effects in the IBL equations cause energy dissipation. To take this into account, such energy conserving schemes should be modified by adding the required dissipation. The derivation of the IBL energy equation is based on the ideas presented in [29].

As a final note, the results of section 3.4 point to another advantage of an entropy formulation: shock curves approximate better to viscous profiles for symmetrized systems, in other words, by a change to entropy variables. This only holds if the viscosity commutes with advection in the entropy variables, since this commutativity is not preserved by a change of variables.

4 Numerical schemes

For conservative systems many (approximate) schemes have been developed and applied (see e.g. [46]). A subcategory thereof which is based on flux difference splitting can essentially be modified to the nonconservative case. The approaches discussed previously in section 3.6 all fall into this category. The resulting schemes form the nonconservative counterpart of several well-known conservative schemes based on (approximate) Riemann problems, and are discussed in this section. Since these schemes are applied in the literature mostly in the context of finite volume methods, their incorporation into the discontinuous Galerkin-framework is treated here as well since it differs from the conservative formulation.

4.1 Approximate Riemann solvers

The construction of a discontinuous Galerkin formulation hinges on the solution of the Riemann problem. For conservation laws, this provides the numerical flux at the cell interfaces. Additionally, nonconservative problems require a numerical nonconservative product to be provided, which depends on the path-definition. For a discontinuous Galerkin formulation the test function $\boldsymbol{\phi}$ is also part of the nonconservative product, hence a 'numerical' path is naturally involved in the weighted residual at the faces. Consider the contribution of the nonconservative product to this residual $\mathcal{F}$:

$$\mathcal{F}(\boldsymbol{\phi}, \boldsymbol{\psi}) = \int_{\mathcal{S}} \int_0^1 \boldsymbol{\phi}(s)\ \mathbf{g}(\boldsymbol{\psi}(s)) \frac{\partial \boldsymbol{\psi}(s)}{\partial s} \mathrm{d}s \cdot \boldsymbol{n} \mathrm{d}\mathcal{S} \tag{48}$$

The Riemann path $\boldsymbol{\psi}(s)$ and the numerical path $\boldsymbol{\phi}(s)$ have been parametrized arbitrarily by a parameter $s \in [0, 1]$. Each test function requires a numerical interpolation between the left and right representations, of which at least one is identically zero in the discontinuous Galerkin approach (for the definition of the test functions, refer to section 4.4). Therefore each face needs two sets of numerical paths:

$$\begin{aligned} \boldsymbol{\phi}_-(0) &= \boldsymbol{\phi}_L & \qquad \boldsymbol{\phi}_+(0) &= \mathbf{0} \\ \boldsymbol{\phi}_-(1) &= \mathbf{0} & \qquad \boldsymbol{\phi}_+(1) &= \boldsymbol{\phi}_R \end{aligned}$$

This leads to a natural splitting of the nonconservative product into $\mathcal{F}_- = \mathcal{F}(\boldsymbol{\phi}_-, \boldsymbol{\psi})$ and $\mathcal{F}_+ = \mathcal{F}(\boldsymbol{\phi}_+, \boldsymbol{\psi})$. Figure 9 shows an example of the paths of the solution and test function for a Riemann problem corresponding to the wave pattern in figure 3 (left shock, contact, right expansion). The cell interface is indicated by the dashed line.

Schemes which are consistent with the paths used to define the weak solutions, are called path-consistent. In other words, adding the split residuals should produce the total residual:

$$\mathcal{F}_-(\boldsymbol{\phi}, \boldsymbol{\psi}) + \mathcal{F}_+(\boldsymbol{\phi}, \boldsymbol{\psi}) = \int_{\mathcal{S}} \boldsymbol{\phi} \cdot \left[\mathbf{g}(\boldsymbol{u}) : \nabla \boldsymbol{u}\right]_{\psi} \cdot \boldsymbol{n} \mathrm{d}\mathcal{S}, \quad \forall \boldsymbol{\phi} \in (L^2(\mathcal{S}))^n \tag{49}$$

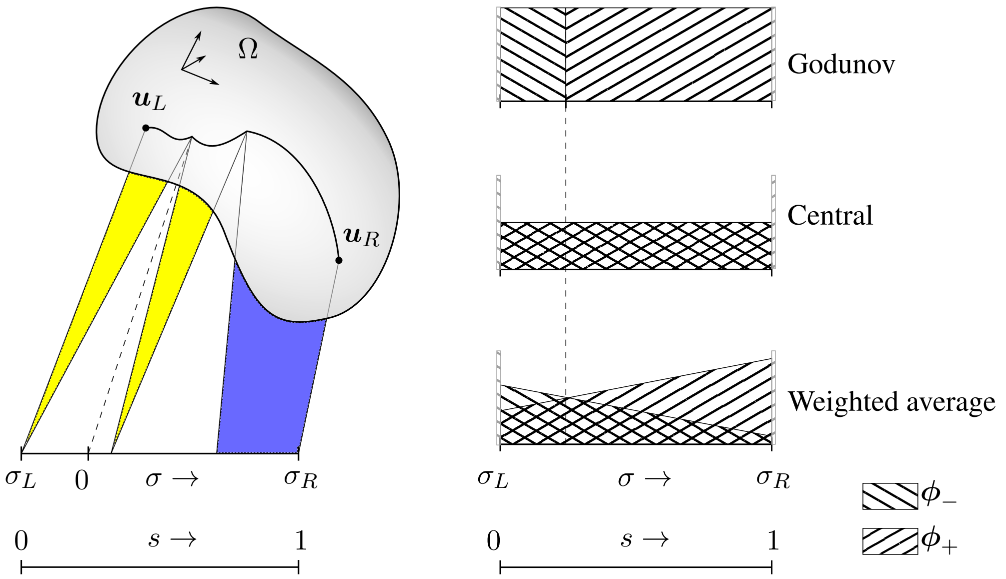

Figure 9: *Mapping of the Riemann and numerical paths to the wave propagation velocity* σ

This is an extension to DG-methods of the definition of path-consistent/path-conservative schemes given in e.g. [6, 7]. Interpreted as a condition on the numerical paths, it states that the resulting stabilization terms are conservative, i.e. they redistribute the residual between $\mathcal{F}_-$ and $\mathcal{F}_+$. The choice of numerical paths used for the splitting will lead to different type of stabilization schemes. Since in general the paths of the test function are different from the (Riemann) solution paths, the discontinuous Galerkin method can be seen as a specific type of Petrov-Galerkin discretization. Path-consistent schemes that fall into this framework are automatically well-balanced for the chosen solution path. Whether the chosen solution path is (an approximation to) the physically relevant one, is a separate issue.

The resulting range of path-consistent nonconservative Riemann solvers can be divided into three main classes:

- Complete Riemann solvers (Godunov, Roe, Osher-Solomon, ECPC)
- Partial Riemann solvers (HLL,HLLC, Rusanov)
- Non-characteristic solvers (Central, Lax-Friedrichs)

The complete solvers are based on the full set of (approximate) characteristics, while the partial variant is limited to a subset. All of these schemes have a conservative variant as well. For the nonconservative variants, convergence depends not only on the path-consistency but on the type of numerical diffusion as well. In addition, the well-balancedness and entropy-satisfying properties of the schemes (see section 3.6) depend on the choice of Riemann path $\psi(s)$ and the numerical paths $\phi(s)$. In the following sections, several schemes are put into this framework and their properties are summarised. Note that these are all first-order solvers, while the framework admits higher order schemes (such as the weighted average flux in figure 9) as well. These will not be considered here however, because in case of smooth solutions, the order of the discontinuous

Galerkin method *is not determined or limited* by the order of the Riemann solvers [28], see also the results in section 5.

4.1.1 Central scheme

The central scheme is obtained for the choice of test functions

$$\boldsymbol{\phi}_{-}(s) = \boldsymbol{\phi}_L/2, \qquad\qquad \boldsymbol{\phi}_{+}(s) = \boldsymbol{\phi}_R/2. \tag{50}$$

The Riemann path can be chosen freely. For hyperbolic equations, this generally provides an unstable method when combined with explicit time integration. The HLLC scheme proposed in [42] falls under this category. An additional diffusive term is needed in order to obtain a stable method, which may have consequences for the convergence and well-balancedness of the resulting scheme. For a conservation law the purely central scheme reduces to

$$\boldsymbol{f}^{\text{central}} = \frac{1}{2}\left(\boldsymbol{f}_L + \boldsymbol{f}_R\right) \tag{51}$$

4.1.2 Godunov's scheme

When the Riemann path equals the exact solution of the Riemann problem for the nonconservative system (assuming the exact family of paths as defined in section 3.2 is known), the exact Godunov flux can be constructed by evaluating the average of this solution at the new time level (see [35]). This is equivalent to choosing the numerical paths equal to

$$\boldsymbol{\phi}_{-}(s) = \begin{cases} \boldsymbol{\phi}_L, & \sigma(s) < 0 \\ 0, & \sigma(s) > 0 \end{cases} \qquad\qquad \boldsymbol{\phi}_{+}(s) = \begin{cases} 0, & \sigma(s) < 0 \\ \boldsymbol{\phi}_R, & \sigma(s) > 0 \end{cases} \tag{52}$$

This means that the test functions contain a jump exactly at the intercell state. For conservation laws, this is consistent with the conservative Godunov scheme as the numerical flux is then evaluated exactly at this state (due to the integration by parts).

Generally, it is expensive (or not even possible) to obtain exact Riemann solutions for nonconservative systems (by considering the viscous limit for example). Therefore this scheme is considered unfeasible for the present application.

4.1.3 Roe's scheme

When a linearization $\tilde{\mathbf{g}}(\boldsymbol{u}_L, \boldsymbol{u}_R)$ of the balance tensor is made, satisfying the following Roe properties [40]

- Hyperbolic: $\tilde{\mathbf{g}}$ has n real, distinct eigenvalues
- Consistent: $\tilde{\mathbf{g}}(\boldsymbol{u}, \boldsymbol{u}) = \mathbf{g}(\boldsymbol{u})$
- Path-conservative: $\tilde{\mathbf{g}}\, [\![\boldsymbol{u}]\!] = [\mathbf{g}(\boldsymbol{u}) : \nabla \boldsymbol{u}]_{\psi}$

then a Roe scheme can be obtained as follows: Replace $\mathbf{g}$ by its linearization $\tilde{\mathbf{g}}$, so that the Riemann problem becomes linear. Hence the path consists of straight segments along the eigenvectors of $\tilde{\mathbf{g}}$; the numerical paths are the same as for the Godunov scheme (52). The resulting Roe scheme can be rewritten without involving path-integrals by using an equivalent formulation based on a linear Riemann path combined with the numerical 'paths'

$$\begin{cases} \boldsymbol{\phi}_-{}^{\mathrm{T}} & = \boldsymbol{\phi}_L{}^{\mathrm{T}} \tilde{\mathbf{R}} \mathbf{I}_{\tilde{\lambda}<0} \tilde{\mathbf{R}}^{-1} \qquad \text{with } \mathbf{I}_{\tilde{\lambda}<0} = \tilde{\mathbf{\Lambda}}_- \tilde{\mathbf{\Lambda}}^{\dagger} \\ \boldsymbol{\phi}_+{}^{\mathrm{T}} & = \boldsymbol{\phi}_R{}^{\mathrm{T}} \tilde{\mathbf{R}} \mathbf{I}_{\tilde{\lambda}>0} \tilde{\mathbf{R}}^{-1} \qquad \text{with } \mathbf{I}_{\tilde{\lambda}>0} = \tilde{\mathbf{\Lambda}}_+ \tilde{\mathbf{\Lambda}}^{\dagger} \end{cases}, \tag{53}$$

i.e. the projections onto the eigenspace for the positive resp. negative eigenvalues of $\tilde{\mathbf{g}}$. Here $\tilde{\mathbf{R}}$ is a matrix containing the right eigenvectors and $\tilde{\mathbf{\Lambda}}$ is a diagonal matrix containing the eigenvalues ($\tilde{\mathbf{\Lambda}}_\pm$ the positive resp. negative part). This leads to the convenient formulation $\mathcal{F}_- = \boldsymbol{\phi}_L \tilde{\mathbf{g}}^{\text{roe}}_- [\![\boldsymbol{u}]\!]$ and $\mathcal{F}_+ = \boldsymbol{\phi}_R \tilde{\mathbf{g}}^{\text{roe}}_+ [\![\boldsymbol{u}]\!]$. Since the linearization depends on the family of paths $\boldsymbol{\psi}$, this method comes with the same difficulties as the Godunov scheme, and additionally an entropy fix may be required. However, if the linearization is available analytically, the computational effort is substantially reduced. For a conservation law, $\tilde{\mathbf{g}} \approx \Delta \mathbf{f} / \Delta \boldsymbol{u}$ and the Roe flux becomes

$$\boldsymbol{f}^{\text{roe}} \approx \frac{1}{2}(\boldsymbol{f}_L + \boldsymbol{f}_R) + \frac{1}{2}(\tilde{\mathbf{g}}^{\text{roe}}_+ - \tilde{\mathbf{g}}^{\text{roe}}_-)(\boldsymbol{u}_L - \boldsymbol{u}_R) \tag{54}$$

4.1.4 Rusanov's scheme

The Rusanov approach falls slightly outside the framework of weighted residuals (equation (48)), because its artificial diffusion can not be obtained by modifying the test function. Instead, its artificial diffusion is directly added to the balance tensor

$$\begin{cases} \mathbf{g}^{\text{rus}}_- & = \tfrac{1}{2}\Big(\mathbf{g} - \alpha \mathbf{I}\Big) \\ \mathbf{g}^{\text{rus}}_+ & = \tfrac{1}{2}\Big(\mathbf{g} + \alpha \mathbf{I}\Big) \end{cases}$$

The only possible numerical path which results in a path-consistent scheme, is then the central path (50). The Lax-Friedrichs scheme is a special case, obtained for $\alpha = \Delta x / \Delta t$, which makes the scheme independent of local characteristic information (subject to a CFL limit). Some properties of the Rusanov scheme are discussed in [8]: it is stable and entropy-satisfying for $\alpha > |\lambda_{\text{max}}|$. The article also develops a slightly modified variant of a Rusanov scheme which does fall into the current framework. This scheme modifies the identity matrix in the artificial diffusion term such that its null space coincides with that of $\mathbf{g}$. This modified scheme corresponds to the numerical path

$$\begin{cases} \boldsymbol{\phi}_-{}^{\mathrm{T}} & = \tfrac{1}{2}\boldsymbol{\phi}_L{}^{\mathrm{T}}\Big(\mathbf{I} - \alpha \mathbf{D}(\boldsymbol{\psi})\Big) \\ \boldsymbol{\phi}_+{}^{\mathrm{T}} & = \tfrac{1}{2}\boldsymbol{\phi}_R{}^{\mathrm{T}}\Big(\mathbf{I} + \alpha \mathbf{D}(\boldsymbol{\psi})\Big) \end{cases} \qquad \text{with } \mathbf{D} = \mathbf{R}(\boldsymbol{\psi})\mathbf{\Lambda}^{\dagger}(\boldsymbol{\psi})\mathbf{R}^{-1}(\boldsymbol{\psi}) \tag{55}$$

This modified scheme is well-balanced and linearly stable, however the entropy-satisfying property is lost [8]. For conservation laws, the scheme reduces to

$$\boldsymbol{f}^{\text{rus}} = \frac{1}{2}\left(\boldsymbol{f}_L + \boldsymbol{f}_R\right) + \frac{1}{2}\alpha\left(\boldsymbol{u}_L - \boldsymbol{u}_R\right) \tag{56}$$

4.1.5 Osher's scheme

For nonconservative systems, this scheme is closely related to the Roe scheme. The splitting is based on the local characteristics along the path, instead of the average characteristics of the linearized problem.

$$\begin{cases} \boldsymbol{\phi}_-{}^{\mathrm{T}} & = \boldsymbol{\phi}_L{}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\psi})\mathbf{I}_{\lambda<0}\mathbf{R}^{-1}(\boldsymbol{\psi}) \\ \boldsymbol{\phi}_+{}^{\mathrm{T}} & = \boldsymbol{\phi}_R{}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\psi})\mathbf{I}_{\lambda>0}\mathbf{R}^{-1}(\boldsymbol{\psi}) \end{cases}, \tag{57}$$

Osher's scheme is entropy satisfying and fully nonlinear [20], comparable to Godunov's scheme. It requires full knowledge of the eigenstructure of the balance tensor, preferably in analytical form. For a linear Riemann path, the scheme simplifies considerably:

$$\begin{cases} \mathcal{F}_- & = \boldsymbol{\phi}_L \cdot \tilde{\mathbf{g}}_-^{\text{osher}} \cdot [\![\boldsymbol{u}]\!] \\ \mathcal{F}_+ & = \boldsymbol{\phi}_R \cdot \tilde{\mathbf{g}}_+^{\text{osher}} \cdot [\![\boldsymbol{u}]\!] \end{cases} \quad \text{with } \tilde{\mathbf{g}}_\pm^{\text{osher}} = \int_0^1 \mathbf{g}_\pm(\boldsymbol{\psi}(s))\mathrm{d}s \tag{58}$$

4.1.6 ECPC scheme

Finally, the entropy-conserving path-consistent ECPC scheme [6] (discussed in section 3.7) is a fully nonlinear and entropy-consistent scheme by design. It does not require any knowledge of the eigenstructure at all. Instead, the entropy variables $\boldsymbol{v}^{\mathrm{T}}$ need to be known, i.e. the transformation that makes the system symmetric (see section 2.5). Then for any Riemann path, a possible ECPC scheme is obtained by the numerical path

$$\begin{cases} \boldsymbol{\phi}_-{}^{\mathrm{T}} & = \frac{1}{2}\boldsymbol{\phi}_L{}^{\mathrm{T}}\Big(\mathbf{I} - \mathbf{V}(\boldsymbol{\psi})\Big) \\ \boldsymbol{\phi}_+{}^{\mathrm{T}} & = \frac{1}{2}\boldsymbol{\phi}_R{}^{\mathrm{T}}\Big(\mathbf{I} + \mathbf{V}(\boldsymbol{\psi})\Big) \end{cases} \quad \text{with } \mathbf{V}(\boldsymbol{\psi}) = \frac{2}{\|[\![\boldsymbol{v}]\!]\|^2}[\![\boldsymbol{v}]\!]\left(\boldsymbol{v}(\boldsymbol{\psi}) - \langle\!\langle \boldsymbol{v}\rangle\!\rangle\right)^{\mathrm{T}} \tag{59}$$

Like Osher's scheme, this simplifies in case of a linear path to

$$\begin{cases} \mathcal{F}_- & = \boldsymbol{\phi}_L \cdot \tilde{\mathbf{g}}_-^{\text{ecpc}} \cdot [\![\boldsymbol{u}]\!] \\ \mathcal{F}_+ & = \boldsymbol{\phi}_R \cdot \tilde{\mathbf{g}}_+^{\text{ecpc}} \cdot [\![\boldsymbol{u}]\!] \end{cases} \quad \text{with } \tilde{\mathbf{g}}_\pm^{\text{ecpc}} = \frac{1}{2}\int_0^1 \Big(\mathbf{I} \pm \mathbf{V}(\boldsymbol{\psi})\Big)\mathbf{g}(\boldsymbol{\psi}(s))\mathrm{d}s \tag{60}$$

4.2 Mesh definition

To discretize the system, the spatial domain $\mathcal{V}$ is partitioned into elements $K_k \in \mathcal{T}_h$, interior faces $S_n^i \in \mathcal{S}_h^i$ and other entities $Q \in \mathcal{Q}_h$. The subscript h is a parameter indicating the maximum cell diameter in the mesh.

$$\mathcal{T}_h : \left\{ K_k \subset \mathcal{V} : \begin{array}{l} K_i \cap K_j = \varnothing \text{ when } i \neq j \\ K_k \text{ is open and } \cup_k \bar{K}_k = \bar{\mathcal{V}} \end{array} \right\}$$

$$\mathcal{S}_h^i : \left\{ S_n \subset \mathcal{V} : \begin{array}{l} \text{There are exactly two elements } K_i, K_j \text{ of } \mathcal{T}_h \\ \text{such that } S_n = \bar{K}_i \cap \bar{K}_j \cap \mathcal{V} \end{array} \right\}$$

$$\mathcal{Q}_h : \left\{ Q \subset \mathcal{V} : \begin{array}{l} \text{There are more than two elements of } \mathcal{T}_h \text{ such that} \\ \text{the intersection of their boundaries and } \mathcal{V} \text{ equals } Q \end{array} \right\}$$

Each of these categories contain a finite number of members or entities. The topological dimension of the entities is equal to the space dimension d for the elements, $d-1$ for the faces and at most $d-2$ for the other entities. An example for a two dimensional open region is shown in figure 10, where the elements are colored green, the faces blue and the other entities red. The boundary is not covered by any of the entities and is shown dashed. The last figure shows a reference cell $\widehat{K}$, corresponding to a piece of the mesh. All entities comprising such a cell are shown in exploded view in figure 11 where the open/closed character of each entity becomes clear. In two dimensions, $\mathcal{Q}_h$ contains only points whereas in three dimensions it contains points and segments.

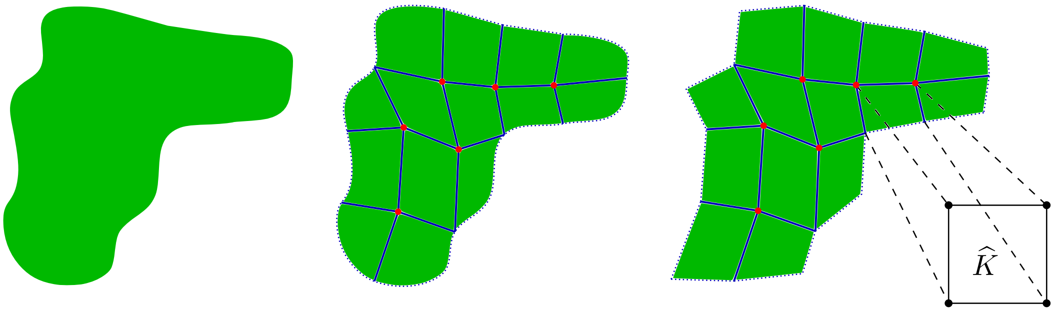

Figure 10: *Making a region (left) into a partition (center) into a mesh (right)*

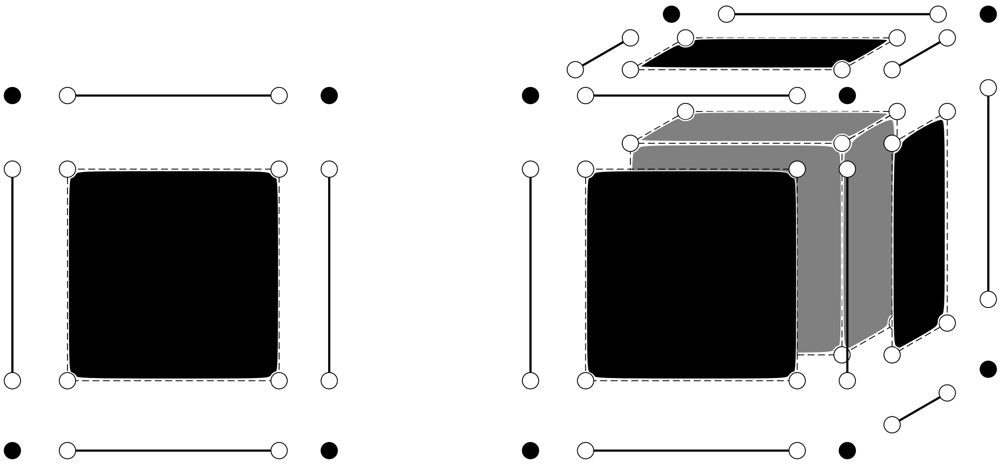

Figure 11: *Exploded view of a 2d square and a 3d cube cell into elements, faces and other entities*

Now that the grid is defined, the solution to the system can be approximated on the broken Sobolev space over $\mathcal{V}$ and $\mathcal{T}_h$, defined as [28]

$$H^1_h : \left\{\phi \in L^2(\mathcal{V}) : \phi|_K \in H^1(K),\ \text{for all } K \in \mathcal{T}_h\right\} \tag{61}$$

Note that the boundary $\partial\mathcal{V}$ is not yet included in any of the entity sets, because its discretization is independent of the discretization of the interior: it can be split arbitrarily into multiple exterior faces S^b_n that together form the boundary. It makes sense to choose the exterior faces equal to the element faces at the boundary, or to consider simple regions of the temporal or spatial boundaries as a single face. The first option will be used here, which results in

$$\mathcal{S}^b_h : \left\{ S_n \subset \partial\mathcal{V} : \begin{array}{l} \text{There is exactly one element } K_k \text{ of } \mathcal{T}_h \\ \text{such that } S_n \cap \bar{K}_k \neq \varnothing \text{ and } \cup_n \bar{S}_n = \partial\mathcal{V} \end{array} \right\}.$$

The set of all faces $\mathcal{S}_h$ is defined as the union of the interior faces $\mathcal{S}^i_h$ and boundary faces $\mathcal{S}^b_h$.

4.3 Path-conservative discontinuous Galerkin method

We will begin by applying the definition of the weak solutions to the domain. The weak form for general nonconservative systems (25) is repeated here,

$$\oint_{\partial\underline{\mathcal{V}}} \underline{\mathbf{f}} \cdot \underline{\boldsymbol{n}} \cdot \boldsymbol{\phi} \mathrm{d}\underline{\mathcal{S}} - \int_{\underline{\mathcal{V}}} \underline{\mathbf{f}}(\boldsymbol{u}) : \underline{\nabla}\boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}} \\ + \int_{\underline{\mathcal{V}}} \underline{\mathbf{g}}(\boldsymbol{u}) : \underline{\nabla}\boldsymbol{u} \cdot \boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}} = \int_{\underline{\mathcal{V}}} \boldsymbol{s}(\boldsymbol{u}) \cdot \boldsymbol{\phi}\mathrm{d}\underline{\mathcal{V}}, \qquad \forall\boldsymbol{\phi} \in (H^1_h(\underline{\mathcal{V}}))^n.$$

The volume integrals need to be split into a sum of integrals over the partitions of $\underline{\mathcal{V}}$ defined by the mesh. The allowed solution space is then used to limit the terms that contributes to each entity. We require that the source term $\boldsymbol{s}(\boldsymbol{u})$ is in $L^2(\underline{\mathcal{V}})$, so that its contribution to the faces and other entities vanishes (since these are of zero measure in $\underline{\mathcal{V}}$). Because $\boldsymbol{\phi}$ and $\boldsymbol{u}$ are in H^1_h, they are allowed a finite jump over a face ($\mathcal{S}_h$) or other entity ($\mathcal{Q}_h$). This means their gradient behaves as a delta distribution times the jump. The integral of this distribution has a finite value over sets of dimension $d - 1$, hence only contributes for the faces. For the conservative flux $\mathbf{f}(\boldsymbol{u})$ at the faces, we choose a numerical flux $\mathbf{h}(\boldsymbol{u}^L, \boldsymbol{u}^R)$ that should approximate the flux for the Riemann problem. The nonconservative product at the faces is split into numerical contibutions to both elements $\tilde{\mathbf{g}}_\pm(\boldsymbol{u}^L, \boldsymbol{u}^R)\, [\![\boldsymbol{u}]\!]$, depending on a path-consistent approximation of the Riemann problem. This results in the following formulation

$$\begin{aligned}\oint_{\partial\mathcal{V}} \mathbf{f}_{li}(\boldsymbol{u})\boldsymbol{n}_l\boldsymbol{\phi}_i \mathrm{d}\mathcal{S} + \sum_{K_k\in\mathcal{T}_h}\int_{K_k} &\mathbf{g}_{lmi}(\boldsymbol{u})\boldsymbol{u}_{m,l}\boldsymbol{\phi}_i - \mathbf{f}_{li}(\boldsymbol{u})\boldsymbol{\phi}_{i,l} - \boldsymbol{s}_i(\boldsymbol{u})\boldsymbol{\phi}_i\mathrm{d}\mathcal{V} \\ + \sum_{S_n\in\mathcal{S}_h^i}\int_{S_n} &\tilde{\mathbf{g}}_{\pm lmi}\left\langle\boldsymbol{\phi}_i\right\rangle_\pm [\![\boldsymbol{u}_m]\!]_{nl} - \mathbf{h}_{li}\,[\![\boldsymbol{\phi}_i]\!]_{nl}\,\mathrm{d}\mathcal{S} = 0, \\ &\forall\boldsymbol{\phi}\in H_h^1(\mathcal{V}). \end{aligned} \tag{62}$$

Indicial notation has been used to distinguish the tensor products. Index i and m corresponds to the i-th (resp. m-th) solution variables, and index l corresponds to the direction in space-time. Subscript k is the element number and n the face number. The Einstein summation convention is used except for subscripts that are explicitly under a summation, and differentiation is indicated by the comma notation. For an overview of all indices, refer to table 1. If the assumption $\langle\boldsymbol{\phi}\rangle_\pm = \langle\!\langle\boldsymbol{\phi}\rangle\!\rangle$ is added, this result equals that of Rhebergen et al. [42], corresponding to a central scheme for the nonconservative product. Here however, no choice has been made on the numerical value of the test functions at the faces, since this depends on the chosen Riemann solver. Hence the formulation (62) allows for non-central approximations of the nonconservative product.

4.4 Approximation space and reference element

All indices used (except k and n) should be read as components, not as numbers. They can only be assigned a number if we define a basis for $\mathcal{V}$, Ω and $\mathcal{L}(\mathcal{V},\Omega)$ respectively. Therefore we approximate the solution in a finite dimensional subspace V_h of H_h^1, consisting of piecewise polynomials of degree p.

$$V_h : \left\{ b:\mathcal{V}\to\mathbb{R} \;\middle|\; \begin{array}{l} b|_K \in P^p(K), \text{ for all } K\in\mathcal{T}_h \\ b|_S \in P^p(S), \text{ for all } S\in\mathcal{S}_h \\ b|_Q = 0 \text{ for all } Q\in\mathcal{Q}_h \end{array} \right\} \tag{63}$$

A basis for V_h can be constructed by combining the polynomial bases on each separate part of the mesh (elements and faces). In order to do so, these parts are mapped to a reference cell that contains exactly one element and its faces, where the basis functions will be defined. The reference cell $\widehat{K}\subset\mathbb{R}^d$ is chosen here as a simple square (or higher dimensional equivalent). This will ease the evaluation of transformations and integrals considerably. Extension to simplices is possible by applying the techniques in [27]. The coordinates on the reference cell are $\boldsymbol{\xi}\in[-1,1]^d$. The transformation from the reference to the physical space and its inverse are given by

$$\begin{aligned} &\boldsymbol{x}_k : \widehat{K}\mapsto K_k && \boldsymbol{\xi}_k : K_k\mapsto\widehat{K} \\ &\boldsymbol{x}_k(\boldsymbol{\xi}) = \boldsymbol{x}_{k0} + \mathbf{J}_k\boldsymbol{\xi} && \boldsymbol{\xi}_k(\boldsymbol{x}) = \mathbf{J}_k^{-1}(\boldsymbol{x}-\boldsymbol{x}_{k0}) \end{aligned} \tag{64}$$

For simplicity, the Jacobians $\mathbf{J}_k$ are constant per element.

The number of linear independent basis functions required to span the space $P^p(\mathbb{R}^d)$ is

$$M(p,d) = \frac{(p+d)!}{p!d!},$$

(a simplified version of the expression in [37]) which differs between the element and the faces. In general, a basis on $\mathcal{T}_h$ can't simply be interpolated to define a basis on $\mathcal{S}_h$ and vice versa. Therefore, restrictions of the approximation space V_h are defined as V_{he} on the element interiors and V_{hf} on the faces. The basis functions for these restricted spaces are defined on the reference cell.

In the element interior, the basis for $\widehat{V}_{he} = P^p(\widehat{K})$ will consist of the basis polynomials b_j on $\widehat{K}$, for which it holds that $\mathrm{span}\{b_0,\ldots,b_M\} = \widehat{V}_{he}$. This can be simple monomials, or orthogonal polynomials such as Chebyshev or Legendre polynomials. Legendre polynomials will make the flux computations more efficient, whereas monomials make the trace operations more efficient. The end result will be independent of the choice of basis, except for numerical roundoff error in the projection step. Admittedly, in some situations the projection error can grow very large, see A.1. After consideration, a monomial basis (in computational coordinates) is chosen for its reduced complexity of implementation.

$$b_j^* = \xi_0^{\alpha_{0j}}\xi_1^{\alpha_{1j}}\ldots\xi_d^{\alpha_{dj}} = \boldsymbol{\xi}^{\boldsymbol{\alpha}_j} \qquad\qquad b_j = \begin{cases} b_j^*, & \xi \in \widehat{K} \\ 0, & \xi \notin \widehat{K} \end{cases} \tag{65}$$

where $\{\boldsymbol{\alpha}_j\}_j$ is a sequence of vectors that determine the order of the monomials. Using the transformation back to the global coordinates leads to a representation of the solution $\boldsymbol{u} \in (V_h)^n$ in terms of the expansion coeffients $\widehat{u}_{ijk}$:

$$\boldsymbol{u}_k = \widehat{u}_{ijk}\boldsymbol{b}_{ijk}. \tag{66}$$

The vectorial basis functions $\boldsymbol{b}_{ijk}$ of $(V_h)^n$ are defined as $\boldsymbol{b}_{ijk} = \boldsymbol{c}_i b_j(\boldsymbol{\xi}_k(\boldsymbol{x}))$, where $\boldsymbol{c}_i$ are orthogonal basis vectors in Ω, each corresponding to one of the independent variables in $\boldsymbol{u}$.

The quantities that 'reside' on the faces such as boundary conditions, numerical flux terms and trace quantities require a unique representation in terms of the approximation space on the faces, V_{hf}. Again, a basis for this space will be defined by mapping the faces S_n to the boundaries of the reference element $\widehat{K}$. In this case, the reference element has two faces normal to each coordinate direction l, that will share the same basis. Defining the basis functions as monomials $\bar{b}_{l\bar{\jmath}}$ for each pair of faces pair of the reference element:

$$\bar{b}_{l\bar{\jmath}}^* = \xi_0^{\beta_{0\bar{\jmath}}}\xi_1^{\beta_{1\bar{\jmath}}}\ldots\xi_d^{\beta_{d\bar{\jmath}}} = \boldsymbol{\xi}^{\boldsymbol{\beta}_{\bar{\jmath}}} \qquad\qquad \bar{b}_{l\bar{\jmath}} = \begin{cases} \bar{b}_{l\bar{\jmath}}^*, & \xi \in \partial\widehat{K} \\ 0, & \xi \notin \partial\widehat{K} \end{cases} \tag{67}$$

where $\{\boldsymbol{\beta}_{l\bar{\jmath}}\}_{\bar{\jmath}}$ is a renumbered subsequence of $\{\boldsymbol{\alpha}_j\}_j$ containing only the $M(p,d-1)$ elements $\{\boldsymbol{\alpha}_j | \alpha_{lj} = 0\}$ (i.e. the polynomials are constant in direction l). This basis for the reference faces $\widehat{V}_{hf}$ can be turned into a global basis for V_{hf} by mapping each actual face S_n to a face of the reference element. Such a map $\widehat{S}$ is found by choosing one of the two adjacent elements as a master element, and applying its coordinate transformation $\boldsymbol{\xi}_k(\boldsymbol{x})$. The resulting map is therefore paremetrized as $\widehat{S} : n \mapsto (k_n, l_n)$ and produces the following global representation for the face quantities $\bar{\boldsymbol{u}}$.

$$\bar{\boldsymbol{u}}_n = \bar{\widehat{u}}_{i\bar{\jmath}n}\bar{\boldsymbol{b}}_{i\bar{\jmath}n}. \tag{68}$$

where the vectorial basis functions $\bar{\boldsymbol{b}}_{i\bar{\jmath}n} = \boldsymbol{c}_i \bar{b}_{l_n\bar{\jmath}}\left(\boldsymbol{\xi}_{k_n}(\boldsymbol{x})\right)$ are used.

Now that a basis for V_{he} and for V_{hf} have been defined, these should be combined into a basis for V_h. Because all functions in V_{hf} are equivalent to the zero function under the broken Sobolev norm $\|\cdot\|_{H^1_h}$, the basis for V_{he} is already a basis for V_h. Using this basis leads to the test functions on the faces in the Galerkin formulation being zero. For a conservative system this has no effect, but for a nonconservative system it ignores the contributions of the nonconservative products over the faces. In other words, the broken Sobolev space is perhaps not the correct setting for (nonconservative) path-consistent discontinuous Galerkin. The basis for V_h should contain a connection to V_{hf} as well: the values of the basis functions on the element K itself and on the element boundary ∂K should be linked. This is taken care of by the numerical paths (see section 4.1) and leads to a path-consistent formulation.

4.5 Connectivity and trace operations

To facilitate a clear and concise notation for the quadrature-free discontinuous Galerkin method on a general mesh, the *connectivity*, *incidence* and *trace* operators are introduced. Since these are all linear, they can be thought of as matrices. First, the trace operation is defined as the limit of a function when approaching a face from the inside of an element.

$$\mathbf{T}_{kn} : P^p(K_k) \to P^p(S_n) \qquad\qquad \mathbf{T}_{kn}(\boldsymbol{u}) \stackrel{\text{def}}{=} \lim_{\boldsymbol{x}\in K_k \to \bar{\boldsymbol{x}} \in S_n} \boldsymbol{u}(\boldsymbol{x}) \tag{69}$$

The connectivity matrices (directed $\mathbf{C}_{\pm\,kn}$ and undirected $\mathbf{C}_{kn}$) and incidence matrix $\mathbf{D}_{kn}$ are defined directly from the mesh topology

Relation between k and n	$\mathbf{C}_{kn}$	$\mathbf{C}_{-\,kn}$	$\mathbf{C}_{+\,kn}$	$\mathbf{D}_{kn}$
Element k is the master element of face n	1	1	0	-1
Element k is the slave element of face n	1	0	1	$+1$
Otherwise	0	0	0	0

Hence the incidence, connectivity and trace matrices are nonzero on at most two elements per face, and the positive flux direction is from the master to the slave element. These matrices, in combination with trace operation can be used to provide several forms of restriction from V_h to V_{hf}: the master trace $\langle\cdot\rangle_-$, slave trace $\langle\cdot\rangle_+$, mean $\langle\!\langle\cdot\rangle\!\rangle$ and jump $[\![\cdot]\!]$. Note that a jump has a magnitude and a direction, e.g. for a scalar ϕ, $[\![\phi]\!] = \phi^+\boldsymbol{n}^+ + \phi^-\boldsymbol{n}^-$.

$$\begin{aligned}
\langle\boldsymbol{u}\rangle_- &\overset{\text{def}}{=} \mathbf{C}_{-\,kn}\mathbf{T}_{kn}(\boldsymbol{u}),\\
\langle\boldsymbol{u}\rangle_+ &\overset{\text{def}}{=} \mathbf{C}_{+\,kn}\mathbf{T}_{kn}(\boldsymbol{u}),\\
\langle\!\langle\boldsymbol{u}\rangle\!\rangle &\overset{\text{def}}{=} \frac{1}{2}\mathbf{C}_{kn}\mathbf{T}_{kn}(\boldsymbol{u}),\\
[\![\boldsymbol{u}]\!] &\overset{\text{def}}{=} \mathbf{D}_{kn}\mathbf{T}_{kn}(\boldsymbol{u})\boldsymbol{n}.
\end{aligned}$$

4.6 Quadrature-free evaluation of integrals

To evaluate the discrete residuals, the basis functions are substituted into the weak formulation (62). Then each term is expressed directly in terms of the set of unknown coefficients, which results in a quadrature free formulation. This approach was developed by Atkins and Shu [2] to reduce the computational effort in evaluation of the discontinuous Galerkin schemes. The present work is derived from an application of this method to the linearized Euler equations by Özdemir [37]. The description from here on assumes that all quantities ($\mathbf{f}$, $\mathbf{h}$, $\mathbf{g}\nabla\boldsymbol{u}$ and $\boldsymbol{s}$) have been projected onto the finite element space $(V_h)^n$. For details on the procedures used to obtain these projections, see Appendix A. The weak formulation is split into the three contributions from the element interiors, $\mathcal{E}_{ijk}$, the interior faces $\mathcal{F}_{ijk}$ and the boundary faces $\mathcal{B}_{ijk}$.

$$\mathcal{E}_{ijk}(\boldsymbol{u}) = \sum_{K_k\in\mathcal{T}_h}\int_{K_k} \mathbf{g}_{lmi}(\boldsymbol{u})\boldsymbol{u}_{m,l}\boldsymbol{b}_{ijk} - \mathbf{f}_{li}(\boldsymbol{u})\boldsymbol{b}_{ijk\,l} - \boldsymbol{s}_i(\boldsymbol{u})\boldsymbol{b}_{ijk}\mathrm{d}\mathcal{V} \tag{70}$$

$$\mathcal{F}_{ijk}(\boldsymbol{u}) = \sum_{S_n\in\mathcal{S}_h^i}\int_{S_n} \tilde{\mathbf{g}}_{\pm\,lmi}\left\langle\boldsymbol{b}_{ijk}\right\rangle_\pm [\![\boldsymbol{u}_m]\!]_{nl} - \mathbf{h}_{li}\,[\![\boldsymbol{b}_{ijk}]\!]_{nl}\,\mathrm{d}\mathcal{S} \tag{71}$$

$$\mathcal{B}_{ijk}(\boldsymbol{u}) = \sum_{S_n\in\mathcal{S}_h^b}\int_{S_n} \mathbf{f}_{iln}(\boldsymbol{u})\boldsymbol{n}_{nl}\boldsymbol{b}_{ijk}\mathrm{d}\mathcal{S} \tag{72}$$

For the definition of the basis functions $\boldsymbol{b}_{ijk}$ and $\bar{\boldsymbol{b}}_{i\bar{\jmath}n}$ refer to equations (65) and (67). In the next three sections, these three integrals will be given an explicit expression. All bookkeeping indices used thereby are summarised in table 1 (indices with bars apply to the faces). Also their upper bounds are given, noting that N_{eq} and N_{dim} is the same for all elements, but $M(p,d)$ may vary between different elements k and unknowns i.

index	meaning	upper bound
i	equation index (fixed direction in Ω)	N_{eq}
$j, \bar{\jmath}$	scalar basis function index	$M(p,d)$
k	element number	N_{elem}
l	space-time unit (fixed direction in $\underline{\mathcal{V}}$)	N_{dim}
m	variable index	N_{eq}
n	face index	N_{faces}
$o, \bar{o}$	flux expansion index	$M_{\text{flux}}(p,d)$

Table 1: *Overview of all bookkeeping indices*

4.6.1 Element integral

The element integral $\mathcal{E}_{ijk}(\boldsymbol{u})$ is particularly simple to evaluate in the DG formulation, because the support of each basis function is confined to a single element. Therefore no distinction is made between the element index in the summation and the index k of the basis function. The individual terms are expanded in terms of the basis functions as

$$\begin{aligned}
\mathbf{f}_{li}(\boldsymbol{u}_k) &= \widehat{A}_{ilmko} b_o(\boldsymbol{\xi}_k) \widehat{u}_{mjk} b_j(\boldsymbol{\xi}_k) \\
\mathbf{g}_{lmi}(\boldsymbol{u}_k) \boldsymbol{u}_{km,\,l} &= \widehat{G}_{ilmko} b_o(\boldsymbol{\xi}_k) \widehat{u}_{mjk} \frac{\partial b_j(\boldsymbol{\xi}_k)}{\partial x_l} \\
\boldsymbol{s}_i(\boldsymbol{u}_k) &= \widehat{s}_{ijk} b_j(\boldsymbol{\xi}_k)
\end{aligned}$$

The transformation of derivatives from the physical space to the reference element is given by

$$\frac{\partial \cdot}{\partial x_l} = \frac{\mathrm{d}\xi_{l'}}{\mathrm{d}x_l} \frac{\partial \cdot}{\partial \xi_{l'}} = \mathbf{J}^{-T}_{ll'} \frac{\partial \cdot}{\partial \xi_{l'}}.$$

Assuming the Jacobian $\mathbf{J}$ is constant over the element, the element integral can be expressed over the reference element as

$$\begin{aligned}
\mathcal{E}_{ijk}(\boldsymbol{u}) &= \sum_{K_k \in \mathcal{T}_h} \int_{\widehat{K}} \left(\widehat{G}_{ilmko} b_o \widehat{u}_{mj'k} \mathbf{J}^{-T}_{ll'} \frac{\partial b_{j'}}{\partial \xi_{l'}} - \widehat{s}_{ij'k} b_{j'} \right) b_j - \widehat{A}_{ilmko} b_o \widehat{u}_{mj'k} b_{j'} \mathbf{J}^{-T}_{kll'} \frac{\partial b_j}{\partial \xi_{l'}} \mathbf{J}_{kl'l} \mathrm{d}\boldsymbol{\xi}_{l'} \\
&= \sum_{K_k \in \mathcal{T}_h} \mathbf{J}^{-T}_{kll'} |\mathbf{J}_k| \left(\widehat{G}_{ilmko} [\mathbf{F}_{l'o}]^{\mathrm{T}}_{jj'} - \widehat{A}_{ilmko} [\mathbf{F}_{l'o}]_{jj'} \right) \widehat{u}_{mj'k} - |\mathbf{J}_k| [\mathbf{M}_e]_{jj'} \widehat{s}_{ij'k}
\end{aligned}$$

where the element mass and flux matrices are defined as

$$[\mathbf{M}_e]_{jj'} \overset{\text{def}}{=} \int_{\widehat{K}} b_{j'} b_j \mathrm{d}\boldsymbol{\xi}, \qquad [\mathbf{F}_{lo}]_{jj'} \overset{\text{def}}{=} \int_{\widehat{K}} b_o b_{j'} \frac{\partial b_j}{\partial \xi_l} \mathrm{d}\boldsymbol{\xi} \tag{73}$$

4.6.2 Face integral

The face integral $\mathcal{F}_{ijk}(\boldsymbol{u})$ is computed with aid of the Riemann solvers. These provide the projections onto the face basis of the conservative flux $\mathbf{h}$ and the nonconservative product $\tilde{\mathbf{g}}_{\pm} [\![\boldsymbol{u}]\!]$:

$$\begin{aligned}
\tilde{\mathbf{g}}_{\pm\,lmi} [\![\boldsymbol{u}_m]\!]_{nl} &= \widehat{g}_{\pm\,i\bar{\jmath}n} \bar{\boldsymbol{b}}_{i\bar{\jmath}n} \\
\mathbf{h}_{li} \boldsymbol{n}_{nl} &= \widehat{h}_{i\bar{\jmath}n} \bar{\boldsymbol{b}}_{i\bar{\jmath}n}
\end{aligned}$$

For straight faces and constant Jacobians, the integral can be expressed in terms of the trace weighting matrix $\mathbf{P}_{kn}$ by transformation to the reference element face

$$\begin{aligned}\mathcal{F}_{ijk}(\boldsymbol{u}) &= \sum_{S_n \in \mathcal{S}_h^i} \int_{\widehat{S}_n} \left(\widehat{g}_{\pm\, i\bar{\jmath}' n} \bar{\boldsymbol{b}}_{i\bar{\jmath}' n} \left\langle \boldsymbol{b}_{ijk} \right\rangle_{\pm} - \widehat{h}_{i\bar{\jmath}' n} \bar{\boldsymbol{b}}_{i\bar{\jmath}' n} \left[\!\left[\boldsymbol{b}_{ijk} \right]\!\right]_{nl} \right) \mathbf{J}_{ll'n} \mathrm{d}\xi_{l'} \\ &= \sum_{S_n \in \mathcal{S}_h^i} |\mathbf{J}_n| |\mathbf{J}_n^{-1} \boldsymbol{n}_n| [\mathbf{P}_{kn}]_{j\bar{\jmath}'} \left(\mathbf{C}_{\pm\, kn} \widehat{g}_{\pm\, i\bar{\jmath}' n} - \mathbf{D}_{kn} \widehat{h}_{i\bar{\jmath}' n} \right)\end{aligned}$$

where the trace weighting matrix is defined by the integral over the face of the reference element for both of the connected elements

$$[\mathbf{P}_{kn}]_{j\bar{\jmath}'} \overset{\text{def}}{=} \int_{\widehat{S}_n} \bar{b}_{\bar{\jmath}'} \mathbf{T}_{kn} b_j \mathrm{d}\xi \tag{74}$$

4.6.3 Boundary integral

The boundary integral $\mathcal{B}_{ijk}(\boldsymbol{u})$, which represents the forcing by the boundary conditions, is very similar to the face integral. The difference is that the numerical flux $\mathbf{h} \cdot \boldsymbol{n}$ is replaced by the boundary condition on $\mathbf{f} \cdot \boldsymbol{n}$. Using the projection coefficients of the boundary (and initial) conditions as $\mathbf{f}_{li} \boldsymbol{n}_{nl} = \widehat{f}^b_{i\bar{\jmath}n} \bar{\boldsymbol{b}}_{i\bar{\jmath}n}$, the boundary integral becomes

$$\mathcal{B}_{ijk}(\boldsymbol{u}) = - \sum_{S_n \in \mathcal{S}_h^b} |\mathbf{J}_n| |\mathbf{J}_n^{-1} \boldsymbol{n}_n| [\mathbf{P}_{kn}]_{j\bar{\jmath}'} \mathbf{D}_{kn} \widehat{f}^b_{i\bar{\jmath}' n}.$$

4.7 Nonlinear solution

The system that needs to be solved can now be summarised as

$$\mathcal{R}_{ijk}(\boldsymbol{u}) = \mathcal{B}_{ijk}(\boldsymbol{u}) + \mathcal{E}_{ijk}(\boldsymbol{u}) + \mathcal{F}_{ijk}(\boldsymbol{u}) = 0, \quad \forall (i,j,k) \tag{75}$$

where

$$\begin{aligned}\mathcal{B}_{ijk}(\boldsymbol{u}) &= - \sum_{S_n \in \mathcal{S}_h^b} |\mathbf{J}_n| |\mathbf{J}_n^{-1} \boldsymbol{n}_n| [\mathbf{P}_{kn}]_{j\bar{\jmath}'} \mathbf{D}_{kn} \widehat{f}^b_{i\bar{\jmath}' n}, \\ \mathcal{E}_{ijk}(\boldsymbol{u}) &= \sum_{K_k \in \mathcal{T}_h} \mathbf{J}^{-T}_{kll'} |\mathbf{J}_k| \left(\widehat{G}_{ilmko} [\mathbf{F}_{l'o}]^{\mathrm{T}}_{jj'} - \widehat{A}_{ilmko} [\mathbf{F}_{l'o}]_{jj'} \right) \widehat{u}_{mj'k} - |\mathbf{J}_k| [\mathbf{M}_e]_{jj'} \widehat{s}_{ij'k} \\ \mathcal{F}_{ijk}(\boldsymbol{u}) &= \sum_{S_n \in \mathcal{S}_h^i} |\mathbf{J}_n| |\mathbf{J}_n^{-1} \boldsymbol{n}_n| [\mathbf{P}_{kn}]_{j\bar{\jmath}'} \left(\mathbf{C}_{\pm\, kn} \widehat{g}_{\pm\, i\bar{\jmath}' n} - \mathbf{D}_{kn} \widehat{h}_{i\bar{\jmath}' n} \right).\end{aligned}$$

This formulation allows to evaluate the residual $\mathcal{R}$ in a modular way, using evaluations of $\widehat{f}$, $\widehat{g}$, $\widehat{h}$, $\widehat{A}$, $\widehat{G}$ and $\widehat{s}$ in terms of the unknowns $\widehat{u}$. To solve this system by Newton iteration, its Jacobian $\mathcal{J}$ should be available as well, preferably in terms of the Jacobians of the individual components (see figure 12). It is not necessary for the Jacobian to be computed explicitly, only its action on a

Newton update $\Delta\widehat{u}$ needs to be computed.

$$\mathcal{J}^{ijk}_{i'j'k'} = \frac{\partial \mathcal{B}_{ijk}}{\partial \widehat{u}_{i'j'k'}} + \frac{\partial \mathcal{E}_{ijk}}{\partial \widehat{u}_{i'j'k'}} + \frac{\partial \mathcal{F}_{ijk}}{\partial \widehat{u}_{i'j'k'}} \tag{76}$$

This leads to the following Newton iteration for each time step:

$$\left\{ \begin{array}{ll} \Delta\widehat{u}^{t,s}_{ijk} & = \left[\mathcal{J}^{ijk}_{i'j'k'}(\widehat{u}^{t,s}_{ijk})\right]^{-1} \mathcal{R}_{i'j'k'}(\widehat{u}^{t,s}_{ijk}, \text{i.c., b.c.}) \\ \widehat{u}^{t,s+1}_{ijk} & = \widehat{u}^{t,s}_{ijk} - \Delta\widehat{u}^{t,s}_{ijk} \end{array} \right.$$

until $\|\Delta\widehat{u}^{t,s}_{ijk}\| \leq \epsilon_{\text{newton}}$, then $\widehat{u}^{t}_{ijk} = \widehat{u}^{t,s+1}_{ijk}$. The initial solution estimate can be based on the previous time step, or taken as zero (if the convergence region allows it). After convergence, the initial and boundary conditions for the next time step are determined, and an estimate for the next solution is made. The Newton method converges quadratically if the initial estimate is 'close enough' to the local solution[11]. If this is not the case, it can be modified by under-relaxation

$$\widehat{u}^{t,s+1}_{ijk} = \widehat{u}^{t,s}_{ijk} - \omega\Delta\widehat{u}^{t,s}_{ijk},$$

where ω is a fixed value between 0 and 1. A more advanced approach similar to under-relaxation is backtracking [41, sec. 9.7] where ω is determined based upon the residual reduction. It can be seen as a combination of a zero-finding algorithm for $\mathcal{R}$ and a minimization algorithm for $\|\mathcal{R}\|$.

$$\omega = \left\{ \begin{array}{ll} 1, & \text{when } \|\mathcal{R}(\widehat{u}^{s+1}_{ijk})\| \leq \|\mathcal{R}(\widehat{u}^{s}_{ijk})\| \\ \left(1 + \frac{\|\mathcal{R}(\widehat{u}^{s+1}_{ijk})\|}{\|\mathcal{R}(\widehat{u}^{s}_{ijk})\|}\right)^{-1} & \text{otherwise} \end{array} \right.$$

If the Jacobian becomes (nearly) singular while propagating the solution in time, this indicates a bifurcation into multiple possible solutions. This type of breakdown corresponds to a physical instability and is not a failure of the numerical method.

[11] more precisely, the inverse Jacobian and the second derivatives should be bounded within the convergence region

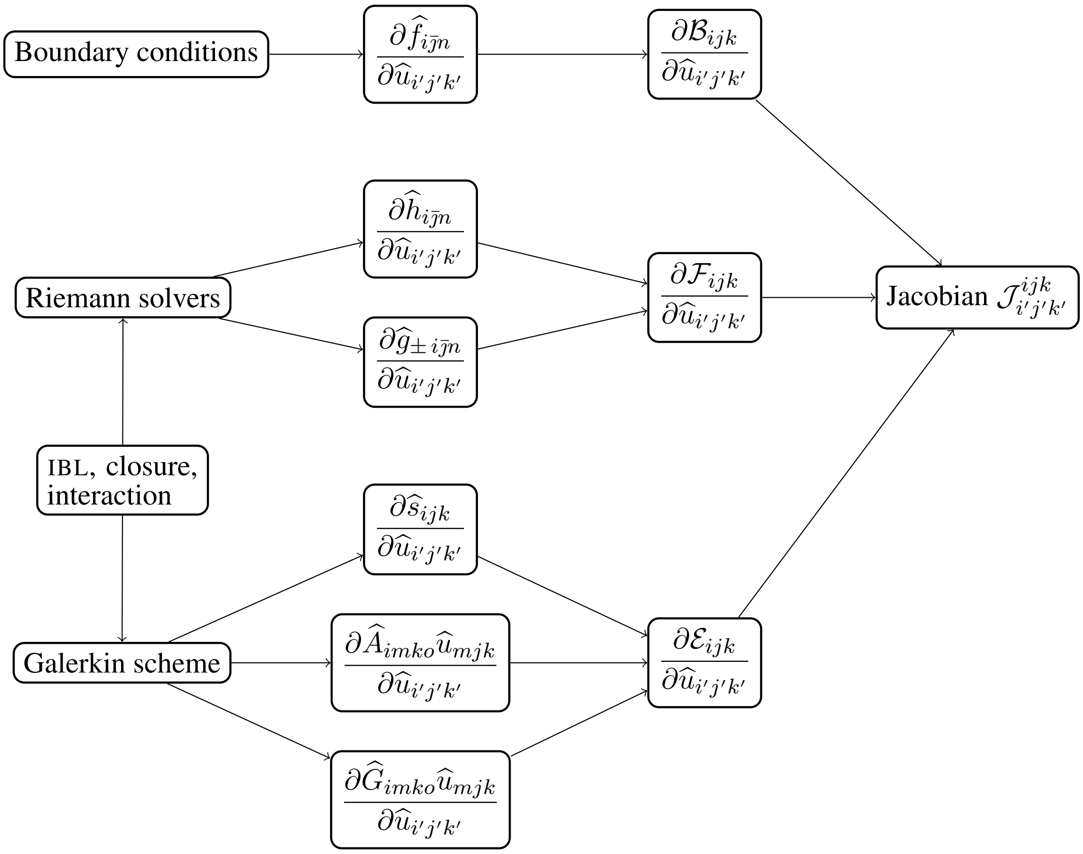
Boundary conditions
$\frac{\partial \widehat{f}_{i\bar{\jmath}n}}{\partial \widehat{u}_{i'j'k'}}$
$\frac{\partial \mathcal{B}_{ijk}}{\partial \widehat{u}_{i'j'k'}}$
$\frac{\partial \widehat{h}_{i\bar{\jmath}n}}{\partial \widehat{u}_{i'j'k'}}$
Riemann solvers
$\frac{\partial \mathcal{F}_{ijk}}{\partial \widehat{u}_{i'j'k'}}$
Jacobian $\mathcal{J}^{ijk}_{i'j'k'}$
$\frac{\partial \widehat{g}_{\pm\, i\bar{\jmath}n}}{\partial \widehat{u}_{i'j'k'}}$
IBL, closure, interaction
$\frac{\partial \widehat{s}_{ijk}}{\partial \widehat{u}_{i'j'k'}}$
Galerkin scheme
$\frac{\partial \widehat{A}_{imko} \widehat{u}_{mjk}}{\partial \widehat{u}_{i'j'k'}}$
$\frac{\partial \mathcal{E}_{ijk}}{\partial \widehat{u}_{i'j'k'}}$
$\frac{\partial \widehat{G}_{imko} \widehat{u}_{mjk}}{\partial \widehat{u}_{i'j'k'}}$

Figure 12: *Construction of the residual Jacobian in a modular implementation*

5 Linear advection problem

Some experience with the classical Runge-Kutta Discontinuous Galerkin (RKDG) method has been obtained by applying it to the linear advection equation in one dimension,

$$\frac{\partial u}{\partial t} + \frac{\partial au}{\partial x} = 0 \quad \text{on } \mathcal{V}, \qquad\qquad \mathcal{V} : (x,t) \in [0,L] \times [0,T] \tag{77}$$

with the initial and boundary conditions

$$u(0,t) = u(L,t) \qquad\qquad u(x,0) = \frac{1}{2} + \sum_{n=1}^{n_{\max}} \cos\left(n\pi(\tfrac{x}{L} - \tfrac{1}{2})\right) \tag{78}$$

The method has been implemented using the quadrature-free approach by Atkins and Shu [2] in combination with a mixed central-upwind flux scheme. A numerical simulation has been performed, propagating the solution over 10 times the domain length. The final approximation is compared to the analytical solution. The convergence in the $L^2(\Omega)$ norm, as well as the dispersion properties are obtained in the following sections.

5.1 Convergence

Let u_{exact} be the exact solution to the problem (in this case equal to the initial condition). The approximate DG solution is

$$u_{\text{approx}}(x,t) = \sum_{k,j} \widehat{u}_{jk}(t) b_j(\xi_k(x)). \tag{79}$$

The best achievable approximation in the finite element space is the projection u_{proj} of the exact solution under the standard L^2 inner product, resulting in the coefficients $\widehat{u}^{\text{opt}}$,

$$\widehat{u}^{\text{opt}}_{jk}(t) = \left[\mathbf{M}_k^{-1}\right]_{ji} |J_k|^{-1} \int_{K_k} u_{\text{exact}}(x,t) b_i(\xi_k(x)) \mathrm{d}x. \tag{80}$$

To evaluate the total error in the solution at the final time, the L^2 norm of the local error is computed $\varepsilon = \|u_{\text{approx}} - u_{\text{exact}}\|_{L^2}$. This can be done conveniently in terms of the coefficients:

$$\begin{aligned}
\varepsilon^2 &= \sum_{K_k \in \mathcal{T}_h} \int_{K_k} |\widehat{u}_{jk} b_j(\xi_k(x)) - u_{\text{exact}}(x)|^2 \,\mathrm{d}x \\
&= \|u_{\text{exact}}\|^2_{L^2} + \|u_{\text{approx}}\|^2_{L^2} - 2 \sum_{K_k \in \mathcal{T}_h} \widehat{u}_{jk} \int_{K_k} u_{\text{exact}}(x) b_j(\xi_k(x)) \mathrm{d}x \\
&= \|u_{\text{exact}}\|^2_{L^2} + \widehat{u}^{\mathrm{T}} \mathbf{MJ} (\widehat{u} - 2\widehat{u}^{\text{opt}}) \\
&= \underbrace{\|u_{\text{exact}}\|^2_{L^2} - (\widehat{u}^{\text{opt}})^{\mathrm{T}} \mathbf{MJ} (\widehat{u}^{\text{opt}})}_{\text{projection error squared}} + \underbrace{(\widehat{u} - \widehat{u}^{\text{opt}})^{\mathrm{T}} \mathbf{MJ} (\widehat{u} - \widehat{u}^{\text{opt}})}_{\text{propagation error squared}}
\end{aligned}$$

Here $\mathbf{M}$ is the block diagonal mass matrix of all elements, and $\mathbf{J}$ is the diagonal matrix of Jacobian determinants. Note that the total error ε can not go below the projection error. For the linear advection problem, the norm of the exact solution is known and the norm of the numerical solution can always be evaluated using the mass matrix. The optimal projection for the given initial condition is described in section A.1, enabling the cross term to be computed as well. Therefore ε can be evaluated exactly without using quadrature or additional projection operations.

The exact projection of the initial condition is numerically limited by the floating point hardware. For the theoretical analysis, see appendix A.1. Both the theoretical result (123) and this numerical experiment show that the point of breakdown depends on the polynomial degree and on the relative frequency (number of waves per cell). This has important consequences, limiting the accuracy of the overall method. See the results in figure 13). This clearly shows the effect of the projection and propagation terms on the total error. The method achieves its theoretical convergence rate of $p+1$ up to $p = 4$. After that, the convergence is limited by the evaluation of the initial condition, which immediately becomes the dominant error source. If the projection error were excluded by measuring the total error with respect to the projected solution, the expected convergence rate is restored. The current presentation is chosen in order not to fool the reader. It is clear that a high order method should only be chosen if the input data can be represented with sufficient accuracy (compared to the dominant error source).

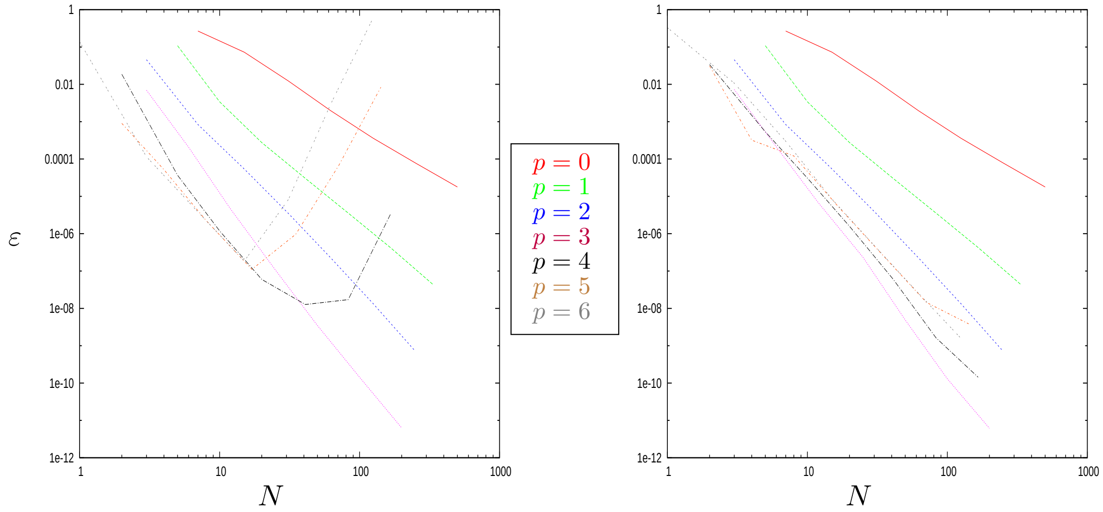

Figure 13: *Convergence of the RKDG method on the linear advection problem, obtained for a sinusoidal initial condition* (78) *with* $n_{max} = 1$. *Global error* ε *due to projection and propagation for one period, against number of elements* N. *Polynomial degrees* 0 *to* 6. *On the left, the numerical breakdown of the projection is observed for higher degrees and lower relative frequencies. On the right, projection was limited to* $p = 4$ *to prevent the breakdown at the cost of accuracy.*

5.2 Dispersion properties

To measure the dispersion properties of the RKDG method, the number of wave modes is increased to $n_{\max} = 8$. A fourth order Runge-Kutta method with a 3rd order DG scheme ($p = 2$) is selected as an example. After propagating the solution on a relatively coarse grid ($N = 15$), its

Fourier transform is computed. Because the problem is linear, the result is independent for each wave mode and the dispersion relation $\omega(k)$ is revealed by computing the amplification factor from the numerical and exact solutions,

$$G(k) = \exp(-i\omega(k)t_{\text{end}}). \tag{81}$$

The damping is given by the absolute value $|G|$ and the phase shift is given by the complex argument $\angle G$. The decay of the solution in the higher frequencies is apparent, but the phase behaviour is excellent. The numerical dispersion properties are well matched by the theoretical results, based on [30]. The difference in phase behaviour for the unresolved case could be due to a high-frequency spurious mode. In this third order experiment, it is seen that 10 to 12 degrees of freedom per wavelength are sufficient for accurate wave-propagation.

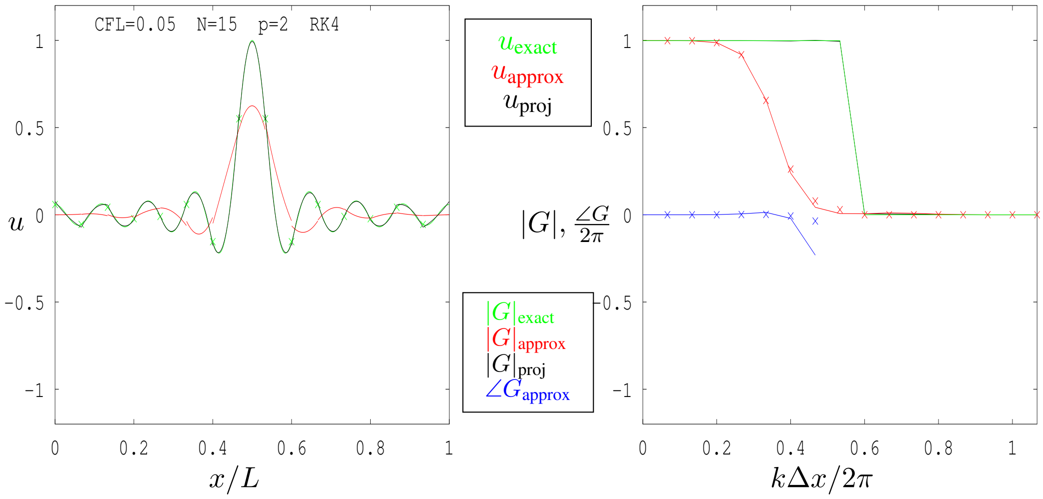

Figure 14: *Dispersion of RKDG method (3rd order in space, 4th order in time). Spatially under-resolved, central scheme at CFL* $= 0.05$ *(*25% *of the limit). Solution after one propagation period on the left, its frequency domain representation on the right. Exact solution (green) and projection (black) are close but discernible. Numerical dispersion (solid) consistent with analytical (crosses).*

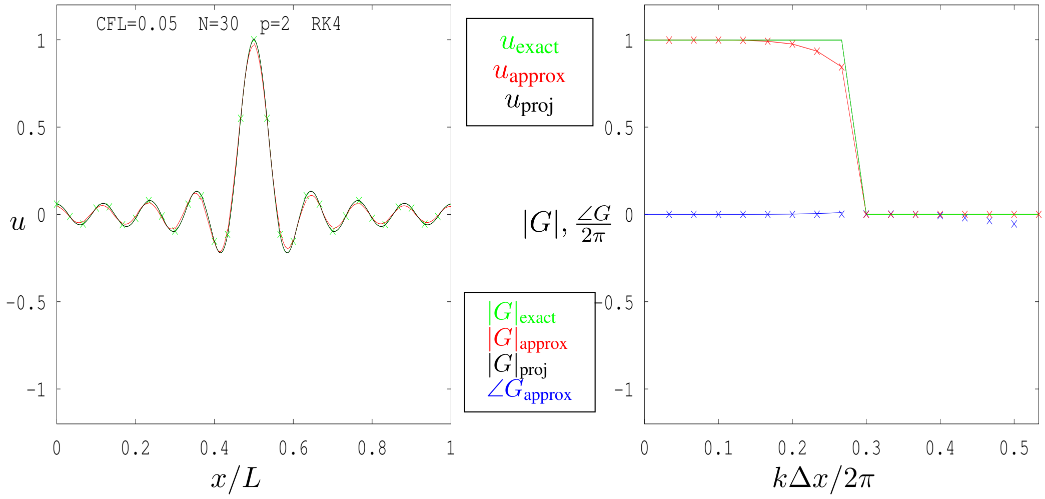

Figure 15: *Dispersion of RKDG method (3rd order in space, 4th order in time). Spatially almost resolved, central scheme at CFL* $= 0.05$ *(*25% *of the limit). Solution after one propagation period on the left, its frequency domain representation on the right. Exact solution (green) and projection (black) overlap. Numerical dispersion (solid) consistent with analytical (crosses).*

6 Unsteady interacting boundary layer

The model for the interacting boundary layer system is derived in this section. A depth-integral average model is obtained from an unsteady control volume approach in 2 dimensions, applied to conservation of mass (§ 6.1), momentum (§ 6.2) and energy (§ 6.3). A local form of the quasi-simultaneous interaction method by Veldman [48] is adapted and included (§ 6.4). The resulting model can be seen as a shear layer averaged Navier-Stokes (SLANS) equation. Neglecting the diffusive terms (§ 6.5), this turns into a diagonalizable first order system of interacting boundary layer equations (IBL). This system of three equations (momentum, energy, interaction) contains six unknowns, and extra closure relations are required. For laminar flow these are given in § 6.6. The analysis in § 6.8 shows that the combined equations form a nonconservative hyperbolic system (NCHS). Finally, § 6.9 adresses the effect of the limiting viscosity on the weak solutions to the system.

In the ordinary approach (e.g. [34, 38, 49]) the derivation of the integral form of the boundary layer equations starts from the differential form of the Navier-Stokes equations. Using order of magnitude estimates, these are approximated by the Prandtl boundary layer equations, then integrated in the boundary normal direction (from 0 to ∞). By a change of dependent variables, this will result in a differential equation where the space dimension is reduced by one.

Instead, most of the assumptions are not required when applying an averaging procedure to integral form of the Navier-Stokes equations, since the boundary layer equations (IBL) are a special case of the shear layer averaged Navier-Stokes equations (SLANS). Only the following assumptions will be made

- The flow in the shear layer is incompressible.
- The shear layer is bounded by two continuous inviscid streamlines / surfaces.
- No body forces are present.
- Pressure work is done by the average pressure through the boundary layer.

The second of these assumptions is the key to this derivation. It implies a.o. that an (unsteady) Bernoulli equation can be used along the outer streamline. Though the incompressible case is assumed, the density ρ will be retained in the conserved variables to assist in the interpretation. The conservation laws will be applied to the control volume shown in figure 16. The control volume is bounded by the wall $z = 0$, the outer streamline $z = \delta_\psi$ and two planes normal to the wall $x = x_a, x = x_b$. Here (x, z) are curvilinear coordinates tangential and normal to the wall, respectively.

6.1 Conservation of mass

Conservation of mass is expressed by the continuity equation (82). Since the outer control volume boundary is an unsteady streamline, the boundary flux includes the effect of the outward normal

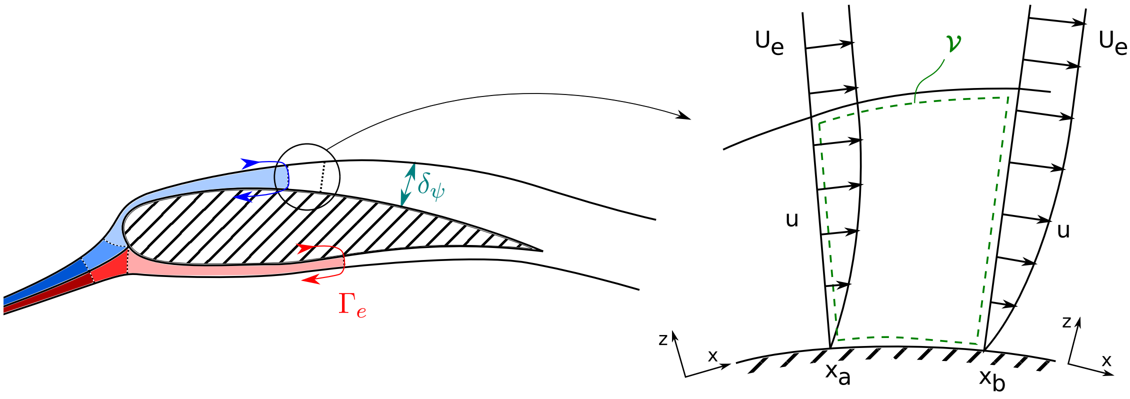

Figure 16: *Definition of the streamline distance δ_ψ, regional circulation Γ_e and control volume $\mathcal{V}$ for the 2d boundary layer. The upstream regions of the upper and lower stream tubes have been colored. The circulation around these regions Γ_e is defined by choosing an orientation for their boundaries as indicated in the figure. Between the upper and lower trailing edge, there will be a jump in Γ_e equal to Γ_{wake}*

speed s of this boundary. The (x, z)-components of the velocity vector $\boldsymbol{u}$ are denoted (u, w).

$$\frac{\partial}{\partial t}\int_{\mathcal{V}(t)} \rho \mathrm{d}\mathcal{V} + \int_{\partial\mathcal{V}} \rho\left(\boldsymbol{u}\cdot\boldsymbol{n} - s\right)\mathrm{d}\mathcal{S} = 0, \tag{82}$$

$$\rho\int_{x_a}^{x_b} \frac{\partial\delta_\psi}{\partial t}\mathrm{d}\tilde{x} + \left[\int_0^{\delta_\psi} \rho u \mathrm{d}z\right]_{x=x_a}^{x=x_b} - \rho\int_{x_a}^{x_b} s\mathrm{d}\tilde{x} = 0.$$

To write this in terms of the well known boundary layer quantities, the displacement thickness δ^* (83), the boundary layer edge velocity $u_e(x) = u(x, z{=}\delta_\psi)$ and edge velocity magnitude $q_e(x) = |\boldsymbol{u}(x, z{=}\delta_\psi)|$ are defined. Using the fact that $s = \frac{\partial\delta_\psi}{\partial t}$, the continuity equation in boundary layer form becomes

$$\delta^* \overset{\text{def}}{=} \int_0^{\delta_\psi}\left(\frac{u_e - u}{q_e}\right)\mathrm{d}z, \tag{83}$$

$$\left[\int_0^{\delta_\psi} \rho u \mathrm{d}z\right]_{x=x_a}^{x=x_b} = \left[\rho(u_e\delta_\psi - q_e\delta^*)\right]_{x=x_a}^{x=x_b} = 0. \tag{84}$$

Since this equation is independent of time and space (only in two dimensional incompressible flow), it can be replaced immediately by the Riemann invariant of mass flow $\dot{m} = $ const:

$$\dot{m} \overset{\text{def}}{=} \rho(u_e\delta_\psi - q_e\delta^*) \tag{85}$$

6.2 Conservation of momentum

A generalization of Von Kármán's momentum integral is obtained by applying the same procedure to the conservation of x-momentum [12] ρu for the control volume (86). Using the assumption

[12] When using the current definitions with the curvilinear coordinates, the 'x-momentum' is not strictly linear momentum, but in fact the angular momentum around the center of curvature.

that the streamline at $z = \delta_\psi$ is continuous in time and outside the shear layer, and accounting for curvature[13] equation (87) is obtained, where subscripts e and w refer to the edge streamline and the wall, respectively.

$$\frac{\partial}{\partial t}\int_{\mathcal{V}(t)} \rho u \mathrm{d}\mathcal{V} + \int_{\partial\mathcal{V}} \rho u \left(\boldsymbol{u}\cdot\boldsymbol{n} - s\right)\mathrm{d}\mathcal{S} + \int_{\partial\mathcal{V}} \boldsymbol{e}_x \cdot \left(p\mathbf{I} - \boldsymbol{\tau}\right)\boldsymbol{n}\mathrm{d}\mathcal{S} = 0, \tag{86}$$

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_a}^{x_b} \rho(u_e\tilde{\delta_\psi} - q_e\delta^*)\mathrm{d}x + \left[\int_0^{\delta_\psi}(\rho u^2 + p - \tau_{xx})\mathrm{d}z\right]_{x=x_a}^{x=x_b} - \int_{x_a}^{x_b}\rho u_e s \mathrm{d}\tilde{x} \\ - \int_{x_a}^{x_b} p_e \frac{\partial\delta_\psi}{\partial\tilde{x}}\mathrm{d}\tilde{x} + \int_{x_a}^{x_b}\tau_w \mathrm{d}x = 0.\end{aligned} \tag{87}$$

Introducing the momentum thickness θ, and postponing the inclusion of the viscous τ_{xx} term to section 6.5, the integral boundary layer momentum equation can be written as

$$\theta \overset{\text{def}}{=} \int_0^{\delta_\psi} \frac{u}{q_e}\left(\frac{u_e - u}{q_e}\right)\mathrm{d}z, \tag{88}$$

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_a}^{x_b} \rho(u_e\tilde{\delta_\psi} - q_e\delta^*)\mathrm{d}x + \left[u_e\dot{m} - \rho q_e^2\theta + p_{\text{avg}}\delta_\psi\right]_{x_a}^{x_b} - \int_{x_a}^{x_b}\left(\rho u_e\frac{\partial\delta_\psi}{\partial t} + p_e\frac{\partial\delta_\psi}{\partial\tilde{x}}\right)\mathrm{d}\tilde{x} \\ + \int_{x_a}^{x_b}\tau_w\mathrm{d}x = 0.\end{aligned} \tag{89}$$

The z-momentum equation can be used to eliminate p_{avg} (the average pressure through the boundary layer) as a variable. For steady, laminar flow over a flat plate, it can be shown that pressure is approximately constant through the boundary layer, and $p_{\text{avg}} = p_e$. For the more general case, an integral variable p^* can be introduced (90). Note that p^* is zero for the steady laminar flat plate flow. Furthermore, the integral term involving $\partial\delta_\psi/\partial x$ and $\partial\delta_\psi/\partial t$ can be rewritten by noting that along the outer streamline, according to Bernoulli's equation $p_e + \frac{1}{2}\rho q_e^2 = p_0 - \rho\,\partial\Gamma_e/\partial t$. Then using $(p_{\text{avg}} - p_0 + \rho\,\partial\Gamma_e/\partial t)\delta_\psi = \frac{1}{2}\rho q_e^2(p^* - \delta_\psi)$, eliminates all explicit pressure terms.

$$p^* \overset{\text{def}}{=} \underbrace{\int_0^{\delta_\psi}\left(1 - 2\frac{p_0 - p}{\rho q_e^2}\right)\mathrm{d}z}_{\text{local vertical pressure gradient}} + \underbrace{\frac{2\delta_\psi}{\rho q_e^2}\frac{\partial\Gamma_e}{\partial t}}_{\text{unsteady upstream circulation}}, \tag{90}$$

$$\frac{\partial\Gamma_e}{\partial t} = \int\limits_{x_{\text{attach}}}^{x} \frac{q_e}{u_e}\frac{\partial q_e}{\partial t}\mathrm{d}\tilde{x} \tag{91}$$

[13] Using radius of curvature $R = \frac{1}{\kappa}$, we define $\mathrm{d}\tilde{x} = (1 + \kappa\delta_\psi)\mathrm{d}x$, $\mathrm{d}\tilde{y} = (1 + \kappa y)\mathrm{d}y$ and $\tilde{\delta_\psi} = (1 + \frac{\kappa\delta_\psi}{2})\delta_\psi$.

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_a}^{x_b}\rho(u_e\tilde{\delta_\psi}-q_e\delta^*)\mathrm{d}x+\left[u_e\dot{m}-\rho q_e^2\theta+\frac{1}{2}\rho q_e^2(p^*-\delta_\psi)\right]_{x=x_a}^{x=x_b}\\+\int_{x_a}^{x_b}\left(\frac{1}{2}\rho q_e^2\frac{\partial\delta_\psi}{\partial\tilde{x}}-\rho u_e\frac{\partial\delta_\psi}{\partial t}-\rho\delta_\psi\frac{q_e}{u_e}\frac{\partial q_e}{\partial t}\right)\mathrm{d}\tilde{x}+\int_{x_a}^{x_b}\tau_w\mathrm{d}x=0.\end{aligned}\tag{92}$$

Integrating the nonconservative term by parts, and rewriting $\frac{\partial\rho q_e}{\partial t}=\left(\frac{\mathrm{d}q_e}{\mathrm{d}u_e}\frac{\partial\rho u_e}{\partial t}+\frac{\mathrm{d}q_e}{\mathrm{d}w_e}\frac{\partial\rho w_e}{\partial t}\right)$ with $u_e^2+w_e^2=q_e^2$, the nonconservative term is transformed

$$\begin{aligned}\int_{x_a}^{x_b}\left(\frac{1}{2}\rho q_e^2\frac{\partial\delta_\psi}{\partial\tilde{x}}-\rho u_e\frac{\partial\delta_\psi}{\partial t}-\rho\delta_\psi\frac{q_e}{u_e}\frac{\partial q_e}{\partial t}\right)\mathrm{d}\tilde{x}=\left[\frac{1}{2}\rho q_e^2\delta_\psi\right]_{x=x_a}^{x=x_b}-\frac{\partial}{\partial t}\int_{x_a}^{x_b}\rho u_e\delta_\psi\mathrm{d}\tilde{x}\\-\int_{x_a}^{x_b}\delta_\psi\left(\frac{\partial\frac{1}{2}\rho q_e^2}{\partial x}+\frac{w_e}{q_e}\frac{\partial\rho w_e}{\partial t}\right)\mathrm{d}\tilde{x}\end{aligned}$$

The wall shear stress scales with the momentum in the outer flow, therefore the friction coefficient is introduced next. Furthermore, moving the mass flux $\dot{m}$ into the nonconservative term eliminates most of the streamline dependence:

$$\frac{\partial}{\partial t}\int_{x_a}^{x_b}-\rho q_e\delta^*\mathrm{d}\tilde{x}+\left[\rho q_e^2\left(\frac{1}{2}p^*-\theta\right)\right]_{x=x_a}^{x=x_b}-\int_{x_a}^{x_b}q_e\delta^*\frac{\partial u_e}{\partial x}+\delta_\psi R_{\text{sep}}\mathrm{d}\tilde{x}+\int_{x_a}^{x_b}\frac{1}{2}C_f\rho u_e^2\mathrm{d}x.\tag{93}$$

where R_{sep} is a term which is negligible except at attachment and (strong) separation points, and C_f is the wall friction coefficient.

$$R_{\text{sep}}\overset{\text{def}}{=}\frac{\partial\frac{1}{2}\rho w_e^2}{\partial x}+\frac{w_e}{q_e}\frac{\partial\rho w_e}{\partial t}\qquad\qquad C_f=\frac{2\tau_w}{\rho u_e^2}\tag{94}$$

Comparing this result to the unsteady boundary layer momentum equation derived by Matsushita et al. [34],

$$\frac{\partial u_e\delta^*}{\partial t}+\frac{\partial u_e^2\theta}{\partial x}+u_e\delta^*\frac{\partial u_e}{\partial x}=\frac{\tau_w}{\rho}$$

it is seen that exact correspondence is obtained by setting $q_e=u_e$, $p^*=0$ and $R_{\text{sep}}=0$.

6.3 Conservation of energy

The momentum and continuity IBL equations can be complemented by an energy equation. Several types of energy equation exist, for total energy, mechanical, internal energy, or turbulent kinetic energy for instance. In incompressible flow, normally the conservation of mass and momentum are sufficient to describe the solution. Conservation of (total) energy can then be solved independently for the temperature. Not all the different energy equations are conservation laws. Some describe a balance between different forms of energy, and may contain sources (production or dissipation terms). For the compressible Euler equations, a conservative mechanical energy equation can be derived from the momentum and continuity equations. For the Navier-Stokes, a dissipation term must be added due to the work done by viscous stresses inside the volume. Both derivations proceed according to the following recipe:

- Finding the expressions for potential and kinetic energy
- Relating their time derivatives to time derivatives of conserved quantities
- Substitute these for the fluxes from the momentum and continuity equation
- Write these as a combined flux and/or source term

The mechanical energy equation then becomes

$$\frac{\partial}{\partial t}\int_{\mathcal{V}} \rho e_{\text{mech}} \mathrm{d}\mathcal{V} + \int_{\partial\mathcal{V}} \rho e_{\text{mech}} \left(\boldsymbol{u}\cdot\boldsymbol{n} - s\right) \mathrm{d}\mathcal{S} + \int_{\partial\mathcal{V}} \left(p\mathbf{I} - \boldsymbol{\tau}\right) \boldsymbol{u}\cdot\boldsymbol{n}\mathrm{d}\mathcal{S} + \int_{\mathcal{V}} S_{\text{me}} \mathrm{d}\mathcal{V} = 0.$$

The dissipation function for mechanical energy S_{me} is defined (cf. Schlichting [44]) as

$$\begin{aligned} S_{\text{me}} &= \operatorname{div}(\boldsymbol{\tau}\boldsymbol{u}) - \boldsymbol{u}\cdot\operatorname{div}(\boldsymbol{\tau}) \\ &= 2\mu\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial y}\right)^2\right] + \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right)^2, \end{aligned}$$

which is always positive. The mechanical energy density e_{mech} is defined as the sum of the kinetic and potential energy densities, cf. [29]:

$$e_{\text{kin}} = \frac{1}{2}\boldsymbol{u}\cdot\boldsymbol{u}, \qquad\qquad e_{\text{pot}} = \phi + \Phi - \frac{p}{\rho},$$

where the energy transport potential Φ is defined below. Note that in the incompressible case, this potential is equal to p/ρ. The potential energy density then equals the potential ϕ, corresponding to a conservative body force.

$$\Phi = \int^{p} \frac{1}{\rho(\bar{p})}\mathrm{d}\bar{p}.$$

Now apply the mechanical energy equation to the boundary layer control volume (figure 16), assuming incompressibility and discarding the body forces,

$$\begin{aligned} \frac{\partial}{\partial t}\int_{\mathcal{V}} \frac{1}{2}\rho\boldsymbol{u}\cdot\boldsymbol{u}\mathrm{d}\mathcal{V} + \left[\int_0^{\delta_\psi} \frac{1}{2}\rho\boldsymbol{u}\cdot\boldsymbol{u}u\mathrm{d}z\right]_{x_a}^{x_b} - \int_{x_a}^{x_b} \frac{1}{2}\rho q_e^2 \frac{\partial\delta_\psi}{\partial t}\mathrm{d}\tilde{x} \\ + \left[\int_0^{\delta_\psi} \boldsymbol{e}_1\cdot(p\boldsymbol{u} - \boldsymbol{\tau}\boldsymbol{u})\mathrm{d}z\right]_{x_a}^{x_b} + \int_{\mathcal{V}} S_{\text{me}}\mathrm{d}\mathcal{V} = 0. \end{aligned}$$

Defining a kinetic energy transport thickness δ_k and a mechanical energy thickness ε,

$$\begin{aligned}\delta_k &\overset{\text{def}}{=} \int_0^{\delta_\psi} \frac{u}{q_e}\left(1 - \frac{\boldsymbol{u}\cdot\boldsymbol{u}}{q_e^2}\right)\mathrm{d}z,\\ \varepsilon &\overset{\text{def}}{=} \int_0^{\delta_\psi} \left(1 - \frac{\boldsymbol{u}\cdot\boldsymbol{u}}{q_e^2}\right)\mathrm{d}\tilde{z},\end{aligned}$$

we can write the mechanical energy equation in boundary layer form as

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_a}^{x_b} \rho q_e^2\,(\delta_\psi - \varepsilon)\,\mathrm{d}\tilde{x} + \left[q_e^2(\dot{m} - \rho q_e \delta_k)\right]_{x_a}^{x_b} - \int_{x_a}^{x_b} \rho q_e^2 \frac{\partial \delta_\psi}{\partial t}\mathrm{d}\tilde{x}\\ +2\left[\int_0^{\delta_\psi} u(p - \tau_{xx}) - w\tau_{xz}\mathrm{d}z\right]_{x_a}^{x_b} + 2\int_{\mathcal{V}} S_{\text{me}}\mathrm{d}\mathcal{V} = 0.\end{aligned}$$

Note that the kinetic energy density is represented in integral form as $\mathcal{E}_\psi = \rho q_e^2(\delta_\psi - \varepsilon)$. The viscous contributions are postponed to section 6.5. The pressure work through the boundary layer is approximated by the average pressure work, $p_{\text{avg}}\dot{m}$. This average is expressed using the definition of p^* as $p_{\text{avg}} = p_0 + \frac{1}{2}\rho q_e^2(p^*/\delta_\psi - 1) - \rho\,\partial\Gamma_e/\partial t$. It is important to note that in the incompressible case, the mechanical energy equation only depends on the pressure up to an arbitrary constant. Since $\dot{m}$ and p_0 do not depend on space or time, this is indeed the case.

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_a}^{x_b} \rho q_e^2(\delta_\psi - \varepsilon)\mathrm{d}\tilde{x} + \left[q_e^2\left(\frac{p^*\dot{m}}{\delta_\psi} - \rho q_e \delta_k\right)\right]_{x_a}^{x_b} - \int_{x_a}^{x_b} 2\dot{m}\frac{q_e}{u_e}\frac{\partial q_e}{\partial t}\mathrm{d}\tilde{x}\\ - \int_{x_a}^{x_b} \rho q_e^2 \frac{\partial \delta_\psi}{\partial t}\mathrm{d}\tilde{x} + 2\int_{\mathcal{V}} S_{\text{me}}\mathrm{d}\mathcal{V} = 0.\end{aligned} \tag{95}$$

The equation can be made streamline-independent by taking together the streamline-dependent terms (involving δ_ψ and $\dot{m}$). In the dissipation integral, $(\partial u/\partial z)^2$ is the dominant term [38], so this is used to scale the dissipation term by defining a dissipation coefficient $\mathcal{D}$

$$\int_0^{\delta_\psi} S_{\text{me}}\mathrm{d}\tilde{z} \approx \int_0^{\delta_\psi} \mu\left(\frac{\partial u}{\partial z}\right)^2 \mathrm{d}\tilde{z} \geq \mu\frac{u_e^2}{\delta_\psi}$$

$$\mathcal{D} \overset{\text{def}}{=} \frac{2\text{Re}_\theta}{\rho q_e^3}\int_0^{\delta_\psi} S_{\text{me}}\mathrm{d}\tilde{z}. \tag{96}$$

The resulting mechanical energy equation is

$$-\frac{\partial}{\partial t}\int_{x_a}^{x_b} \rho q_e^2 \varepsilon \mathrm{d}\tilde{x} + \left[q_e^2\left(\frac{p^*\dot{m}}{\delta_\psi} - \rho q_e \delta_k\right)\right]_{x_a}^{x_b} + \int_{x_a}^{x_b} \rho q_e^3 \text{Re}_\theta^{-1}\mathcal{D} + 2q_e\delta^*\frac{q_e}{u_e}\frac{\partial \rho q_e}{\partial t}\mathrm{d}\tilde{x} = 0. \tag{97}$$

6.4 Local interaction

The edge velocity equation or interaction equation is not a conservation law. Furthermore, its domain of definition is not the control volume as depicted in figure 16, but the bounding streamline (stream surface) only. In non-interacting boudary layers, the edge velocity would therefore be treated as a boundary condition or forcing term. However, when strong interaction occurs, it is necessary to treat this ‘boundary condition’ instead as a simultaneous or quasi-simultaneous process in order to obtain a solution [10]. An adapted version of the localized quasi-simultaneous approach [48] is chosen here for its speed and accuracy properties. Also, this approach provides more physical insight than the blind coupling of the systems in the discretized domain.

The effect of the boundary layer on the edge velocity can be modelled by considering the equivalent inviscid flow (EIV) over a flat plate. Taking the difference between the true conservation of mass (84) and the inviscid conservation of mass for a flat plate $\partial_x \rho u_e \delta_\psi = 0$, it is apparent that the inviscid equation requires a source term σ in order to be equivalent:

$$\sigma = \frac{\partial q_e \delta^*}{\partial x}.$$

This procedure can also be applied to a generic inviscid flow, and is not limited to a flat plate, see [19]. The effect of these sources on the edge velocity $u_e(x)$ at a certain x location is found by superimposing the fundamental potential flow solutions scaled by the local source strengths:

$$\Delta u_e(x) = \frac{1}{\pi} \int_{\text{wall}} \frac{(x-\xi)\sigma(\xi)}{(x-\xi)^2 + {\delta_\psi}^2} \mathrm{d}\xi \tag{98}$$

The contribution due to a finite subregion $\mathcal{W}(x)$ of the wall around point x is found through integration by parts

$$\Delta u_e(x) = \frac{1}{\pi} \left[\frac{(x-\xi) q_e \delta^*}{(x-\xi)^2 + {\delta_\psi}^2} \right]_{\partial \mathcal{W}(x)} - \frac{1}{\pi} \int_{\mathcal{W}(x)} q_e \delta^* \frac{\partial}{\partial \xi} \left(\frac{(x-\xi)}{(x-\xi)^2 + {\delta_\psi}^2} \right) \mathrm{d}\xi.$$

For a small region $\mathcal{W}$, say $(x-\xi)|_{\partial \mathcal{W}} = \pm \delta_\psi$, this is equal to

$$\Delta u_e(x) = \frac{1}{\pi} \left(-\frac{[q_e \delta^*]_{x-\delta_\psi}}{2\delta_\psi} + \frac{(q_e \delta^*)_{\text{central}}}{\delta_\psi} - \frac{[q_e \delta^*]_{x+\delta_\psi}}{2\delta_\psi} \right) \approx -\frac{\delta_\psi}{2\pi} \frac{\partial^2 q_e \delta^*}{\partial x^2} \tag{99}$$

The central term is the result of the integral, which acts as a convolution with the negative part of a bump function (see figure 17). Equation (99) looks very much like a discretized second derivative (although no discretization has been applied at all). In contrast to similar methods (by Coenen [13], Veldman [48] and Bijleveld [4]), the current formulation is independent of any grid size. A further simplification is obtained by only retaining only the central term. As described by the above authors, this approximates the full interaction matrix with a diagonal matrix. It provides a less accurate but more stable interaction method. A more physical interpretation is given here: this approach essentially cuts the ‘tails’ of the convolution function describing the effect of the boundary layer on the edge velocity. For the first order system, it is sufficient to know the time derivative $\partial_t u_e$. Considering only the localized interaction effect due to this small region,

$$\frac{\partial u_e}{\partial t} = \frac{\partial \Delta u_e}{\partial t} \approx \frac{1}{\pi} \frac{\partial}{\partial t} \frac{q_e \delta^*}{\delta_\psi} \tag{100}$$

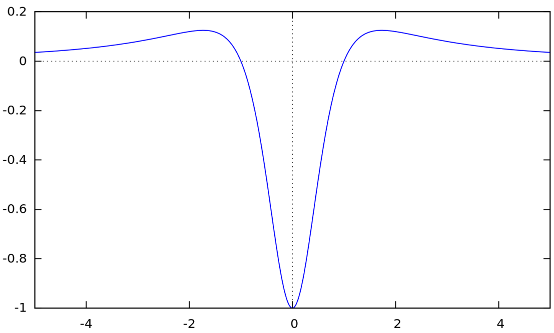

Figure 17: *Effect of the local momentum defect on the edge velocity*

This equation is mostly convective, which can be seen by expressing the right hand side in terms of the momentum conservation equation. Using a constant mass flux $\dot{m}$, the right hand side is related to the evolution of the momentum defect $q_e\delta^*$ by

$$\frac{\partial}{\partial t}\left(\frac{q_e\delta^*}{\delta_\psi}\right) = \frac{1}{\delta_\psi}\left(1 - \frac{\delta^*}{\delta_\psi}\right)\frac{\partial q_e\delta^*}{\partial t} + \frac{\delta^*}{\delta_\psi}\frac{\partial u_e}{\partial t},$$

hence the convective character of this term is inherited from the momentum equation. The interaction equation therefore becomes

$$\begin{aligned}(\pi - H_\psi)\frac{\partial u_e}{\partial t} &= \frac{1}{\delta_\psi}(1 - H_\psi)\frac{\partial q_e\delta^*}{\partial t} \\ &= \frac{1}{\delta_\psi}(1 - H_\psi)\left[S_{\text{mo}} - G_{\text{mo}}\frac{\partial \boldsymbol{u}}{\partial x} - \frac{\partial}{\partial x}\left(F_{\text{mo}} - B_{\text{mo}}\frac{\partial \boldsymbol{u}}{\partial x}\right)\right]. \end{aligned} \tag{101}$$

The source S, flux F and nonconservative and diffusive terms G and B from the generic momentum equation have been used. The additional nondimensional factor $H_\psi = \delta^*/\delta_\psi$ is typically smaller than one (since the streamline is assumed outside the shear layer). The full expansion in momentum terms serves only to separate the convective and diffusive terms. In practice, equation (101) is expressed more easily without expanding the momentum term.

6.5 Diffusive terms

Since the second order (diffusive) terms are important in defining solutions to a nonconservative first-order system (see section 3.4), these terms are given here. The continuity equation contains no diffusion; the diffusion in the momentum and mechanical energy equations is due to viscosity and in the interaction equation due to the elliptic nature of the inviscid flow. All diffusive terms are of the form

$$\frac{\partial}{\partial x}\left(B(\boldsymbol{u})\frac{\partial \boldsymbol{u}}{\partial x}\right)$$

The momentum diffusion term is due to the viscous normal stress τ_{xx}. In laminar flow, it is equal to $\tau_{xx} = \mu\partial_x u$. Integrating this over the control volume and using the continuity equation to substitute $\partial_x u = -\partial_z w$,

$$-B_{\text{mo}}\frac{\partial \boldsymbol{u}}{\partial x} = -\int_0^{\delta_\psi} \mu \frac{\partial u}{\partial x}\mathrm{d}z = \mu w_e = \mu u_e \frac{\partial \delta_\psi}{\partial x}$$

Using the conservation of mass $\partial_x \dot{m} = 0$,

$$= \mu \left[\frac{\partial q_e \delta^*}{\partial x} - \delta_\psi \frac{\partial u_e}{\partial x}\right].$$

The mechanical energy diffusion, or viscous work term is rewritten in a similar way, using $\tau_{xz} \approx \mu \partial_z u$

$$\begin{aligned} B_{\text{me}}\frac{\partial \boldsymbol{u}}{\partial x} &= \int_0^{\delta_\psi} (w\tau_{xz} + u\tau_{xx})\mathrm{d}z = \mu w_e u_e + \int_0^{\delta_\psi} 2u\mu \frac{\partial u}{\partial x}\mathrm{d}z \\ &= \mu \left[w_e u_e + \frac{\partial}{\partial x}\int_0^{\delta_\psi} u^2 \mathrm{d}z - u_e^2 \frac{\partial \delta_\psi}{\partial x}\right] \end{aligned}$$

Canceling the first and last term, and recognising the expression for momentum transport

$$= \nu \frac{\partial}{\partial x}\left[u_e \dot{m} - \rho q_e^2 \theta\right]$$

The edge velocity diffusion has already been found in terms of the momentum equation. The explicit expression in terms of the unknowns is found as

$$-B_{\text{in}}\frac{\partial \boldsymbol{u}}{\partial x} = \text{Re}_\theta^{-1}\frac{H_\psi}{H}(1 - H_\psi)q_e \left(\frac{\partial q_e \delta^*}{\partial x} - \delta_\psi \frac{\partial u_e}{\partial x}\right)$$

Making the boundary layer assumptions $q_e \approx u_e$ and $\theta = \varepsilon - \delta^*$, the total diffusion matrix in terms of the unknowns $\boldsymbol{u} = -[q_e\delta^*, q_e^2\varepsilon, u_e]^{\text{T}}$ becomes

$$\mathbf{B} = \begin{bmatrix} B_{\text{mo}} \\ B_{\text{me}} \\ B_{\text{in}} \end{bmatrix} = \rho \begin{bmatrix} \nu & 0 & -\nu\delta_\psi \\ -\nu u_e & \nu & -\nu u_e \delta_\psi \\ k q_e & 0 & -k q_e \delta_\psi \end{bmatrix} \tag{102}$$

where $k = \text{Re}_\theta^{-1}\frac{H_\psi}{H}(1 - H_\psi)$ is positive. Without further analysis, an order of magnitude estimate for this diffusion can be made. For solutions without shocks or oscillations, the derivative $\partial/\partial x$ corresponds to a length scale $1/L$. As a result, the ratio inertial to viscous terms is of order Re_L for all three equations.

6.6 Closure

Since there are additional unknowns θ, δ_k, p^*, C_f and $\mathcal{D}$, whose behaviour is not prescribed by the conservation of mass, momentum and energy, additional modeling will be required to obtain actual results. This can be done by either:

- Adding additional equations
  - Conservation of total energy, angular momentum
  - Balance equations (quasi-steady, quasi-local assumptions)

- Assuming (parametrized) velocity and pressure profiles
  - Consistent with boundary conditions at wall and edge
  - Exact profiles according to analytical solutions for special flows
  - Curve fits or assimilations of experimentally obtained data
  - Combination of the above

Note that when an analytical or approximate solution is known for a specific case, all the additional unknowns can be evaluated. If empirical relations can be deduced between the additional unknowns and the problem variables, this can be parametrized and used as an additional equation, called a *closure relation*. The problem variables include the set of conserved variables, boundary conditions, sources etc. Note that the additional unknowns can depend on anything, (not only the problem variables) and inferring a closure relation dependent on the problem variables only will require further (statistical) modelling and/or approximation.

6.6.1 Analytical solutions

Several special boundary layer flows have analytical solutions. Four of these are summarised here: Stokes' first and second problems (impulsively started/oscillating flat plate), the steady stagnation flow and Blasius' flat plate boundary layer. For these analytic solutions, a characteristic length c and boundary layer 'thickness' δ_{99} are given in the table below. All flows are self-similar, the first two on an infinite domain, the last two on a semi-infinite domain.

Flow description	u_e	c	δ_{99}/c	δ^*/c
Impulsively started flat plate	$H(t)$	$\sqrt{\nu t_0/\pi}$	$6.46\sqrt{t/t_0}$	$2\sqrt{t/t_0}$
Oscillating flat plate	$\sin(\omega t)$	$\sqrt{\nu/\omega}$	6.5	$\sec(\omega t)\sin(\omega t - \pi/4)$
Steady stagnation flow	kx	$\sqrt{\nu/kU_\infty}$	2.4	0.6479
Blasius boundary layer	1	$\sqrt{\nu x_0/U_\infty}$	$5.0\sqrt{x/x_0}$	$1.721\sqrt{x/x_0}$

The (nondimensionalized) values for θ, p^* and C_f can be computed from these solutions, and used as closure relations.

Flow description	θ/c	p^*/δ_ψ	C_f
Impulsively started flat plate	$2(\sqrt{2}-1)\sqrt{t/t_0}$	0	$\frac{2c}{U_\infty t}\sqrt{t/t_0}$
Oscillating flat plate	$\csc^2(\omega t)\sin(2\omega t + \pi/4)/4$	0	$\frac{\nu}{U_\infty c}\sin(\omega t + \pi/4)$
Steady stagnation flow	0.292	$2/3$	$2.465\frac{c}{x}$
Blasius boundary layer	$0.664\sqrt{x/x_0}$	0	$0.664\frac{c}{x}\sqrt{x/x_0}$

Table 2: *Analytical closure relations for selected flows.*

Most of the values from the above table can simply be found in the literature (e.g. White [49], Schlichting [44]) or they be derived from a known velocity profile. The pressure defect term, p^* which measures the pressure variation over the boundary layer, is normally assumed zero in boundary layer theory. It can in fact be computed for any viscous flow solution, in particular for the Falkner-Skan flows. This is done here for the steady stagnation flow. For each Reynolds number, the outer streamline according to (108) is used. The pressure behaviour near the stagnation point is well approximated by the exact inviscid solution, $p^*/\delta_\psi = 2/3$.

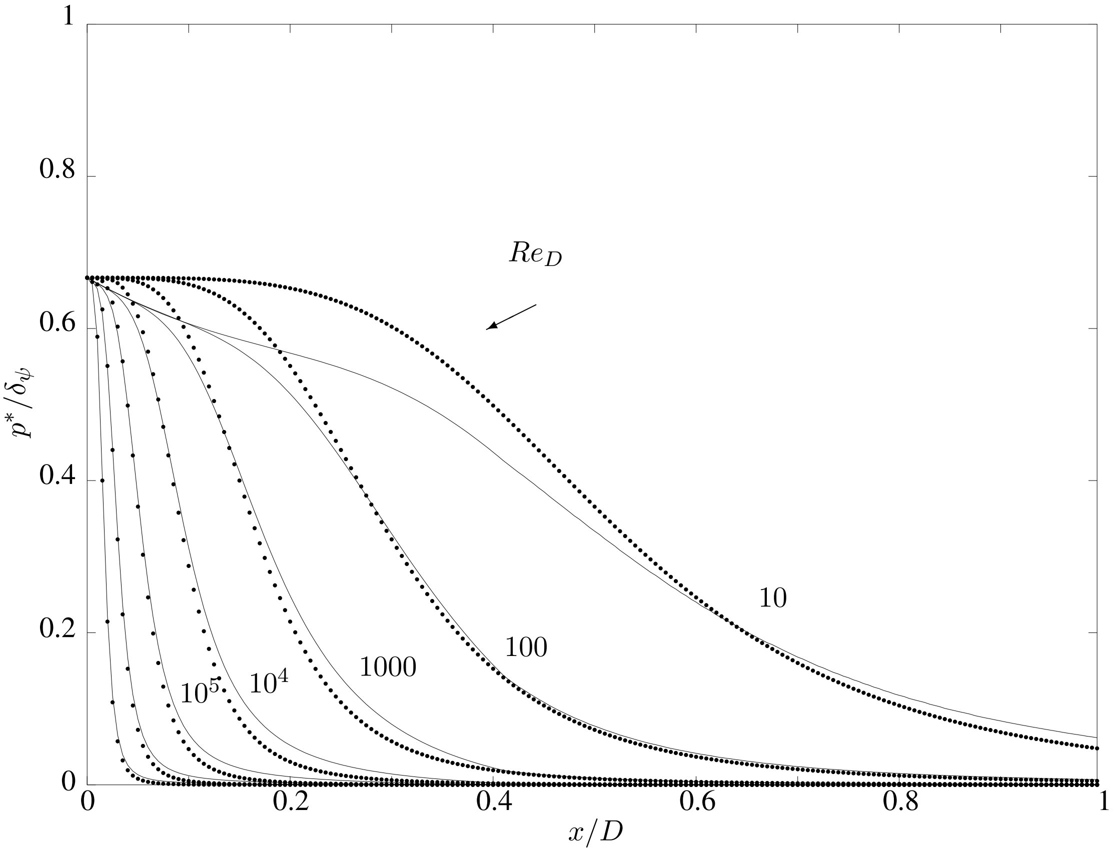

Figure 18: *Behaviour of p^* in stagnation flow for different Reynolds numbers, based on the streamline just outside shear layer at $x = D$. Inviscid solution (dotted), Falkner-Skan solution (solid)*

6.6.2 Nondimensional quantities

To apply the closure relations from literature, some commonly defined variables must be related to the set of independent variables. A number of nondimensional quantities is introduced to limit the range of possible closure relations, and to obtain proper scaling. These quantities are defined in the table below. Next to the two well known factors (H, H^*), a new one is introduced: the budget factor B is defined the ratio of x-momentum transport to the kinetic energy, as observed when moving with the outer flow. From this definition it is clear that $B \in [0, 1]$ and is usually close to one in attached flow. In the traditional integral boundary layer approximation $B = 1$, so that the missing energy density becomes $\varepsilon = \delta^* + \theta$. In three dimensions, the budget factor becomes more important for modelling cross flow.

quantity	description	definition
p^*/δ_ψ	Pressure defect ratio	(90)
$\mathcal{D}$	Dissipation coefficient	(96)
C_f	Skin friction coefficient	(94)
H	Shape factor	δ^*/θ
H^*	Transport factor	δ_k/θ
B	Budget factor	(103)

$$B \overset{\text{def}}{=} \frac{\displaystyle\int_0^{\delta_\psi} (u - u_e)^2 \mathrm{d}z}{\displaystyle\int_0^{\delta_\psi} |\boldsymbol{u} - u_e \boldsymbol{e}_1|^2 \mathrm{d}z} = \frac{u_e \delta^* - q_e \theta}{2u_e \delta^* - q_e \varepsilon} \tag{103}$$

The Hartree solutions to the Falkner-Skan equations provide a one-parameter profile family, which can be used to relate all nondimensional quantities to a single parameter. Choosing H as an independent parameter, authors such as Drela [17] or Nishida [36] have produced curve fits for the other parameters. The laminar closure described by Nishida will be used, as also given by Van den Boogaard [5], with $\mathcal{D} = \mathrm{Re}_\theta C_D$.

$$H^*(H) = \begin{cases} 1.528 + 0.0111\frac{(H-4.35)^2}{H+1} - 0.0278\frac{(H-4.35)^3}{H+1} & \\ \quad -0.0002[(H-4.35)H]^2 & \text{if } H < 4.35, \\ 1.528 + 0.015\frac{(H-4.35)^2}{H} & \text{if } H > 4.35, \end{cases} \tag{104}$$

$$\mathrm{Re}_\theta C_f(H) = \begin{cases} -0.07 + 0.0727\frac{(5.5-H)^3}{(H+1)} & \text{if } H < 5.5, \\ -0.07 + 0.015\left(1 - \frac{1}{(H-4.5)}\right)^2 & \text{if } H > 5.5, \end{cases} \tag{105}$$

$$\mathrm{Re}_\theta \frac{C_D}{H^*}(H) = \begin{cases} 0.207 + 0.00205(4-H)^{5.5} & \text{if } H < 4, \\ 0.207 - 0.0016\frac{(H-4)^2}{1+0.02(H-4)^2} & \text{if } H > 4. \end{cases} \tag{106}$$

6.6.3 Laminar flow: Thwaites' method

A simple approximation to the laminar boundary layer is provided by Thwaites' method [49], an empirical method is based on statistical correlations. Here it is shown that in special cases, it can be obtained as well from the mechanical energy equation with the dissipation closure of Nishida (106). Comparing Thwaites with the steady mechanical energy equation:

$$\frac{\partial u_e^6 \theta^2}{\partial x} \approx 0.45 \nu u_e^5 \quad \text{(Thwaites)} \qquad \frac{\partial}{\partial x}(\rho q_e^3 H^* \theta) = \rho q_e^3 \mathrm{Re}_\theta^{-1} \mathcal{D} \quad \text{(mechanical energy)}$$

Now assume H^* varies very slowly with x compared to $q_e^3\theta$, such that it can be taken out of the derivative. Then multiply the mechanical energy equation with $q_e^3\theta/\rho H^*$:

$$q_e^3 \theta \frac{\partial q_e^3 \theta}{\partial x} = \frac{1}{2}\frac{\partial q_e^6 \theta^2}{\partial x} = \mathcal{D} H^{*-1} \nu q_e^5$$

Thwaites' approximation is obtained when taking $2\mathcal{D} = 0.45H^*$, which is consistent with the relation given by Nishida (106).

6.6.4 Turbulent flow

For turbulent flows, Reynolds averaging can be applied to the momentum and energy equations before the shear layer averaging. The additional terms due to turbulence in the (mean) energy and momentum equations should then be modelled. For this purpose, the Boussinesq hypothesis introduces a turbulent viscosity, which can be related to the turbulent kinetic energy. The transport of the turbulent kinetic energy can be added as an additional equation, for instance using the k-ϵ (two equation) model or the Smagorinsky (one-equation) model. When one intends to implement turbulence modelling, compare also the approach used by Drela [19]. Alternatively, the dissipation can simply be adjusted empirically for the turbulent case. The other closure relations (such as skin friction) should be adapted as well.

6.7 Initial and boundary conditions

The boundary conditions to the interacting boundary layer system are described here. The term *boundary* can cause some confusion in this context, since the boundaries of the integral domain do not include all boundaries of the control volume (see figure 4). Boundary conditions refer only to the integral domain boundaries. Initial conditions are prescribed for the integral domain.

Common types of domain boundary are inflow, outflow and (vertical) wall boundaries. The last type does not make sense in the 2d IBL, since prescribing it yields only trivial solutions. Note that the stagnation and/or separation points are ordinary points in the interior of the domain, and do not require a boundary condition. For flows around sharp corners (such as the trailing edge of an airfoil), boundary conditions may be applied there to explicitly induce separation. For a well-posed hyperbolic problem, the characteristic variables should be prescribed where their characteristics enter the domain.

For the initial conditions, the same can be applied to the characteristic formulation in space-time. Naively, when starting the boundary layer from rest ($u_e = 0$), δ^* and ε are undefined. However, the conserved variables are well defined in terms of the streamline (δ_ψ) plus the momentum and energy densities ($\dot{m}$, $\mathcal{E}_\psi$). Therefore we should specify an initial distribution for δ_ψ which satisfies the following conditions:

- It will correspond to a continuous streamline of the outer flow for $t = 0^+$
- This streamline will remain outside the shear layer for all $t > 0$

After specifying δ_ψ, the initial mass flux and energy density can be chosen. Note that the mass flux should be constant in 2 dimensions, or divergence-free in 3 dimensions. Together, these conditions give a complete initial condition from which the unknowns $[q_e\delta^*, q_e^2\varepsilon, u_e]$ can be found.

Now it is time to show how this works in practice, by finding an appropriate initial streamline for the analytical flow solutions of table 2. Since the first two flows are independent of x and parallell to the wall, $\delta_\psi = \delta_{99}$ is a good choice for an initial streamline. The mass flux follows from $\dot{m} = \rho(u_e\delta_\psi - q_e\delta^*)$. In stagnation flow, the streamlines converge asymptotically towards the (displaced) surface according to (107), see figure 19. The chosen streamline equals an inviscid streamline displaced by the distance δ^*, and must remain outside the shear layer. This can be achieved by setting a minimum value for $(\delta_\psi/\delta_{99})$. The mass flux then follows from equation (108). Here D is a specific diameter. The product kD is known analytically for several types of stagnation flow ($kD = 4$ for a cylinder in incompressible flow, for instance[49]).

$$\overset{\text{streamline position}}{\delta_\psi} \quad = \quad \overset{\text{inviscid streamline}}{\frac{\dot{m}}{kx}} \quad + \quad \overset{\text{local displacement}}{\delta^*} \quad . \tag{107}$$

$$\begin{aligned} \text{Stagnation:} \quad & \frac{\dot{m}}{Du_{max}} = \frac{2.4\,(\,\delta_\psi/\delta_{99}\,)_{\text{min}} - 0.6479}{\sqrt{kD\text{Re}_D}} \\ \text{Blasius:} \quad & \frac{\dot{m}}{x_{max}} = \frac{5.0\,(\,\delta_\psi/\delta_{99}\,)_{\text{min}} - 1.721}{\sqrt{Re_{x_{\text{max}}}}} \end{aligned} \quad . \tag{108}$$

For the Blasius boundary layer a similar procedure can be followed. In fact, an initial condition can be obtained for any flow if the edge velocity u_e along a streamline is known.

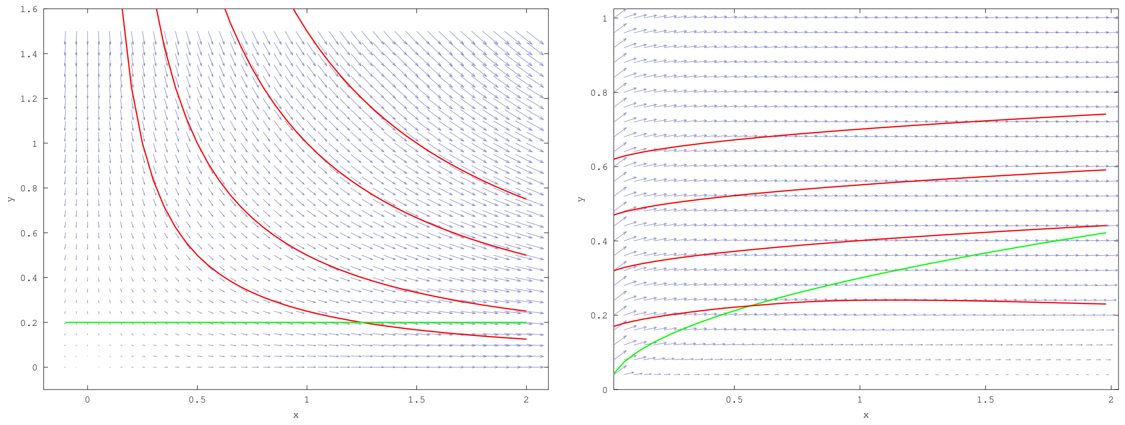

Figure 19: *Boundary layer thickness δ_{99} (green) and outer flow streamlines δ_ψ (red) for steady stagnation (left) and Blasius boundary layer (right)*

6.8 Nonconservative Hyperbolic System

Combining the momentum, energy and interaction equations into a nonconservative system of the form (22), the unknowns $\boldsymbol{u}$, flux vector $\mathbf{f}$, balance tensor $\mathbf{g}$ and source vector $\boldsymbol{s}$ are defined as

$$\boldsymbol{u} = \begin{bmatrix} -q_e\delta^* \\ -q_e^2\varepsilon \\ -u_e \end{bmatrix} \qquad \mathbf{f}^x(\boldsymbol{u}) = \begin{bmatrix} u_e(1-2B) & B & 0 \\ q_e H^* u_e(1-2B) & q_e H^* B & 0 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{u} + q_e^2 p^* \begin{bmatrix} \frac{1}{2} \\ \frac{\dot{m}}{\delta_\psi} \\ 0 \end{bmatrix}$$

$$\mathbf{g}^x(\boldsymbol{u}) = \begin{bmatrix} 0 & 0 & q_e\delta^* \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{g}^t(\boldsymbol{u}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -2q_e\delta^* \\ \frac{1}{\delta_\psi}(1-H_\psi) & 0 & -H_\psi \end{bmatrix}$$

$$\boldsymbol{s}(\boldsymbol{u}) = \begin{bmatrix} \delta_\psi R_{\text{sep}} - \frac{1}{2}C_f u_e^2 \\ -\text{Re}_\theta^{-1}\mathcal{D}q_e^3 \\ 0 \end{bmatrix}.$$

To obtain the characteristics for this system, the flux Jacobians are found by differentiation $A = \partial_{\boldsymbol{u}}\mathbf{f} + \mathbf{g}$. To simplify the procedure, the boundary layer assumptions are used.

$$A^x = \begin{bmatrix} -q_e & 1 & 0 \\ q_e^2(\frac{\partial H^*}{\partial H} - \alpha) & q_e\alpha & q_e^2\beta\delta^* \\ 0 & 0 & 0 \end{bmatrix} \qquad A^t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2q_e\delta^* \\ (\pi - H_\psi)\,\gamma/\delta^* & 0 & \pi - H_\psi \end{bmatrix}$$

with

$$\alpha = H^* - H\frac{\partial H^*}{\partial H} \qquad \beta = \frac{H^*}{H}(1-H) \qquad \gamma = H_\psi\frac{1-H_\psi}{\pi - H_\psi}$$

The resulting eigenvalue problem is $A^x - \lambda A^t = 0$. One of the eigenvalues for this system is zero (corresponding to the interaction equation). The other two are easily determined by the trace and reduced determinant:

$$\text{Tr} = \alpha + 2\gamma - 1 \qquad \text{Det}^* = \frac{\partial H^*}{\partial H} + \beta\gamma \qquad \frac{\lambda^\pm}{q_e} = \frac{\text{Tr} \pm \sqrt{\text{Tr}^2 + 4\text{Det}^*}}{2}$$

Comparing to the eigenvalues obtained by Van den Boogaard [5], equality is obtained for $\gamma = 0$, wich corresponds to the limit without interaction: the bounding streamline lies at infinity. The eigenvectors are

$$\boldsymbol{r}_\pm = \begin{bmatrix} 1 \\ q_e + \lambda_\pm \\ \gamma/\delta^* \end{bmatrix}, \qquad \boldsymbol{r}_0 = \begin{bmatrix} 1 \\ q_e \\ -\frac{\partial H^*}{\partial H}/\beta\delta^* \end{bmatrix}.$$

These can be used in the formulation of the Riemann solvers, and in the determination of the correct viscous limit of the IBL system. The effect of the interaction can be seen by increasing

γ (which corresponds to moving the outer streamline closer to the boundary layer). For the reference flows, the maximum allowable H_ψ is about 0.3, so the interaction effect γ is limited. The effect of interaction on the eigenvalues of the system can be seen in figure 20.

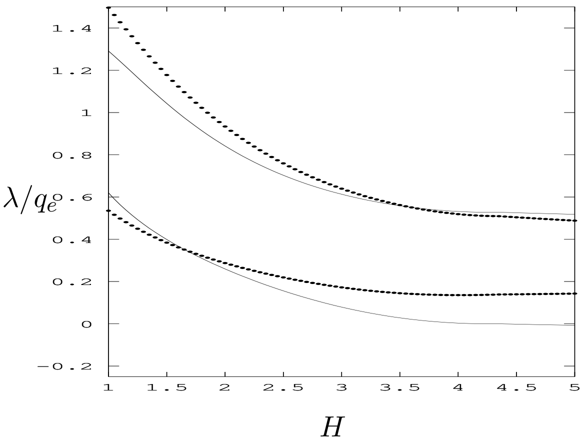

Figure 20: *Eigenvalues of the* IBL *system for* $\gamma = 0$ *(solid) and* $\gamma = 0.06$ *(dotted).*

6.9 Viscous limit

By simple computation it is found that the eigenvectors of the hyperbolic part of the IBL equations are different from those of the diffusive part (102). In other words, the advection and diffusion matrices do not commute. According to the results of section 3.4 this means that the shock curves are influenced substantially by the diffusive terms: the shock profiles are not tangent to the characteristics, causing a deviation (46) from the Rankine-Hugoniot curve of first order in the shock strength ϵ.

Since the separation process is associated with converging characteristics and even 'shocks' (see [33] and [19]), this makes it hard to obtain physically correct predictions of separation *by any discretization scheme*. Indeed, even in the non-discretized form there are significant effects of the second order terms, even though the system is hyperbolic up to order Re_L^{-1} for smooth solutions (see section 6.5). This observation could not be made if the IBL equations were derived in the traditional way from the Prandtl equations (in which case the assumptions exclude separation). Of course, the resulting error should be compared to other sources of error such as the empirical closure relations. Conversely, the presence of the error should be taken into account when calibrating the closure model. Finally, the ESPC scheme [6] is suggested as a possible remedy to the convergence error.

7 Conclusions

It has been shown that the interacting boundary layer equations do not form a set of conservation laws, but rather a more general nonconservative hyperbolic system. The difficulties in defining and modelling these type of systems originate from their dependence on the limit of the physical process from which they are derived; in contrast with conservation laws, whose definition and modelling do not require such additional information. In case of the interacting boundary layer, this information is contained in the averaged Navier-Stokes formulation for the shear layer, derived this thesis. The second order diffusive terms, neglected in the hyperbolic system, can be used to determine the correct behaviour.

A more important contribution of this thesis, is the understanding how this additional physical information affecting the solution is to be implemented in a numerical approximation, more specifically in a space-time implicit discontinuous Galerkin framework. The limiting behaviour of the numerical approximation is determined by two aspects: the choice of physical paths in the solution space Ω (as introduced by Dal Maso, LeFloch and Murat), and the choice of numerical test function paths in Ω^* (as introduced in this thesis). Existing numerical approaches can be divided into two categories, based on converge to physically relevant solutions is obtained. The *well-balanced* schemes focus on the use of correct solution paths where available. These schemes preserve equilibrium solutions. The *entropy conserving* schemes impose an additional conservation principle by construction of a special artificial diffusion term. The relation between the artificial diffusion terms and the test function paths (for central, entropy-conserving, Godunov, Roe, modified Rusanov, and Osher solvers) has been established in section 4.1. The existence of such a relation is a requirement for well-balancing, which is attained when using the physically correct solution paths. Since the exact paths for the interacting boundary layer system are not available analytically, simplified linear solution paths can be used to obtain convergence to the physical solution, as long as the numerical stabilization is consistent with the higher order diffusive terms in the limiting physical process, as in the ESPC scheme.

An implicit space-time path-consistent quadrature-free discontinuous Galerkin framework for solving nonlinear, nonconservative hyperbolic systems has been created and implemented in Fortran. This framework is suitable to the interacting boundary layer system, as it provides low numerical diffusion and accounts for the nonconservative product in time direction, not found in other solvers. It is well-documented, modular and expandable to allow for easy integration of models developed in future research. The implementation reflects the clear separation in this thesis between physical model and numerical scheme. To allow the same solver to be used for both 2-dimensional and 3-dimensional boundary layers, the framework and numerical scheme are implemented in an n-dimensional way. Ideally, only the model requires modification when the step to three dimensions will be made.

The unsteady integral boundary layer equations are well known in their original form as derived from the Prandtl equations. The current approach of applying a control volume averaging to the full Navier-Stokes equations in a streamline-enclosed region, provides a more complete model. This is useful in a number of ways: by reducing the required assumptions, it is noted that the system is valid even for separated flow and at stagnation, albeit including some terms not present in the original model. Secondly, the second order regularization of the hyperbolic system is obtained directly, which is needed for the definition of weak solutions to the system.

8 Recommendations

A nonconservative system should be tested to see if it is a conservation law, and if possible reformulated as such for easier and more reliable numerical approximation. For any genuinely nonconservative system, path-consistency should be enforced for any numerical scheme, which corresponds to a conservative numerical stabilization.

The implications of the nonconservative properties of the unsteady interacting boundary layer equations for the actual solutions of unsteady attached and separated flows have only been considered theoretically, and will have to be investigated further by numerical experiments. The occurence of shock-like solutions in unsteady separating flow is expected (examples exist in literature), providing a big challenge for the numerical schemes.

Before applying simple models, such as the blade element momentum (BEM) method for wind turbines on a given situation, one should check if the investigated situation falls within the assumptions of the method. If this is not the case, instead of introducing empirical corrections into an invalid model to account for the observed or expected effects, one should turn to the next better model that includes such effects based on first principles. On the other hand, when creating a model, reduce as much as possible the number of tuneable coefficients. Boundary layer correlations should be no exception.

While quadrature-free techniques can provide efficiency gains in discontinuous Galerkin solvers, they are most effective for linear problems, due to the precomputing of operators. The disadvantages include the lack of standard implementations, and involved projection operations for non-polynomial functions.

Finally, it is advisable for research of such multi-faceted problems to develop and maintain a software framework in which new methods can be tested, without needing to build everything from the ground up. Implementation of improvements and new ideas from research should be possible by adding a new module (e.g. Riemann solver, turbulence model, boundary condition, pre-conditioner) to the existing framework. This way, research efforts can remain focused on the problem at hand.

References

[1] F. Alouges and B. Merlet. Approximate shock curves of non-conservative hyperbolic equations in one space dimension. *Journal of Hyperbolic Differential Equations*, 1(4):769–788, December 2004.

[2] H. Atkins and C. Shu. Quadrature-free implementation of discontinuous Galerkin method for hyperbolic equations. *AIAA*, 36(5):775–782, May 1998.

[3] P. Bakker and B. van Leer. Lecture notes on gasdynamics, AE4-140, February 2005.

[4] H. A. Bijleveld. *A quasi-simultaneous interaction method for the determination of the aerodynamic forces on wind turbine blades*. PhD thesis, University of Groningen, 2013.

[5] E. van den Boogaard. Method for unsteady integral boundary layer equations. Master's thesis, Delft University of Technology, Kluyverweg 1, December 2010.

[6] M. J. Castro, U. S. Fjordholm, S. Mishra, and C. Parés. Entropy conservative and entropy stable schemes for nonconservative hyperbolic systems. *SIAM J. Numerical Analysis*, 51(3):1371–1391, May 2013.

[7] M. J. Castro, P. G. LeFloch, M. L. Muñoz-Ruiz, and C. Parés. Why many theories of shock waves are necessary: Convergence error in formally path-consistent schemes. *Journal of Computational Physics*, 227(17):8107 – 8129, 2008.

[8] M. J. Castro, A. Pardo, C. Parés, and E. F. Toro. On some fast well-balanced first order solvers for nonconservative systems. *Math. Comput.*, 79(271):1427–1472, 2010.

[9] J. J. Cauret, J. F. Colombeau, and A. Y. L. Roux. Discontinuous generalized solutions of nonlinear nonconservative hyperbolic equations. *Journal of Mathematical Analysis and Applications*, 139(2):552 – 573, 1989.

[10] T. Cebeci and J. Cousteix. *Modeling and computation of boundary-layer flows*. Springer, Berlin Heidelberg, 2 edition, 2005. ISBN-3-540-65010-5.

[11] T. Cebeci, M. F. Platzer, H. M. Jang, and H. H. Chen. Inviscid-viscous interaction approach to the calculation of dynamic stall initiation on airfoils. *Journal of Turbomachinery*, 115(4):714–723, 1993.

[12] N. Chalmers and E. Lorin. On the numerical approximation of one-dimensional nonconservative hyperbolic systems. *Journal of Computational Science*, 4(1-2):111–124, 2013. dx.doi.org/10.1016/j.jocs.2012.08.002.

[13] E. G. M. Coenen. *Viscous-Inviscid Interaction with the Quasi-Simultaneous Method for 2D and 3D Aerodynamic Flow*. PhD thesis, Rijksuniversiteit Groningen, 2001. ISBN: 90-367-1472-9.

[14] J. F. Colombeau. Multiplication of distributions. *Bulletin of the American Mathematical Society*, 23(2):251 – 268, 1990.

[15] G. Dal Maso, P. G. Lefloch, and F. Murat. Definition and weak stability of nonconservative products. *Journal de mathématiques pures et appliquées*, 74(6):483–548, 1995.

[16] L. van Dommelen and S. Shen. The spontaneous generation of the singularity in a separating laminar boundary layer. *Journal of Computational Physics*, 38(2):125 – 140, 1980.

[17] M. Drela. *Two-Dimensional Transonic Aerodynamic Design and Analysis using the Euler Equations*. PhD thesis, Massachusetts Institute of Technology, 1984.

[18] M. Drela. XFOIL: An analysis and design system for low Reynolds number airfoils, 1989.

[19] M. Drela. Three-dimensional integral boundary layer formulation for general configurations. In *Fluid Dynamics and Co-located Conferences*. American Institute of Aeronautics and Astronautics, June 2013.

[20] M. Dumbser and E. F. Toro. A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. *Journal of Scientific Computing*, 48(1-2):70–88, 2011.

[21] A. van Garrel. Development of a wind turbine aerodynamics simulation module. ECN rapport ECN-C–03-079, ECN, August 2003.

[22] A. van Garrel. Development of a wind turbine rotor flow panel method. Report ECN-E–11-071, ECN, january 2012.

[23] A. Gawel, C. Tam, and K. Breen. Energy technology perspectives 2012: pathways to a clean energy system. Technical report, International Energy Agency (IEA), 9 rue de la Fédération, 75739 Paris Cedex 15, France, 2012.

[24] M. I. Gerritsma. Computational fluid and structucal dynamics, September 2006.

[25] E. Godlewski and P. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer, 1st edition, 1996. ISBN 0-387-94529-6.

[26] S. Goldstein. On laminar boundary-layer flow near a position of separation. *Quarterly Journal of Mechanics and Applied Mathematics*, 1(1):43–69, 1948.

[27] I. Good and T. Tideman. Integration over a simplex, truncated cubes, and eulerian numbers. *Numerische Mathematik*, 30:355–367, 1978.

[28] J. Hesthaven and T. Warburton. *Nodal discontinuous Galerkin methods, algorithms, analysis and applications*. Springer, 1 edition, 2008.

[29] B. van 't Hof and A. E. P. Veldman. Mass, momentum and energy conserving (MAMEC) discretizations on general grids for the compressible Euler and shallow water equations. *Journal of Computational Physics*, 231(1):4723–4744, 2012.

[30] F. Q. Hu and H. L. Atkins. Eigensolution analysis of the discontinuous Galerkin method with nonuniform grids: I. one space dimension. *Journal of Computational Physics*, 182(2):516 – 545, 2002.

[31] M.-S. Liou, C.-H. Chang, and T. G. Theofanous. How to solve compressible multifluid equations: A simple, robust, and accurate method. *AIAA Journal*, 46(9):2345–2356, september 2008.

[32] J. R. Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, 1:179–191, 1985.

[33] M. Matsushita and T. Akamatsu. Numerical computation of unsteady laminar boundary layers with separation using one-parameter integral method. *JSME*, 28(240):1044–1048, June 1985.

[34] M. Matsushita, S. Murata, and T. Akamatsu. Studies on boundary-layer separation in unsteady flows using an integral method. *J.Fluid Mech.*, 149:477–501, 1984.

[35] M. L. Muñoz-Ruiz and C. Parés. Godunov method for nonconservative hyperbolic systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41:169–185, 1 2007.

[36] B. Nishida. *Fuly simultaneous coupling of the full potential equation and the integral boundary layer equations in three dimensions*. PhD thesis, Massachusetts Institute of Technology, February 1996.

[37] H. Özdemir. *High-order Discontinuous Galerkin Method on Hexahedral elements for aeroacoustics*. PhD thesis, University of Twente, 2006.

[38] H. Özdemir. Unsteady three-dimensional integral boundary layer equations. Confidential report ECN-X–09-044, ECN, 2009.

[39] H. Özdemir. RotorFlow II: Status report. Confidential report ECN-X–12-029, ECN, 2012.

[40] C. Parés and M. Castro. On the well-balance property of Roe's method for nonconservative hyperbolic systems. applications to shallow-water systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 38(5):821–852, 2004.

[41] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes in Fortran 77*, volume 1 of *Fortran Numerical Recipes*. Cambridge University Press, 2 edition, 1992. ISBN 0-521-43064-X.

[42] S. Rhebergen, O. Bokhove, and J. J. W. v/d Vegt. Discontinuous Galerkin finite element methods for hyperbolic nonconservative partial differential equations. *Journal of Computational Physics*, 227:1887–1922, 2008.

[43] R. v. Rooij. Modification of the boundary layer in XFOIL for improved stall prediction. Report IW-96087R, Delft University of Technology, Delft, The Netherlands, September 1996.

[44] H. Schlichting. *Boundary layer theory*. Springer, Berlin, 8th edition, 2000. ISBN 3-540-66270-7.

[45] G. Stewart. On the continuity of the generalized inverse. *SIAM Journal on Applied Mathematics*, 17(1):33–45, 1969.

[46] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, 3rd edition, 2009. ISBN 978-3-540-25202-3.

[47] A. E. P. Veldman. New, quasi-simultaneous method to calculate interacting boundary layers. *AIAA journal*, 19(1):79–85, 1981.

[48] A. E. P. Veldman. A simple interaction law for viscous–inviscid interaction. *Journal of Engineering Mathematics*, 65(4):367–383, 2009.

[49] F. White. *Viscous fluid flow*. McGraw-Hill, 3rd edition, 2006. ISBN 0-07-124493-X.

[50] Z. Zhang, F. Liu, and D. Schuster. Calculations of unsteady flow and flutter by an Euler and integral boundary-layer method on cartesian grids. In *Proc. 22nd Applied Aerodynamics Conference*. AIAA, August 2004.

[51] J. van Zwieten, 2013. Personal Communication.

A Projection of nonlinear terms

The flux and source terms will be represented by their projections on the finite element space. Since the residuals are only tested in this space, any part orthogonal to the finite element space inherently does not contribute to the equation. The divergence of the flux, the source and the nonconservative product together form the residual, so it must be possible to compute their projection on the finite element space accurately. Therefore the flux should be computed up to order $p+1$, the source and nonconservative product up to order p. The representation of the unknowns is not specified, but they should allow to compute the residual terms up to the required accuracy. Choosing the temporal flux $(q_e\delta^*, q_e^2\varepsilon)$ as the unknowns, their order should be $p+1$ to achieve the design order. The required accuracy of the other variables is then determined by the computational model. For both conservative and nonconservative product and source terms, per vector component there are two cases to distinguish:

- The component consists of sums and/or products of the primitive variables
- The component includes rational and/or transcendental functions

In the first case, the primitive variables can be represented as polynomials of order $p+1$ for the conservative flux and order p for the nonconservative product and source terms: this will make the projection of the residual onto the finite element basis exact. In the second case, the projection is not determined by a finite subspace in terms of the primitive variables. In the IBL equations, the closure relations are usually of this form [14]. In order to deal with this, some simplifications are made:

- Rational functions are projected using a limited number of terms in the numerator and denominator
- Transcendental functions are approximated by limited taylor series

The limits should be chosen as low as possible without the incurred error becoming dominant. The required order for the remaining variables (q_e, p^*, and the closure set $\{C_f, H_k, B_u, \mathrm{Re}_\delta^{-1}, \mathcal{D}\}$) can be inferred from these limits.

A.1 Projection of Fourier series onto polynomials

An often used basis for approximating functions on a finite domain is the Fourier or harmonic basis. The space used for the finite element approximation consists of polynomials, so a transformation from harmonic to polynomial representation is sought. Since both harmonics and polynomials can approximate any continuous function, there should be a relation between these approximate representations. However, because the approximation spaces do not overlap, there does not exist an exact transformation (except for trivial cases) and a 'closest match' is the best to be expected. Such a closest match relation (defined by projection) exists, is unique and is symmetric.

[14] i.e. the nondimensionalization involved in the closure set leads to rational functions such as the shape factor H.

To find an explicit form of this relation, it is sufficient to consider the projection of a Fourier-cosine series F^N of order N onto the space of piecewise polynomials P^p of order p.

$$f(x) \in F^N : \quad f(x) = c_0 + \sum_{n=1}^{N} c_n \cos(n\pi x) \tag{109}$$

The projection onto the space of polynomials can be computed by considering the projection onto basis functions for the polynomial space. This can be done analytically for any type of polynomial basis, but the derivation is most straightforward using monomials. A transformation can be used if a change of basis is desired. The basis monomials are defined on a reference element $\hat{K} : [-1, 1]$.

$$\phi_n(\xi) = \xi^n, \quad \xi \in \hat{K} \tag{110}$$

Using this definition, we project $f(x)$ onto the space $P^p(\hat{K})$ using the basis 110. This projection will yield the coefficients of the local projection for the reference element. To make this derivation valid for all elements, the transformation from reference to physical space is defined in 111, where J is the Jacobian of this transformation.

$$x = J\xi + r \tag{111}$$

The projection is performed using the L^2 inner product:

$$\boldsymbol{u}_{\text{proj}} = \mathbf{M}^{-1}\boldsymbol{u}_\phi \tag{112}$$

$$[\boldsymbol{u}_\phi]_j = \langle\, f(J\xi + r), \phi_j(\xi)\,\rangle_{\hat{K}} \tag{113}$$

$$[\mathbf{M}]_{ij} = \langle\, \phi_j, \phi_i\,\rangle_{\hat{K}} \tag{114}$$

Instead of considering the whole summation for (109), the complex exponential $p_0(x) = \mathrm{e}^{ikx}$ will be projected. The summation is then just a linear combination of complex exponentials, and the result can be substituted term by term into the summation. So we start by projecting $p_0(x)$ onto $P^p(\hat{K})$ using basis $\phi_n(\xi) = \xi^n$, and write out the integrals from the inner product.

$$\begin{aligned}\langle\, p_0(J\xi + r), \phi_n(\xi)\,\rangle_{\hat{K}} &= \int_{\hat{K}} \mathrm{e}^{ik(J\xi+r)}\xi^n d\xi \\ &= \frac{-i}{kJ}\left(\left[\xi^n \mathrm{e}^{ik(J\xi+r)}\right]_{\xi=-1}^{\xi=1} - \int_{\hat{K}} \mathrm{e}^{ik(J\xi+r)} n\xi^{n-1} d\xi\right)\end{aligned} \tag{115}$$

For $n = 0$ this expression is readily evaluated, since the integral term in 115 drops out. For $n > 0$, algebraic expressions can be obtained by applying partial integration recursively until the integral term drops out. To do so, we give names to the factors under the integral after m partial integrations: p_m for the exponential factor and q_m for the monomial factor. To see what happens to these factors, we do another partial integration.

$$\begin{aligned}\langle\, p_0(J\xi+r), \phi_n(\xi)\,\rangle_{\hat{K}} &= \int_{\hat{K}} p_0 q_0 d\xi \\ &= [p_1 q_0]_{\xi=-1}^{\xi=1} - \int_{\hat{K}} p_1 q_1 d\xi \\ &= [p_1 q_0]_{\xi=-1}^{\xi=1} - [p_2 q_1]_{\xi=-1}^{\xi=1} + \int_{\hat{K}} p_2 q_2 d\xi\end{aligned}$$

The formulae for p_m and q_m are

$$p_m = \int_{\hat{K}} p_{m-1} d\xi = \left(\tfrac{-i}{kJ}\right)^m \mathrm{e}^{ik(J\xi+r)} \tag{116}$$

$$q_m = \frac{\partial q_{m-1}}{\partial \xi} \quad = \begin{cases} \frac{n!}{(n-m)!}\xi^{n-m}, & m \le n \\ 0, & m > n \end{cases} \tag{117}$$

Now it is apparent that the integral term will drop out after $m = n+1$ partial integrations. The result will then be given by (118).

$$\begin{aligned}\langle\, p_0(J\xi+r), \phi_n(\xi)\,\rangle_{\hat{K}} &= \sum_{m=0}^{n} (-1)^m \, [p_{m+1} q_m]_{\xi=-1}^{\xi=1} \\ &= \sum_{m=0}^{n} (-1)^m \left(\tfrac{-i}{kJ}\right)^{m+1} \mathrm{e}^{ikr} \tfrac{n!}{(n-m)!} \left(\mathrm{e}^{ikJ} - (-1)^{n-m}\mathrm{e}^{-ikJ}\right) \\ &= -2i\mathrm{e}^{ikr}\sin(kJ) \sum_{\substack{m=0 \\ m-n \text{ even}}}^{n} i^{m+1} \left(\tfrac{1}{kJ}\right)^{m+1} \tfrac{n!}{(n-m)!} \\ &\quad - 2\mathrm{e}^{ikr}\cos(kJ) \sum_{\substack{m=0 \\ m-n \text{ odd}}}^{n} i^{m+1} \left(\tfrac{1}{kJ}\right)^{m+1} \tfrac{n!}{(n-m)!}\end{aligned} \tag{118}$$

This is in fact a pair of waves times polynomials in $(kJ)^{-1}$ of order $n+1$. Both of the polynomials will be either purely real or imaginary, due to splitting the summation in odd and even parts. Introducing the real-valued polynomials P_n allows us to obtain a more convenient formulation of (118) in the form of (119).

$$P_n^{\text{odd}}(kJ) = \sum_{\substack{m=1 \\ m \text{ odd}}}^{n+1} (-1)^{\frac{m-1}{2}} \left(\tfrac{1}{kJ}\right)^m \tfrac{n!}{(n-m+1)!}, \quad P_n^{\text{even}}(kJ) = \sum_{\substack{m=2 \\ m \text{ even}}}^{n+1} (-1)^{\frac{m-2}{2}} \left(\tfrac{1}{kJ}\right)^m \tfrac{n!}{(n-m+1)!}$$

$$\begin{aligned}&\langle\, \exp(ik(J\xi+r)), \phi_n(\xi)\,\rangle_{\hat{K}} \\ &\quad = \begin{cases} 2\mathrm{e}^{ikr}\left(P_n^{\text{even}}(kJ)\cos(kJ) - iP_n^{\text{odd}}(kJ)i\sin(kJ)\right) & \text{for even } n \\ 2\mathrm{e}^{ikr}\left(P_n^{\text{even}}(kJ)i\sin(kJ) - iP_n^{\text{odd}}(kJ)\cos(kJ)\right) & \text{for odd } n \end{cases}\end{aligned} \tag{119}$$

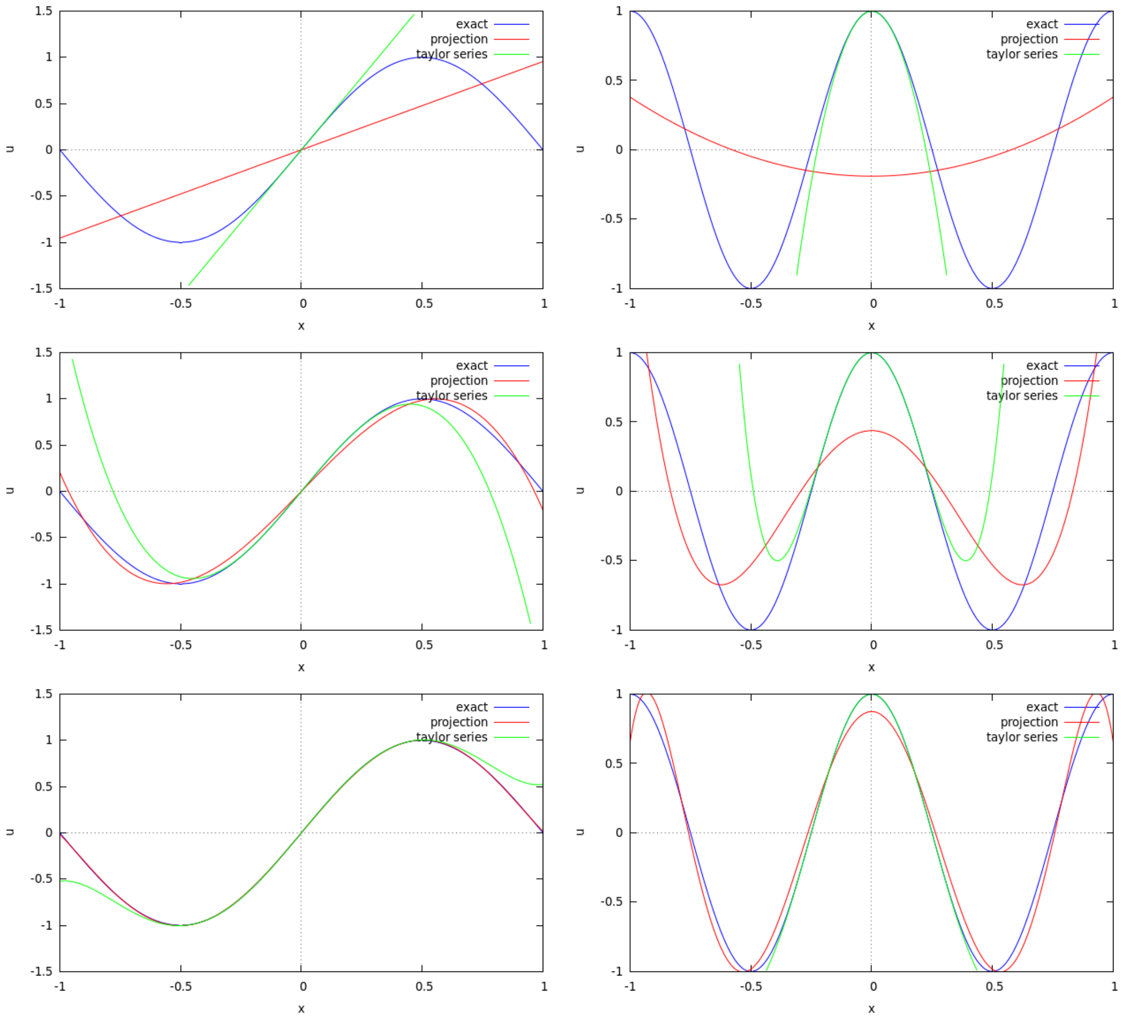

Figure 21: *Left: Projections of sin(πx) of order (1,3,5) Right: Projections of cos$(2\pi x)$ of order (2,4,6)*

To arrive at the projection of the cosine terms in (109), we take the real part of (119).

$$\begin{aligned}&\langle \cos(k(J\xi + r)), \phi_n(\xi) \rangle_{\hat{K}} \\ &\qquad = \begin{cases} 2\cos(kr)\left(P_n^{\text{odd}}(kJ)\sin(kJ) + P_n^{\text{even}}(kJ)\cos(kJ)\right) & \text{for even } n \\ 2\sin(kr)\left(P_n^{\text{odd}}(kJ)\cos(kJ) - P_n^{\text{even}}(kJ)\sin(kJ)\right) & \text{for odd } n \end{cases}\end{aligned} \tag{120}$$

Evaluating this expression for $n = 0 \ldots p$ gives the vector $\boldsymbol{u}_\phi$, from which the projection coefficients are obtained by solving with the mass matrix.

A.2 Special/limiting cases

As a special case of the preceding projection operation, one can consider the transformation from Fourier basis to monomial basis (for a single element). To obtain a convenient formulation let $J = \pi$ and $r = 0$. Then (119) becomes the Fourier series of $\phi_n(\xi)$ (121). This special case is used to study the convergence and behaviour of the high frequency end ($kJ > 1$) end of the projection. The results are shown in figure 22. Note that the projection error is a decreasing sequence, whereas the taylor expansion error is not. When evaluating $\varepsilon_{\text{proj}}$ (122) using $\boldsymbol{u}_\phi =$

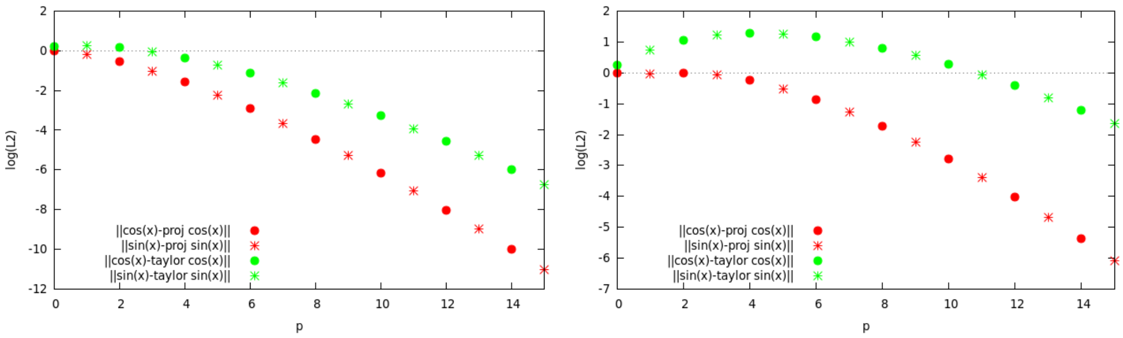

Figure 22: *Convergence of the projection (red) and taylor series (green) for increasing p, with $k = 1$ (left) and $k = 2$ (right)*

$\langle \cos(k\pi\xi), \phi_n(\xi) \rangle_{\hat{K}}$ (the real part of 121), for a specific order p an analytical expression for $\varepsilon_{\text{proj}}(k)$ is obtained. To evaluate this expression numerically for orders $p > 8$ requires more than the 52 bits of accuracy provided by double precision floating point hardware[15].

$$\langle \exp(ik\pi\xi), \phi_n(\xi) \rangle_{\hat{K}} = \begin{cases} 2P_n^{\text{even}}(k\pi)(-1)^k & \text{for even } n \\ -2iP_n^{\text{odd}}(k\pi)(-1)^k & \text{for odd } n \end{cases}, k \in \mathbb{Z} \tag{121}$$

$$\begin{aligned} \varepsilon_{\text{proj}}^2 &= \|u_{\text{exact}}\|_{L^2(\Omega)}^2 - \boldsymbol{u}_{\text{proj}}^T \mathbf{M} \boldsymbol{u}_{\text{proj}} \\ &= \|u_{\text{exact}}\|_{L^2(\Omega)}^2 - \boldsymbol{u}_{\phi}^T \mathbf{M}^{-1} \boldsymbol{u}_{\phi} \end{aligned} \tag{122}$$

$$\varepsilon_{\text{proj}}^2(\text{k=1}) = \begin{cases} 1 & p = 0 \\ 1 - \frac{90}{\pi^4} & p = 2 \\ 1 - \frac{1890\pi^4 - 37800\pi^2 + 198450}{\pi^8} & p = 4 \\ 1 - \frac{13356\pi^8 - 1413720\pi^6 + 52827390\pi^4 - 681080400\pi^2 + 2809456650}{\pi^{12}} & p = 6 \\ 1 - \frac{57420\pi^{12} - 18378360\pi^{10} + 2347294950\pi^8 - 133005272400\pi^6}{\pi^{16}} & \\ \quad - \frac{+3441584396250\pi^4 - 37253395179000\pi^2 + 139700231921250}{\pi^{16}} & p = 8 \end{cases} \tag{123}$$

The low frequency limit $kJ \to 0$ also has a small problem: the polynomials P_n^{odd} and P_n^{even} contain larger and larger terms. This makes the result (120) invalid for the case $k = 0$. However, the result for that case can be obtained in a simpler way (124).

$$\begin{aligned} \langle 1, \phi_n(\xi) \rangle_{\hat{K}} &= \int_{-1}^{1} \xi^n d\xi \\ &= \begin{cases} \frac{2}{n+1} & \text{for even } n \\ 0 & \text{for odd } n \end{cases} \end{aligned} \tag{124}$$

[15] The largest integer in this expression is $\mathcal{O}(1 \cdot 10^{15}) = \mathcal{O}(2^{50})$. Since ε is obtained by taking differences between these large numbers, catastrophic cancellation will occur when insufficient significant digits are used. This results in a sudden major loss of precision after this limit.

A.3 Projection of rational functions onto polynomials

It is also useful to project a different class of functions, formed by a ratio of polynomials up to order n/m. At first, only polynomials in one variable x are considered. For a rational function $f(x) = c(x)/d(x)$, the abscissa onto a single basis function $\phi(x)$ is

$$f_\phi = \langle\, c(x)/d(x)\, , \phi(x)\, \rangle_K \, .$$

The components of f are found by multiplying by the inverse mass matrix. To evaluate this integral, long division is applied to $f(x)\phi(x)$ if the order of $c(x)\phi(x)$ equals or exceeds the order of $d(x)$, resulting in

$$f_\phi = \int_K q(x)\phi(x) + r(x)\phi(x)/d(x)\ \mathrm{d}x$$

where the order of $r(x)\phi(x)$ is less than the order of $d(x)$. The polynomials $q(x)$ and $r(x)$ are the quotient and the remainder, respectively. Note that q and r depend on the basis function ϕ. Then the polynomial $d(x)$ is factorised into first-order components based on its (complex) roots z_i. By a partial fraction decomposition, the residues a_i at each root are found, such that

$$f_\phi = \int_K \left(q(x)\phi(x) + \sum_i \frac{a_i}{x - z_i} \right) \mathrm{d}x.$$

The factorization of $d(x)$ is usually the most resource-demanding.

Then for each zero z_i (or complex pair of zeros z_i, z_{i+1}), the contribution to f_ϕ can be expressed analytically using

$$\int \frac{1}{x-z} dx = \ln|x - z| + C \tag{125}$$

$$\int \frac{1}{(x-z-yi)(x-z+yi)} dx = \frac{1}{|y|} \arctan \frac{x-z}{|y|} + C \tag{126}$$

$$\int \frac{x-z}{(x-z-yi)(x-z+yi)} dx = \frac{1}{2} \ln|x^2 + y^2 + z^2 - 2zx| + C \tag{127}$$

The first integral is singular at $x = z$, therefore the value is undefined when the region of integration includes this point. However, the limit of the integral (excluding a vanishingly small region around the singular point) does exist in this case. It is called the Cauchy Principal Value (CPV) of this integral, and is obtained by setting the constant of integration C for both regions ($x > z$ and $x < z$) equal.

So the algorithm for determining a projection from a rational function $f(x) = c_n(x)/d_m(x)$ to a polynomial $p_p(x)$ is:

1 Factor the divisor d_m, obtaining the m zeros z_k, $k \in \{1..m\}$.

2 For each basis function ϕ_j, $j \in \{1..p\}$:

2.1 Multiply the numerator c_n with the basis function ϕ_j.

2.2 Divide the result by d_m to obtain the quotient q_{n+j-m} and the remainder r_{m-1}.

2.3 Perform partial fraction decomposition on r_{m-1}/d_m, obtaining the m residues a_k.

2.4 Compute:

- the simple integral over the element of the polynomial q_{n+j-m}.
- the contribution of single poles (z_k, a_k) using equation 125.
- the contribution of complex pole pairs using equation 126.

2.5 Add the above contributions to obtain the integral $(f_\phi)_j$

3 Multiply by the inverse mass matrix $[\mathbf{M}]_{ij}$ to obtain the coefficients p_i, $i \in \{1..p\}$.

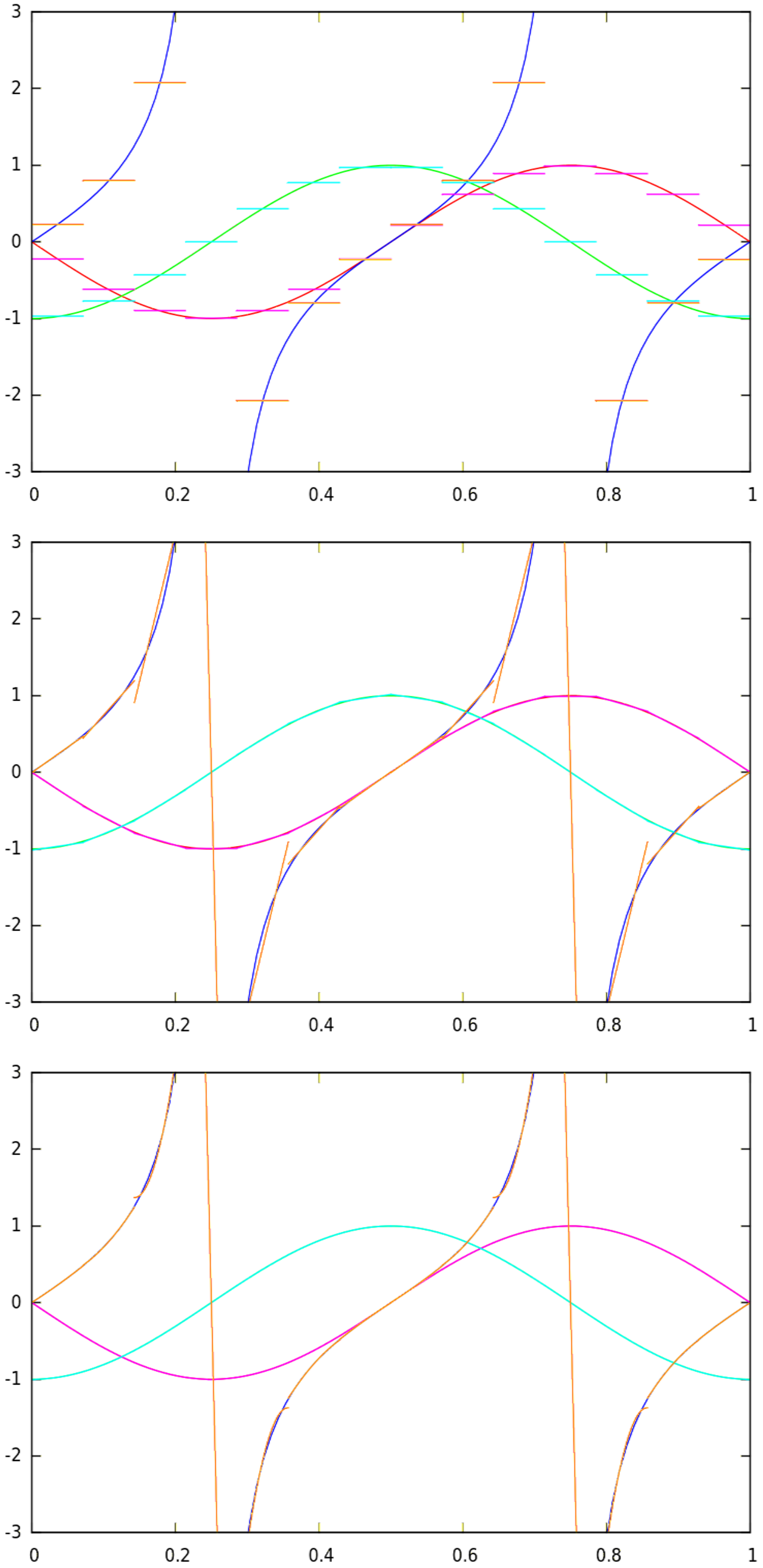

Figure 23: *Projection of sine, cosine and of the ratio of polynomials for* $p = 0, 1, 2$*.*

B Eigenvalue problems

Since not everyone seems familiar with the generalized eigenvalue problem, the following sections will provide the background. The generalized eigenvalue problem is treated completely parallel to the regular eigenvalue problem.

B.1 Regular eigenvalue problems

The regular eigenvalue problem for a square n by n matrix A is to find nontrivial scalar-vector pairs $(\lambda_k, \boldsymbol{v}_k)$ that solve

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

The trivial solution $\boldsymbol{v} = 0$ is excluded. To find a nontrivial solution, neccessarily $(A - \lambda I)$ must be singular, leading to an equation for the determinant,

$$(A - \lambda I)\boldsymbol{v} = \mathbf{0} \qquad \Longrightarrow \qquad \det(A - \lambda I) = 0.$$

By inspection of the the above condition, $\boldsymbol{v}$ is in the null space of $(A - \lambda I)$ and orthogonal to the row space of $(A - \lambda I)$. Only if A is diagonalisable, there exist n distinct solutions to the eigenvalue problem, whose eigenvectors are linearly independent.

B.2 Generalized eigenvalue problems

The generalized eigenvalue problem for a pair of square matrices A and B is to find nontrivial scalar-scalar-vector triples $(\alpha_k, \beta_k, \boldsymbol{r}_k)$ that solve

$$\alpha A\boldsymbol{r} = \beta B\boldsymbol{r}.$$

The trivial solution $\boldsymbol{r}$=0 is exluded, as well as α=β=0. To find a nontrivial solution, neccessarily $(\alpha A - \beta B)$ must be singular, leading to any of the following equations for the determinant,

$$\det(\alpha A - \beta B) = 0 \quad \text{or} \quad \det(A - \lambda B) = 0 \quad \text{or} \quad \det(\mu A - B) = 0.$$

Hence the generalized eigenvalues can be uniquely identified by $\lambda_k = \beta_k/\alpha_k$ when α_k is nonzero, or by $\mu_k = \alpha_k/\beta_k$ when β_k is nonzero. To have n distinct solutions to the generalized eigenvalue problem (with linearly independent eigenvectors), it is sufficient that A and B are simultaneously diagonalizable. In that case, the matrices A and B also commute. This is not required; another sufficient condition is A and B symmetric, with at least one of the two positive definite.

The generalized eigenvector $\boldsymbol{r}_k$ is called a right eigenvector, since it multiplies the matrices from the right. Effectively this means it is orthogonal to the row space of $(\alpha A - \beta B)$. The problem can also be given in terms of the column space,

$$\alpha \boldsymbol{l}^{\mathrm{T}} A = \beta \boldsymbol{l}^{\mathrm{T}} B.$$

in which case $\boldsymbol{l}^{\mathrm{T}}$ are the left generalized eigenvectors. They correspond to the same eigenvalues, since the singularity of $(\alpha A - \beta B)$ makes both the row rank and the column rank are strictly less than n. Only if both matrices A and B are symmetric, the left and right eigenvectors are the same $\boldsymbol{l}_k = \boldsymbol{r}_k$, and moreover they become pairwise orthogonal $\frac{\boldsymbol{r}_i}{\|\boldsymbol{r}_i\|} \cdot \frac{\boldsymbol{r}_j}{\|\boldsymbol{r}_j\|} = \delta_{ij}$.

B.3 Proofs

Given that two n by n matrices A and B are simultaneously diagonalisable, i.e.

$$\begin{aligned} A &= RDR^{-1} \\ B &= RER^{-1} \end{aligned}$$

with D and E diagonal matrices, R an invertible matrix. It is claimed that there are n independent solutions to the generalised eigenvalue problem for A and B, and that A and B commute. The last property is straightforward,

$$AB = RDER^{-1} = REDR^{-1} = BA.$$

Given the two diagonal matrices, a sequence $(\alpha_k, \beta_k)_{k=1...n}$ can be chosen that satisfies

$$D_{kk}\alpha_k = E_{kk}\beta_k \quad \forall k \in \{1...n\} \quad \text{i.e.} \quad D\ \mathrm{diag}(\alpha) = E\ \mathrm{diag}(\beta).$$

where at least one of α_k or β_k is nonzero for each k. After premultiplying with R, it is found that the individual columns of R contain the n independent solutions

$$AR\ \mathrm{diag}(\alpha) = BR\ \mathrm{diag}(\beta) \quad \text{i.e.} \quad \alpha_k A\boldsymbol{r}_k = \beta_k B\boldsymbol{r}_k \quad \forall k \in \{1...n\} \quad \square$$

Given two n by n matrices: A symmetric and B symmetric positive definite, it is claimed that there are n independent solutions to the generalised eigenvalue problem. Because B is s.p.d. it can be decomposed as $B = LL^{\mathrm{T}}$ (Cholesky) with a unique orthogonal matrix L. Applying a symmetry-preserving transformation to A, a diagonalizable matrix G is obtained, with a full set of eigensolutions $(\lambda_k, \boldsymbol{y}_k)$, $k \in \{1...n\}$:

$$G \overset{\text{def}}{=} L^{-1}AL^{-\mathrm{T}} \qquad\qquad G\boldsymbol{y}_k = \lambda_k \boldsymbol{y}_k \quad \forall k \in \{1...n\}$$

Substituting $\boldsymbol{y}_k = L^{\mathrm{T}}\boldsymbol{r}_k$ and premultiplying the equations with L,

$$A\boldsymbol{r}_k = \lambda_k B\boldsymbol{r}_k \quad \forall k \in \{1...n\} \quad \square$$