UNIVERSITETET I BERGEN

Det matematisk-naturvitenskapelige fakultet

On the energy stability of high-order finite volume schemes for initial-boundary value problems

Master thesis in Applied and Computational Mathematics

by

Thomas Bjarne Hestvik

University of Bergen,
Department of Mathematics,

June 2023

Abstract

We examine the energy stability of high-order finite volume schemes approximating linear hyperbolic initial-boundary value problems. In particular, we consider schemes obtained by the k-exact method and the spectral volume method using the central numerical flux. To determine the stability of the schemes we use the energy method, and investigate the resulting terms. Finally, we compute numerical results verifying the accuracy of the schemes.

Acknowledgements

I wish to thank my supervisor, professor Magnus Svärd for his guidance throughout the project and for helping me find an interesting research topic. I would also like to thank him for being available to discuss mathematics at any time. Further, I would like to thank all the people at the Department of Mathematics at University of Bergen for making my time here interesting and enjoyable. Last but not least a special thanks goes out to my friends and family.

Notation

$\frac{\partial}{\partial \xi} u = u_\xi = \partial_\xi u$.
$\mathbf{x} = [x_1, \ldots, x_n]^T, \qquad d\mathbf{x} = dx_1 dx_2 \ldots dx_n$.
$\Omega \subset \mathbb{R}^n$ and $\partial\Omega$ denotes the boundary of Ω. $\overline{\Omega} = \Omega \cup \partial\Omega$.
$\mathbf{u} \cdot \mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \sum_i u_i v_i$.
$\hat{n}$ = outward pointing unit normal vector on a closed curve. If ∂C_i is a closed (simple) curve and $\Gamma_{ij} \subset \partial C_i$ is a simple curve, then $\hat{n}$ on Γ_{ij} is outward pointing w.r.t. ∂C_i.
dS = infinitesimal arc length of a curve, i.e. $\sqrt{dx^2 + dy^2}$.
$\Gamma_{ij} = \partial C_i \cap \partial C_j$ where C_i, C_j are control volumes.
N_i = the set of indices j such that $\Gamma_{ij} \neq \emptyset$.
w^d_{ijq} = quadrature weights of the d-element flux over Γ_{ij}.
If $u : \Omega \times [0, T) \to \mathbb{R}$, then $\|u(\cdot, t)\|_2 = \|u(\cdot, t)\|_{L^2} = \sqrt{\int_\Omega |u(\mathbf{x}, t)|^2 d\mathbf{x}}$.
If $\mathbf{x} \in \mathbb{R}^2$, then $\|\mathbf{x}\|_{\mathbb{R}^2} = \sqrt{x_1^2 + x_2^2}$.
C^k = the linear space of k-times continuously differentiable functions. If $k = 0$ it is the space of continuous functions. In this case we drop the superscript and simply write C.
L^p = the linear space of measurable functions bounded in the L^p norm.

Contents

1 Introduction

1.1 Motivation

Recently, great progress has been made in the development of high-order accurate, conservative, and provably energy stable schemes for linear and linearized problems. For instance, energy stable weigthed essentially non-oscillatory (WENO) finite difference schemes were developed in [1, 2, 3, 4], energy stable discontinuous Galerkin (DG) spectral element schemes were developed in [5], and a series of papers [6, 7, 8, 9] prove energy stability for flux reconstruction (FR), or correction procedure via reconstruction (CPR) schemes. These papers show that schemes obtained from the different numerical methods can be viewed as summation-by-parts (SBP) schemes with simultaneous approximation terms (SAT). The summation-by-parts simultaneous approximation terms (SBP-SAT) framework is highly effective at obtaining provably stable schemes for initial-boundary value problems (IBVP) [10, 11, 12, 13, 14]. To achieve this, SBP schemes utilize discrete difference operators which satisfy the summation-by-parts property. In short, this property is the discrete counterpart to the integration by parts property which the continuous derivative operator satisfies. As integration by parts is the key to proving stability for the continuous problems, so is summation-by-parts key to proving stability for schemes.

In [15, 16] it is shown that certain finite volume methods can be formulated in the SBP-SAT framework. Further research on SBP-SAT FVM is detailed in [17, 18, 19, 20, 21, 22, 23]. In these papers, only methods giving low-order schemes are studied. Therefore, we would like to find similar results for high-order finite volume methods and their corresponding schemes. In particular, we consider schemes obtained by the k-exact method [24, 25, 26, 27, 28] and the spectral volume (SV) method [29, 30, 31, 32, 33, 34, 35].

We aim in this thesis to examine the energy stability of schemes obtained from the k-exact method and the spectral volume method. In particular we look at schemes approximating 1D and 2D linear hyperbolic IBVP. We attempt to discover k-exact schemes and spectral volume schemes which satisfy the SBP property.

1.2 Outline

The thesis is organized as follows. Section 2 presents a short introduction to the theory of linear hyperbolic problems. In section 3 we introduce finite volume methods and define stability. Section 4 presents two high-order finite volume methods, the k-exact method and the spectral volume method. Section 5 details our energy stability analysis of k-exact schemes. In section 6 we analyze the energy stability of spectral volume schemes. Section 7 presents some numerical results verifying that the schemes are high-order accurate. Finally, in section 8 we give concluding remarks and suggestions for future work.

2 Linear hyperbolic problems

In this section we recap the elementary theory of linear hyperbolic initial value problems and initial-boundary value problems.

2.1 Linear hyperbolic equations

Definition 2.1 ([36, 37]). *Let $A \in \mathbb{R}^{m\times m}$ and $u(x,t) = [u_1(x,t), ..., u_m(x,t)]^T$. The system of equations*

$$u_t + Au_x = 0,$$

is said to be hyperbolic if A is diagonalizable with real eigenvalues.

Remark. It is common to define several notions of hyperbolicity, e.g. weakly, strongly, strictly, symmetric hyperbolic. We keep the text simple and use only the notion described above.

Consider the simplest hyperbolic system, consisting of one scalar equation

$$u_t + au_x = 0, \qquad (a \in \mathbb{R}). \tag{2.1}$$

The equation (2.1) is called the advection equation or transport equation. Note that functions of the form $u(x,t) \equiv \phi(x - at)$ satisfy the equation. Observe that for some point (ξ, τ)

$$u(\xi,\tau) = \phi(\xi - a\tau) = \phi(x_0) = u(x_0, 0),$$

for $x_0 = \xi - a\tau$. Hence the value of u at any point in the xt-plane is determined by ϕ at some corresponding point x_0 on the x-line. Put in other words, u is constant along the characteristic lines $(x(t), t)$ satisfying $x'(t) = a$, $x(0) = x_0$. The initial data $u(x, 0) = \phi(x)$ is moved in the positive $(a > 0)$ or negative $(a < 0)$ x-direction as t increases. Note that (2.1) can be written as

$$u_t + f(u)_x = 0,$$

with $f(u) = au$, and we say that the equation is in conservation form. Equations in conservation form are called conservation laws, and f is called the flux function.

Next, consider a hyperbolic system of m equations in one spatial variable:

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_t + \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_x = 0. \tag{2.2}$$

Definition 2.1 tells us that the coefficient matrix is diagonalizable, and that the eigenvalues are real. Therefore, there exists matrices R, Λ, where Λ is diagonal with real elements such that $A = R\Lambda R^{-1}$. Hence (2.2) can be written as

$$u_t + R\Lambda R^{-1} u_x = 0.$$

Apply R^{-1} on the left and substitute $w = R^{-1}u$ to find

$$w_t + \Lambda w_x = 0, \qquad \text{or equivalently} \qquad \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}_t + \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_m \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}_x = 0. \tag{2.3}$$

Finding solutions for (2.3) is the same as finding solutions for (2.1); $w_i(x,t) = \phi(x - \lambda_i t)$. Thus functions $u(x,t)$ of the form

$$u(x,t) = Rw(x,t) = \sum_{i=1}^{m} \phi(x - \lambda_i t) r_i,$$

where $w_i(x,0) = \phi(x)$ and r_i is the eigenvector corresponding to λ_i, satisfy (2.2). Next we look at the variable coefficient scalar equation.

$$u_t + a(x,t)u_x = 0, \qquad (a(x,t) \in \mathbb{R}). \tag{2.4}$$

The characteristic curves satisfy $x'(t) = a(x,t)$, $x(0) = x_0$, as shown by

$$\frac{d}{dt}u(x(t),t) = u_t + x'(t)u_x = 0,$$

(using (2.4)). Note that the variable coefficient equation satisfies definition 2.1, since the eigenvalue $a(x,t) \in \mathbb{R}$ for any choice of (x,t). If $a(x,t)$ is bounded and smooth, we can solve the characterisitc ODEs backwards in time to obtain the relation $\psi(x,t) = x_0$. Then the unique solution of the variable coefficient initial value problem consisting of (2.4) with the initial value $u(x,0) = u_0(x)$ is given by

$$u(x,t) = u(x_0,0) = u(\psi(x,t),0) = u_0(\psi(x,t)).$$

Example 1 ([36, 38]). Let $u_t + (2x)u_x = 0$. The characteristic curves are the solutions of

$$\frac{d}{dt}x(t) = 2x, \qquad x(0) = x_0.$$

In other words, they are given by $(x_0 e^{2t}, t)$ ranging x_0 over the x-domain. Given a point (ξ, τ) we can find $u(\xi,\tau) = u(\psi(\xi,\tau),0)$ by solving

$$\frac{d}{dt}x(t) = -2x, \qquad x(0) = \xi,$$

up to $t = \tau$. It follows that $u(\xi,\tau) = u_0(\xi e^{-2\tau})$.

Moving on, consider the variable coefficient system.

Definition 2.2 ([36]). *Let $A(x,t) : \mathbb{R} \times [0,T] \to \mathbb{R}^{m\times m}$ and $u = [u_1(x,t), \dots, u_m(x,t)]^T$. The system of equations*

$$u_t + A(x,t)u_x = 0,$$

is said to be hyperbolic if for any pair $(x,t) \in \mathbb{R} \times [0,T]$ the matrix A is diagonalizable with real eigenvalues.

Recall how we found solutions of the constant coefficient system (2.2). Attempting the same approach, i.e. $A(x,t) = R(x,t)\Lambda(x,t)R^{-1}(x,t)$, $w = R^{-1}u$, we find

$$\begin{aligned} Rw_t + R_t w + R\Lambda R^{-1}(Rw)_x = 0 \\ \iff w_t + R^{-1}R_t w + \Lambda R^{-1}(R_x w + Rw_x) = 0 \\ \iff w_t + \Lambda w_x = -R^{-1}(R_t + \Lambda R^{-1}R_x)w. \end{aligned}$$

If the right hand side reduces to zero we recover a decoupled system $w + \Lambda(x,t)w_x = 0$ which we can solve by finding the characteristic curves as described previously.

Note that we can extend definition 2.1 to accomodate for equations where the spatial variable is of higher dimension.

Definition 2.3 ([39]). *Let $A_i \in \mathbb{R}^{m\times m}$, $u(\mathbf{x},t) = [u_1(\mathbf{x},t),\dots,u_m(\mathbf{x},t)]^T$ and $\mathbf{x} \in \mathbb{R}^n$. The system of equations*

$$u_t + \sum_{i=1}^{n} A_i u_{x_i} = 0,$$

is said to be hyperbolic if all linear combinations of the coefficient matrices A_i is diagonalizable with real eigenvalues.

Consider the simplest multi-dimensional hyperbolic system, consisting of one scalar equation in two dimensions

$$u_t + au_x + bu_y = 0, \qquad (a,b \in \mathbb{R}).$$

Note that functions $u(x,y,t) = \phi(x - at, y - bt)$ satisfy the equation, since

$$\phi_t + a\phi_x + b\phi_y = \phi_x \frac{\partial x}{\partial t} + \phi_y \frac{\partial y}{\partial t} + a\phi_x + b\phi_y = -a\phi_x - b\phi_y + a\phi_x + b\phi_y = 0.$$

Further, given some point (ξ,η,τ) we have

$$u(\xi,\eta,\tau) = \phi(\xi - a\tau, \eta - b\tau) = \phi(\xi_0,\eta_0) = u(\xi_0,\eta_0,0),$$

for $\xi_0 = \xi - a\tau$ and $\eta_0 = \eta - b\tau$. Just like the 1D case, the value of u at any point is determined by the value of $u\big|_{t=0}$ at some corresponding point. In other words, u is constant along the characteristic curves $(\xi(t),\eta(t),t)$ given by

$$\frac{d}{dt}\xi(t) = a, \qquad \xi(0) = \xi_0, \qquad \frac{d}{dt}\eta(t) = b, \qquad \eta(0) = \eta_0.$$

Now consider a two-dimensional hyperbolic system consisting of m equations

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_t + \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_x + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ b_{m1} & \cdots & \cdots & b_{mm} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_y = 0. \tag{2.5}$$

Let the coefficient matrices be denoted by A and B. Suppose that the initial data is given by $u(x,y,0) = f(x,y) = [f_1(x,y),\dots,f_m(x,y)]^T$ where $f_i \in L^2(\mathbb{R}^2)$ and f is 2π-periodic. Moreover, suppose for the moment that the Fourier series of f is given by a single wave

$$f = \frac{1}{2\pi} e^{i\langle \omega, \mathbf{x} \rangle} \hat{f}(\omega),$$

where $\omega = [\omega_1, \omega_2]^T$. We make the guess that

$$u(x,t) = \frac{1}{2\pi} e^{i\langle \omega, \mathbf{x} \rangle} \hat{u}(\omega, t), \qquad \text{where} \qquad \hat{u}(\omega, 0) = \hat{f}(\omega).$$

Substituting this into the PDE we obtain

$$u_t + Au_x + Bu_y = \frac{1}{2\pi} e^{i\langle \omega, \mathbf{x} \rangle} \hat{u}_t(\omega, t) + \frac{1}{2\pi} Ai\omega_1 e^{i\langle \omega, \mathbf{x} \rangle} \hat{u}(\omega, t) + \frac{1}{2\pi} Bi\omega_2 e^{i\langle \omega, \mathbf{x} \rangle} \hat{u}(\omega, t) = 0.$$

Define the operator $\hat{P}(i\omega)$ by

$$\hat{P}(i\omega) = i(A\omega_1 + B\omega_2),$$

to obtain

$$\hat{u}_t(\omega, t) = \hat{P}(i\omega)\hat{u}(\omega, t).$$

Now using the initial condition $\hat{u}(\omega, 0) = \hat{f}(\omega)$ we find the unique solution of the problem (2.5) with periodic initial data f,

$$u(x,t) = \frac{1}{2\pi} e^{i\langle \omega, \mathbf{x} \rangle} e^{\hat{P}(i\omega)t} \hat{f}(\omega).$$

Note that if f is given by

$$f = \frac{1}{2\pi} \sum_{\omega} e^{i\langle \omega, \mathbf{x} \rangle} \hat{f}(\omega),$$

then by the above arguments and the superposition principle we obtain the unique solution

$$u(x,t) = \frac{1}{2\pi} \sum_{\omega} e^{i\langle \omega, \mathbf{x} \rangle} e^{\hat{P}(i\omega)t} \hat{f}(\omega).$$

2.2 Well-posedness and stability

Consider the constant coefficient hyperbolic initial value, or Cauchy problem

$$u_t + P\left(\frac{\partial}{\partial \mathbf{x}}\right) u = 0, \tag{2.6}$$

$$u(\mathbf{x}, 0) = f(\mathbf{x}). \tag{2.7}$$

Let $\mathbf{x} = [x_1, \ldots, x_m]^T$ and assume $f \in L^2$ is 2π-periodic. Let $\omega = [\omega_1, \ldots, \omega_m]^T$ and suppose that

$$f(\mathbf{x}) = (2\pi)^{-(m/2)} e^{i\langle \omega, \mathbf{x} \rangle} \hat{f}(\omega).$$

Following the discussion at the end of section 2.1, the unique solution of (2.6-2.7) is given by

$$\hat{u}(\omega, t) = e^{\hat{P}(i\omega)t} \hat{f}(\omega), \qquad u(\mathbf{x}, t) = (2\pi)^{-(m/2)} e^{i\langle \omega, \mathbf{x} \rangle} e^{\hat{P}(i\omega)t} \hat{u}(\omega, t).$$

Definition 2.4 ([36, 40])**.** *The problem (2.6-2.7) is said to be stable if there exists positive constants K, α, which are independent of t, ω such that*

$$|e^{\hat{P}(i\omega)t}| \le Ke^{\alpha t}. \tag{2.8}$$

Now we show that the inequality (2.8) is equivalent with the following:

$$\|u(\cdot,t)\|_2 \leq Ke^{\alpha t}\,\|f(\cdot)\|_2\,,$$

where u is the solution of the problem and the L^2 norm is defined by

$$\|v\|_2 = \left(\int_{-\infty}^{\infty} |v(s)|^2 ds\right)^{1/2}.$$

Assume (2.8) holds, by Parseval's relation we find

$$\|u(\cdot,t)\|_2^2 = \sum_{\omega} |\hat{u}(\omega,t)|^2 = \sum_{\omega} |e^{\hat{P}(i\omega)t}\hat{f}(\omega)|^2 \leq \sum_{\omega} |e^{\hat{P}(i\omega)t}|^2|\hat{f}(\omega)|^2 \leq \sum_{\omega} |Ke^{\alpha t}|^2|\hat{f}(\omega)|^2 = |Ke^{\alpha t}|^2\,\|f(\cdot)\|_2^2\,.$$

Stability implies the solution is bounded by the initial data together with an exponential factor. Often the term *energy* is used to describe the squared L^2 norm of a function. We will therefore describe stability in the L^2 norm as energy stability.

Definition 2.5 (Hadamard, [39]). *The problem (2.6-2.7) is said to be well-posed if there exists an unique solution which depends continuously on the initial data f.*

By what we have just shown, well-posedness means that there exists an unique solution and that the problem is stable. Consider the variable coefficient initial value problem

$$u_t + P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u = 0, \tag{2.9}$$

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \tag{2.10}$$

with periodic boundary conditions and initial data. We define

$$(f,g) = \int_0^{2\pi} \cdots \int_0^{2\pi} \langle f(\mathbf{x}), g(\mathbf{x})\rangle d\mathbf{x},$$

where $d\mathbf{x} = dx_1 \ldots dx_m$.

Definition 2.6 ([36]). *The differential operator $P(\mathbf{x}, t, \partial/\partial\mathbf{x})$ is said to be semibounded if for any interval $t_p \leq t \leq T$, there is a constant α such that for all sufficiently smooth functions w,*

$$-(w, Pw) - (Pw, w) \leq 2\alpha\,\|w\|_2^2$$

Theorem 2.1 ([36]). *If the operator $P(\mathbf{x}, t, \partial/\partial\mathbf{x})$ is semibounded, then the problem (2.9-2.10) is stable.*

Proof. Let P be semibounded and let u be a solution of (2.9-2.10). Then

$$u_t + Pu = 0, \qquad \text{and} \qquad -(u, Pu) - (Pu, u) \leq 2\alpha\,\|u\|_2^2\,.$$

Note that multiplying the PDE by u and integrating over the spatial domain yields

$$(u, u_t) + (u, Pu) = 0.$$

Adding the transposed equation to the above we obtain

$$\frac{d}{dt}\|u(\cdot,t)\|_2^2 = -(u,Pu)-(Pu,u) \leq 2\alpha\|u(\cdot,t)\|_2^2,$$

for some α, which implies

$$\|u(\cdot,t)\|_2^2 \leq e^{2\alpha t}\|f(\cdot)\|_2^2.$$

□

Moving on, we give a result concerning the well-posedness of inhomogenous Cauchy problems. Consider

$$u_t + P\left(\mathbf{x},t,\frac{\partial}{\partial\mathbf{x}}\right)u = F(\mathbf{x},t), \tag{2.11}$$

$$u(\mathbf{x},0) = f(\mathbf{x}), \tag{2.12}$$

where f is 2π periodic in $\mathbf{x}$ as before and likewise for F . Assume the problem

$$u_t + P\left(\mathbf{x},t,\frac{\partial}{\partial\mathbf{x}}\right)u = 0, \qquad t\geq\tau, \tag{2.13}$$

$$u(\mathbf{x},\tau) = F(\mathbf{x},\tau), \tag{2.14}$$

is well posed for all τ and u is the unique solution. Define the solution operator $S(t,\tau)$ by

$$u(\mathbf{x},t) = S(t,\tau)u(\mathbf{x},\tau), \qquad t\geq\tau.$$

Theorem 2.2 (Duhamel's Principle, [36])**.** *Let $S(t,\tau)$ denote the solution operator of (2.13-2.14). Then the solution of the problem (2.11-2.12) can be written in the form*

$$u(\mathbf{x},t) = S(t,0)f(\mathbf{x}) + \int_0^t S(t,\tau)F(\mathbf{x},\tau)d\tau, \tag{2.15}$$

and

$$\|u(\cdot,t)\|_2 \leq K\left(e^{\alpha t}\|f(\cdot)\|_2 + \phi^*(\alpha,t)\max_{0\leq\tau\leq t}\|F(\cdot,\tau)\|_2\right),$$

where

$$\phi^*(\alpha,t) = \begin{cases}\frac{1}{\alpha}(e^{\alpha t}-1), & \alpha\neq 0\\ t, & \alpha = 0\end{cases}.$$

Next, we briefly discuss the related definitions and results for initial-boundary value problems. Consider the following problem with $f\in L^2$

$$u_t + P\left(x,t,\frac{\partial}{\partial x}\right)u = 0, \qquad 0\leq x\leq 1, \qquad 0\leq t\leq T, \tag{2.16}$$

$$u(x,0) = f(x), \tag{2.17}$$

$$L_0\left(t,\frac{\partial}{\partial x}\right)u(0,t) = 0, \tag{2.18}$$

$$L_1\left(t,\frac{\partial}{\partial x}\right)u(1,t) = 0. \tag{2.19}$$

We define stability as before

Definition 2.7 ([36]). *The problem (2.16-2.19) is said to be stable if there exists constants K, α not dependent on f such that*

$$\|u(\cdot,t)\|_2 \leq Ke^{\alpha t}\,\|u(\cdot,0)\|_2$$

Further, we define the problem to be well-posed if it has an unique solution and it is stable. We can once again use Duhamel's principle to prove well-posedness of the inhomogenous problem given that the family of homogenous problems is well-posed. In other words, if (2.16-2.18) is well-posed for $f \in L^2$ then so is

$$\begin{aligned} u_t + P\left(x,t,\frac{\partial}{\partial x}\right)u &= F(x,t), \qquad 0 \leq x \leq 1, \qquad 0 \leq t \leq T,\\ u(x,0) &= f(x),\\ L_0\left(t,\frac{\partial}{\partial x}\right)u(0,t) &= 0,\\ L_1\left(t,\frac{\partial}{\partial x}\right)u(1,t) &= 0. \end{aligned}$$

for $F \in L^2$. Now consider the problem with inhomogenous boundary data,

$$u_t + P\left(x,t,\frac{\partial}{\partial x}\right)u = F(x,t) \tag{2.20}$$

$$u(x,t_0) = f(x) \tag{2.21}$$

$$L_0\left(t,\frac{\partial}{\partial x}\right)u(0,t) = g(t) \tag{2.22}$$

Suppose $L_0\phi(0,t) = g(t)$. Let $\tilde{u} = u - \phi$ such that

$$\begin{aligned} \tilde{u}_t + P\left(x,t,\frac{\partial}{\partial x}\right)\tilde{u} &= F(x,t) - \phi_t - P\left(x,t,\frac{\partial}{\partial x}\right)\phi\\ \tilde{u}(x,t_0) &= f(x) - \phi(x,t_0)\\ L_0\left(t,\frac{\partial}{\partial x}\right)\tilde{u}(0,t) &= 0 \end{aligned}$$

Given the existence of such a $\phi(x,t)$ we see that the inhomogenous IBVP with inhomogenous boundary data is well-posed if the corresponding homogenous IBVP is well-posed (by Duhamel's principle). However, as pointed out in [36], the energy estimate for u will now depend on g_t. Hence the boundary data must be differentiable to obtain an energy estimate. To avoid this we define a notion of strong stability independent of g_t.

Definition 2.8 ([36]). *The problem (2.20-2.22) is said to be strongly stable if there exists a bounded functional $K(t,t_0)$ independent of the inital and boundary data such that*

$$\|u(\cdot,t)\|_2^2 \leq K(t,t_0)\left(\|u(\cdot,t_0)\|_2^2 + \int_{t_0}^{t}\|F(\cdot,\tau)\|_2^2 + |g(\tau)|^2 d\tau\right)$$

If the problem is strongly stable and there exists an unique solution, we say that it is strongly well-posed.

2.3 Riemann problems

A Riemann problem is an initial value problem in one spatial dimension where the initial data is given by

$$f(x) = \begin{cases} f_L, & x < p_0 \\ f_R, & x > p_0 \end{cases},$$

where p_0 is some point of discontinuity and f_L, f_R are some constants. Consider the following Riemann problem

$$u_t + au_x = 0, \qquad u(x,0) = \begin{cases} u_L, & x < 0 \\ u_R, & x > 0 \end{cases}.$$

We recall that the advection equation simply moves the initial data with the wave speed a. In other words, the unique solution is given by $u(x,t) = u(x-at,0)$. Note that

$$u(x-at,0) = \begin{cases} u_L, & x-at < 0 \\ u_R, & x-at > 0 \end{cases}.$$

Hence the solution of the Riemann problem can be visualized by dividing the (x,t)-plane into two parts seperated by the characteristic line emenating from the point of discontinuity, $x(t) = at$. This is illustrated in Fig. 1.

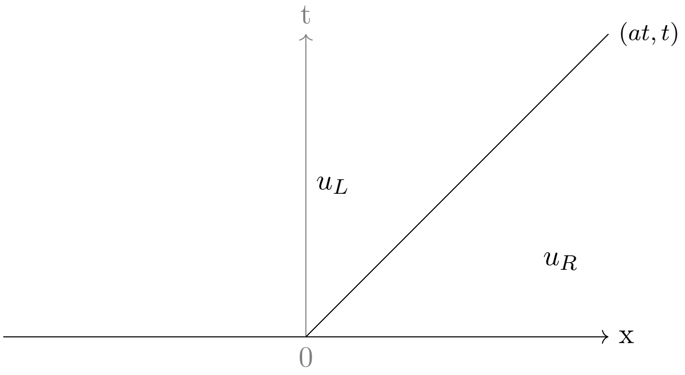


Figure 1: Solution of the simplest Riemann problem.

Consider the Riemann problem for the hyperbolic constant coefficient system,

$$u_t + Au_x = 0, \qquad u(x,0) = \begin{cases} u_L, & x < 0 \\ u_R, & x > 0 \end{cases},$$

where $u = [u_1(x,t), \ldots, u_m(x,t)]^T$, $A \in \mathbb{R}^{m\times m}$ and $u_L, u_R \in \mathbb{R}^m$. Let $A = R\Lambda R^{-1}$ for some matrices R, Λ, where Λ is diagonal. Put $w = R^{-1}u$ to find

$$w_t + \Lambda w_x = 0, \qquad w(x,0) = \begin{cases} w_L, & x < 0 \\ w_R, & x > 0 \end{cases}, \tag{2.23}$$

where w_L, w_R are defined by

$$u_L = \sum_{i=1}^{m} (w_L)_i r_i, \qquad \text{and} \qquad u_R = \sum_{i=1}^{m} (w_R)_i r_i,$$

and r_i is the eigenvector associated with the i-th eigenvalue of A. Since (2.23) is decoupled we have that

$$w_i(x,t) = \begin{cases} (w_L)_i, & x - \lambda_i t < 0 \\ (w_R)_i, & x - \lambda_i t > 0 \end{cases},$$

where $w_i r_i = u_i$. Organize the eigenvalues of A increasingly, i.e. $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$ to find

$$u(x,t) = \sum_{i=1}^{m} w_i r_i = \begin{cases} \sum_{i=1}^{m} (w_L)_i r_i, & x \leq \lambda_1 t \\ \sum_{j=1}^{i} (w_R)_i r_i + \sum_{j=i+1}^{m} (w_L)_j r_j, & \lambda_i t \leq x \leq \lambda_{i+1} t \\ \sum_{i=1}^{m} (w_R)_i r_i, & x \geq \lambda_m t \end{cases}.$$

For instance, if $m = 3$ then $u(x,t)$ is given by

$$u(x,t) = \sum_{i=1}^{3} w_i r_i = \begin{cases} \sum_{i=1}^{3} (w_L)_i r_i, & x \leq \lambda_1 t \\ (w_R)_1 r_1 + \sum_{j=2}^{3} (w_L)_j r_j, & \lambda_1 t \leq x \leq \lambda_2 t \\ \sum_{j=1}^{2} (w_R)_i r_i + (w_L)_3 r_3, & \lambda_2 t \leq x \leq \lambda_3 t \\ \sum_{i=1}^{3} (w_R)_i r_i, & x \geq \lambda_3 t \end{cases}.$$

Once again we can visualize the solution $u(x,t)$ as constant states seperated by characteristic waves, as shown in Fig. 2. Here we have shown the case for $\lambda_1 < 0, \lambda_2, \lambda_3 > 0$.

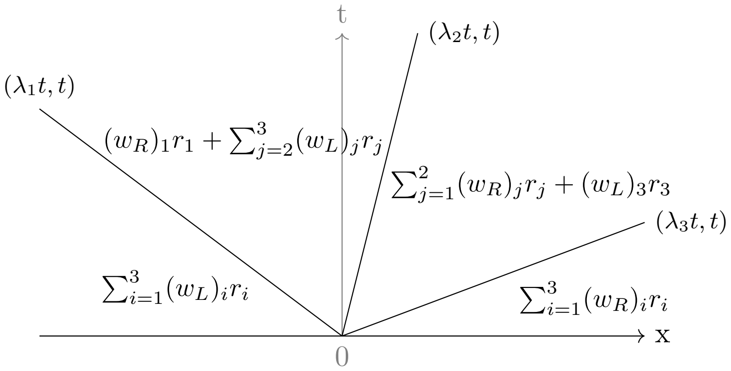

t
$(\lambda_2 t, t)$
$(\lambda_1 t, t)$
$(w_R)_1 r_1 + \sum_{j=2}^{3} (w_L)_j r_j$
$\sum_{j=1}^{2} (w_R)_j r_j + (w_L)_3 r_3$
$(\lambda_3 t, t)$
$\sum_{i=1}^{3} (w_L)_i r_i$
$\sum_{i=1}^{3} (w_R)_i r_i$
x
0

Figure 2: Solution of the constant coefficient system Riemann problem.

The variable coefficient Riemann problem behaves as expected. Consider

$$u_t + a(x,t) u_x = 0, \qquad u(x,0) = \begin{cases} u_L, & x < 0 \\ u_R, & x > 0 \end{cases}.$$

If the problem is well-posed, we know the unique solution can be written $u(x,t) = u(\psi(x,t),0)$. That is,

$$u(x,t) = \begin{cases} u_L, & \psi(x,t) < 0 \\ u_R, & \psi(x,t) > 0 \end{cases}.$$

Once again the solution is constant on either side of the characteristic wave emenating from the point of discontinuity.

Riemann problems for nonlinear hyperbolic equations are considerably more interesting. However, this topic is outside the scope of our text. We suggest [38, 41, 42, 43] for further reading.

2.4 Test problems

Here we describe the linear hyperbolic problems which we will use to analyze the schemes obtained from the numerical methods of section 4. In particular, we derive energy estimates and show that the problems are strongly well-posed by definition 2.8. We remark that these problems are very simple and commonly used to analyze schemes, see for instance [15, 5, 44].

Problem 1. *Let $x \in \Omega = [0,1] \subset \mathbb{R}$ and $t \in [0,T] \subset \mathbb{R}^+$. Find the function $u : \Omega \times [0,T) \to \mathbb{R}$ satisfying*

$$u_t + u_x = 0, \qquad u(x,0) = f(x), \qquad u(0,t) = g(t).$$

where $f(x) = sin(2\pi x)$ and $g(t) = \sin(-2\pi t)$.

We multiply the PDE with u and integrate over Ω:

$$\int_\Omega uu_t + uu_x dx = 0 \iff \int_0^1 uu_t dx + \int_0^1 uu_x dx = 0 \iff \int_0^1 \frac{1}{2}\frac{\partial}{\partial t}(u^2)dx + \int_0^1 \frac{1}{2}\frac{\partial}{\partial x}(u^2)dx = 0.$$

Note that

$$\int_0^1 \frac{1}{2}\frac{\partial}{\partial t}(u^2)dx = \frac{1}{2}\frac{d}{dt}\int_0^1 u^2 dx = \frac{1}{2}\frac{d}{dt}\left\|u(\cdot,t)\right\|_2^2, \qquad \text{and} \qquad \frac{1}{2}\int_0^1 \frac{\partial}{\partial x}(u^2)dx = \frac{1}{2}(u^2(1,t) - u^2(0,t)).$$

Hence

$$\frac{d}{dt}\left\|u(\cdot,t)\right\|_2^2 = u^2(0,t) - u^2(1,t) \leq u^2(0,t) = g^2(t).$$

Integrating in time from 0 to T we obtain

$$\int_0^T \frac{d}{dt}\left\|u(\cdot,t)\right\|_2^2 dt \leq \int_0^T g^2(t)dt,$$
$$\iff \left\|u(\cdot,T)\right\|_2^2 \leq \left\|u(\cdot,0)\right\|_2^2 + \int_0^T g^2(t)dt.$$

By definition 2.8, this shows that the problem is strongly stable. Further, inserting $f(x) = \sin(2\pi x)$ and $g(t) = \sin(-2\pi t)$ we find that

$$u(x,t) = \sin(2\pi(x-t)),$$

is the unique solution of the problem, making the problem is strongly well-posed.

Problem 2. *Let $\mathbf{x} = [x,y]^T \in \Omega = [0,1]^2$ and $0 \leq t \leq T < \infty$. Find the function $u : \Omega \times [0,T] \to \mathbb{R}$*

satisfying

$$u_t + u_x + u_y = 0, \qquad u(x,y,0) = \sin\left(2\pi\left(\frac{x}{2}+\frac{y}{2}\right)\right), \qquad u(0,y,t) = g_1(y,t), \qquad u(x,0,t) = g_2(x,t),$$

where

$$g_1(y,t) = \sin\left(2\pi\left(\frac{y}{2}-t\right)\right), \qquad \textit{and} \qquad g_2(x,t) = \sin\left(2\pi\left(\frac{x}{2}-t\right)\right).$$

We multiply the PDE by u and integrate over the spatial domain to find

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\|u(\cdot,t)\right\|_2^2 &= -\frac{1}{2}\int_0^1\int_0^1 \frac{\partial}{\partial x}u^2 dxdy - \frac{1}{2}\int_0^1\int_0^1 \frac{\partial}{\partial y}u^2 dydx,\\ &= -\frac{1}{2}\int_0^1 u^2(1,y,t) - g_1^2(y,t)dy - \frac{1}{2}\int_0^1 u^2(x,1,t) - g_2^2(x,t)dx,\\ &= \frac{1}{2}\int_0^1 g_1^2(y,t) - u^2(1,y,t)dy + \frac{1}{2}\int_0^1 g_2^2(x,t) - u^2(x,1,t)dx,\\ &\leq \frac{1}{2}\int_0^1 g_1^2(y,t)dy + \frac{1}{2}\int_0^1 g_2^2(x,t)dx.\end{aligned}$$

Hence

$$\left\|u(\cdot,T)\right\|_2^2 \leq \left\|u(\cdot,0)\right\|_2^2 + \int_0^T\left(\int_0^1 g_1^2(y,t)dy + \int_0^1 g_2^2(x,t)dx\right)dt,$$

so the problem is strongly stable. Moreover, the function

$$u(x,y,t) = \sin\left(2\pi\left(\frac{x}{2}+\frac{y}{2}-t\right)\right),$$

is the unique solution. Therefore the problem is strongly well-posed.

3 Finite volume methods

3.1 Introduction

Consider the initial-boundary value problem

$$\begin{cases} u_t + \nabla \cdot f(u) = 0, \qquad \mathbf{x} \in \Omega \subset \mathbb{R}^n, \quad 0 \leq t \leq T < \infty, \\ u(\mathbf{x}, 0) = \phi(\mathbf{x}), \\ u(\partial\Omega, t) = g(\partial\Omega, t). \end{cases} \tag{3.1}$$

which we will assume is well-posed. A finite volume method aims to determine $u(\mathbf{x}, t)$ numerically. The general procedure can be broken down as follows:

1. Partition the spatial domain Ω into a set of N control volumes $\{C_i\}_{i=1}^N$, and let V_i denote the measure of C_i. We require that $C_i \cap C_j = \emptyset$, $\forall j \neq i$ and $\bigcup_{i=1}^N \overline{C_i} = \overline{\Omega}$.

2. Let $u_i(t)$ denote an approximation of the volume-averaged value of the conserved quantity $u(\mathbf{x}, t)$ in C_i.

$$u_i(t) \approx \frac{1}{V_i} \int_{C_i} u(\mathbf{x}, t) d\mathbf{x}.$$

3. Integrate the PDE over C_i to obtain

$$\frac{d}{dt} u_i(t) = -\frac{1}{V_i} \int_{C_i} \nabla \cdot f(u(\mathbf{x}, t)) d\mathbf{x} = -\frac{1}{V_i} \int_{\partial C_i} f(u(\mathbf{x}, t)) \cdot \hat{n} dS, \tag{3.2}$$

where dS is the infinitesimal arc length of ∂C_i, $\hat{n}$ is the outward pointing unit normal vector, and the second equality follows from Gauss' theorem. Denote $\Gamma_{ij} = \partial C_i \cap \partial C_j$ and define N_i to be the set of all indices j such that $\Gamma_{ij} \neq \emptyset$. Suppose that $\partial C_i = \cup_{j \in N_i} \Gamma_{ij}$, then

$$\frac{d}{dt} u_i(t) = -\frac{1}{V_i} \sum_{j \in N_i} \int_{\Gamma_{ij}} f(u(\mathbf{x}, t)) \cdot \hat{n} dS. \tag{3.3}$$

We will refer to such a control volume as an interior volume. If $\partial C_i \cap \partial\Omega = \Gamma_{i\partial\Omega} \neq \emptyset$ we write $\partial C_i = \bigcup_{j \in N_i} \Gamma_{ij} \cup \Gamma_{i\partial\Omega}$ and obtain

$$\frac{d}{dt} u_i(t) = -\frac{1}{V_i} \sum_{j \in N_i} \int_{\Gamma_{ij}} f(u(\mathbf{x}, t)) \cdot \hat{n} dS - \frac{1}{V_i} \int_{\Gamma_{i\partial\Omega}} f(u(\mathbf{x}, t)) \cdot \hat{n} dS. \tag{3.4}$$

In this case we say that C_i is a boundary volume.

4. Let $t = t^*$ be fixed and write $u(\mathbf{x}, t^*) = u(\mathbf{x})$, $u_i(t^*) = u_i$. Denote by F_{ij} an approximation of $f(u(\mathbf{x}))$ restricted to Γ_{ij}. If C_i is an interior volume we find that

$$\frac{d}{dt} u_i = -\frac{1}{V_i} \sum_{j \in N_i} \int_{\Gamma_{ij}} F_{ij} \cdot \hat{n} dS, \tag{3.5}$$

and we approximate the integral using some quadrature rule. The way we obtain F_{ij} and the quadrature rule we choose is what seperates one finite volume method from another.

5. Note that (3.5) for $i = 1, \ldots, N$ is a system of N ordinary differential equations. Therefore, we can evolve u_i in time by some numerical ODE method.

The procedure outlined above gives us a scheme which approximates the integral form of (3.1). If the problem is linear and well-posed, and if the method is consistent and stable, the solution of the approximation will converge to $u(\mathbf{x}, t)$ as $V_i \to 0$ (see [45]). We did not give any details on how to implement the boundary condition, but will do so in section 3.5.

3.2 Different types of grids and volumes

Let $\Omega \subset \mathbb{R}^n$ and consider the problem of partitioning Ω into a collection of subsets C_i satisfying

$$C_i \cap C_j = \emptyset, \quad \forall j \neq i, \qquad \bigcup_{i=1}^{N} \overline{C}_i = \overline{\Omega}. \tag{3.6}$$

Consider first the case where $\Omega = [\Omega^-, \Omega^+] \subset \mathbb{R}$. Define a grid, or node set, $\{x_i\}_{i=1}^{N+1}$ where $x_i \in \Omega$ for $i = 1, \ldots, N+1$, and assume that there exists indices $L, R \in \{1, \ldots, N+1\}$ such that $x_L = \Omega^-$ and $x_R = \Omega^+$.

Definition 3.1. *Let $\{x_i\}_{i=1}^{N+1}$ be some grid contained in $\overline{\Omega}$. If*

$$x_i \in \partial\Omega,$$

then x_i is said to be a boundary (grid) point, or boundary node.

Clearly in the above case, x_L, x_R are the boundary nodes.

Definition 3.2. *If*

$$x_i < x_{i+1}, \qquad i = 1, \ldots, N,$$

then the grid is said to be structured.

Definition 3.3. *Assume $\{x_i\}_{i=1}^{N+1}$ is a structured grid. If*

$$|x_i - x_{i+1}| = h, \qquad i = 1, \ldots, N,$$

where $h \in \mathbb{R}$ is some constant, then the grid is said to be regular.

If a grid is not structured, we say that it is unstructured. If a grid is not regular, we say that it is irregular. Returning to the problem of partitioning $\Omega = [\Omega^-, \Omega^+]$ into control volumes, assume $\{x_i\}_{i=1}^{N}$ is a structured grid in $\overline{\Omega}$ and $\partial\Omega \subset \{x_i\}_{i=1}^{N}$. Then the control volumes

$$C_i = (x_i, x_{i+1}), \qquad i = 1, \ldots, N, \tag{3.7}$$

satisfy (3.6). Given some control volume $C_i = (C_i^-, C_i^+)$ we will use V_i to denote its measure. We give some elementary definitions:

Definition 3.4. *Let $\{C_i\}_{i=1}^{N}$ be a set of control volumes satisfying (3.6). If*

$$C_i^+ = C_{i+1}^-, \qquad i = 1, \ldots, N-1,$$

then the volumes are said to be structured.

Definition 3.5. *Let $\{C_i\}_{i=1}^{N}$ be a set of control volumes satisfying (3.6). If*

$$V_i = h, \qquad i = 1, \ldots, N,$$

where $h \in \mathbb{R}$ is some constant, then the volumes are said to be regular.

Definition 3.6. *Let $\{C_i\}_{i=1}^N$ be a set of control volumes satisfying (3.6). If*

$$\partial C_i \cap \partial\Omega = \emptyset,$$

then C_i is an interior volume. If

$$\partial C_i \cap \partial\Omega \neq \emptyset,$$

then C_i is a boundary volume.

Further, we will use the notation Γ_{ij} to denote $\partial C_i \cap \partial C_j$, and we generalize the notation such that $\Gamma_{i\partial\Omega} = \partial C_i \cap \partial\Omega$. Moreover, we let N_i denote the set of indices j such that $\Gamma_{ij} \neq \emptyset$. Thus, if C_i is an interior volume then $\partial C_i = \cup_{j\in N_i}\Gamma_{ij}$, and if C_i is a boundary volume then $\partial C_i = \cup_{j\in N_i}\Gamma_{ij} \cup \Gamma_{i\partial\Omega}$.

In (3.7) we gave an example of a set of volumes satisfying (3.6). Another common set of volumes partitioning $\Omega \subset \mathbb{R}$ can be found as follows. Assume $\{x_i\}_{i=1}^{N+1}$ is a structured grid. Define the dual grid $\{x_{i+1/2}\}_{i=1}^N$ by $x_{i+1/2} = (x_i + x_{i+1})/2$. Then

$$\begin{aligned} C_1 &= (x_1, x_{1+1/2}), \\ C_i &= (x_{i-1/2}, x_{i+1/2}), \qquad i = 2, \dots, N-1, \\ C_N &= (x_{N-1/2}, x_N), \end{aligned}$$

satisfy (3.6). Moving on, consider $\Omega \subset \mathbb{R}^2$.

Definition 3.7. *Let $\Omega \subset \mathbb{R}^2$ and $L = [0,1]^2$. If there exists a bijective continuous linear transformation $\psi : \Omega \to L$ with a continuous inverse ψ^{-1}, then Ω is a logically rectangular domain.*

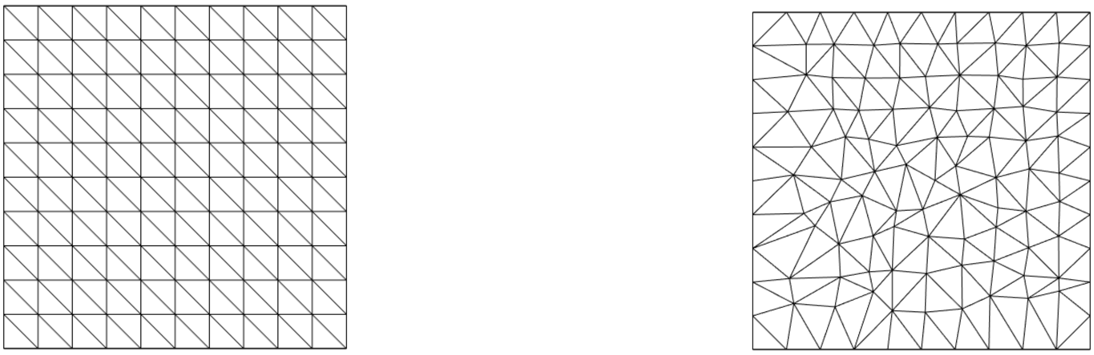
If Ω is logically rectangular then it is sufficient to find control volumes in L, as $\{\psi^{-1}(C_i)\}_{i=1}^N$ gives us the desired partition. Therefore we consider first the case $\Omega = L = [0,1]^2$. Recall that for $\Omega \subset \mathbb{R}$ we found control volumes induced by structured grids. This approach also works for the 2D case. Let $\{x_i\}_{i=1}^{m+1}$ and $\{y_j\}_{j=1}^{l+1}$ be two structured grids contained in $[0,1]$ and assume $x_1 = y_1 = 0$, $x_{m+1} = y_{l+1} = 1$. Define

$$\begin{aligned} C_i &= (x_i, x_{i+1}) \times (y_1, y_2), \qquad i = 1, \dots, m, \\ C_i &= (x_{i-m}, x_{i-m+1}) \times (y_2, y_3), \qquad i = m+1, \dots, 2m, \\ &\vdots \\ C_i &= (x_{i-(l-1)m}, x_{i-(l-1)m+1}) \times (y_l, y_{l+1}), \qquad i = (l-1)m+1, \dots, lm, \end{aligned}$$

and put $N = lm$. Then $\{C_i\}_{i=1}^N$ satisfies (3.6). Once again, the volumes are said to be regular if $V_i = V_j = h$ for all i, j. If we define the volumes using one index for each dimension, we can also talk about structured volumes in the 2D case. Note that dual grid volumes can also be defined by a similar procedure.

Suppose $\Omega \neq [0,1]^2$, or that the volumes $\{\psi^{-1}(C_i)\}_{i=1}^N$ are somehow difficult to work with. There exists a simple method to partition any path-connected domain $\Omega \subset \mathbb{R}^2$ into control volumes $\{C_i\}_{i=1}^N$ satisfying (3.6). Let $\{p_i\}_{i=1}^m$ be a set of distinct points in Ω, and consider the Deluanay triangulation generated by the points. Let each triangle generated by the triangulation be a control volume C_i. Under some mild conditions on the points p_i, the control volumes generated by the triangulation will satisfy (3.6). We illustrate some control volume partitions of $\Omega = [0,1]^2$ in Fig. 3.

Similarly to how a structured grid induces a dual grid, unstructured triangular grids also induce dual grids. Further, triangular volumes induce dual volumes. These are defined by connecting each grid edge midpoint to the center of mass of the volumes whose boundaries they intersect, see Fig 4.



(a) Regular grid and volumes.

(b) Irregular grid and volumes.

Figure 3: Illustration of common grids and volumes.

3.3 Consistency, accuracy and stability

Consider some well-posed problem $\mathbf{x} \in \Omega \subset \mathbb{R}^n$, $0 \leq t \leq T < \infty$,

$$u_t + \nabla \cdot f(u) = 0, \qquad u(\mathbf{x}, 0) = \phi(\mathbf{x}), \qquad u(\partial\Omega, t) = g(\partial\Omega, t), \tag{3.8}$$

and let $\{C_i\}_{i=1}^{N}$ be a set of control volumes satisfying

$$C_i \cap C_j = \emptyset, \qquad (i \neq j), \qquad \text{and} \qquad \bigcup_{i=1}^{N} \overline{C_i} = \overline{\Omega}.$$

Further, define $\{u_i(t)\}_{i=1}^{N}$ and F by

$$u_i(t) \approx \frac{1}{V_i} \int_{C_i} u(\mathbf{x}, t) d\mathbf{x}, \qquad F_i \approx \int_{\partial C_i} f(u(\mathbf{x}, t)) \cdot \hat{n} dS,$$

such that

$$\frac{d}{dt}\mathbf{u} + V^{-1}F = 0, \tag{3.9}$$

is a finite volume scheme approximating (3.8), where $\mathbf{u} = [u_1, ..., u_N]^T$ and $V = \text{diag}(V_1, ..., V_N)$.

Definition 3.8. *Assume $u(\mathbf{x}, t)$ is the unique solution of the problem (3.8). Then the local truncation error τ_i of the finite volume scheme (3.9) is given by*

$$\frac{d}{dt} \int_{C_i} u(\mathbf{x}, t) dx + F_i = \tau_i.$$

Definition 3.9. *If the local tructation error τ_i satisfies*

$$\tau_i = \mathcal{O}(h^r), \qquad \text{where} \qquad h = \max_{C_i} \sup_{a,b \in C_i} \|a - b\|_{\mathbb{R}^n},$$

for $i = 1, \ldots, N$, then the scheme is r-th order accurate.

An r-th order accurate scheme with $r > 0$ is said to be *consistent.* Note that the finite volume scheme approximates the integrated in space PDE if

$$F_i = \int_{\partial C_i} f(u(\mathbf{x}, t)) \cdot \hat{n} dS + \mathcal{O}(h^r), \qquad r > 0, \qquad (i = 1, \dots, N),$$

i.e., if the scheme is consistent. Define any r-th order scheme with $r \geq 3$ to be a *high-order* scheme. Define any finite volume method generating high-order schemes to be a high-order method. Suppose that we modify the problem (3.8) such that $\Omega \subset \mathbb{R}^2$ and the flux function $f(u)$ is linear. Then by definition 2.7, the IBVP is stable if for $g \equiv 0$ there exists constants K, α independent of ϕ and t such that

$$\|u(\cdot, t)\|_2^2 \leq K e^{\alpha t} \|\phi(\cdot)\|_2^2 .$$

Further, the problem is strongly stable if there exists a bounded functional $K(t)$ independent of ϕ and g such that

$$\|u(\cdot, t)\|_2^2 \leq K(t) \left(\|\phi(\cdot)\|_2^2 + \int_0^t \int_{\partial\Omega} g^2(\mathbf{x}, \tau) d\mathbf{x} d\tau \right) .$$

Assuming we have a consistent scheme for the problem, we would like to have similar estimates for the numerical solution. As the numerical solution is not a function of the spatial domain in the traditional sense, we must consider a discrete version of the L^2 norm. Recall that if the exact solution is continuous over C_i then there exists some point $\mathbf{x}^* \in C_i$ such that

$$u(\mathbf{x}^*, t) = \frac{1}{V_i} \int_{C_i} u(\mathbf{x}, t) d\mathbf{x}.$$

Thus, if the error between $u_i(t)$ and the volume-averaged value of the solution $u(\mathbf{x}, t)$ over C_i is $\mathcal{O}(h^r)$ with $r > 0$, then

$$u_i^2(t) \to u^2(\mathbf{x}^*, t) \qquad \text{as} \qquad h \to 0,$$

and

$$V_i u_i^2(t) \to \int_{C_i} u^2(\mathbf{x}, t) d\mathbf{x} \qquad \text{as} \qquad h \to 0.$$

To summarize, we have found that

$$\langle \mathbf{u}(t), V\mathbf{u}(t) \rangle = \sum_{i=1}^{N} V_i u_i^2(t) = \|\mathbf{u}(t)\|_V^2 \to \|u(\cdot, t)\|_2^2 .$$

This motivates the following definitions.

Definition 3.10 ([10, 36])**.** *A semidiscrete scheme*

$$\frac{d}{dt}\mathbf{u} + V^{-1} F \mathbf{u} = 0,$$

approximating some well-posed linear problem

$$u_t + P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u = 0, \qquad \mathbf{x} \in \Omega, \qquad t \in [0, T],$$
$$u(\mathbf{x}, 0) = \phi(\mathbf{x}),$$
$$u(\partial\Omega, t) = g(\partial\Omega, t),$$

is said to be (energy) stable, if for $g \equiv 0$ we have the estimate

$$\|\mathbf{u}(T)\|_V \leq K e^{\alpha_d T} \|\mathbf{u}(0)\|_V ,$$

where $\|\mathbf{u}\|_V = \langle \mathbf{u}, V\mathbf{u} \rangle$ and $K, \alpha_d \in \mathbb{R}$ are independent of $\mathbf{u}(0)$ and T.

Note that

$$\frac{d}{dt} \|\mathbf{u}\|_V^2 \leq 0,$$

implies stability.

Definition 3.11 ([10, 36])**.** *A semidiscrete scheme*

$$\frac{d}{dt}\mathbf{u} + V^{-1} F \mathbf{u} = 0,$$

approximating some well-posed problem

$$\begin{aligned} u_t + P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u = 0, \qquad \mathbf{x} \in \Omega, \qquad t \in [0, T], \\ u(\mathbf{x}, 0) = \phi(\mathbf{x}), \\ u(\partial\Omega, t) = g(\partial\Omega, t), \end{aligned}$$

is said to be strongly stable if we have the estimate

$$\|\mathbf{u}(T)\|_V^2 \leq K(T) \left(\|\mathbf{u}(0)\|_V^2 + \max_{\tau \in [0,T]} \|g(\tau)\|_V^2 \right),$$

where $K(t) \in L^\infty[0, T]$ is independent of $\mathbf{u}(0)$ and g.

Note that

$$\frac{d}{dt} \|\mathbf{u}\|_V^2 \leq g^2(t),$$

implies strong stability.

Definition 3.12 ([10])**.** *A semidiscrete scheme*

$$\frac{d}{dt}\mathbf{u} + V^{-1} F \mathbf{u} = 0,$$

approximating some well-posed problem

$$\begin{aligned} u_t + P\left(\mathbf{x}, t, \frac{\partial}{\partial \mathbf{x}}\right) u = 0, \qquad \mathbf{x} \in \Omega, \qquad t \in [0, T], \\ u(\mathbf{x}, 0) = \phi(\mathbf{x}), \\ u(\partial\Omega, t) = g(\partial\Omega, t), \end{aligned}$$

is said to be strictly stable if we have the estimate

$$\|\mathbf{u}(T)\|_V^2 \leq \|u(\cdot, T)\|_2^2 .$$

There are also other notions of stability for schemes. An especially popular ([33, 35, 7]) stability analysis tool is due to von Neumann. However, this approach is only valid for initial value problems and for schemes on structured, regular grids/volumes. We will show how one can prove stability for IBVP schemes in section 3.5.

3.4 Godunov's method

The quintessential finite volume method is Godunov's method, introduced in [46]. Following the procedure described in section 3.1, consider the problem of approximating the flux over the boundaries:

$$\sum_{j\in N_i}\int_{\Gamma_{ij}} f(u(\mathbf{x},t))\cdot\hat{n}dS.$$

Recall that $u(\mathbf{x},t)\big|_{\Gamma_{ij}}$ is unknown. Consider creating a piecewise constant reconstruction $R(\mathbf{x})$ of $u(\mathbf{x},t)$ at the current time step using the u_i's. That is, we define

$$R(\mathbf{x})\big|_{\mathbf{x}\in C_i} = u_i.$$

At each boundary $\Gamma_{ij} = \partial C_i \cap \partial C_j \neq \emptyset$ we have that $R(\mathbf{x}) = u_i$ on one side and $R(\mathbf{x}) = u_j$ on the other side. In other words, R is discontinuous accross the boundary. Godunov's method proceeds by solving solving the Riemann problem on $p_b \in \Gamma_{ij}$ given by the PDE and R. Formally, we imagine a local (x,t)-like coordinate system (ξ,τ) at each point $p_b \in \Gamma_{ij}$ with $\xi = 0 = p_b$. In this plane we have the Riemann problem

$$u_\tau + f(u)_\xi = 0, \qquad u(\xi,0) = \begin{cases} u_i, & \xi < 0 \\ u_j, & \xi > 0 \end{cases}.$$

Here the ξ coordinate is parallel to and has the same direction as the unit normal vector $\hat{n}$ of the boundary Γ_{ij} at p_b. Further, let $\tau = 0$ at the current time step. Using $RP(u_i,u_j)$ to denote the exact or approximate solution of the problem at $\xi = 0, \tau > 0$, we obtain

$$\sum_{j\in N_i}\int_{\Gamma_{ij}} f(u(\mathbf{x}))\cdot\hat{n}dS \approx \sum_{j\in N_i}\int_{\Gamma_{ij}} f(R(\mathbf{x}))\cdot\hat{n}dS = \sum_{j\in N_i}\int_{\Gamma_{ij}} f(RP(u_i,u_j))\cdot\hat{n}dS = \sum_{j\in N_i} f(RP(u_i,u_j))\int_{\Gamma_{ij}}\hat{n}dS,$$

where we tacitly assumed the Riemann problem was identical at every point p_b on Γ_{ij}.

Remark. Godunov's method is sometimes described in the fully discrete setting. In this case, the waves given by the solution of the Riemann problem are used to find the update of the unknown over the time interval $[t, t+\Delta t]$.

An alternative derivation is to consider the idea of a numerical flux flunction F. The numerical flux function is always dependent on the two states at either side of the discontinuity, $F \equiv F(u_L,u_R)$. To illustrate, we would have that

$$\sum_{j\in N_i}\int_{\Gamma_{ij}} f(u(\mathbf{x},t))\cdot\hat{n}dS \approx \sum_{j\in N_i}\int_{\Gamma_{ij}} f(R(\mathbf{x},t))\cdot\hat{n}dS = \sum_{j\in N_i}\int_{\Gamma_{ij}} F(u_i,u_j)\cdot\hat{n}dS = \sum_{j\in N_i} F(u_i,u_j)\int_{\Gamma_{ij}}\hat{n}dS.$$

We see that the two approximations are equivalent if $F(u_i,u_j) = f(RP(u_i,u_j))$, i.e. using Godunov's numerical flux [47]. Throughout this thesis we will often consider the central numerical flux

$$F(u_L,u_R) = f\left(\frac{u_L+u_R}{2}\right).$$

In general, most numerical fluxes in the scientifc literature are of the form

$$F(u_L, u_R) = f\left(\frac{u_L + u_R}{2}\right) - \lambda(u_R - u_L),$$

where $\lambda \in \mathbb{R}$ is sometimes called the upwinding parameter.

3.5 SBP-SAT schemes

In [15, 16] it was shown that certain finite volume schemes satisfy the summation-by-parts (SBP) property. These schemes could then be modified by adding simultaneous approximation terms (SAT) to make them strongly stable in a discrete L^2 norm. Here we introduce the concept of SBP-SAT schemes and illustrate how to prove energy stability of schemes. Consider problem 1 from section 2.4:

Let $x \in \Omega = [0,1] \subset \mathbb{R}$ and $t \in [0,T] \subset \mathbb{R}^+$. Find the function $u : \Omega \times [0,T] \to \mathbb{R}$ satisfying

$$u_t + u_x = 0, \qquad u(x,0) = f(x), \qquad u(0,t) = g(t).$$

where $f(x) = \sin(2\pi x)$ and $g(t) = \sin(-2\pi t)$.

Recall that the energy rate was found to be

$$\frac{d}{dt}\|u(\cdot,t)\|_2^2 = u^2(0,1) - u^2(1,t) \leq g^2(t).$$

We will demonstrate that SBP-SAT schemes obtain the same energy rate in a discrete L^2 norm. Let $\{C_i\}_{i=1}^N$ denote a set of structured control volumes satisfying the conditions

$$C_i \cap C_j = \emptyset, \qquad (i \neq j), \qquad \text{and} \qquad \bigcup_{i=1}^N \overline{C_i} = \overline{\Omega}.$$

Let $u_i(t)$ denote an approximation of the volume-averaged value of $u(x,t)$ over C_i, i.e.

$$u_i(t) \approx \frac{1}{V_i}\int_{C_i} u(x,t)dx, \qquad i = 1, \dots, N,$$

where V_i denotes the measure of C_i. Integrating the PDE over C_i and using integration by parts we obtain the finite volume approximation

$$V_i \frac{d}{dt} u_i = u(C_i^-, t) - u(C_i^+, t)$$

where $C_i = (C_i^-, C_i^+)$. As $u(C_i^\pm, t)$ is unknown, we want to approximate these values using the data $\{u_i(t)\}_{i=1}^N$. Consider a fixed time $t = t^*$ and ignore the time dependence of the variables. Following Godunov's method, we create the piecewise constant function $R(x)$ given by $R(x) = u_i$ for $x \in C_i$. Then we determine $u(C_i^+)$ by solving the Riemann problems

$$u_t + u_x = 0, \qquad u(x,0) = \begin{cases} u_i, & x < C_i^+ \\ u_{i+1}, & x > C_i^+ \end{cases}$$

either exactly or approximately. Note that we used the structure of the control volumes to know that $C_i^+ = C_{i+1}^-$. Denote by $RP(u_i, u_{i+1})$ either the exact or approximate solution of the Riemann problem evaluated at $x = C_i^+$, $t > 0$. Then, approximate $u(C_i^+)$ by $RP(u_i, u_{i+1})$ and $u(C_i^-)$ by $RP(u_{i-1}, u_i)$.

Following the papers [15, 16] we let $RP(a,b) = (a+b)/2$. The scheme for interior volumes becomes

$$V_i \frac{d}{dt} u_i = \frac{u_{i-1}+u_i}{2} - \frac{u_i+u_{i+1}}{2} = \frac{u_{i-1}-u_{i+1}}{2}.$$

Since u_0 and u_{N+1} are undefined, we will use $u(C_1^-) \approx u_1$ and $u(C_N^+) \approx u_N$. Then

$$V_1 \frac{d}{dt} u_1 = u_1 - \frac{u_1+u_2}{2} = \frac{u_1-u_2}{2}, \qquad \text{and} \qquad V_N \frac{d}{dt} u_N = \frac{u_{N-1}+u_N}{2} - u_N = \frac{u_{N-1}-u_N}{2}.$$

Define matrices Q, V by

$$Q = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & \dots & \dots \\ -\frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \dots & \dots & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad V = \operatorname{diag}(V_1, V_2, \dots, V_N),$$

to write the scheme as

$$V \frac{d}{dt} \mathbf{u} = -Q\mathbf{u},$$

where $\mathbf{u} = [u_1, u_2, \dots, u_N]^T$.

Definition 3.13. *A discrete differential operator $D = V^{-1}Q$ is said to be a summation-by-parts operator, if V is symmetric positive definite and Q satisfies $Q + Q^T = \operatorname{diag}(-1, 0, \dots, 0, 1)$.*

We see that $D = V^{-1}Q$ for the given matrices Q, V, is a summation-by-parts operator. Suppose that we modify the scheme by adding the SAT,

$$V \frac{d}{dt} \mathbf{u} = -Q\mathbf{u} + \text{SAT} = -Q\mathbf{u} + \tau(u_1 - g(t))e_1, \tag{3.10}$$

where $e_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^N$.

Proposition 3.1. *The scheme (3.10) with $\tau = -1$ is strongly stable.*

Proof. Multiply by $\mathbf{u}^T$ on the left and add the transposed equation to obtain

$$\frac{d}{dt} \|\mathbf{u}\|_V^2 = -\mathbf{u}^T(Q+Q^T)\mathbf{u} + 2u_1\tau(u_1 - g(t)) = u_1^2 - u_N^2 + 2u_1\tau(u_1 - g(t)) \le (1+2\tau)u_1^2 - 2u_1\tau g(t).$$

Inserting $\tau = -1$ gives

$$\frac{d}{dt} \|\mathbf{u}\|_V^2 \le -u_1^2 + 2u_1 g(t) = -(u_1 - g(t))^2 + g^2(t) \le g^2(t).$$

□

Instead of adding a SAT to implement the boundary condition, suppose we implement it directly into the approximation of $u(C_1^-)$ in the flux calculation. The scheme for u_1 becomes

$$V_1 \frac{d}{dt} u_1 = g(t) - \frac{u_1+u_2}{2},$$

and the matrix Q becomes

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \dots & \dots \\ -\frac{1}{2} & 0 & \frac{1}{2} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \dots & \dots & -\frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

giving the scheme

$$V\frac{d}{dt}\mathbf{u} = -Q\mathbf{u} + g(t)e_1.$$

Further, we have the discrete energy estimate

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -u_1^2 - u_N^2 + 2u_1 g(t) \leq -(u_1 - g(t))^2 + g^2(t) \leq g^2(t).$$

We see that directly implementing the boundary condition in the flux approximation for this problem and finite volume method also gives a strongly stable scheme. In fact, the scheme modified by the SAT with $\tau = -1$ and the scheme in which we injected the boundary condition into the flux directly are completely equivalent. To see this, simply compare the first equation in the two schemes and note that they are the same:

$$-\frac{u_1}{2} - \frac{u_2}{2} + g(t) = \frac{u_1}{2} - \frac{u_2}{2} - u_1 + g(t).$$

Next, consider problem 2 from section 2.4:

Let $\mathbf{x} = (x, y) \in \Omega = [0,1]^2$, $0 \leq t \leq T < \infty$ and $u = u(\mathbf{x}, t)$ be a real-valued function. Consider

$$u_t + u_x + u_y = 0, \qquad u(x,y,0) = \sin\left(2\pi\left(\frac{x}{2} + \frac{y}{2}\right)\right), \qquad u(0,y,t) = g_1(y,t), \qquad u(x,0,t) = g_2(x,t),$$

where g_1 and g_2 are given by

$$g_1(y,t) = \sin(2\pi(y/2 - t)), \qquad g_2(x,t) = \sin(2\pi(x/2 - t)).$$

In section 2.4 we found the energy rate to be

$$\frac{d}{dt}\|u\|_2^2 = \int_0^1 g_1^2(y,t) - u^2(1,y,t)dy + \int_0^1 g_2^2(x,t) - u^2(x,1,t)dx \leq \int_0^1 g_1^2(y,t)dy + \int_0^1 g_2^2(x,t)dx.$$

Once again we will demonstrate a finite volume scheme which satisfies the same energy rate in a discrete L^2 norm. Let Ω be discretized by an unstructured triangular grid with N nodes $\{p_i\}_{i=1}^N$, and construct N node-centered control volumes $\{C_i\}_{i=1}^N$ defined by the midpoints of the lines connecting p_i with its neighbours and the triangle centroids as shown in Fig. 4. Let $u_i(t)$ denote an approximation of the volume-averaged value of $u(\mathbf{x}, t)$ over C_i, i.e.

$$u_i(t) \approx \frac{1}{V_i}\int_{C_i} u(\mathbf{x}, t)d\mathbf{x}.$$

Integrating the PDE over $C_i \subset \Omega$ and applying Gauss' theorem we find

$$\int_{C_i} u_t dxdy = -\int_{C_i} u_x + u_y dxdy = -\int_{C_i} \nabla \cdot [u,u]^T dxdy = -\oint_{\partial C_i} [u,u] \cdot \hat{n} dS = -\oint_{\partial C_i} udy + \oint_{\partial C_i} udx,$$

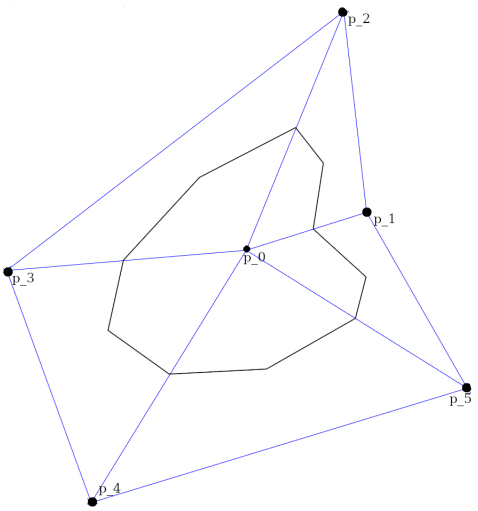


Figure 4: Illustration of node-centered control volume on an unstructured irregular triangular grid.

and the finite volume approximation

$$V_i \frac{d}{dt} u_i = -\oint_{\partial C_i} u dy + \oint_{\partial C_i} u dx,$$

where we used the fact that $\hat{n} dS = [dy, -dx]^T$. Suppose that C_i is an interior volume, i.e.

$$\partial C_i = \bigcup_{j \in N_i} \Gamma_{ij}.$$

We can obtain the flux over each Γ_{ij} independently. As $u(\mathbf{x}, t)\big|_{\Gamma_{ij}}$ is unknown we must somehow approximate these values. Following Godunov's method, define N polynomials $\{R_i(\mathbf{x})\}_{i=1}^N$ by $R_i(\mathbf{x}) = u_i$ for $\mathbf{x} \in C_i$. Consider the Riemann problem

$$u_t + u_\xi = 0, \qquad u(\xi, 0) = \begin{cases} R_i(\xi) = u_i, & \xi < 0 \\ R_j(\xi) = u_j, & \xi > 0 \end{cases},$$

where $\xi \in \mathbb{R}^2$ is parallel to the normal vector on Γ_{ij} pointing outward w.r.t. ∂C_i, and $\xi = 0$ on Γ_{ij}. Let $RP(u_i, u_j)$ denote either the exact or approximate solution of the problem, evaluated at $\xi = 0$ and $t > 0$. We will use the approximation $u(\mathbf{x}, t)\big|_{\Gamma_{ij}} = RP(u_i, u_j)$. Further, we will use the central flux approximation

$$RP(a, b) = \frac{a + b}{2},$$

and the midpoint quadrature rule given by

$$\int_a^b \psi(x) dx = (b - a)\psi((a + b)/2).$$

Thus, the scheme for interior volumes becomes

$$\begin{aligned}
V_i \frac{d}{dt} u_i &= -\oint_{\partial C_i} u dy + \oint_{\partial C_i} u dx \\
&= -\sum_{j \in N_i} \int_{\Gamma_{ij}} u dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} u dx \\
&\approx -\sum_{j \in N_i} \int_{\Gamma_{ij}} RP(u_i, u_j) dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} RP(u_i, u_j) dx \\
&\approx -\sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{u_i + u_j}{2} dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{u_i + u_j}{2} dx \\
&\approx -\sum_{j \in N_i} \left(\frac{u_i + u_j}{2} \right) \Delta y_{ij} + \sum_{j \in N_i} \left(\frac{u_i + u_j}{2} \right) \Delta x_{ij} \\
&= -\sum_{j \in N_i} u_i \frac{\Delta y_{ij}}{2} - \sum_{j \in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j \in N_i} u_i \frac{\Delta x_{ij}}{2} + \sum_{j \in N_i} u_j \frac{\Delta x_{ij}}{2},
\end{aligned}$$

where $\Delta x_{ij}, \Delta y_{ij}$ are the changes in x, y over Γ_{ij} respectively.

Remark. For interior volumes,

$$\sum_{j \in N_i} \Delta x_{ij} = \sum_{j \in N_i} \Delta y_{ij} = 0.$$

Next we consider the case where C_i is a boundary volume, i.e.

$$\partial C_i = \bigcup_{j \in N_i} \Gamma_{ij} \cup \Gamma_{i\partial\Omega}, \qquad \text{where} \qquad \Gamma_{i\partial\Omega} = \partial C_i \cap \partial\Omega.$$

Let $\Delta y_{i\partial\Omega}$ and $\Delta x_{i\partial\Omega}$ denote the y, x lengths respectively along the boundary $\Gamma_{i\partial\Omega}$. As there is no reconstruction on the other side of $\Gamma_{i\partial\Omega}$, we will proceed as for the 1D problem and approximate $u(\mathbf{x}, t)\big|_{\Gamma_{i\partial\Omega}}$ by $R_i(\mathbf{x}) = u_i$. The scheme for boundary volumes becomes

$$\begin{aligned}
V_i \frac{d}{dt} u_i &= -\oint_{\partial C_i} u dy + \oint_{\partial C_i} u dx \\
&= -\sum_{j \in N_i} \int_{\Gamma_{ij}} u dy - \int_{\Gamma_{i\partial\Omega}} u dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} u dx + \int_{\Gamma_{i\partial\Omega}} u dx \\
&\approx -\sum_{j \in N_i} \int_{\Gamma_{ij}} RP(u_i, u_j) dy - \int_{\Gamma_{i\partial\Omega}} u_i dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} RP(u_i, u_j) dx + \int_{\Gamma_{i\partial\Omega}} u_i dx \\
&\approx -\sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{u_i + u_j}{2} dy - \int_{\Gamma_{i\partial\Omega}} u_i dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} \frac{u_i + u_j}{2} dx + \int_{\Gamma_{i\partial\Omega}} u_i dx \\
&\approx -\sum_{j \in N_i} \left(\frac{u_i + u_j}{2} \right) \Delta y_{ij} + \sum_{j \in N_i} \left(\frac{u_i + u_j}{2} \right) \Delta x_{ij} - u_i \Delta y_{i\partial\Omega} + u_i \Delta x_{i\partial\Omega} \\
&= -\sum_{j \in N_i} u_i \frac{\Delta y_{ij}}{2} - \sum_{j \in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j \in N_i} u_i \frac{\Delta x_{ij}}{2} + \sum_{j \in N_i} u_j \frac{\Delta x_{ij}}{2} - u_i \Delta y_{i\partial\Omega} + u_i \Delta x_{i\partial\Omega}.
\end{aligned}$$

Since ∂C_i is a closed simple curve it follows that

$$\sum_{j \in N_i} \Delta x_{ij} = -\Delta x_{i\partial\Omega}, \qquad \sum_{j \in N_i} \Delta y_{ij} = -\Delta y_{i\partial\Omega},$$

and the boundary volume scheme can be written as

$$V_i \frac{d}{dt} u_i = -u_i \frac{\Delta y_{i\partial\Omega}}{2} - \sum_{j \in N_i} u_j \frac{\Delta y_{ij}}{2} + u_i \frac{\Delta x_{i\partial\Omega}}{2} + \sum_{j \in N_i} u_j \frac{\Delta x_{ij}}{2}.$$

In order to determine the energy stability of the scheme we want to write it in the spatially global form

$$V \frac{d}{dt} \mathbf{u} = -K$$

where $V = \mathrm{diag}(V_1, \dots, V_N)$, $\mathbf{u} = [u_1, \dots, u_N]^T$ and $-K$ is the vector containing the flux approximations previously described. Suppose that we define matrices Q_x, Q_y by

$$(Q_x)_{ij} = \frac{\Delta y_{ij}}{2}, \qquad (Q_x)_{ii} = \begin{cases} \frac{\Delta y_{i\partial\Omega}}{2}, & i \in N_{\partial\Omega} \\ 0, & \text{otherwise} \end{cases},$$

$$(Q_y)_{ij} = -\frac{\Delta x_{ij}}{2}, \qquad (Q_y)_{ii} = \begin{cases} -\frac{\Delta x_{i\partial\Omega}}{2}, & i \in N_{\partial\Omega} \\ 0, & \text{otherwise} \end{cases},$$

where $N_{\partial\Omega}$ is the set of indices i such that $\partial C_i \cap \partial\Omega \neq \emptyset$. Then we can write the finite volume scheme as

$$V \frac{d}{dt} \mathbf{u} = -Q_x \mathbf{u} - Q_y \mathbf{u}. \tag{3.11}$$

Consider the fact that we are always integrating in a counter-clockwise manner. This implies that $\Delta y_{ij} = -\Delta y_{ji}$ and $\Delta x_{ij} = -\Delta x_{ji}$. It follows that

$$\mathbf{1}^T (Q_x + (Q_x)^T) \mathbf{1} = \sum_{i \in N_{\partial\Omega}} \Delta y_{i\partial\Omega}, \qquad \text{and} \qquad \mathbf{1}^T (Q_y + (Q_y)^T) \mathbf{1} = \sum_{i \in N_{\partial\Omega}} -\Delta x_{i\partial\Omega}.$$

Here we used $\mathbf{1} = [1, 1, \dots, 1]^T \in \mathbb{R}^N$. Let us examine the energy estimate of the scheme. Multiply (3.11) by $\mathbf{u}^T$ and add the transpose to obtain

$$\begin{aligned} \mathbf{u}^T V \mathbf{u}_t + \mathbf{u}_t^T V \mathbf{u} &= -\mathbf{u}^T Q_x \mathbf{u} - \mathbf{u}^T Q_x^T \mathbf{u} - \mathbf{u}^T Q_y \mathbf{u} - \mathbf{u}^T Q_y^T \mathbf{u} \\ \iff \frac{d}{dt} \|\mathbf{u}\|_V^2 &= -\mathbf{u}^T (Q_x + Q_x^T) \mathbf{u} - \mathbf{u}^T (Q_y + Q_y^T) \mathbf{u} \\ &= -\sum_{i \in N_{\partial\Omega}} (u_i^2 \Delta y_{i\partial\Omega} - u_i^2 \Delta x_{i\partial\Omega}). \end{aligned}$$

Due to the counter-clockwise integration orientation, $\Delta y_{i\partial\Omega}$ is positive at $x = 1$ and negative at $x = 0$. Likewise, $\Delta x_{i\partial\Omega}$ is positive at $y = 0$ and negative at $y = 1$. Let B_1 denote the set $\{(x, y) \in \overline{\Omega} : x = 0\}$ and B_2 denote the set $\{(x, y) \in \overline{\Omega} : y = 0\}$. Then,

$$\frac{d}{dt} \|\mathbf{u}\|_V^2 \leq -\sum_{i : \Gamma_{i\partial\Omega} \subset B_1} u_i^2 \Delta y_{i\partial\Omega} + \sum_{i : \Gamma_{i\partial\Omega} \subset B_2} u_i^2 \Delta x_{i\partial\Omega}.$$

Now we can add SAT to implement the boundary conditions in a stable manner. In particular, let (x_i, y_i) denote the midpoint of $\Gamma_{i\partial\Omega}$ and consider the scheme

$$V \frac{d}{dt} \mathbf{u} = -Q_x \mathbf{u} - Q_y \mathbf{u} + \tau_1 \left(u_i - g_1(y_i, t)\right) \Delta y_{i\partial\Omega} e_{B_1} + \tau_2 \left(u_i - g_2(x_i, t)\right) \Delta x_{i\partial\Omega} e_{B_2} \tag{3.12}$$

where e_{B_1}, $e_{B_2} \in \mathbb{R}^N$ are given by

$$(e_{B_1})_i = \begin{cases} 1, & \Gamma_{i\partial\Omega} \subset B_1 \\ 0, & \text{otherwise} \end{cases}, \qquad (e_{B_2})_i = \begin{cases} 1, & \Gamma_{i\partial\Omega} \subset B_2 \\ 0, & \text{otherwise} \end{cases}$$

Proposition 3.2. *If g_1, g_2 are integrated exactly by the midpoint quadrature rule then the scheme (3.12) with $\tau_1 = 1$ and $\tau_2 = -1$ is strongly stable*

Proof. Note that

$$\begin{aligned} \frac{d}{dt}\left\|\mathbf{u}\right\|_V^2 &\leq - \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \Delta y_{i\partial\Omega}\left(u_i^2 - 2\tau_1 u_i(u_i - g_1(y_i,t))\right) + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega}\left(u_i^2 + 2\tau_2 u_i(u_i - g_2(x_i,t))\right) \\ &= - \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \Delta y_{i\partial\Omega}\left((1-2\tau_1)u_i^2 + 2\tau_1 u_i g_1(y_i,t)\right) + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega}\left((1+2\tau_2)u_i^2 - 2\tau_2 u_i g_2(x_i,t)\right). \end{aligned}$$

Inserting $\tau_1 = 1$ and $\tau_2 = -1$ we obtain

$$\begin{aligned} \frac{d}{dt}\left\|\mathbf{u}\right\|_V^2 &\leq - \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \Delta y_{i\partial\Omega}\left(-u_i^2 + 2u_i g_1(y_i,t)\right) + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega}\left(-u_i^2 + 2u_i g_2(x_i,t)\right) \\ &= - \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \Delta y_{i\partial\Omega}\left(-(u_i - g_1(y_i,t))^2 + g_1^2(y_i,t)\right) \\ &\quad + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega}\left(-(u_i - g_2(x_i,t))^2 + g_2^2(x_i,t)\right) \\ &\leq - \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \Delta y_{i\partial\Omega} g_1^2(y_i,t) + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega} g_2^2(x_i,t). \end{aligned}$$

If g_1, g_2 are integrated exactly by the midpoint quadrature rule then

$$\sum_{i:\Gamma_{i\partial\Omega}\subset B_1} -\Delta y_{i\partial\Omega} g_1^2(y_i,t) + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \Delta x_{i\partial\Omega} g_2^2(x_i,t) = \int_0^1 g_1^2(y,t)dy + \int_0^1 g_2^2(x,t)dx.$$

□

Consider implementing the boundary conditions directly into the flux approximation instead as we did for problem 1. Let B_3 denote the set $\{(x,y) \in \overline{\Omega} : y = 1\}$ and B_4 denote the set $\{(x,y) \in \overline{\Omega} : x = 1\}$. Recall that

$$-\sum_{j\in N_i} \frac{\Delta y_{ij}}{2} = \frac{\Delta y_{i\partial\Omega}}{2}, \qquad \text{and} \qquad \sum_{j\in N_i} \frac{\Delta x_{ij}}{2} = -\frac{\Delta x_{i\partial\Omega}}{2}.$$

The boundary volume schemes become

$$
\begin{aligned}
&\Gamma_{i\partial\Omega} \subset B_1 : u_i \sum_{j\in N_i} \Delta x_{ij} = 0 \\
&\qquad \implies V_i \frac{d}{dt} u_i = -\sum_{j\in N_i} u_i \frac{\Delta y_{ij}}{2} - \sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} - g_1(y_i,t)\Delta y_{i\partial\Omega} \\
&\qquad = u_i \frac{\Delta y_{i\partial\Omega}}{2} - \sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} - g_1(y_i,t)\Delta y_{i\partial\Omega} \\
&\Gamma_{i\partial\Omega} \subset B_2 : u_i \sum_{j\in N_i} \Delta y_{ij} = 0 \\
&\qquad \implies V_i \frac{d}{dt} u_i = -\sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_i \frac{\Delta x_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} + g_2(x_i,t)\Delta x_{i\partial\Omega} \\
&\qquad = -\sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} - u_i \frac{\Delta x_{i\partial\Omega}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} + g_2(x_i,t)\Delta x_{i\partial\Omega} \\
&\Gamma_{i\partial\Omega} \subset B_3 : u_i \sum_{j\in N_i} \Delta x_{ij} = 0 \\
&\qquad \implies V_i \frac{d}{dt} u_i = -\sum_{j\in N_i} u_i \frac{\Delta y_{ij}}{2} - \sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} - u_i \Delta y_{i\partial\Omega} \\
&\qquad = -u_i \frac{\Delta y_{i\partial\Omega}}{2} - \sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} \\
&\Gamma_{i\partial\Omega} \subset B_4 : u_i \sum_{j\in N_i} \Delta y_{ij} = 0 \\
&\qquad \implies V_i \frac{d}{dt} u_i = -\sum_{j\in N_i} u_i \frac{\Delta y_{ij}}{2} - \sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + \sum_{j\in N_i} u_i \frac{\Delta x_{ij}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2} + u_i \Delta x_{i\partial\Omega} \\
&\qquad = -\sum_{j\in N_i} u_j \frac{\Delta y_{ij}}{2} + u_i \frac{\Delta x_{i\partial\Omega}}{2} + \sum_{j\in N_i} u_j \frac{\Delta x_{ij}}{2}.
\end{aligned}
$$

Hence Q_x, Q_y, become

$$
(Q_x)_{ij} = \frac{\Delta y_{ij}}{2}, \qquad (Q_x)_{ii} = \begin{cases} \frac{\Delta y_{i\partial\Omega}}{2}, & \Gamma_{i\partial\Omega} \subset B_3 \\ -\frac{\Delta y_{i\partial\Omega}}{2}, & \Gamma_{i\partial\Omega} \subset B_1 \\ 0, & \text{otherwise} \end{cases},
$$

$$
Q_y)_{ij} = -\frac{\Delta x_{ij}}{2}, \qquad (Q_y)_{ii} = \begin{cases} -\frac{\Delta x_{i\partial\Omega}}{2}, & \Gamma_{i\partial\Omega} \subset B_4 \\ \frac{\Delta x_{i\partial\Omega}}{2}, & \Gamma_{i\partial\Omega} \subset B_2 \\ 0, & \text{otherwise} \end{cases},
$$

and the scheme can be written as

$$
V \frac{d}{dt} \mathbf{u} = -Q_x \mathbf{u} - Q_y \mathbf{u} - g_1(y_i,t)\Delta y_{i\partial\Omega} e_{B_1} + g_2(x_i,t)\Delta x_{i\partial\Omega} e_{B_2}. \tag{3.13}
$$

We find the energy estimate for (3.13) to be

$$
\begin{aligned}
\frac{d}{dt}\left\|\mathbf{u}\right\|_V^2 = &- \sum_{i:\Gamma_{i\partial\Omega}\subset B_3} u_i^2 \Delta y_{i\partial\Omega} + \sum_{i:\Gamma_{i\partial\Omega}\subset B_4} u_i^2 \Delta x_{i\partial\Omega} \\
&- \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \left(-u_i^2 + 2u_i g_1(y_i,t)\right)\Delta y_{i\partial\Omega} + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \left(-u_i^2 + 2u_i g_2(x_i,t)\right)\Delta x_{i\partial\Omega} \\
\leq &- \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} \left(-(u_i - g_1(y_i,t))^2 + g_1^2(y_i,t)\right)\Delta y_{i\partial\Omega} + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} \left(-(u_i - g_2(x_i,t)^2 + g_2^2(x_i,t)\right)\Delta x_{i\partial\Omega} \\
\leq &- \sum_{i:\Gamma_{i\partial\Omega}\subset B_1} g_1^2(y_i,t)\Delta y_{i\partial\Omega} + \sum_{i:\Gamma_{i\partial\Omega}\subset B_2} g_2^2(x_i,t)\Delta x_{i\partial\Omega}.
\end{aligned}
$$

That is, the energy estimate is the same as for the scheme where we implemented the BC using the SAT.

To summarize, in this subsection we have illustrated that Godunov's method with the piecewise constant reconstructions and the central flux is able to produce schemes satisfying the SBP property. Since the schemes satisfy the SBP property we were able to derive energy estimates for them mimicking the energy estimates for the continuous problems. We must remark that the examples shown here are not novel, and refer the reader to [15, 16] for more indepth discussion.

4 High-order finite volume methods

In the previous section we saw that creating degree 0 reconstructions and applying the central flux together with the midpoint rule to approximate the fluxes yielded energy stable schemes for the simple linear hyperbolic problems of section 2.4. A natural idea to increase the accuracy order is to create a higher order reconstruction of u, and use a higher-order quadrature rule. In fact, this idea is quite old, and seems to originate in [48, 49]. The high-order Godunov method was further developed in [50] and [51]. It seems the k-exact method introduced in [24] is one of the earliest extensions to 2D problems. More recently, the spectral volume method [31] can be seen as an improvement on the k-exact method.

In this section we describe the k-exact finite volume method in the context of 2D problems.

4.1 k-exact method

Here we describe the k-exact method introduced in [24]. Suppose we follow the procedure in section 3.1 and obtain

$$\frac{d}{dt}u_i(t) + \frac{1}{V_i}\int_{\partial C_i} f(u(\mathbf{x},t))\cdot\hat{n}dS = 0, \tag{4.1}$$

where

$$u_i(t) \approx \frac{1}{V_i}\int_{C_i} u(\mathbf{x},t)d\mathbf{x}, \qquad \mathbf{u} = [u_1,..,u_N]^T.$$

Our goal is to find a high-order accurate approximation of

$$\int_{\partial C_i} f(u(\mathbf{x},t))\cdot\hat{n}dS.$$

Let $t = t^*$ be fixed and write $u(\mathbf{x},t^*) = u(\mathbf{x})$, $u_i(t^*) = u_i$. In the k-exact method we aim to find a degree k polynomial reconstruction $R_i^k(\mathbf{x}-p_i)$ of $u(\mathbf{x})$ using the data $\mathbf{u}$. Here p_i denotes the center of mass of C_i. This polynomial must satisfy the k-exactness property

$$R_i^k(\mathbf{x}-p_i) - u(\mathbf{x}) = \mathcal{O}(h^{k+1}), \qquad (\mathbf{x}\in C_i), \tag{4.2}$$

and the conservation property

$$\frac{1}{V_i}\int_{C_i} R_i^k(\mathbf{x}-p_i)d\mathbf{x} = u_i, \tag{4.3}$$

where $h = \max_{C_i}\sup_{a,b\in C_i}\|a-b\|_{\mathbb{R}^2}$ as usual. Assuming $u(\mathbf{x})$ is smooth we have

$$\begin{aligned} u(\mathbf{x}) &= u\big|_{p_i} + u_x\big|_{p_i}(x-x_i) + u_y\big|_{p_i}(y-y_i) + \frac{1}{2}u_{xx}\big|_{p_i}(x-x_i)^2 + u_{xy}\big|_{p_i}(x-x_i)(y-y_i) + ... \\ &= u\big|_{p_i} + \sum_{1\le n+m} \frac{1}{n!m!}\frac{\partial^{n+m}u}{\partial x^n\partial y^m}\bigg|_{p_i}(x-x_i)^n(y-y_i)^m. \end{aligned}$$

Consider representing R_i^k by its Taylor series as well:

$$\begin{aligned} R_i^k(\mathbf{x}-p_i) &= R_i^k\big|_{p_i} + (R_i^k)_x\big|_{p_i}(x-x_i) + (R_i^k)_y\big|_{p_i}(y-y_i) + \frac{1}{2}(R_i^k)_{xx}\big|_{p_i}(x-x_i)^2 + (R_i^k)_{xy}\big|_{p_i}(x-x_i)(y-y_i) + ... \\ &= R_i^k\big|_{p_i} + \sum_{1\le n+m\le k} \frac{1}{n!m!}\frac{\partial^{n+m}R_i^k}{\partial x^n\partial y^m}\bigg|_{p_i}(x-x_i)^n(y-y_i)^m. \end{aligned}$$

It follows that (4.2) is satisfied if

$$\frac{1}{n!m!}\frac{\partial^{n+m}R_i^k}{\partial x^n \partial y^m}\bigg|_{p_i} = \frac{1}{n!m!}\frac{\partial^{n+m}u}{\partial x^n \partial y^m}\bigg|_{p_i} + \mathcal{O}(h^{k+1-(m+n)}), \qquad (0 \le n+m \le k). \tag{4.4}$$

(Cf. [27]). Note that if

$$\left| u_i - \frac{1}{V_i}\int_{C_i} u(\mathbf{x})d\mathbf{x} \right| = \mathcal{O}(h^{k+1}),$$

then

$$\begin{aligned} u_i &= u\big|_{p_i} + \sum_{1\le n+m\le k} \frac{1}{n!m!}\frac{\partial^{n+m}u}{\partial x^n \partial y^m}\bigg|_{p_i} \frac{1}{V_i}\int_{C_i}(x-x_i)^n(y-y_i)^m + \mathcal{O}(h^{k+1}), \\ &= u\big|_{p_i} + \sum_{1\le n+m\le k} \frac{1}{n!m!}\frac{\partial^{n+m}u}{\partial x^n \partial y^m}\bigg|_{p_i} \overline{x^n y^m}_i + \mathcal{O}(h^{k+1}), \end{aligned}$$

where

$$\overline{x^n y^m}_i \equiv \frac{1}{V_i}\int_{C_i}(x-x_i)^n(y-y_i)^m dxdy,$$

are the polynomial basis moments over C_i. Further, if (4.3) is to be satisfied we must have that

$$\begin{aligned} u_i &= R_i^k\big|_{p_i} + \sum_{1\le n+m\le k} \frac{1}{n!m!}\frac{\partial^{n+m}R_i^k}{\partial x^n \partial y^m}\bigg|_{p_i} \frac{1}{V_i}\int_{C_i}(x-x_i)^n(y-y_i)^m + \mathcal{O}(h^{k+1}), \\ &= R_i^k\big|_{p_i} + \sum_{1\le n+m\le k} \frac{1}{n!m!}\frac{\partial^{n+m}R_i^k}{\partial x^n \partial y^m}\bigg|_{p_i} \overline{x^n y^m}_i. \end{aligned}$$

Thus, we have obtained one equation for $(k+2)(k+1)/2$ unknowns. To obtain the remaining equations we will proceed much in the same way, by requiring that

$$\left| \frac{1}{V_j}\int_{C_j} R_i^k(\mathbf{x}-p_i)d\mathbf{x} - u_j \right| = 0.$$

Note that

$$\begin{aligned} \frac{1}{V_j}\int_{C_j} R_i(\mathbf{x}-p_i)d\mathbf{x} =& R_i^k\big|_{p_i} + (R_i^k)_x\big|_{p_i}\frac{1}{V_j}\int_{C_j}(x-x_i)dxdy + (R_i^k)_y\big|_{p_i}\frac{1}{V_j}\int_{C_j}(y-y_i)dxdy \\ &+ (R_i^k)_{xx}\big|_{p_i}\frac{1}{2V_j}\int_{C_j}(x-x_i)^2dxdy + (R_i^k)_{xy}\big|_{p_i}\frac{1}{V_j}\int_{C_j}(x-x_i)(y-y_i)dxdy \\ &+ (R_i^k)_{yy}\big|_{p_i}\frac{1}{2V_j}\int_{C_j}(y-y_i)^2dxdy + ... \end{aligned}$$

To avoid integrating $(x-x_i)^n(y-y_i)^m$ over C_j for various j, we may substitute $(x-x_i)$ with $(x-x_j)+(x_j-x_i)$ and likewise for $(y-y_i)$. Then as shown in [25, 26] we obtain

$$\frac{1}{V_j}\int_{C_j} R_i(\mathbf{x}-p_i)d\mathbf{x} = u\big|_{p_i} + u_x\big|_{p_i}\hat{x}_{ij} + u_y\big|_{p_i}\hat{y}_{ij} + u_{xx}\big|_{p_i}\frac{\widehat{x^2}_{ij}}{2} + u_{xy}\big|_{p_i}\widehat{xy}_{ij} + u_{yy}\big|_{p_i}\frac{\widehat{y^2}_{ij}}{2} + ...,$$

where

$$\begin{aligned}\widehat{x^ny^m}_{ij} &\equiv \frac{1}{V_j}\int_{C_j}((x-x_j)+(x_j-x_i))^n((y-y_j)+(y_j-y_i))^m dxdy\\ &= \sum_{l=0}^{m}\sum_{k=0}^{n}\binom{m}{l}\binom{n}{k}(x_j-x_i)^k(y_j-y_i)^l\overline{x^{n-k}y^{m-l}}_j,\end{aligned}$$

(cf. Binomial Theorem). Now the idea is to determine the coefficients $\alpha_{mn} = (\partial_x^n\partial_y^m R_i^k\big|_{p_i})/(n!m!)$ by solving the linear system

$$M_i\alpha = [u_j]_{j=i}^{j_{\sigma-1}} \iff \begin{bmatrix} 1 & \overline{x}_i & \overline{y}_i & \overline{x^2}_i & \overline{xy}_i & \overline{y^2}_i & \cdots\\ 1 & \hat{x}_{ij_1} & \hat{y}_{ij_1} & \widehat{x^2}_{ij_1} & \widehat{xy}_{ij_1} & \widehat{y^2}_{ij_1} & \cdots\\ 1 & \hat{x}_{ij_2} & \hat{y}_{ij_2} & \widehat{x^2}_{ij_2} & \widehat{xy}_{ij_2} & \widehat{y^2}_{ij_2} & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots\\ 1 & \hat{x}_{ij_{\sigma-1}} & \hat{y}_{ij_{\sigma-1}} & \widehat{x^2}_{ij_{\sigma-1}} & \widehat{xy}_{ij_{\sigma-1}} & \widehat{y^2}_{ij_{\sigma-1}} & \cdots \end{bmatrix} \begin{bmatrix} R_i^k\big|_{p_i}\\ (R_i^k)_x\big|_{p_i}\\ (R_i^k)_y\big|_{p_i}\\ \frac{1}{2}(R_i^k)_{xx}\big|_{p_i}\\ (R_i^k)_{xy}\big|_{p_i}\\ \frac{1}{2}(R_i^k)_{yy}\big|_{p_i}\\ \vdots\\ \vdots \end{bmatrix} = \begin{bmatrix} u_i\\ u_{j_1}\\ u_{j_2}\\ u_{j_3}\\ \vdots\\ \vdots\\ \vdots\\ u_{j_{\sigma-1}} \end{bmatrix}, \tag{4.5}$$

where the stencil is $\{C_i\}\cup\{C_j\}_{j=j_1}^{j_{\sigma-1}}$ and

$$\sigma = \frac{(k+2)(k+1)}{2}.$$

If M_i is nonsingular, the system (4.5) can be solved and $R_i^k(\mathbf{x}-p_i)$ is obtained. To be precise, $\alpha = M_i^{-1}[u_j]_{j=i}^{j_{\sigma-1}}$ leads to

$$R_i^k(\mathbf{x}-p_i) = \sum_{m+n\leq k}\left(M_i^{-1}[u_j]_{j=i}^{j_{\sigma-1}}\right)_{mn}(x-x_i)^m(y-y_i)^n.$$

If we define a linear map

$$S_i : \mathbb{R}^N \to \mathbb{R}^\sigma, \qquad \mathbf{u} \mapsto u_ie_1 + u_{j_1}e_2 + \cdots + u_{j_{\sigma-1}}e_\sigma,$$

then

$$R_i^k(\mathbf{x}-p_i) = \sum_{m+n\leq k}\left(M_i^{-1}S_i\mathbf{u}\right)_{mn}(x-x_i)^m(y-y_i)^n.$$

Further, we factorize R_i^k as an inner product

$$R_i^k(\mathbf{x}-p_i) = \langle\alpha, e(\mathbf{x}-p_i)\rangle = \left\langle M_i^{-1}S_i\mathbf{u}, e(\mathbf{x}-p_i)\right\rangle, \tag{4.6}$$

where

$$e(\mathbf{x}-p_i) = \begin{bmatrix} 1\\ (x-x_i)\\ (y-y_i)\\ (x-x_i)^2\\ (x-x_i)(y-y_i)\\ \vdots \end{bmatrix}. \tag{4.7}$$

The expression (4.6) will be useful in section 5.

After obtaining reconstructions $\{R_i^k\}_{i=1}^N$, we follow the idea of Godunov's method (see section 3.4). That is, we will assume that

$$R_i^k(\mathbf{x}-p_i) \neq R_j^k(\mathbf{x}-p_j), \qquad (\mathbf{x}\in\Gamma_{ij}),$$

holds for all i, j. Then $u(\mathbf{x}, t^*)$ at Γ_{ij} is approximated by solving the Riemann problems

$$u_t + f(u)_\xi = 0, \qquad u(\xi, 0) = \begin{cases} R_i^k(\mathbf{x}-p_i), & \xi < 0 \\ R_j^k(\mathbf{x}-p_j), & \xi > 0 \end{cases} \tag{4.8}$$

either exactly or approximately. Here ξ is parallel to the normal vector on Γ_{ij} and $\xi = 0$ lies on Γ_{ij}. Using the notation

$$RP\left(R_i^k(\mathbf{x}-p_i), R_j^k(\mathbf{x}-p_j)\right)$$

to denote the exact or approximate solution of (4.8) at $\xi = 0$ and $t > 0$, the k-exact approximation of (4.1) is given by

$$\frac{d}{dt}u_i + \frac{1}{V_i}\sum_{j\in N_i}\int_{\Gamma_{ij}} f\left(RP\left(R_i^k(\mathbf{x}-p_i), R_j^k(\mathbf{x}-p_j)\right)\right)\cdot\hat{n}dS = 0,$$

where $\partial C_i = \cup_{j\in N_i}\Gamma_{ij}$. After applying some quadrature rule, we obtain

$$\frac{d}{dt}u_i + \frac{1}{V_i}\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq} f\left(RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right)\right) = 0,$$

where m is the number of quadrature points on Γ_{ij}, w_{ijq} are the quadrature weights, and $\mathbf{x}_{ijq}$ are the quadrature points.

Remark. As noted in [26], we can use Gauss' theorem to write

$$\begin{aligned}\overline{x^n y^m}_i &\equiv \frac{1}{V_i}\int_{C_i}(x-x_i)^n(y-y_i)^m dxdy = \frac{1}{(n+1)V_i}\int_{C_i}\nabla\cdot[(x-x_i)^{n+1}(y-y_i)^m, 0]dxdy \\ &= \frac{1}{(n+1)V_i}\int_{\partial C_i}(x-x_i)^{n+1}(y-y_i)^m\cdot\hat{n}_1 dS\end{aligned}$$

where $\hat{n}_1$ is the first component of the outward normal vector. Now we can use Gaussian quadrature to find the moments with accuracy order $k+1$ by $m = \lceil (k+1)/2 \rceil$ quadrature points on each subset Γ_{ij} of ∂C_i. Suppose each Γ_{ij} is a straight line and let $(x_{ij,0}, y_{ij,0}), (x_{ij,1}, y_{ij,1})$ denote their endpoints. Then (A.3) implies

$$\begin{aligned}&\overline{x^n y^m}_i = \frac{1}{(n+1)V_i}\sum_{j\in N_i}\int_{\Gamma_{ij}}(x-x_i)^{n+1}(y-y_i)^m dy \\ &= \frac{1}{(n+1)V_i}\sum_{j\in N_i}\frac{y_{ij,1}-y_{ij,0}}{2}\int_{-1}^{1}\left(\frac{x_{ij,1}-x_{ij,0}}{2}\xi + \frac{x_{ij,1}+x_{ij,0}}{2} - x_i\right)^{n+1}\left(\frac{y_{ij,1}-y_{ij,0}}{2}\xi + \frac{y_{ij,1}+y_{ij,0}}{2} - y_i\right)^m d\xi \\ &= \frac{1}{(n+1)V_i}\sum_{j\in N_i}\frac{\Delta y_{ij}}{2}\int_{-1}^{1}\left(\frac{\Delta x_{ij}}{2}\xi + \frac{x_{ij,1}+x_{ij,0}}{2} - x_i\right)^{n+1}\left(\frac{\Delta y_{ij}}{2}\xi + \frac{y_{ij,1}+y_{ij,0}}{2} - y_i\right)^m d\xi,\end{aligned}$$

where $\Delta x_{ij} = x_{ij,1} - x_{ij,0}$ and similarly for Δy_{ij}.

Remark. The moments $\overline{x^n y^m}, \widehat{x^n y^m}$ over each control volume need only be calculated in the initialization of the numerical method. The stored values are then to be used in the reconstructions in each time step.

It might occur that we choose some stencil $\{C_i\} \cup \{C_j\}_{j=j_1}^{j_{\sigma-1}}$ leading to a singular coefficient matrix M_i. To avoid this problem there are various stencil selection techniques, or reconstruction methods, to choose from. The different reconstruction methods also attempt to solve some other downsides of the k-exact FVM. We describe some of these in the following text.

4.1.1 Least-squares reconstruction

Consider determining α using a least-squares approximation, minimizing the difference between u_j and the volume-averaged value of R_i^k over C_j for all C_j in the stencil. The main motivation for using a least-squares approximation is that the linear system obtained from a stencil of $\sigma = (k+2)(k+1)/2$ elements might be singular. Hence we increase the stencil to guarantee existence of (least-squares) solutions. To guarantee that the reconstructions satisfy the conservation condition, we must set the first equation as a constraint. Sometimes geometric weights are applied such that the data closer to p_i is prioritized in the approximation. In short, the reconstruction coefficients are found by solving the constrained least-squares problem

$$\left[\begin{array}{ccccccc} 1 & \overline{x}_i & \overline{y}_i & \overline{x^2}_i & \overline{xy}_i & \overline{y^2}_i & \cdots \\ \hline \gamma_{ij_1} & \gamma_{ij_1}\hat{x}_{ij_1} & \gamma_{ij_1}\hat{y}_{ij_1} & \gamma_{ij_1}\widehat{x^2}_{ij_1} & \gamma_{ij_1}\widehat{xy}_{ij_1} & \gamma_{ij_1}\widehat{y^2}_{ij_1} & \cdots \\ \gamma_{ij_2} & \gamma_{ij_2}\hat{x}_{ij_2} & \gamma_{ij_2}\hat{y}_{ij_2} & \gamma_{ij_2}\widehat{x^2}_{ij_2} & \gamma_{ij_2}\widehat{xy}_{ij_2} & \gamma_{ij_2}\widehat{y^2}_{ij_2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \gamma_{ij_{\sigma_i}} & \gamma_{ij_{\sigma_i}}\hat{x}_{ij_{\sigma_i}} & \gamma_{ij_{\sigma_i}}\hat{y}_{ij_{\sigma_i}} & \gamma_{ij_{\sigma_i}}\widehat{x^2}_{ij_{\sigma_i}} & \gamma_{ij_{\sigma_i}}\widehat{xy}_{ij_{\sigma_i}} & \gamma_{ij_{\sigma_i}}\widehat{y^2}_{ij_{\sigma_i}} & \cdots \end{array}\right] \begin{bmatrix} R_i^k\big|_{p_i} \\ (R_i^k)_x\big|_{p_i} \\ (R_i^k)_y\big|_{p_i} \\ \frac{1}{2}(R_i^k)_{xx}\big|_{p_i} \\ (R_i^k)_{xy}\big|_{p_i} \\ \frac{1}{2}(R_i^k)_{yy}\big|_{p_i} \\ \vdots \\ \vdots \end{bmatrix} = \left[\begin{array}{c} u_i \\ \hline \gamma_{ij_1}u_{j_1} \\ \gamma_{ij_2}u_{j_2} \\ \gamma_{ij_3}u_{j_3} \\ \vdots \\ \vdots \\ \vdots \\ \gamma_{ij_{\sigma_i}}u_{j_{\sigma_i}} \end{array}\right],$$

where the equation above the line is the constraint. Here the stencil consists of $\{C_i\} \cup \{C_j\}_{j=j_1}^{j_{\sigma_i}}$ for some $\sigma_i \geq \sigma = (k+2)(k+1)/2$. The geometric weights γ_{ij} can for instance be defined by

$$\gamma_{ij} = \frac{1}{\|p_i - p_j\|_{\mathbb{R}^2}}.$$

Recall that the least-squares problem can be solved using the QR factorization. To make sure that the constraint is satisfied exactly, we proceed as in [25]. Begin by eliminating the first column by subtracting $\gamma_{ij} \cdot$ [constraint] to obtain

$$\left[\begin{array}{cccccc} 1 & \overline{x}_i & \overline{y}_i & \overline{x^2}_i & \overline{xy}_i & \cdots \\ \hline 0 & \gamma_{ij_1}(\hat{x}_{ij_1} - \overline{x}_i) & \gamma_{ij_1}(\hat{y}_{ij_1} - \overline{y}_i) & \gamma_{ij_1}(\widehat{x^2}_{ij_1} - \overline{x^2}_i) & \gamma_{ij_1}(\widehat{xy}_{ij_1} - \overline{xy}_i) & \cdots \\ 0 & \gamma_{ij_2}(\hat{x}_{ij_2} - \overline{x}_i) & \gamma_{ij_2}(\hat{y}_{ij_2} - \overline{y}_i) & \gamma_{ij_2}(\widehat{x^2}_{ij_2} - \overline{x^2}_i) & \gamma_{ij_2}(\widehat{xy}_{ij_2} - \overline{xy}_i) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ 0 & \gamma_{ij_{\sigma_i}}(\hat{x}_{ij_{\sigma_i}} - \overline{x}_i) & \gamma_{ij_{\sigma_i}}(\hat{y}_{ij_{\sigma_i}} - \overline{y}_i) & \gamma_{ij_{\sigma_i}}(\widehat{x^2}_{ij_{\sigma_i}} - \overline{x^2}_i) & \gamma_{ij_{\sigma_i}}(\widehat{xy}_{ij_{\sigma_i}} - \overline{xy}_i) & \cdots \end{array}\right] \begin{bmatrix} R_i^k\big|_{p_i} \\ (R_i^k)_x\big|_{p_i} \\ (R_i^k)_y\big|_{p_i} \\ \frac{1}{2}(R_i^k)_{xx}\big|_{p_i} \\ (R_i^k)_{xy}\big|_{p_i} \\ \frac{1}{2}(R_i^k)_{yy}\big|_{p_i} \\ \vdots \\ \vdots \end{bmatrix} = \left[\begin{array}{c} u_i \\ \hline \gamma_{ij_1}(u_{j_1} - u_i) \\ \gamma_{ij_2}(u_{j_2} - u_i) \\ \gamma_{ij_3}(u_{j_3} - u_i) \\ \vdots \\ \vdots \\ \vdots \\ \gamma_{ij_{\sigma_i}}(u_{j_{\sigma_i}} - u_i) \end{array}\right].$$

Next, use householder reflectors to obtain the QR factorization of the matrix. Since the $\sigma \times \sigma$ submatrix $\hat{R}_i$ of R_i is upper triangular, we obtain α by back-substitution. Since $R_i^k\big|_{p_i}$ remains a free variable up to the last equation (the constraint), it is chosen such that the constraint is satisfied exactly. If we define the linear map

$$S_i : \mathbb{R}^N \to \mathbb{R}^{\sigma_i+1}, \qquad \mathbf{u} \mapsto u_i e_1 + \gamma_{ij_1}(u_{j_1} - u_i)e_2 + \cdots + \gamma_{ij_{\sigma_i}}(u_{j_{\sigma_i}} - u_i)e_{\sigma_i+1},$$

then

$$\alpha = \hat{R}_i^{-1} Q_i^T S_i \mathbf{u}, \qquad \text{and} \qquad R_i^k(\mathbf{x} - p_i) = \left\langle \hat{R}_i^{-1} Q_i^T S_i \mathbf{u}, e(\mathbf{x} - p_i) \right\rangle,$$

where $e(\mathbf{x} - p_i)$ is given by (4.7).

4.1.2 ENO reconstruction

Consider the case where the exact solution $u(\mathbf{x}, t)$ contains a discontinuity at the current time step. If the reconstruction stencil $\{C_i\} \cup \{C_j\}_{j=j_1}^{j_{\sigma_i}}$ contains the discontinuity we may find that the reconstruction behaves oscillatory (cf. Gibbs phenomenon). With the goal of producing more "physically correct" numerical solutions we would therefore prefer to choose a different stencil which does not contain the discontinuity. This is the goal of the essentially non-oscillatory (ENO) reconstruction procedure. To define the term essentially non-oscillatory, suppose R is a k-exact ENO reconstruction of some function u. Then

$$TV(R) \leq TV(u) + \mathcal{O}(h^k),$$

where TV denotes the total variation. As noted in [28], the reconstruction procedure described in the seminal paper [51] cannot be directly extended to the multidimensional case. Therefore we will present the two ENO reconstruction methods given in [27]:

1. Let the reconstruction in cell C_i be given by

   $$R_i(\mathbf{x} - p_i) = \theta_i u_i + (1 - \theta_i) R_i^k(\mathbf{x} - p_i).$$

   Here $R_i^k(\mathbf{x} - p_i)$ is the usual k-exact reconstruction obtained using the possibly "bad" stencil. The parameter θ_i is given by

   $$0 \leq \theta_i \leq 1, \qquad \theta_i \approx 1 \quad \text{when the stencil contains the discontinuity,}$$

   and $\theta_i = \mathcal{O}(h^{k+1})$ otherwise. We see that this approach reduces the reconstruction to first order accuracy when the stencil is "bad", and retains the $k+1$ order accuracy otherwise. It seems possible that we may quickly lose the overall accuracy of the approximation as more and more reconstructions reduce to first order accuracy.

2. Instead of reducing the accuracy to first order, we choose to find a new stencil which does not contain the discontinuity. There might be many such stencilcs, so we must determine which one to use. Suppose $\mathbf{S}$ is the set of stencils available to give a k-exact reconstruction. Compute the reconstruction using each stencil and define a measure

   $$\sigma_r = \sum_{m+n \leq k} |\alpha_{mn}^r|,$$

   where α^r is the coefficient vector for the reconstruction obtained using stencil $S_r \in \mathbf{S}$. We recall that the components in α are approximations of the partial derivatives of u evaluated at p_i. Choose the stencil $S_r \in \mathbf{S}$ minimizing σ_r.

4.1.3 Other reconstruction methods

We have very briefly described two common reconstruction methods. Other notable reconstruction methods are the compact least squares reconstruction method recently developed in [52, 53, 54] and the various weighted ENO, or simply WENO, reconstruction methods developed in for instance [55, 56, 57]. Further, we mention that one can consider the reconstruction problem in some other basis than the standard polynomial basis.

4.2 Spectral volume method

The spectral volume method aims to simplify the reconstruction problem. Let $\{p_i\}_{i=1}^N$ be some set of points in Ω generating some unstructured triangular grid. Following the literature [29, 30, 31, 32] we will consider each triangle to be a spectral volume SV_i.

As before we need $(k+2)(k+1)/2$ degrees of freedom in order to find a k-exact polynomial reconstruction of $u(\mathbf{x})$ inside SV_i. Unlike the k-exact method we partition SV_i into $\sigma = (k+2)(k+1)/2$ control volumes $C_{i,j}$ $(j = 1, ..., \sigma)$ and define

$$u_{i,j}(t) \approx \frac{1}{V_{i,j}} \int_{C_{i,j}} u(\mathbf{x}, t) d\mathbf{x}.$$

We require that the reconstruction P_i of u in SV_i conserves the volume-average value of u in each control volume contained in SV_i. In other words,

$$\frac{1}{V_{i,j}} \int_{C_{i,j}} P_i(x, y) dxdy = u_{i,j}.$$

Further, we begin by obtaining a reconstruction on some reference element SV_r. In particular, suppose the center of mass for SV_r is $\mathbf{x}_r = (0, 0)$. Then the reconstruction problem reads: Find the coefficient vector α satisfying

$$\frac{1}{V_{r,j}} \sum_{l=1}^{\sigma} \alpha_l \int_{C_{r,j}} e_l(x, y) dxdy = u_{r,j}$$

Put $u_r = [u_{r,1}, u_{r,2}, \dots, u_{r,\sigma}]^T$ so that the system of equations can be written as

$$M\alpha = u_r$$

where M is the matrix defined by

$$M = \begin{bmatrix} \frac{1}{V_{r,1}} \int_{C_{r,1}} e_1(x, y) dxdy & \dots & \frac{1}{V_{r,1}} \int_{C_{r,1}} e_\sigma(x, y) dxdy \\ \vdots & \vdots & \vdots \\ \frac{1}{V_{r,\sigma}} \int_{C_{r,\sigma}} e_1(x, y) dxdy & \dots & \frac{1}{V_{r,\sigma}} \int_{C_{r,\sigma}} e_\sigma(x, y) dxdy \end{bmatrix}$$

Note that the reconstruction stencil is given by the control volumes inside the spectral volume. Since the control volumes can be defined however we see fit, the idea is that we can guarantee a nonsingular matrix M. Hence we do not have to increase the stencil above σ elements and use a least-squares approximation, for instance. Further, since each spectral volume is geometrically similar (here: triangles), we can use geometrically similar control volumes for each of them, and hence the same reconstruction stencil. Assuming

M is nonsingular we have $\alpha = M^{-1}u_r$ and we obtain the reconstruction inside SV_r,

$$P_r(x,y) = \sum_{j=1}^{\sigma} L_j(x,y)u_{r,j} = Lu_r,$$

where $L \equiv e(x,y)M^{-1}$. After obtaining P_r, we can transform it into P_i for $i = 1,\dots,N$. Noting that M might be ill-conditioned, it was recommended in [31] to perform analytical inversion via Mathematica. Finally, since the reconstruction P_i is continuous inside the spectral volume, we do not obtain Riemann problems on the boundaries between control volumes lying in the same spectral volume, which simplifes the flux calculations. Of course, at the boundaries between spectral volumes we still need to apply a numerical flux or approximate Riemann solver. The scheme updates the control volume averages at each time step

$$\frac{d}{dt}u_{i,j} = -\frac{1}{V_{i,j}} \sum_{k\in N_{i,j}} \int_{\Gamma_{i,j,k}} f(u(\mathbf{x},t)) \cdot \hat{n} dS,$$

where $u_{i,j}$ denotes the volume averaged value in $C_{i,j}$, and $\partial C_{i,j} = \cup_{k\in N_{i,j}}\Gamma_{i,j,k}$. Assuming the boundary faces $\Gamma_{i,j,k}$ are contained inside SV_i for $k \in N_{i,j}^{\delta} \subset N_{i,j}$ we obtain

$$\frac{d}{dt}u_{i,j} = -\frac{1}{V_{i,j}} \sum_{k\in N_{i,j}^{\delta}} \int_{\Gamma_{i,j,k}} f(P_i(\mathbf{x})) \cdot \hat{n} dS - \frac{1}{V_{i,j}} \sum_{k\in N_{i,j}\setminus N_{i,j}^{\delta}} \int_{\Gamma_{i,j,k}} f(RP\left(P_i(\mathbf{x}), P_k(\mathbf{x})\right)) \cdot \hat{n} dS.$$

Applying quadrature changes the scheme in the familiar way.

Remark. The control volumes inside any given spectral volume can be general polygons.

5 Stability analysis of k-exact schemes

In this section we examine the energy stability of schemes obtained via the k-exact method.

5.1 Problem 1: 1D linear advection

Inspired by [15, 6, 5, 9] we consider problem 1 from section 2.4.

Let $x \in \Omega = [0,1] \subset \mathbb{R}$ and $t \in [0,T] \subset \mathbb{R}^+$. Find the function $u : \Omega \times [0,T] \to \mathbb{R}$ satisfying

$$u_t + u_x = 0, \qquad u(x,0) = f(x), \qquad u(0,t) = g(t),$$

where $f(x) = \sin(2\pi x)$ and $g(t) = \sin(-2\pi t)$.

In section 2.4 we found the energy rate to be

$$\frac{d}{dt}\|u(\cdot,t)\|_2^2 = u^2(0,t) - u^2(1,t) \le g^2(t). \tag{5.1}$$

We want to determine if schemes obtained by the k-exact method satisfisy a discrete equivalent energy rate. In order to do so, we begin by finding the general k-exact scheme approximating the PDE. Recall that schemes approximating the PDE can be obtained via the k-exact method as follows: We partition Ω into N control volumes $\{C_i\}_{i=1}^N$ satisfying

$$C_i \cap C_j = \emptyset \quad (i \neq j), \qquad \text{and} \qquad \bigcup_{i=1}^{N} \overline{C_i} = \overline{\Omega}.$$

Let C_i^-, C_i^+ denote the lower and upper bounds of C_i respectively, and let V_i denote the measure of C_i. Further, we will use the notation $u_i(t)$ to denote an approximation of the volume-averaged value of $u(x,t)$ over C_i. That is,

$$u_i(t) \approx \frac{1}{V_i}\int_{C_i} u(x,t)dx, \qquad i = 1,\dots,N.$$

Integrating the PDE over C_i gives us

$$\int_{C_i} u_t(x,t)dx + \int_{C_i} u_x(x,t)dx = 0 \qquad \Longleftrightarrow \qquad \frac{d}{dt}\int_{C_i} u(x,t)dx + u(C_i^+,t) - u(C_i^-,t) = 0,$$

and the finite volume approximation

$$V_i \frac{d}{dt} u_i(t) = -u(C_i^+,t) + u(C_i^-,t). \tag{5.2}$$

Since $u(x,t)$ is the unknown solution of the continuous problem, we must somehow approximate the control volume boundary evaluations $u(C_i^{\pm},t)$. Considering a fixed time $t = t^*$, we ignore the time dependence of $u(x,t)$ and $u_i(t)$, writing $u(x,t^*) = u(x)$ and $u_i(t^*) = u_i$. The k-exact method proceeds by finding polynomial functions $R_i^k(x - x_i)$ for $i = 1,\dots,N$ satisfying

1.
$$R_i^k(x - x_i) - u(x) = \mathcal{O}(h^{k+1}), \qquad x \in C_i,$$

2.
$$\frac{1}{V_i}\int_{C_i} R_i^k(x-x_i)dx = u_i.$$

Here $h = \max_{C_i}\sup_{x,y\in C_i}|x-y|$ and x_i is the center of mass of C_i. We say that $R_i^k(x-x_i)$ is a (k-exact) reconstruction of $u(x)$ in C_i. As shown in section 4, the reconstructions can be written as inner products,

$$R_i^k(x-x_i) = \langle e(x-x_i), L_i\mathbf{u}\rangle, \tag{5.3}$$

where $e(x-x_i)$, $L_i\mathbf{u}\in\mathbb{R}^{k+1}$ are given by

$$(e(x-x_i))_j = (x-x_i)^{j-1}, \qquad (L_i\mathbf{u})_j = \frac{1}{(j-1)!}\frac{d^{j-1}}{dx^{j-1}}R_i^k(x_i). \tag{5.4}$$

In general, the reconstructions will not match at the control volume boundaries. Hence, following the idea of Godunov's method (cf. section 3.4), we approximate the values $u(C_i^+)$ by solving the Riemann problems

$$u_t + u_x = 0, \qquad u(x,0) = \begin{cases} R_i^k(C_i^+ - x_i), & x < C_i^+ \\ R_{j_1}^k(C_{j_1}^- - x_{j_1}), & x > C_i^+ \end{cases}, \tag{5.5}$$

either exactly or approximately, where C_{j_1} is the control volume adjacent to C_i such that $C_{j_1}^- = C_i^+$. If we use the notation

$$RP\left(R_i^k(C_i^+ - x_i), R_{j_1}^k(C_{j_1}^- - x_{j_1})\right),$$

to denote the exact or approximate solution of (5.5) at $x = C_i^+$ for $t > 0$, we obtain

$$u(C_i^+) \approx RP\left(R_i^k(C_i^+ - x_i), R_{j_1}^k(C_{j_1}^- - x_{j_1})\right).$$

Likewise, if C_{j_2} is the control volume satisfying $C_{j_2}^+ = C_i^-$, we obtain

$$u(C_i^-) \approx RP\left(R_{j_2}^k(C_{j_2}^+ - x_{j_2}), R_i^k(C_i^- - x_i)\right).$$

Note that $C_i^+ - x_i = V_i/2$ and $C_i^- - x_i = -V_i/2$. Combining this with (5.3) we obtain cleaner expressions for $u(C_i^\pm)$,

$$u(C_i^+) \approx RP\left(\langle e(V_i/2), L_i\mathbf{u}\rangle, \langle e(-V_{j_1}/2), L_{j_1}\mathbf{u}\rangle\right), \qquad u(C_i^-) \approx RP\left(\langle e(V_{j_2}/2), L_{j_2}\mathbf{u}\rangle, \langle e(-V_i/2), L_i\mathbf{u}\rangle\right).$$

Consider the case where C_i is the left boundary volume, meaning that some control volume $C_{j_2}\in\{C_i\}_{i=1}^N$ satisfying $C_{j_2}^+ = C_i^-$ does not exist. In this case we use the approximation

$$u(C_i^-) \approx R_i^k(C_i^- - x_i) = \langle e(-V_i/2), L_i\mathbf{u}\rangle.$$

Similarly, if C_i is the right boundary volume, meaning that some $C_{j_1}\in\{C_i\}_{i=1}^N$ satifying $C_{j_1}^- = C_i^+$ does not exist, we use the approximation

$$u(C_i^+) \approx R_i^k(C_i^+ - x_i) = \langle e(V_i/2), L_i\mathbf{u}\rangle.$$

By (5.2) and the above it follows that the general k-exact scheme approximating $u_t + u_x = 0$ is given by

$$V_i \frac{d}{dt} u_i = \begin{cases} -RP\left(\langle e(V_i/2), L_i \mathbf{u}\rangle, \langle e(-V_{j_1}/2), L_{j_1}\mathbf{u}\rangle\right) + \langle e(-V_i/2), L_i \mathbf{u}\rangle, & C_i \text{ is the left boundary volume} \\ -RP\left(\langle e(V_i/2), L_i \mathbf{u}\rangle, \langle e(-V_{j_1}/2), L_{j_1}\mathbf{u}\rangle\right) & \\ \qquad +RP\left(\langle e(V_{j_2}/2), L_{j_2} \mathbf{u}\rangle, \langle e(-V_i/2), L_i\mathbf{u}\rangle\right), & C_i \text{ is an interior volume} \\ -\langle e(V_i/2), L_i \mathbf{u}\rangle + RP\left(\langle e(V_{j_2}/2), L_{j_2} \mathbf{u}\rangle, \langle e(-V_i/2), L_i\mathbf{u}\rangle\right), & C_i \text{ is the right boundary volume.} \end{cases} \tag{5.6}$$

Remark. In the above, it is understood that the control volumes C_{j_1}, C_{j_2} are adjacent to C_i, i.e. the indices j_1, j_2 are dependent on the index i.

In order to analyze the stability of the scheme, we will write it in the (spatially) global form

$$V \frac{d}{dt}\mathbf{u} = -K, \tag{5.7}$$

where $V = \operatorname{diag}(V_1, V_2, \ldots, V_N)$, $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$ and K is the vector such that $-K_i =$ the right hand side of (5.6). Recall from section 3.3 that V induces a discrete L^2 norm: $\langle \mathbf{u}, V\mathbf{u}\rangle = \|\mathbf{u}\|_V^2$. Taking the inner product of (5.7) with $\mathbf{u}$ and adding the transposed equation we obtain

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle. \tag{5.8}$$

Our goal is to determine if, or when, (5.8) mimics (5.1). That is, if or when

$$-\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle \leq u^2(C_L^-) - u^2(C_R^+).$$

where C_L is the left boundary volume and C_R is the right boundary volume. At this stage it is not possible to determine the above, as we have not specified the function RP which resides in K. Suppose that we apply the approximation

$$RP(a, b) = \frac{a+b}{2}.$$

Then (5.6) becomes

$$V_i \frac{d}{dt} u_i = \begin{cases} -\left\langle \frac{e(V_i/2)}{2}, L_i \mathbf{u}\right\rangle - \left\langle \frac{e(-V_{j_1}/2)}{2}, L_{j_1}\mathbf{u}\right\rangle + \langle e(-V_i/2), L_i \mathbf{u}\rangle, & C_i \text{ is the left boundary volume} \\ -\left\langle \frac{e(V_i/2)}{2}, L_i \mathbf{u}\right\rangle - \left\langle \frac{e(-V_{j_1}/2)}{2}, L_{j_1}\mathbf{u}\right\rangle & \\ \qquad +\left\langle \frac{e(V_{j_2}/2)}{2}, L_{j_2} \mathbf{u}\right\rangle + \left\langle \frac{e(-V_i/2)}{2}, L_i\mathbf{u}\right\rangle, & C_i \text{ is an interior volume} \\ -\langle e(V_i/2), L_i \mathbf{u}\rangle + \left\langle \frac{e(V_{j_2}/2)}{2}, L_{j_2} \mathbf{u}\right\rangle + \left\langle \frac{e(-V_i/2)}{2}, L_i\mathbf{u}\right\rangle, & C_i \text{ is the right boundary volume,} \end{cases} \tag{5.9}$$

where the indices j_1, j_2 are dependent on i as before. Let $\mathbf{Lu} = [L_1\mathbf{u}, L_2\mathbf{u}, \ldots, L_N\mathbf{u}]^T$ and define a matrix Q by

$$Q_{ii} = \frac{e(V_i/2) - e(-V_i/2)}{2}, \qquad Q_{ij} = \begin{cases} \frac{e(-V_j/2)}{2}, & C_j^- = C_i^+ \\ -\frac{e(V_j/2)}{2}, & C_j^+ = C_i^- \\ 0, & \text{otherwise,} \end{cases}$$

for all indices i corresponding to interior volumes. If C_L denotes the left boundary volume, and C_R denotes the right boundary volume, we define

$$Q_{LL} = -e(-V_L/2) + \frac{e(V_L/2)}{2} = -\frac{e(-V_L/2)}{2} + \frac{e(V_L/2) - e(-V_L/2)}{2}, \qquad Q_{Lj} = \begin{cases} \frac{e(-V_j/2)}{2}, & C_j^- = C_L^+ \\ 0, & \text{otherwise}, \end{cases}$$

$$Q_{RR} = e(V_R/2) - \frac{e(-V_R/2)}{2} = \frac{e(V_R/2)}{2} + \frac{e(V_R/2) - e(-V_R/2)}{2}, \qquad Q_{Rj} = \begin{cases} -\frac{e(V_j/2)}{2}, & C_j^+ = C_R^- \\ 0, & \text{otherwise}. \end{cases}$$

Now the scheme (5.9) can be written as

$$V\frac{d}{dt}\mathbf{u} = -Q\mathbf{L}\mathbf{u}, \tag{5.10}$$

and the energy rate is

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle - \langle Q\mathbf{L}\mathbf{u}, \mathbf{u}\rangle.$$

Remark. By (5.4) it is clear that

$$\left(\frac{e(V_i/2) - e(-V_i/2)}{2}\right)_j = \frac{1}{2}\left(\frac{V_i^{j-1}}{2^{j-1}} - \frac{(-V_i)^{j-1}}{2^{j-1}}\right) = \frac{1}{2^j}\left(V_i^{j-1} - (-V_i)^{j-1}\right) = \begin{cases} \frac{1}{2^{j-1}}(V_i^{j-1}), & j \text{ is even} \\ 0, & j \text{ is odd} \end{cases}, \tag{5.11}$$

$$\iff \frac{e(V_i/2) - e(-V_i/2)}{2} = \left[0, \frac{V_i}{2}, 0, \frac{V_i^3}{8}, 0, \dots, \frac{V_i^k}{2^k}\right]^T \tag{5.12}$$

Remark. If we use the 0-exact reconstruction $R_i^0 = u_i$ we recover the SBP scheme examined in section 3.5.

To simplify the explicit form of $-\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle - \langle Q\mathbf{L}\mathbf{u}, \mathbf{u}\rangle$ we begin by considering the case where the volumes are structured and regular: Let the control volumes be structured and regular, meaning that

1. $C_{i-1}^+ = C_i^-$,
2. $V_i = V_j = h$,
3. $C_i^+ - x_i = -(C_i^- - x_i) = h/2$,

for all i, j. Property 1 implies C_1 is the left boundary volume and C_N is the right boundary volume. In this case Q becomes

$$Q_{ii} = \begin{cases} -\frac{e(-h/2)}{2} + \frac{e(h/2) - e(-h/2)}{2}, & i = 1 \\ \frac{e(h/2) - e(-h/2)}{2}, & i = 2, \dots, N-1 \\ \frac{e(h/2)}{2} + \frac{e(h/2) - e(-h/2)}{2}, & i = N \end{cases}, \qquad Q_{ij} = \begin{cases} \frac{e(-h/2)}{2}, & j = i+1 \\ -\frac{e(h/2)}{2}, & j = i-1 \\ 0, & \text{otherwise}. \end{cases}$$

Note that the definition of Q implies

$$\begin{aligned} \langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle = &-u_1\left\langle \frac{e(-h/2)}{2}, L_1\mathbf{u}\right\rangle + u_1\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_1\mathbf{u}\right\rangle + u_1\left\langle \frac{e(-h/2)}{2}, L_2\mathbf{u}\right\rangle \\ &+ \sum_{i=2}^{N-1} -u_i\left\langle \frac{e(h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle + u_i\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_i\mathbf{u}\right\rangle + u_i\left\langle \frac{e(-h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle \\ &- u_N\left\langle \frac{e(h/2)}{2}, L_{N-1}\mathbf{u}\right\rangle + u_N\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_N\mathbf{u}\right\rangle + u_N\left\langle \frac{e(h/2)}{2}, L_N\mathbf{u}\right\rangle. \end{aligned}$$

Next, recall that the inner product is defined as $\langle Q\mathbf{L}\mathbf{u},\mathbf{u}\rangle = \mathbf{L}\mathbf{u}^T Q^T \mathbf{u}$. We determine Q^T to be

$$Q^T_{ii} = \begin{cases} -\frac{e(-h/2)}{2} + \frac{e(h/2)-e(-h/2)}{2}, & i=1 \\ \frac{e(h/2)-e(-h/2)}{2}, & i=2,\dots,N-1 \\ \frac{e(h/2)}{2} + \frac{e(h/2)-e(-h/2)}{2}, & i=N \end{cases}, \qquad Q^T_{ij} = \begin{cases} -\frac{e(h/2)}{2}, & j=i+1 \\ \frac{e(-h/2)}{2}, & j=i-1 \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} \langle Q\mathbf{L}\mathbf{u},\mathbf{u}\rangle &= -u_1\left\langle \frac{-e(-h/2)}{2}, L_1\mathbf{u}\right\rangle + u_1\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_1\mathbf{u}\right\rangle - u_1\left\langle \frac{e(h/2)}{2}, L_2\mathbf{u}\right\rangle \\ &+ \sum_{i=2}^{N-1} u_i\left\langle \frac{e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle + u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u}\right\rangle - u_i\left\langle \frac{e(h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle \\ &+ u_N\left\langle \frac{e(-h/2)}{2}, L_{N-1}\mathbf{u}\right\rangle + u_N\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_N\mathbf{u}\right\rangle + u_N\left\langle \frac{e(h/2)}{2}, L_N\mathbf{u}\right\rangle. \end{aligned}$$

Further, we find

$$\begin{aligned} \langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle + \langle Q\mathbf{L}\mathbf{u},\mathbf{u}\rangle &= -u_1\left\langle e(-h/2), L_1\mathbf{u}\right\rangle + 2u_1\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_1\mathbf{u}\right\rangle - u_1\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_2\mathbf{u}\right\rangle \\ &+ \sum_{i=2}^{N-1} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle + 2u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u}\right\rangle - u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle \\ &- u_N\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{N-1}\mathbf{u}\right\rangle + 2u_N\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_N\mathbf{u}\right\rangle + u_N\langle e(h/2), L_N\mathbf{u}\rangle. \end{aligned}$$

Note that

$$-u_1\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_2\mathbf{u}\right\rangle + \sum_{i=2}^{N-1} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle = \sum_{i=1}^{N-1} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle,$$

and

$$-u_N\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{N-1}\mathbf{u}\right\rangle + \sum_{i=2}^{N-1} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle = \sum_{i=2}^{N} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle.$$

Therefore

$$\begin{aligned} &\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle + \langle Q\mathbf{L}\mathbf{u},\mathbf{u}\rangle = -u_1\left\langle e(-h/2), L_1\mathbf{u}\right\rangle \\ &+ \sum_{i=2}^{N} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle + \sum_{i=1}^{N} 2u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u}\right\rangle + \sum_{i=1}^{N-1} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i+1}\mathbf{u}\right\rangle \\ &+ u_N\left\langle e(h/2), L_N\mathbf{u}\right\rangle. \end{aligned}$$

Observe that swapping the index in the first summation by $i+1$ gives

$$\sum_{i=2}^{N} -u_i\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i-1}\mathbf{u}\right\rangle = \sum_{i=1}^{N-1} -u_{i+1}\left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u}\right\rangle.$$

Further, swapping the index in the third summation by $i-1$ gives

$$\sum_{i=1}^{N-1} -u_i \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_{i+1}\mathbf{u} \right\rangle = \sum_{i=2}^{N} -u_{i-1} \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle .$$

Hence,

$$\begin{aligned}
&\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle + \langle Q\mathbf{L}\mathbf{u}, \mathbf{u}\rangle = -u_1 \langle e(-h/2), L_1\mathbf{u}\rangle \\
&+ \sum_{i=1}^{N-1} -u_{i+1} \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle + \sum_{i=1}^{N} 2u_i \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle + \sum_{i=2}^{N} -u_{i-1} \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle \\
&+ u_N \langle e(h/2), L_N\mathbf{u}\rangle ,
\end{aligned}$$

or equivalently,

$$\begin{aligned}
&\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle + \langle Q\mathbf{L}\mathbf{u}, \mathbf{u}\rangle = -u_1 \langle e(-h/2), L_1\mathbf{u}\rangle + u_N \langle e(h/2), L_N\mathbf{u}\rangle \\
&+ \sum_{i=2}^{N-1} (-u_{i+1} + 2u_i - u_{i-1}) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle \\
&+ (2u_1 - u_2) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_1\mathbf{u} \right\rangle + (2u_N - u_{N-1}) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_N\mathbf{u} \right\rangle .
\end{aligned}$$

Finally, we have shown that

$$\frac{d}{dt}\|\mathbf{u}\|_H^2 = -\langle \mathbf{u}, Q\mathbf{L}\mathbf{u}\rangle - \langle Q\mathbf{L}\mathbf{u}, \mathbf{u}\rangle = u_1 \langle e(-h/2), L_1\mathbf{u}\rangle - u_N \langle e(h/2), L_N\mathbf{u}\rangle + M, \tag{5.13}$$

where $H = \operatorname{diag}(h, h, \ldots, h)$, $\langle \mathbf{u}, H\mathbf{u}\rangle = \|\mathbf{u}\|_H^2$ and

$$\begin{aligned}
M = &\sum_{i=2}^{N-1} (u_{i+1} - 2u_i + u_{i-1}) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_i\mathbf{u} \right\rangle \\
&+ (u_2 - 2u_1) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_1\mathbf{u} \right\rangle + (u_{N-1} - 2u_N) \left\langle \frac{e(h/2)-e(-h/2)}{2}, L_N\mathbf{u} \right\rangle .
\end{aligned}$$

Remark. By (5.12) we know

$$\frac{e(h/2)-e(-h/2)}{2} = \left[0, \frac{h}{2}, 0, \frac{h^3}{8}, 0, \frac{h^5}{2^5}, \ldots, \frac{h^k}{2^k}\right]^T .$$

Suppose that we define matrices D and B by

$$D = \frac{1}{h^2}\begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & \ddots \\ & & & -1 \end{bmatrix}, \qquad B = \operatorname{diag}\left(\frac{e(h/2)-e(-h/2)}{2}, \frac{e(h/2)-e(-h/2)}{2}, \ldots\right),$$

Proposition 5.1. *The term M in (5.13) can satisfies*

$$M = \langle (D + D^T)\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle$$

Proof. Note that

$$\langle (D + D^T)\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle = \langle D\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle + \langle D^T\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle,$$

and

$$\langle D\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle = \sum_{i=1}^{N-1}(-u_i + u_{i+1})\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_i\mathbf{u}\right\rangle - u_N\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_N\mathbf{u}\right\rangle$$
$$\langle D^T\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle = -u_1\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_1\mathbf{u}\right\rangle + \sum_{i=2}^{N}(-u_i + u_{i-1})\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_i\mathbf{u}\right\rangle.$$

Thus,

$$\begin{aligned}\langle (D + D^T)\mathbf{u}, B\mathbf{L}\mathbf{u}\rangle &= (-2u_1 + u_2)\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_1\mathbf{u}\right\rangle \\ &+ \sum_{i=2}^{N-1}(u_{i+1} - 2u_i + u_{i-1})\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_i\mathbf{u}\right\rangle \\ &+ (-2u_N + u_{N-1})\left\langle \frac{e(h/2) - e(-h/2)}{2}, L_N\mathbf{u}\right\rangle \\ &= M\end{aligned}$$

□

Note $D\mathbf{u} \approx u_x$ and $D^T\mathbf{u} \approx -u_x$ ([36]). Thus, we have a rough idea that M is an approximation of αu_x integrated over the domain for some α,

$$M \approx \int_0^1 2u_x\alpha dx.$$

However, it seems we cannot determine the sign of M unless the reconstruction method is specified. We conclude that the k-exact scheme using the central flux is not energy stable in the general case, even on structured regular grids.

5.2 Problem 2: 2D linear advection

We consider problem 2 in section 2.4:

Let $\mathbf{x} = (x, y) \in \Omega = [0, 1]^2$, $0 \leq t \leq T < \infty$ and $u = u(\mathbf{x}, t)$ be a real-valued function. Consider

$$u_t + u_x + u_y = 0, \qquad u(x, y, 0) = \sin\left(2\pi\left(\frac{x}{2} + \frac{y}{2}\right)\right), \qquad u(0, y, t) = g_1(y, t), \qquad u(x, 0, t) = g_2(x, t),$$

where g_1 and g_2 are given by

$$g_1(y, t) = \sin(2\pi(y/2 - t)), \qquad g_2(x, t) = \sin(2\pi(x/2 - t)).$$

In section 2.4 we found the energy rate to be

$$\frac{d}{dt}\|u(\cdot, t)\|_2^2 \leq \int_0^1 g_1^2(y, t)dy + \int_0^1 g_2^2(x, t)dx. \tag{5.14}$$

As for the 1D problem, we attempt to determine if schemes obtained by the k-exact method will mimic this energy rate. Schemes approximating the PDE can be obtained by the k-exact method as follows: We partition Ω into N control volumes $\{C_i\}_{i=1}^N$ satisfying

$$C_i \cap C_j = \emptyset \quad (i \neq j), \qquad \text{and} \qquad \bigcup_{i=1}^N \overline{C_i} = \overline{\Omega}.$$

We let $u_i(t)$ denote an approximation of the volume-averaged value of $u(\mathbf{x}, t)$ over C_i. Formally,

$$u_i(t) \approx \frac{1}{V_i}\int_{C_i} u(\mathbf{x}, t)d\mathbf{x},$$

where V_i is the measure of C_i. Integrating the PDE over C_i we obtain

$$\begin{aligned} \frac{d}{dt}\frac{1}{V_i}\int_{C_i} ud\mathbf{x} &= -\frac{1}{V_i}\int_{C_i}\nabla\cdot[u, u]d\mathbf{x} = -\frac{1}{V_i}\int_{\partial C_i}[u, u]\cdot([dy, -dx]/\sqrt{dx^2 + dy^2})\cdot\sqrt{dx^2 + dy^2} && (5.15)\\ &= -\frac{1}{V_i}\int_{\partial C_i} udy + \frac{1}{V_i}\int_{\partial C_i} udx. && (5.16)\end{aligned}$$

Let p_i denote the center of mass of C_i. Considering a fixed time $t = t^*$, we will write $u(\mathbf{x}, t^*) = u(\mathbf{x})$ and $u_i(t^*) = u_i$. We proceed by finding reconstructions $\{R_i^k\}_{i=1}^N$ of $u(\mathbf{x})$ in $\{C_i\}_{i=1}^N$, satisfying

$$R_i^k(\mathbf{x} - p_i) - u(\mathbf{x})\big|_{C_i} = \mathcal{O}(h^{k+1}),$$

and

$$\frac{1}{V_i}\int_{C_i} R_i^k(\mathbf{x} - p_i)d\mathbf{x} = u_i,$$

where $h = \max_{C_i}\sup_{a,b\in C_i}\|a - b\|_{\mathbb{R}^2}$. Suppose C_i is an interior volume, meaning that $\partial C_i \cap \partial\Omega = \emptyset$. If we let N_i be the set of indices j such that $\partial C_i \cap \partial C_j \neq \emptyset$, then $\partial C_i = \cup_{j\in N_i}\Gamma_{ij}$ where $\Gamma_{ij} = \partial C_i \cap \partial C_j$. Using (5.16) we obtain the finite volume approximation

$$V_i\frac{d}{dt}u_i = -\sum_{j\in N_i}\int_{\Gamma_{ij}} u(\mathbf{x})dy + \sum_{j\in N_i}\int_{\Gamma_{ij}} u(\mathbf{x})dx. \tag{5.17}$$

Following the idea of Godunov's method, we approximate $u(\mathbf{x})$ at each point $\mathbf{x} \in \Gamma_{ij}$ by solving the Riemann problems

$$u_t + u_\xi = 0, \qquad u(\xi, 0) = \begin{cases} R_i^k(\mathbf{x} - p_i), & \xi < 0 \\ R_j^k(\mathbf{x} - p_j), & \xi > 0 \end{cases}, \tag{5.18}$$

either exactly or approximately. Here ξ is parallel to the normal vector on Γ_{ij} and $(\xi = 0) \in \Gamma_{ij}$. Using the notation

$$RP\left(R_i^k(\mathbf{x} - p_i), R_j^k(\mathbf{x} - p_j)\right),$$

to denote the exact or approximate solution of (5.18) at $\xi = 0$ and $t > 0$, the finite volume approximation (5.17) becomes

$$V_i \frac{d}{dt} u_i = -\sum_{j \in N_i} \int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x} - p_i), R_j^k(\mathbf{x} - p_j)\right) dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x} - p_i), R_j^k(\mathbf{x} - p_j)\right) dx.$$

Recall that m-point Gaussian quadrature has an accuracy order of $2m$. To preserve the accuracy order $k+1$ of the reconstructions R_i^k we use $m = \lceil (k+1)/2 \rceil$ quadrature points. Then

$$V_i \frac{d}{dt} u_i = -\sum_{j \in N_i} \sum_{q=1}^{m} w_{ijq}^2 RP\left(R_i^k(\mathbf{x}_{ijq} - p_i), R_j^k(\mathbf{x}_{ijq} - p_j)\right) + \sum_{j \in N_i} \sum_{q=1}^{m} w_{ijq}^1 RP\left(R_i^k(\mathbf{x}_{ijq} - p_i), R_j^k(\mathbf{x}_{ijq} - p_j)\right), \tag{5.19}$$

where the quadrature weights w_{ijq}^d are defined by

$$w_{ijq}^2 = y_{ij}'(\xi_q) \int_{-1}^{1} l_q(\xi) d\xi, \qquad w_{ijq}^1 = x_{ij}'(\xi_q) \int_{-1}^{1} l_q(\xi) d\xi, \tag{5.20}$$

and the quadrature points $\mathbf{x}_{ijq}$ are given by

$$\mathbf{x}_{ijq} = \mathbf{r}_{ij}(\xi_q) = \left(x_{ij}(\xi_q), y_{ij}(\xi_q)\right),$$

where ξ_q is the q-th root of the degree m Legendre polynomial and $\Gamma_{ij} = \{\mathbf{r}_{ij}(\xi) : -1 \le \xi \le 1\}$. See Appendix A for more details.

Now suppose C_i is a boundary volume, meaning that $\partial C_i \cap \partial\Omega \neq \emptyset$. In particular, $\partial C_i = \cup_{j \in N_i} \Gamma_{ij} \cup \Gamma_{i\partial\Omega}$ where $\Gamma_{i\partial\Omega} = \partial C_i \cap \partial\Omega$. Using (5.16) we obtain the finite volume approximation

$$V_i \frac{d}{dt} u_i = -\sum_{j \in N_i} \int_{\Gamma_{ij}} u(\mathbf{x}) dy - \int_{\Gamma_{i\partial\Omega}} u(\mathbf{x}) dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} u(\mathbf{x}) dx + \int_{\Gamma_{i\partial\Omega}} u(\mathbf{x}) dx.$$

As in the 1D case, we will approximate $u(\mathbf{x})$ at $\partial\Omega$ by the reconstruction corresponding to the boundary volume. Following the same approach as for the interior volume scheme, the boundary volume scheme becomes

$$\begin{aligned} V_i \frac{d}{dt} u_i = &-\sum_{j \in N_i} \int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x} - p_i), R_j^k(\mathbf{x} - p_j)\right) dy + \sum_{j \in N_i} \int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x} - p_i), R_j^k(\mathbf{x} - p_j)\right) dx \\ &- \int_{\Gamma_{i\partial\Omega}} R_i^k(\mathbf{x} - p_i) dy + \int_{\Gamma_{i\partial\Omega}} R_i^k(\mathbf{x} - p_i) dx. \end{aligned}$$

Applying the quadrature rule we obtain

$$V_i\frac{d}{dt}u_i = -\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^2 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) + \sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^1 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) \tag{5.21}$$

$$-\sum_{q=1}^{m} w_{i\partial\Omega q}^2 R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dy + \sum_{q=1}^{m} w_{i\partial\Omega q}^1 R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dx. \tag{5.22}$$

Thus, we have found that the general k-exact scheme approximating $u_t + u_x + u_y = 0$ is given by

$$V_i\frac{d}{dt}u_i = \begin{cases} \begin{cases} -\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^2 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) \\ \quad +\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^1 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) \end{cases}, & C_i \text{ is an interior volume} \\ \begin{cases} -\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^2 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) \\ \quad +\sum_{j\in N_i}\sum_{q=1}^{m} w_{ijq}^1 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) \\ \quad -\sum_{q=1}^{m} w_{i\partial\Omega q}^2 R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dy + \sum_{q=1}^{m} w_{i\partial\Omega q}^1 R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dx \end{cases}, & C_i \text{ is a boundary volume.} \end{cases} \tag{5.23}$$

To determine the energy stability of the scheme, we want to write it in the spatially global form

$$V\frac{d}{dt}\mathbf{u} = -K, \tag{5.24}$$

where $V = \text{diag}(V_1, V_2, \dots, V_N)$, $\mathbf{u} = [u_1, u_2, \dots, u_N]^T$ and K is the vector such that $-K_i =$ the right hand side of (5.23). Recall from section 3.3 that V induces a discrete L^2 norm: $\langle \mathbf{u}, V\mathbf{u}\rangle = \|\mathbf{u}\|_V^2$. Taking the inner-product of (5.24) with $\mathbf{u}$ and adding the transposed equation we obtain

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle. \tag{5.25}$$

Our goal is to determine if, or when, (5.25) mimics (5.14). That is, if or when

$$-\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle \leq \int_0^1 g_1^2(y,t)dy + \int_0^1 g_2^2(x,t)dx.$$

In order to do so we must specify the function RP residing in K. Suppose that we use the central flux,

$$RP\left(R_i^k(\mathbf{x}-p_i), R_j^k(\mathbf{x}-p_j)\right) = \frac{R_i^k(\mathbf{x}-p_i) + R_j^k(\mathbf{x}-p_j)}{2}. \tag{5.26}$$

Applying (5.26) to (5.19) yields

$$V_i\frac{d}{dt}u_i = -\sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}R_i^k(\mathbf{x}_{ijq}-p_i) - \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}R_j^k(\mathbf{x}_{ijq}-p_j) \tag{5.27}$$

$$+\sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}R_i^k(\mathbf{x}_{ijq}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}R_j^k(\mathbf{x}_{ijq}-p_j). \tag{5.28}$$

Recall from section 4 that the reconstructions can be factorized as

$$R_i^k(\mathbf{x}-p_i) = \langle e(\mathbf{x}-p_i), L_i\mathbf{u}\rangle\,,$$

and in the 2D case, $e(\mathbf{x}-p_i)$, $L_i\mathbf{u} \in \mathbb{R}^{(k+2)(k+1)/2}$ are given by

$$(e(\mathbf{x}-p_i))_l = (x-x_i)^{m_l}(y-y_i)^{n_l}, \qquad (L_i\mathbf{u})_l = \frac{1}{m_l!n_l!}\frac{\partial^{m_l+n_l}}{\partial x^{m_l}\partial y^{n_l}}R_i^k(p_i).$$

Further, observe that

$$\frac{w_{ijq}^2}{2}R_i^k(\mathbf{x}_{ijq}-p_i) = \frac{w_{ijq}^2}{2}\left\langle e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle = \left\langle \frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle,$$

and

$$\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}R_i^k(\mathbf{x}_{ijq}-p_i) = \sum_{q=1}^{m}\left\langle \frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle = \left\langle \sum_{q=1}^{m}\frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle.$$

Therefore, the scheme for interior volumes (5.27-5.28) can be written as

$$\begin{aligned} V_i\frac{d}{dt}u_i = &-\left\langle \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle - \sum_{j\in N_i}\left\langle \sum_{q=1}^{m}\frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u}\right\rangle \\ &+\left\langle \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle + \sum_{j\in N_i}\left\langle \sum_{q=1}^{m}\frac{w_{ijq}^1}{2}e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u}\right\rangle. \end{aligned}$$

Similarly, applying (5.26) to the boundary volume scheme (5.21-5.22) we obtain

$$\begin{aligned} V_i\frac{d}{dt}u_i = &-\sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}R_i^k(\mathbf{x}_{ijq}-p_i) - \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}R_j^k(\mathbf{x}_{ijq}-p_j) \\ &+\sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}R_i^k(\mathbf{x}_{ijq}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}R_j^k(\mathbf{x}_{ijq}-p_j) \\ &-\sum_{q=1}^{m}w_{i\partial\Omega q}^2R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dy + \sum_{q=1}^{m}w_{i\partial\Omega q}^1R_i^k(\mathbf{x}_{i\partial\Omega q}-p_i)dx. \end{aligned}$$

Once again using the inner product factorization of R_i^k, we obtain the boundary volume scheme

$$\begin{aligned} V_i\frac{d}{dt}u_i = &-\left\langle \sum_{q=1}^{m}w_{i\partial\Omega q}^2e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle - \sum_{j\in N_i}\left\langle \sum_{q=1}^{m}\frac{w_{ijq}^2}{2}e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u}\right\rangle \\ &+\left\langle \sum_{q=1}^{m}w_{i\partial\Omega q}^1e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m}\frac{w_{ijq}^1}{2}e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u}\right\rangle + \sum_{j\in N_i}\left\langle \sum_{q=1}^{m}\frac{w_{ijq}^1}{2}e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u}\right\rangle. \end{aligned}$$

If we define matrices Q_x, Q_y by

$$(Q_x)_{ii} = \begin{cases} \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is an interior volume} \\ \sum_{q=1}^{m} w^2_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is a boundary volume,} \end{cases}$$

$$(Q_x)_{ij} = \begin{cases} \sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), & j\in N_i \\ 0, & \text{otherwise,} \end{cases}$$

$$(Q_y)_{ii} = \begin{cases} -\sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is an interior volume} \\ -\sum_{q=1}^{m} w^1_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) - \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is a boundary volume,} \end{cases}$$

$$(Q_y)_{ij} = \begin{cases} -\sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), & j\in N_i \\ 0, & \text{otherwise,} \end{cases}$$

then the k-exact scheme approximating $u_t + u_x + u_y = 0$ with the central numerical flux (5.26) is given by

$$V\frac{d}{dt}\mathbf{u} = -Q_x\mathbf{L}\mathbf{u} - Q_y\mathbf{L}\mathbf{u} \tag{5.29}$$

where $\mathbf{Lu} = [L_1\mathbf{u}, L_2\mathbf{u}, \dots, L_N\mathbf{u}]^T$. The discrete energy rate of the scheme (5.29) is given by

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -\langle \mathbf{u}, Q_x\mathbf{Lu}\rangle - \langle \mathbf{u}, Q_y\mathbf{Lu}\rangle - \langle Q_x\mathbf{Lu}, \mathbf{u}\rangle - \langle Q_y\mathbf{Lu}, \mathbf{u}\rangle$$

Remark. If we use the 0-exact reconstructions $R_i^0 = u_i$, then we recover the SBP scheme examined in section 3.5. To see this, note that w^d_{ijq} would reduce to Δy_{ij} or Δx_{ij}. Further, we would have that $e(\mathbf{x}) \equiv 1$.

If we let $N_{\partial\Omega}$ denote the set of indices i satisfying $i \in N_{\partial\Omega} \implies \{C_i \text{ is a boundary volume}\}$, then

$$\begin{aligned} -\langle \mathbf{u}, Q_x\mathbf{Lu}\rangle = &-\sum_{i\notin N_{\partial\Omega}} u_i \left(\left\langle \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u} \right\rangle \right) \\ &- \sum_{i\in N_{\partial\Omega}} u_i \left(\left\langle \sum_{q=1}^{m} w^2_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^{m} \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u} \right\rangle \right) \\ -\langle \mathbf{u}, Q_y\mathbf{Lu}\rangle = &\sum_{i\notin N_{\partial\Omega}} u_i \left(\left\langle \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u} \right\rangle \right) \\ &+ \sum_{i\in N_{\partial\Omega}} u_i \left(\left\langle \sum_{q=1}^{m} w^1_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^{m} \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_j), L_j\mathbf{u} \right\rangle \right). \end{aligned}$$

Using the defniitions of Q_x and Q_y we determine the transposed matrices;

$$(Q_x^T)_{ii} = \begin{cases} \sum_{j\in N_i}\sum_{q=1}^m \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is an interior volume} \\ \sum_{q=1}^m w^2_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^m \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is a boundary volume,} \end{cases}$$

$$(Q_x^T)_{ij} = \begin{cases} \sum_{q=1}^m \frac{w^2_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), & j\in N_i \\ 0, & \text{otherwise,} \end{cases}$$

$$(Q_y^T)_{ii} = \begin{cases} -\sum_{j\in N_i}\sum_{q=1}^m \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is an interior volume} \\ -\sum_{q=1}^m w^1_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) - \sum_{j\in N_i}\sum_{q=1}^m \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), & C_i \text{ is a boundary volume,} \end{cases}$$

$$(Q_y^T)_{ij} = \begin{cases} -\sum_{q=1}^m \frac{w^1_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), & j\in N_i \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} -\langle Q_x\mathbf{Lu},\mathbf{u}\rangle = &-\sum_{i\notin N_{\partial\Omega}} u_i \left(\left\langle \sum_{j\in N_i}\sum_{q=1}^m \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^m \frac{w^2_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), L_j\mathbf{u} \right\rangle \right) \\ &- \sum_{i\in N_{\partial\Omega}} u_i \left(\left\langle \sum_{q=1}^m w^2_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^m \frac{w^2_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^m \frac{w^2_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), L_j\mathbf{u} \right\rangle \right) \\ -\langle Q_y\mathbf{Lu},\mathbf{u}\rangle = &\sum_{i\notin N_{\partial\Omega}} u_i \left(\left\langle \sum_{j\in N_i}\sum_{q=1}^m \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^m \frac{w^1_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), L_j\mathbf{u} \right\rangle \right) \\ &+ \sum_{i\in N_{\partial\Omega}} u_i \left(\left\langle \sum_{q=1}^m w^1_{i\partial\Omega q} e(\mathbf{x}_{i\partial\Omega q}-p_i) + \sum_{j\in N_i}\sum_{q=1}^m \frac{w^1_{ijq}}{2} e(\mathbf{x}_{ijq}-p_i), L_i\mathbf{u} \right\rangle + \sum_{j\in N_i} \left\langle \sum_{q=1}^m \frac{w^1_{jiq}}{2} e(\mathbf{x}_{jiq}-p_i), L_j\mathbf{u} \right\rangle \right). \end{aligned}$$

We see that for $k > 0$ the operators $D_x = V^{-1}Q_x$, $D_y = V^{-1}Q_y$ will not satisfy a generalized SBP property. Thus, we conclude that the scheme is not stable in the general case.

6 Stability analysis of spectral volume schemes

In this section we examine the energy stability of schemes obtained via the spectral volume method.

6.1 Problem 1: 1D linear advection

Consider problem 1 from section 2.4.

Let $x \in \Omega = [0,1] \subset \mathbb{R}$ and $t \in [0,T) \subset \mathbb{R}^+$. Find the function $u : \Omega \times [0,T) \to \mathbb{R}$ satisfying

$$u_t + u_x = 0, \qquad u(x,0) = \sin(2\pi(x)), \qquad u(0,t) = \sin(-2\pi t).$$

Recall that the energy rate was found to be

$$\frac{d}{dt}\left\|u(\cdot,t)\right\|_2^2 = u^2(0,t) - u^2(1,t) \leq g^2(t) \tag{6.1}$$

and we wish to determine if spectral volume schemes approximating the PDE will satisfy a discrete equivalent energy rate. Recall that spectral volume schemes approximating the PDE can be found as follows: We begin by partitioning the spatial domain Ω into N spectral volumes $\{SV_i\}_{i=1}^N$ satisfying

$$SV_i \cap SV_j = \emptyset, \quad (i \neq j), \qquad \text{and} \qquad \bigcup_{i=1}^{N} \overline{SV_i} = \overline{\Omega}.$$

Next, each spectral volume SV_i is partitioned into $(k+1)$ control volumes $\{C_{i,j}\}_{j=1}^{k+1}$, and we let $u_{i,j}(t)$ denote an approximation of the volume-averaged value of $u(x,t)$ over $C_{i,j}$. That is,

$$u_{i,j}(t) \approx \frac{1}{V_{i,j}} \int_{C_{i,j}} u(x,t)dx,$$

where $V_{i,j}$ is the measure of $C_{i,j}$. Let t^* denote a fixed time, and write $u(x,t^*) = u(x)$, $u_{i,j}(t^*) = u_{i,j}$. We proceed by finding reconstructions $\{P_i\}_{i=1}^N$ satisfying

$$P_i(x) - u(x)\big|_{SV_i} = \mathcal{O}(h^{k+1}), \qquad \text{and} \qquad \frac{1}{V_{i,j}} \int_{C_{i,j}} P_i(x)dx = u_{i,j},$$

where $h = \max_{C_{i,j} \in SV_i} \sup_{x,y \in C_{i,j}} |x-y|$. Recall from section 4 that if we define $u_i = [u_{i,1}, u_{i,2}, \ldots, u_{i,k+1}]^T$ then P_i can factorized as

$$P_i(x) = \langle L_i(x), u_i \rangle, \tag{6.2}$$

where $L_i(x)$ is a linear transformation of the polynomial basis function, dependent on the transformation between SV_i and the reference spectral volume SV_r. Integrating the PDE over $C_{i,j}$ we obtain

$$\frac{d}{dt}\int_{C_{i,j}} u(x,t^*)dx = -u(C_{i,j}^+, t^*) + u(C_{i,j}^-, t^*),$$

and the finite volume approximation

$$V_{i,j}\frac{d}{dt}u_{i,j} = -u(C_{i,j}^+) + u(C_{i,j}^-). \tag{6.3}$$

Without loss of generality, we may assume that the control volumes inside each spectral volume are structured.

In other words, we will assume that $C^+_{i,j} = C^-_{i,j+1}$. It follows that the lower bound SV^-_i and upper bound SV^+_i of SV_i satisfy $SV^-_i = C^-_{i,1}$ and $SV^+_i = C^+_{i,k+1}$. Since the reconstruction P_i is continuous inside SV_i, we use the approximations

$$\begin{aligned} u(C^+_{i,j}) &= P_i(C^+_{i,j}) = \langle L_i(C^+_{i,j}), u_i\rangle, \qquad (j = 1, \ldots, k),\\ u(C^-_{i,j}) &= P_i(C^-_{i,j}) = \langle L_i(C^-_{i,j}), u_i\rangle, \qquad (j = 2, \ldots, k+1), \end{aligned}$$

where we applied (6.2). At the spectral volume boundaries we will follow the approach of Godunov's method and consider the Riemann problems

$$u_t + u_x = 0, \qquad u(x,0) = \begin{cases} P_i(SV^+_i), & x < SV^+_i \\ P_{j_1}(SV^-_{j_1}), & x > SV^+_i \end{cases}, \tag{6.4}$$

where SV_{j_1} is the spectral volume adjacent to SV_i satisfying $SV^-_{j_1} = SV^+_i$. If we use the notation

$$RP\left(P_i(SV^+_i), P_{j_1}(SV^-_{j_1})\right)$$

to denote an exact or approximate solution of (6.4) at $x = SV^+_i$ for $t > 0$, then our approximation of $u(SV^+_i)$ is given by

$$u(SV^+_i) = RP\left(P_i(SV^+_i), P_{j_1}(SV^-_{j_1})\right).$$

Likewise, if SV_{j_2} is the spectral volume adjacent to SV_i satisfying $SV^+_{j_2} = SV^-_i$, then we will approximate $u(SV^-_i)$ by

$$u(SV^-_i) = RP\left(P_{j_2}(SV^+_{j_2}), P_i(SV^-_i)\right).$$

Note that by (6.2), we may write

$$u(SV^+_i) = RP\left(\langle L_i(SV^+_i), u_i\rangle, \langle L_{j_1}(SV^-_{j_1}), u_{j_1}\rangle\right), \qquad u(SV^-_i) = RP\left(\langle L_{j_2}(SV^+_{j_2}), u_{j_2}\rangle, \langle L_i(SV^-_i), u_i\rangle\right).$$

If SV_i is the left boundary spectral volume, meaning that some $SV_j \in \{SV_i\}_{i=1}^N$ satisfying $SV^+_j = SV^-_i$ does not exist, then we will use the approximation

$$u(SV^-_i) = P_i(SV^-_i) = \langle L_i(SV^-_i), u_i\rangle.$$

Similarly, if SV_i is the right boundary spectral volume, meaning that some $SV_j \in \{SV_i\}_{i=1}^N$ satisfying $SV^-_j = SV^+_i$ does not exist, then we will use the approximation

$$u(SV^+_i) = P_i(SV^+_i) = \langle L_i(SV^+_i), u_i\rangle.$$

If we let SV_L and SV_R denote the left and right boundary spectral volumes respectively, we obtain the general spectral volume scheme approximating $u_t + u_x = 0$:

$$V_{i,j}\frac{d}{dt}u_{i,j} = \begin{cases} -\langle L_i(C^+_{i,j}), u_i\rangle + \langle L_i(C^-_{i,j}), u_i\rangle, & j = 2, \ldots k \\ \begin{cases} -\langle L_i(C^+_{i,j}), u_i\rangle + \langle L_i(C^-_{i,j}), u_i\rangle, & i = R \\ -RP\left(\langle L_i(SV^+_i), u_i\rangle, \langle L_{j_1}(SV^-_{j_1}), u_{j_1}\rangle\right) + \langle L_i(C^-_{i,j}), u_i\rangle, & \text{otherwise} \end{cases}, & j = k+1 \\ \begin{cases} -\langle L_i(C^+_{i,j}), u_i\rangle + \langle L_i(C^-_{i,j}), u_i\rangle, & i = L \\ -\langle L_i(C^+_{i,j}), u_i\rangle + RP\left(\langle L_{j_2}(SV^+_{j_2}), u_{j_2}\rangle, \langle L_i(SV^-_i), u_i\rangle\right), & \text{otherwise} \end{cases}, & j = 1. \end{cases} \tag{6.5}$$

Remark. In the above it is understood that the spectral volumes SV_{j_1}, SV_{j_2} are adjacent to SV_i.

By the linearity of the inner product, we may write (6.5) as

$$V_{i,j}\frac{d}{dt}u_{i,j} = \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & j = 2,\dots k \\ \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & i = R \\ -RP\left(\langle L_i(SV_i^+), u_i\rangle, \langle L_{j_1}(SV^-_{j_1}), u_{j_1}\rangle\right) + \langle L_i(C^-_{i,j}), u_i\rangle, & \text{otherwise} \end{cases}, & j = k+1 \\ \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & i = L \\ -\langle L_i(C^+_{i,j}), u_i\rangle + RP\left(\langle L_{j_2}(SV^+_{j_2}), u_{j_2}\rangle, \langle L_i(SV_i^-), u_i\rangle\right), & \text{otherwise} \end{cases}, & j = 1. \end{cases} \tag{6.6}$$

To analyze the discrete energy stability of the scheme we will write it in the spatially global form

$$V\frac{d}{dt}\mathbf{u} = -K. \tag{6.7}$$

Here K is the vector such that $-K_i =$ the right hand side of (6.6), and V, $\mathbf{u}$ are defined as

$$V = \operatorname{diag}(V_{1,1},\dots,V_{1,k+1},\dots,V_{N,1},\dots,V_{N,k+1}), \qquad \text{and} \qquad \mathbf{u} = [u_{1,1},\dots,u_{1,k+1},\dots,u_{N,1},\dots,u_{N,k+1}]^T.$$

As for the k-exact method, the matrix V induces a discrete L^2 norm. Taking the inner-product of (6.7) with $\mathbf{u}$ we obtain

$$\frac{d}{dt}\|\mathbf{u}\|_V^2 = -\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle. \tag{6.8}$$

We are interested in determining if, or when, (6.8) mimics (6.1). That is, we want to determine if or when

$$-\langle \mathbf{u}, K\rangle - \langle K, \mathbf{u}\rangle \le u^2(SV_L^-) - u^2(SV_R^+) \approx P_L^2(SV_L^-) - P_R^2(SV_R^+) = \langle L_L(SV_L^-), u_L\rangle^2 - \langle L_R(SV_R^+), u_R\rangle^2,$$

where SV_L denotes the left boundary volume and SV_R denotes the right boundary volume. To proceed we specify the function RP which resides in K.

$$RP(a,b) = \frac{a+b}{2}.$$

Now (6.6) becomes

$$V_{i,j}\frac{d}{dt}u_{i,j} = \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & j = 2,\dots k \\ \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & i = R \\ -\langle \frac{L_i(SV_i^+)}{2}, u_i\rangle - \langle \frac{L_{j_1}(SV^-_{j_1})}{2}, u_{j_1}\rangle + \langle L_i(C^-_{i,j}), u_i\rangle, & \text{otherwise} \end{cases}, & j = k+1 \\ \begin{cases} -\langle L_i(C^+_{i,j}) - L_i(C^-_{i,j}), u_i\rangle, & i = L \\ -\langle L_i(C^+_{i,j}), u_i\rangle + \langle \frac{L_{j_2}(SV^+_{j_2})}{2}, u_{j_2}\rangle + \langle \frac{L_i(SV_i^-)}{2}, u_i\rangle, & \text{otherwise} \end{cases}, & j = 1. \end{cases} \tag{6.9}$$

Due to time constraints we did not have the opportunity to continue the analysis any further.

7 Numerical results

In this section we present numerical results for k-exact schemes approximating problem 1 in section 2.4. The results are obtained by implementing the method in Fortran. The code is made by the author, and it is freely available to download at [58].

Consider problem 1 in section 2.4

Let $x \in \Omega = [0,1] \subset \mathbb{R}$ and $t \in [0,T] \subset \mathbb{R}^+$. Find the function $u : \Omega \times [0,T] \to \mathbb{R}$ satisfying

$$u_t + u_x = 0, \qquad u(x,0) = \sin(2\pi(x)), \qquad u(0,t) = \sin(-2\pi t).$$

We consider k-exact schemes with polynomial reconstructions of degree $k = 0,1,2,3$ obtained by the least-squares reconstruction method. For the temporal discretization we use the standard 4th-order Runge-Kutta method and the time step Δt satisfying $CFL = \Delta t/h = 1/2$ where $h = \max_{C_i} V_i$. We obtain results using the central flux. The spatial domain is discretized by control volumes $C_i = (x_i, x_{i+1})$ where $\{x_i\}_{i=1}^{N+1}$ is a structured grid such that $C_1^- = 0$. The boundary condition is implemented by substituting $R_1^k(0 - x_1) = \sin(-2\pi t)$ in the flux evaluation. The initial data $\mathbf{u}(t=0)$ is obtained by 5th-order Gaussian quadrature of the initial function $u(x,0)$ over the control volumes. The L^2 error is measured at $t = 1$ using

$$E = L^2 \text{ error } = (\langle \mathbf{u}(0) - \mathbf{u}(1), V\mathbf{u}(0) - \mathbf{u}(1)\rangle)^{1/2} = \|\mathbf{u}(0) - \mathbf{u}(1)\|_V ,$$

(recall that the analytical solution is periodic in time with a period of 1). The convergence rate μ is calculated using

$$\mu_j = \frac{\ln(E_j/E_{j-1})}{\ln(h_j/h_{j-1})},$$

where the index j corresponds to the partition P_j obtained using N_j control volumes. Consider first the case where the grid is regular, giving regular control volumes. For the regular volume partitions we have that $\ln(h_j/h_{j-1}) = \ln(1/2)$. Results for the structured and regular volumes are shown in table 1.

N	L^2 error	μ
50	4.45E-02	-
100	2.22E-02	1.01
200	1.11E-02	0.99
400	5.55E-03	1.00

(a) $k = 0$.

N	L^2 error	μ
50	3.50E-03	-
100	1.05E-03	1.78
200	2.72E-04	1.90
400	6.94E-05	2.00

(b) $k = 1$.

N	L^2 error	μ
50	1.12E-03	-
100	1.36E-04	3.04
200	1.68E-05	3.02
400	2.08E-06	3.01

(c) $k = 2$.

N	L^2 error	μ
50	1.49E-04	-
100	6.32E-06	4.56
200	2.93E-07	4.70
400	2.34E-08	4.35

(d) $k = 3$.

Table 1: L^2 error at $t = 1$ using N structured regular volumes and reconstruction polynomials of degree k.

Next, we define a structured irregular grid $\{x_i\}_{i=1}^{N+1}$ by

$$x_i = \frac{\exp(\xi_i) - 1}{\exp(1) - 1}, \qquad \xi_i = (i-1)\frac{1}{N},$$

and the control volumes $C_i = (x_i, x_{i+1})$ for $i = 1, \ldots, N$. We perform the same convergence analysis as for the structured regular volumes. Results for these structured irregular volumes are shown in table 2.

We see that the k-exact method with the central flux produces numerical results with a high convergence order. However, we stress that the stability of the schemes cannot be determined by numerical results, as they correspond to specific problem data, specific reconstruction methods, and specific control volumes.

N	h	L^2 error	μ
50	3.13E-02	6.98E-02	-
100	1.57E-02	3.49E-02	1.00
200	7.89E-03	1.75E-02	1.00
400	3.95E-03	8.77E-03	1.00

(a) $k = 0$.

N	h	L^2 error	μ
50	3.13E-02	4.29E-03	-
100	1.57E-02	8.77E-04	2.30
200	7.89E-03	2.09E-04	2.08
400	3.95E-03	5.24E-05	2.00

(b) $k = 1$.

N	h	L^2 error	μ
50	3.13E-02	2.87E-03	-
100	1.57E-02	3.70E-04	2.97
200	7.89E-03	4.67E-05	3.01
400	3.95E-03	5.86E-06	3.00

(c) $k = 2$.

N	h	L^2 error	μ
50	3.13E-02	1.23E-04	-
100	1.57E-02	4.46E-06	4.81
200	7.89E-03	2.01E-07	4.50
400	3.95E-03	1.17E-08	4.11

(d) $k = 3$.

Table 2: L^2 error at $t = 1$ using N structured irregular volumes and reconstruction polynomials of degree k.

8 Conclusion

We have studied the energy stability of high-order finite volume schemes obtained by the k-exact method and the spectral volume method. In particular, we have looked at schemes using the central numerical flux. We found that schemes are not stable in the general case, and that they do not satisfy the summation-by-parts property.

8.1 Suggestions for future work

As we have only considered schemes employing the central numerical flux, we believe a study considering a more general numerical flux would be interesting. Further, we believe the approach to stability analysis done in [6, 59] could produce interesting the results for the schemes we have discussed here as well.

A Gauss-Legendre quadrature over Γ_{ij}

Let Γ_{ij} be a smooth curve in $\mathbb{R}^2$ parametrized by the variable $\xi \in [-1,1]$. Let $f = [f_1, f_2]$ be our flux function and consider

$$\int_{\Gamma_{ij}} f(u(x,y)) \cdot \hat{n} dS$$

Suppose

$$\Gamma_{ij} = \{\mathbf{r}_{ij}(\xi) : -1 \leq \xi \leq 1\}, \qquad \mathbf{r}_{ij}(\xi) = x_{ij}(\xi)e_1 + y_{ij}(\xi)e_2$$

Then

$$\begin{aligned}\int_{\Gamma_{ij}} f(u(x,y)) \cdot \hat{n} dS &= \int_{-1}^{1} f(u(\mathbf{r}_{ij}(\xi))) \cdot \hat{n}_{ij}(\xi) d\xi = \int_{-1}^{1} f(u(\mathbf{r}_{ij}(\xi))) \cdot \begin{bmatrix} dy_{ij}(\xi)/d\xi \\ -dx_{ij}(\xi)/d\xi \end{bmatrix} d\xi \\ &= \int_{-1}^{1} f_1(u(\mathbf{r}_{ij}(\xi)))y'_{ij}(\xi)d\xi - \int_{-1}^{1} f_2(u(\mathbf{r}_{ij}(\xi)))x'_{ij}(\xi)d\xi \\ &= \int_{\Gamma_{ij}} f_1(u(x,y))dy - \int_{\Gamma_{ij}} f_2(u(x,y))dx\end{aligned}$$

Now we are ready to apply the quadrature rule. Let ξ_q denote the q-th root of the degree m Legendre polynomial defined by

$$p_m(x) = \frac{1}{2^m m!}\frac{d^m}{dx^m}[(x^2-1)^m]$$

Then

$$\begin{aligned}\int_{-1}^{1} f_1(u(\mathbf{r}_{ij}(\xi)))y'_{ij}(\xi)d\xi &\approx \int_{-1}^{1} f_1(u(\mathbf{r}_{ij}(\xi_q)))y'_{ij}(\xi_q)l_q(\xi)d\xi \\ &= \sum_{q=1}^{m} w^2_{ijq} f_1(u(\mathbf{r}_{ij}(\xi_q)))\end{aligned}$$

where the weights w^2_{ijq} are given by

$$w^2_{ijq} = y'_{ij}(\xi_q)\int_{-1}^{1} l_q(\xi)d\xi, \qquad l_q(\xi) = \prod_{j=1, j\neq q}^{m} \frac{\xi - \xi_j}{\xi_q - \xi_j}$$

Note that

$$c_q = \int_{-1}^{1} l_q(\xi)d\xi$$

corresponding to the q-th root of the degree m Legendre polynomial is a very common parameter in numerical analysis, and the values are usually stored preemptively. We give the values for $m = 2, 3, 4$ in table 3.

Remark. If the roots ξ_q are ordered monotonically, then $\xi_q = -\xi_{q^*}$ for $q^* = m - q + 1$. Moreover, the coefficients satisfy $c_q = c_{q^*}$.

Similarly,

$$\begin{aligned}\int_{-1}^{1} f_2(u(\mathbf{r}_{ij}(\xi)))x'_{ij}(\xi)d\xi &\approx \int_{-1}^{1} f_2(u(\mathbf{r}_{ij}(\xi_q)))x'_{ij}(\xi_q)l_q(\xi)d\xi \\ &= \sum_{q=1}^{m} w^1_{ijq} f_2(u(\mathbf{r}_{ij}(\xi_q)))\end{aligned}$$

m	roots ξ_q	coefficients c_q
2	$-\sqrt{1/3}$	1
	$\sqrt{1/3}$	1
3	$-\sqrt{3/5}$	$5/9$
	0	$8/9$
	$\sqrt{3/5}$	$5/9$
4	$-\sqrt{\frac{15+2\sqrt{30}}{35}}$	$\frac{90-5\sqrt{30}}{180}$
	$-\sqrt{\frac{15-2\sqrt{30}}{35}}$	$\frac{90+5\sqrt{30}}{180}$
	$\sqrt{\frac{15-2\sqrt{30}}{35}}$	$\frac{90+5\sqrt{30}}{180}$
	$\sqrt{\frac{15+2\sqrt{30}}{35}}$	$\frac{90-5\sqrt{30}}{180}$

Table 3: Legendre roots and integral of the basis functions.

where

$$w_{ijq}^1 = x_{ij}'(\xi_q)\int_{-1}^{1} l_q(\xi)d\xi$$

If we use the shorthand $\mathbf{r}_{ij}(\xi_q) = \mathbf{x}_{ijq}$ then we have obtained

$$\int_{\Gamma_{ij}} f(u(x,y))\cdot\hat{n}dS \approx \sum_{q=1}^{m} w_{ijq}^2 f_1(u(\mathbf{x}_{ijq})) - \sum_{q=1}^{m} w_{ijq}^1 f_2(u(\mathbf{x}_{ijq}))$$

Remark. Often in our topic of interest, the curves Γ_{ij} will be piecewise line segments. Then in each segment, the coordinate function derivatives will be constant, which simplifies the calculations. In particular, if Γ_{ij} is a straight line in $\mathbb{R}^2$, then

$$\begin{aligned}\Gamma_{ij} &= \{\mathbf{r}_{ij}(\xi) : -1 \le \xi \le 1\} = \{x_{ij}(\xi)e_1 + y_{ij}(\xi)e_2 : -1 \le \xi \le 1\} &\text{(A.1)}\\ &= \left\{\frac{(x_1,y_1)-(x_0,y_0)}{2}\xi + \frac{(x_1,y_1)+(x_0,y_0)}{2} : -1 \le \xi \le 1\right\} &\text{(A.2)}\\ &= \left\{\left(\frac{x_1-x_0}{2}\xi + \frac{x_1+x_0}{2}\right)e_1 + \left(\frac{y_1-y_0}{2}\xi + \frac{y_1+y_0}{2}\right)e_2 : -1 \le \xi \le 1\right\} &\text{(A.3)}\end{aligned}$$

where x_0, y_0 are the lower coordinates and x_1, y_1 are the upper coordinates w.r.t. to some direction choice. Clearly in this case, $x_{ij}' = (x_1 - x_0)/2 = \Delta x_{ij}/2$ and $y_{ij}' = (y_1 - y_0)/2 = \Delta y_{ij}/2$.

Now consider

$$\int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x}-p_i), R_j^k(\mathbf{x}-p_j)\right)dy - \int_{\Gamma_{ij}} RP\left(R_i^k(\mathbf{x}-p_i), R_j^k(\mathbf{x}-p_j)\right)dx \tag{A.4}$$

By our previous discussion,

$$\sum_{q=1}^{m} w_{ijq}^2 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right) - \sum_{q=1}^{m} w_{ijq}^1 RP\left(R_i^k(\mathbf{x}_{ijq}-p_i), R_j^k(\mathbf{x}_{ijq}-p_j)\right)$$

is the m-point Gauss-Legendre quadrature approximation of (A.4).

References

[1] N. K. Yamaleev and M. H. Carpenter, “Third-order energy stable weno scheme,” *Journal of Computational Physics*, vol. 228, no. 8, pp. 3025–3047, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002199910900014X

[2] ——, “A systematic methodology for constructing high-order energy stable weno schemes,” *Journal of Computational Physics*, vol. 228, no. 11, pp. 4248–4272, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999109001132

[3] T. C. Fisher, M. H. Carpenter, N. K. Yamaleev, and S. H. Frankel, “Boundary closures for fourth-order energy stable weighted essentially non-oscillatory finite-difference schemes,” *Journal of Computational Physics*, vol. 230, no. 10, pp. 3727–3752, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999111000684

[4] M. H. Carpenter, T. C. Fisher, and N. K. Yamaleev, *Boundary Closures for Sixth-Order Energy-Stable Weighted Essentially Non-Oscillatory Finite-Difference Schemes.* Boston, MA: Springer US, 2013, pp. 117–160. [Online]. Available: https://doi.org/10.1007/978-1-4614-5389-5_6

[5] G. J. Gassner, “A skew-symmetric discontinuous galerkin spectral element discretization and its relation to sbp-sat finite difference methods,” *SIAM Journal on Scientific Computing*, vol. 35, no. 3, pp. A1233–A1253, 2013. [Online]. Available: https://doi.org/10.1137/120890144

[6] P. Vincent, P. Castonguay, and A. Jameson, “A new class of high-order energy stable flux reconstruction schemes,” *Journal of Scientific Computing*, 2011.

[7] ——, “Insights from von neumann analysis of high-order flux reconstruction schemes,” *Journal of Computational Physics*, vol. 230, no. 22, pp. 8134–8154, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999111004323

[8] ——, “A new class of high-order energy stable flux reconstruction schemes for triangular elements,” *Journal of Scientific Computing*, 2012.

[9] H. Ranocha, P. Öffner, and T. Sonar, “Summation-by-parts operators for correction procedure via reconstruction,” *Journal of Computational Physics*, vol. 311, pp. 299–328, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999116000632

[10] M. Svärd and J. Nordström, “Review of summation-by-parts schemes for initial–boundary-value problems,” *Journal of Computational Physics*, vol. 268, pp. 17–38, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002199911400151X

[11] O. O’Reilly, T. Lundquist, E. M. Dunham, and J. Nordström, “Energy stable and high-order-accurate finite difference methods on staggered grids,” *Journal of Computational Physics*, vol. 346, pp. 572–589, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999117304758

[12] O. O’Reilly and N. A. Petersson, “Energy conservative sbp discretizations of the acoustic wave equation in covariant form on staggered curvilinear grids,” *Journal of Computational Physics*, vol. 411, p. 109386, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999120301601

[13] L. Gao, D. C. Del Rey Fernández, M. Carpenter, and D. Keyes, “Sbp–sat finite difference discretization of acoustic wave equations on staggered block-wise uniform grids,” *Journal of Computational and Applied Mathematics*, vol. 348, pp. 421–444, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377042718305272

[14] D. Del Rey Fernández, J. Hicken, and D. Zingg, “Review of summation-by-parts operators with simultaneous approximation terms for the numerical solution of partial differential equations,” *Computers & Fluids*, vol. 95, p. 171–196, 05 2014.

[15] J. Nordström and M. Björck, “Finite volume approximations and strict stability for hyperbolic problems,” *Applied Numerical Mathematics*, vol. 38, no. 3, pp. 237–255, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168927401000277

[16] J. Nordström, K. Forsberg, C. Adamsson, and P. Eliasson, “Finite volume methods, unstructured meshes and strict stability for hyperbolic problems,” *Applied Numerical Mathematics*, vol. 45, no. 4, pp. 453–473, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168927402002398

[17] M. Svärd and J. Nordström, “Stability of finite volume approximations for the laplacian operator on quadrilateral and triangular grids,” *Applied Numerical Mathematics*, vol. 51, no. 1, pp. 101–125, 2004.

[18] F. Ham, K. Mattsson, and G. Iaccarino, “Accurate and stable finite volume operators for unstructured flow solvers,” *Annual Research Briefs*, pp. 243–261, 2006.

[19] M. Svärd, J. Gong, and J. Nordström, “An accuracy evaluation of unstructured node-centred finite volume methods,” *Applied Numerical Mathematics - APPL NUMER MATH*, vol. 58, pp. 1142–1158, 08 2008.

[20] ——, “Stable artificial dissipation operators for finite volume schemes on unstructured grids,” *Applied Numerical Mathematics*, vol. 56, no. 12, pp. 1481–1490, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168927405002060

[21] A. Gjesteland, “Sbp-sat schemes for hyperbolic problems,” 2019. [Online]. Available: https://hdl.handle.net/1956/19980

[22] F. Ham, K. Mattsson, G. Iaccarino, and P. Moin, *Towards Time-Stable and Accurate LES on Unstructured Grids*, 07 2007, vol. 56, pp. 235–249.

[23] Q. Abbas, E. van der Weide, and J. Nordström, “Energy stability of the muscl scheme,” in *Numerical Mathematics and Advanced Applications 2009*, G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 61–68.

[24] T. Barth and P. Frederickson, “High order solution of the euler equations on unstructured grids quadratic reconstruction,” 02 1990.

[25] C. Ollivier-Gooch and M. Van Altena, “A high-order-accurate unstructured mesh finite-volume scheme for the advection–diffusion equation,” *Journal of Computational Physics*, vol. 181, no. 2, pp. 729–752, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999102971597

[26] C. Ollivier-Gooch, A. Nejat, and K. Michalak, “On obtaining high-order finite-volume solutions to the euler equations on unstructured meshes,” 06 2007.

[27] A. Harten and S. R. Chakravarthy, “Multi-dimensional eno schemes for general geometries,” INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING HAMPTON VA, Tech. Rep., 1991.

[28] R. Abgrall, “On essentially non-oscillatory schemes on unstructured meshes: Analysis and implementation,” *Journal of Computational Physics*, vol. 114, no. 1, pp. 45–58, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002199918471148X

[29] Z. Wang, “Spectral (finite) volume method for conservation laws on unstructured grids. basic formulation: Basic formulation,” *Journal of Computational Physics*, vol. 178, no. 1, pp. 210–251, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999102970415

[30] Z. J. Wang and Y. Liu, “Spectral (finite) volume method for conservation laws on unstructured grids: Ii. extension to two-dimensional scalar equation,” *Journal of Computational Physics*, vol. 179, no. 2, pp. 665–697, 2002.

[31] ——, “Spectral (finite) volume method for conservation laws on unstructured grids iii: One dimensional systems and partition optimization,” *Journal of Scientific Computing*, vol. 20, pp. 137–157, 2004.

[32] Z. J. Wang, L. Zhang, and Y. Liu, “Spectral (finite) volume method for conservation laws on unstructured grids iv: extension to two-dimensional systems,” *Journal of Computational Physics*, vol. 194, no. 2, pp. 716–741, 2004.

[33] K. Van den Abeele and C. Lacor, “An accuracy and stability study of the 2d spectral volume method,” *Journal of Computational Physics*, vol. 226, no. 1, pp. 1007–1026, 2007.

[34] K. Van den Abeele, C. Lacor, and Z. Wang, “On the connection between the spectral volume and the spectral difference method,” *Journal of Computational Physics*, vol. 227, no. 2, pp. 877–885, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999107003889

[35] K. Van den Abeele, G. Ghorbaniasl, M. Parsani, and C. Lacor, “A stability analysis for the spectral volume method on tetrahedral grids,” *Journal of Computational Physics*, vol. 228, no. 2, pp. 257–265, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999108005263

[36] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods, Second Edition.* John Wiley & Sons, Ltd, 2013.

[37] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, ser. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.

[38] H. Holden and N. Risebro, *Front Tracking for Hyperbolic Conservation Laws.* Springer New York, 2002. [Online]. Available: https://books.google.no/books?id=cvqMMgEACAAJ

[39] L. C. Evans, *Partial differential equations.* Providence, R.I.: American Mathematical Society, 2010.

[40] B. Gustafsson, *High Order Difference Methods for Time Dependent PDE*, 01 2008, vol. 38.

[41] G. B. Whitham, *Linear and Nonlinear Waves*, ser. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley-Interscience, 1999.

[42] J. S. Hesthaven, *Numerical Methods for Conservation Laws.* Philadelphia, PA: Society for Industrial and Applied Mathematics, 2018. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611975109

[43] E. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction.* Springer Berlin, Heidelberg, 01 2009.

[44] J. S. Hesthaven and T. Warburton, *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications.* Springer Science & Business Media, 2007.

[45] P. Lax and R. Richtmyer, "Survey of the stability of linear finite difference equations," *Communications on Pure and Applied Mathematics*, vol. 9, pp. 267–293, 01 1956. [Online]. Available: https://math.berkeley.edu/~wilken/228B.S07/LaxRichtmyer.pdf

[46] S. K. Godunov and I. Bohachevsky, "Finite difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics," *Matematičeskij sbornik*, vol. 47, no. 3, pp. 271–306, 1959.

[47] B. van Leer, J. Thomas, P. Roe, and R. Newsome, "A comparison of numerical flux formulas for the euler and navier-stokes equations," 06 1987.

[48] B. Van Leer, "Towards the ultimate conservative difference scheme. iv. a new approach to numerical convection," *Journal of Computational Physics*, vol. 23, no. 3, pp. 276–299, 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/002199917790095X

[49] B. van Leer, "Towards the ultimate conservative difference scheme. v. a second-order sequel to godunov's method," *Journal of Computational Physics*, vol. 32, no. 1, pp. 101–136, 1979. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021999179901451

[50] P. Colella and P. R. Woodward, "The piecewise parabolic method (ppm) for gas-dynamical simulations," *Journal of Computational Physics*, vol. 54, no. 1, pp. 174–201, 1984. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021999184901438

[51] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, "Uniformly high order accurate essentially non-oscillatory schemes, iii," *Journal of Computational Physics*, vol. 71, no. 2, pp. 231–303, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021999187900313

[52] Q. Wang, Y.-X. Ren, and W. Li, "Compact high order finite volume method on unstructured grids i: Basic formulations and one-dimensional schemes," *Journal of Computational Physics*, vol. 314, pp. 863–882, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999116000784

[53] ——, "Compact high order finite volume method on unstructured grids ii: Extension to two-dimensional euler equations," *Journal of Computational Physics*, vol. 314, pp. 883–908, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002199911600200X

[54] Q. Wang, Y.-X. Ren, J. Pan, and W. Li, "Compact high order finite volume method on unstructured grids iii: Variational reconstruction," *Journal of Computational Physics*, vol. 337, pp. 1–26, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999117301249

[55] G.-S. Jiang and C.-W. Shu, "Efficient implementation of weighted eno schemes," *Journal of Computational Physics*, vol. 126, no. 1, pp. 202–228, 1996. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999196901308

[56] C.-W. Shu and S. Osher, “Efficient implementation of essentially non-oscillatory shock-capturing schemes,” *Journal of Computational Physics*, vol. 77, no. 2, pp. 439–471, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0021999188901775

[57] C. Hu and C.-W. Shu, “Weighted essentially non-oscillatory schemes on triangular meshes,” *Journal of Computational Physics*, vol. 150, no. 1, pp. 97–127, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999198961654

[58] T. B. Hestvik. [Online]. Available: https://github.com/thomasbjarne/master_degree_project

[59] A. Jameson, “A proof of the stability of the spectral difference method for all orders of accuracy,” *J. Sci. Comput.*, vol. 45, pp. 348–358, 10 2010.