DRAFT/ ENERGY ESTIMATE FOR ACTIVE FLUX DISCRETIZATION OF LINEAR ADVECTION IBVP

THOMAS BJARNE HESTVIK

Abstract. We derive an energy estimate for the third-order Active Flux discretization of the 1D scalar linear advection initial-boundary value problem. Our estimate is valid for problem data in L^2, whenever the CFL number $\lambda := a\Delta t/h$ satisfies $\lambda \in [0, 1]$. Of notable consequence, the estimate proves the scheme is energy stable, i.e. stable w.r.t. perturbations of the problem data in a discrete L^2 norm.

Contents

1. Introduction

The Active Flux method is a new finite volume method [1, 2, 3], inspired by van Leer's "Scheme V" [4]. The schemes obtained by the Active Flux method have been shown to hold several desirable properties, including: stationarity preservation [5]; truly multidimensional flux evaluations [2, 6]; well-balancing for the shallow water equations [7]; local stencils and high-order accuracy. Results on the stability of Active Flux schemes for linear problems have as of this writing been limited to spectral analysis of the discrete evolution operator, see e.g. [8]. In particular, to the best of our knowledge there are no stability results for Active Flux schemes approximating linear initial-boundary value problems (IBVP).

To determine the stability of schemes for linear IBVP we use the (discrete) energy method. The procedure aims to prove that the numerical solution of a scheme, consistent with some problem, satisfies a discrete L^2, or energy estimate at any discrete time t^n which mimics the L^2 estimate of the exact solution of the problem. This is achieved by obtaining an appropriate upper bound on the increase of the discrete L^2 estimate of the numerical solution between arbitrary time steps t^{n+1} and t^n. To read more about the energy method see e.g. [9, 10, 11, 12] and the references therein.

Unlike traditional finite volume methods, the Active Flux method saves and updates both approximate mean values and approximate point values in each time step. This allows for a continuous reconstruction of the solution, enabling flux evaluation at the boundaries without the use of Riemann solvers. The additional degrees of freedom complicates the energy analysis

Date: 28. February 2024.

2020 *Mathematics Subject Classification.* 65M12, 65M08.

Key words and phrases. Active Flux, finite volume methods, energy stability, initial-boundary value problem, linear advection.

by adding inner-products of approximate mean values and finite-difference operators applied to approximate point values. As the sign of these products is indeterminate unless we make strict assumptions on the problem data, we are only able to give a rough estimate using the typical inequalities.

In this paper, we apply the Active Flux method to the 1D linear advection IBVP:

$$\begin{cases} \partial_t u + a\partial_x u = 0, & x \in \Omega = (x_L, x_R) \subset \mathbb{R}, \quad t \in (0,T], \\ u(x,0) = u_0(x), & x \in \Omega, \\ u(x_L, t) = g(t), & t \in (0,T], \end{cases} \tag{1.1}$$

with $u_0 \in L^2(\Omega)$, $g \in L^2([0,T])$ and positive advection speed $a > 0$. It is well-known that the solution $u \in L^\infty([0,T], L^2(\Omega))$ of (1.1) satisfies

$$\|u(\cdot,t)\|^2_{L^2(\Omega)} \leq Ce^{\eta t}\left(\|u_0\|^2_{L^2(\Omega)} + a\int_0^t |g(\tau)|^2 d\tau\right), \tag{1.2}$$

for all $\eta \geq 0$ and $C \geq 1$. Our goal is to prove that the numerical solution $\overline{\mathbf{u}}^n$ of the third-order Active Flux scheme approximating (1.1) satisfies

$$\|\overline{\mathbf{u}}^n\|^2_h \leq Ke^{\theta n\Delta t}\left(\left\|\overline{\mathbf{u}}^0\right\|^2_h + a\Delta t\sum_{j=1}^n\sum_{k=0}^2 w_k|g(t^{j-1} + k\Delta t/2)|^2\right),$$

whenever Δt satisfies some CFL condition. Here K and θ are constants in $\mathbb{R}$ independent of u_0 and g. Further, $\|\cdot\|_h$ denotes our discrete L^2-bounded norm (cf. Lemma 1), and $\sum_{k=0}^2 w_k\Delta t|g(t^{j-1} + k\Delta t/2)|^2$ denotes Simpson's rule applied to $\int_{t^{j-1}}^{t^j} |g(\tau)|^2 d\tau$. This result is obtained in the following form:

Theorem 1 (Main result)**.** *Let $\overline{\mathbf{u}}^n$ be the numerical solution of the third-order Active Flux scheme approximating (1.1). If the CFL number λ satisfies $0 \leq \lambda \leq 1$, then*

$$\|\overline{\mathbf{u}}^n\|^2_h \leq 2e^{8c\frac{a}{h}n\Delta t}\left(\left\|\overline{\mathbf{u}}^0\right\|^2_h + a\Delta t\sum_{j=1}^n\sum_{k=0}^2 w_k|g(t^{j-1} + k\Delta t/2)|^2\right),$$

for all $n \geq 1$. Here $c \to 1$ as $h \to 0$.

The remaining text is organized as follows. In Section 2.1 we introduce the Active Flux method. In Section 2.2 we motivate stability in the norm $\|\cdot\|_h$ for finite volume schemes approximating scalar linear hyperbolic problems. In Section 3 we apply the Active Flux method to (1.1), and show how the resulting scheme can be written as $\overline{\mathbf{u}}^{n+1} = \mathcal{A}(\overline{\mathbf{u}}^n, \mathbf{u}^n)$ where $\mathcal{A}$ is an affine map and $\mathbf{u}^n$ the approximate point values. In Section 4 we prove Theorem 1. Finally, in Section 5 we give some concluding remarks.

2. Preliminaries

2.1. **Active Flux method.**

In this subsection we briefly summarize the Active Flux method [2] with the assumption of regular (i.e. equal measure) control volumes. Consider the 1D conservation law problem

$$\partial_t u + \partial_x f(u) = 0, \qquad x \in \Omega = (x_L, x_R), \qquad t \in (0,T], \tag{2.1}$$

with some initial and boundary conditions. Let Ω be partitioned into N control volumes $\{C_i\}_{i=1}^N$ given by $C_i = (x_{i-1/2}, x_{i+1/2})$ with $x_{1-1/2} = x_L$ and $x_{N+1/2} = x_R$ such that:

- $C_i \cap C_j = \emptyset$ for $i \neq j$;
- $\bigcup_{i=1}^N \overline{C_i} = \overline{\Omega}$.

Next, define a temporal grid $t^n = n\Delta t$ with $n \in \{0, 1, \dots, N_T\}$ and $t^{N_T} \in (T - \Delta t, T + \Delta t)$. We will use the notation $x_i = \frac{1}{2}(x_{i-1/2} + x_{i+1/2})$, $h = x_{i+1/2} - x_{i-1/2}$,

$$\overline{u}_i^n \approx \frac{1}{h}\int_{C_i} u(x, t^n)dx, \qquad \text{and} \qquad u_{i\pm1/2}^n \approx u(x_{i\pm1/2}, t^n).$$

In the Active Flux method, we store approximations of both control-volume-averaged values and control-volume-boundary values at the current time step,

$$\overline{\mathbf{u}}^n = [\overline{u}_1^n, \overline{u}_2^n, \dots]^T \in \mathbb{R}^N, \qquad \mathbf{u}^n = [u_{1-1/2}^n, u_{2-1/2}^n, \dots]^T \in \mathbb{R}^{N+1}.$$

Suppose that $\overline{\mathbf{u}}^n$ and $\mathbf{u}^n$ are given. To find $\overline{\mathbf{u}}^{n+1}$ we begin by calculating $\mathbf{u}^{n+1/2}$ and $\mathbf{u}^{n+1}$ as follows: Define $R \in \mathcal{C}^0(\overline{\Omega}) \cap \mathcal{C}^\infty(\cup_{i=1}^N C_i)$ by

$$R(x) = R_i(x), \qquad x \in \overline{C_i}, \qquad i = 1, \dots, N,$$

where R_i denotes the Taylor series of u centered at x_i, with the derivatives replaced by finite differences. The approximation of $u(x_i)$ is chosen such that R_i satisfies

$$R_i(x_{i\pm1/2}) = u_{i\pm1/2}^n, \qquad \frac{1}{h}\int_{C_i} R_i(x)dx = \overline{u}_i^n, \qquad i = 1, \dots, N.$$

In other words,

$$\text{(2.2)} \quad R_i(x) = \frac{1}{4}\left(6\overline{u}_i^n - u_{i+1/2}^n - u_{i-1/2}^n\right) + \frac{1}{h}(x - x_i)\left(u_{i+1/2}^n - u_{i-1/2}^n\right) + \frac{3}{h^2}(x - x_i)^2\left(-2\overline{u}_i^n + u_{i+1/2}^n + u_{i-1/2}^n\right).$$

Next, let S denote the solution operator of the problem (2.1), i.e. $S(t, t')u(x, t') = u(x, t)$ for $t > t'$. Then for $k = 1, 2$ we obtain $u_{i\pm1/2}^{n+k/2}$ by applying S to R evaluated at $x_{i\pm1/2}$. That is,

$$u_{i\pm1/2}^{n+k/2} = S(t^n + k\Delta t/2, t^n)R(x_{i\pm1/2}) \approx S(t^n + k\Delta t/2, t^n)u(x_{i\pm1/2}, t^n) = u(x_{i\pm1/2}, t^{n+k/2}).$$

Example. Consider (2.1) with $f(u) = au$ and periodic boundary. Clearly

$$S(t, t')u(x, t') = u(x - a(t - t'), t'),$$

and we update the elements in $\mathbf{u}^n$ by

$$u_{i\pm1/2}^{n+k/2} = R(x_{i\pm1/2} - ak\Delta t/2).$$

After obtaining $\mathbf{u}^{n+1/2}$ and $\mathbf{u}^{n+1}$, the mean values $\overline{\mathbf{u}}^n$ are updated by the fully discretized finite volume form of (2.1),

$$\overline{u}_i^{n+1} = \overline{u}_i^n - \frac{\Delta t}{h}\left(\overline{f}_{i+1/2}^n - \overline{f}_{i-1/2}^n\right),$$

where

$$\overline{f}_{i\pm1/2}^n \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(x_{i\pm1/2}, \tau))d\tau.$$

Following [2], we use Simpson's rule to approximate the average flux in time,

$$\overline{f}_{i\pm1/2}^n = \frac{1}{6}\left(f(u_{i\pm1/2}^n) + 4f(u_{i\pm1/2}^{n+1/2}) + f(u_{i\pm1/2}^{n+1})\right).$$

Now we repeat the process, obtaining $\mathbf{u}^{(n+1)+1/2}$ and $\mathbf{u}^{(n+1)+1}$ first, and then $\overline{\mathbf{u}}^{(n+1)+1}$ and so on until $\overline{\mathbf{u}}^{N_T}$ is obtained.

Remark. For nonlinear problems, an exact solution operator is seldom available. In this case, an approximate solution operator (see [13]) can be used to update the point values $\mathbf{u}^n$. Alternatively, one could use a modification of the Active Flux method, see e.g. [3, 14].

2.2. **Energy stability.** We want the numerical solution of a scheme consistent with (1.1) to exhibit continuity with respect to the problem data, similarly to (1.2). To formalize this idea, we follow the standard theory for stability of finite difference schemes (see [9]). Write $H = \operatorname{diag}(h, \dots, h)$,

$$\mathbf{F} = \left[\overline{f}^n_{1+1/2} - \overline{f}^n_{1-1/2}, \dots, \overline{f}^n_{N+1/2} - \overline{f}^n_{N-1/2}\right]^T,$$

and suppose that

$$\overline{\mathbf{u}}^{n+1} = \overline{\mathbf{u}}^n - \Delta t H^{-1}\mathbf{F}$$

is consistent with (1.1). Since

$$\overline{u}^n_i \approx \frac{1}{h}\int_{C_i} u(x,t^n)dx,$$

it seems natural to identify the numerical solution $\overline{\mathbf{u}}^n$ with the simple function $U^n : \Omega \to \mathbb{R}$ given by $U^n(x) = \sum_{i=1}^N \overline{u}^n_i \chi_{C_i}(x)$. We have

$$\|U^n\|^2_{L^2(\Omega)} = \int_\Omega |U^n|^2 = \sum_{i=1}^N \int_{C_i} |\overline{u}^n_i \chi_{C_i}|^2 = \sum_{i=1}^N |\overline{u}^n_i|^2 \int_{C_i} \chi_{C_i} = \sum_{i=1}^N h|\overline{u}^n_i|^2 = \langle \overline{\mathbf{u}}^n, H\overline{\mathbf{u}}^n\rangle = \|\overline{\mathbf{u}}^n\|^2_h,$$

which motivates the following definition.

Definition 2.1. *A finite volume scheme*

$$\overline{\mathbf{u}}^{n+1} = \overline{\mathbf{u}}^n - \Delta t H^{-1}\mathbf{F}$$

consistent with (1.1) is energy stable if there exist K, $\theta \in \mathbb{R}$ independent of u_0 and g such that

$$\|\overline{\mathbf{u}}^n\|^2_h \le Ke^{\theta n \Delta t}\left(\left\|\overline{\mathbf{u}}^0\right\|^2_h + a\Delta t \sum_{j=1}^n \sum_{k=0}^q w_k |g(t^{j-1} + \xi_k \Delta t)|^2\right), \qquad \forall n \ge 1,$$

where

$$\sum_{k=0}^q w_k \Delta t \phi(t^{j-1} + \xi_k \Delta t) \to \int_{t^{j-1}}^{t^j} \phi(\tau)d\tau,$$

as $\Delta t \to 0$ for all $\phi \in L^1((t^{j-1}, t^j))$.

Remark. We will use the following implication in section 4.

$$\overline{\mathbf{u}}^{n+1} = \overline{\mathbf{u}}^n - \Delta t H^{-1}\mathbf{F} \implies \left\|\overline{\mathbf{u}}^{n+1}\right\|^2_h - \|\overline{\mathbf{u}}^n\|^2_h = -2\Delta t\langle \overline{\mathbf{u}}^n, \mathbf{F}\rangle + \frac{(\Delta t)^2}{h}\langle \mathbf{F}, \mathbf{F}\rangle. \tag{2.3}$$

Next, we prove that if $\overline{\mathbf{u}}^n$ is consistent with the control-volume-averaged data of some function u, then the h norm is bounded by the L^2 norm. We write $f = \mathcal{O}(h)$ if there exists $c \in \mathbb{R}^+$ such that $|f| \le ch$ as $h \to 0$.

Lemma 1. *Let Ω be a bounded real interval, $u(\cdot, t^n) \in L^2(\Omega)$ and $\bigcup_{i=1}^N \overline{C_i} = \overline{\Omega}$. If*

$$\overline{u}^n_i = \frac{1}{h}\int_{C_i} u(x,t^n)dx + \mathcal{O}(h^r), \qquad i = 1, \dots, N,$$

then there exists $M \in \mathbb{R}^+$ depending on h such that

$$\|\overline{\mathbf{u}}^n\|^2_h \le M\,\|u(\cdot,t^n)\|^2_{L^2(\Omega)}.$$

Proof. By assumption,

$$h(\overline{u}^n_i)^2 = \frac{1}{h}\left(\int_{C_i} u(x,t^n)dx\right)^2 + \mathcal{O}(h^r)\int_{C_i} u(x,t^n)dx + \mathcal{O}(h^{2r+1}), \qquad i = 1, \dots, N.$$

Summing over $i = 1, \dots, N$ and noting that $Nh = x_R - x_L = |\Omega|$,

$$\|\overline{\mathbf{u}}^n\|^2_h = \sum_{i=1}^N \frac{1}{h}\left(\int_{C_i} u(x,t^n)dx\right)^2 + \mathcal{O}(h^r)\int_\Omega u(x,t^n)dx + |\Omega|\mathcal{O}(h^{2r}).$$

Using the Cauchy-Schwarz inequality,

$$\int_\Omega u(x,t^n)dx \leq |\Omega|^{1/2} \left\| u(\cdot,t^n) \right\|_{L^2(\Omega)}, \qquad \frac{1}{h}\left(\int_{C_i} u(x,t^n)dx\right)^2 \leq \left\| u(\cdot,t^n) \right\|_{L^2(C_i)}^2.$$

Since $|\Omega|$ and $\left\| u(\cdot,t^n) \right\|_{L^2(\Omega)}$ are bounded above, we have

$$\mathcal{O}(h^r)|\Omega|^{1/2} \left\| u(\cdot,t^n) \right\|_{L^2(\Omega)} = \mathcal{O}(h^r), \qquad \text{and} \qquad |\Omega|\mathcal{O}(h^{2r}) = \mathcal{O}(h^{2r}).$$

Therefore, writing $\mathcal{O}(h^r) + \mathcal{O}(h^{2r}) = \mathcal{O}(h^r)$ gives

$$\|\overline{\mathbf{u}}^n\|_h^2 \leq \|u(\cdot,t^n)\|_{L^2(\Omega)}^2 + \mathcal{O}(h^r) \leq \|u(\cdot,t^n)\|_{L^2(\Omega)}^2 + ch^r,$$

for some $c \in \mathbb{R}^+$. Let $M = 1 + ch^r / \|u(\cdot,t^n)\|_{L^2(\Omega)}^2$ to complete the proof. □

Using Lemma 1, we can prove that the approximate solution of (1.1) obtained by an energy stable finite volume scheme will satisfy the estimate (1.2) in the limit $h \to 0$ and $\Delta t \to 0$. More precisely, we can prove the following proposition.

Proposition 2.1. *Let $\overline{\mathbf{u}}^n$ be the numerical solution of an energy stable finite volume scheme consistent with (1.1), and $U^n : \Omega \to \mathbb{R}$ the simple function $U^n(x) = \sum_{i=1}^N \overline{u}_i^n \chi_{C_i}(x)$. Then*

$$\lim_{\Delta t,\, h \to 0} \|U^n\|_{L^2(\Omega)}^2 \leq Ce^{\eta n \Delta t}\left(\|u_0\|_{L^2(\Omega)}^2 + a\int_0^{t^n} |g(\tau)|^2 d\tau\right), \qquad \forall n \geq 1,$$

for some bounded $C \in \mathbb{R}^+$.

Proof. We may assume without loss of generality that the initial numerical data $\overline{\mathbf{u}}^0$ satisfies

$$\overline{u}_i^0 = \frac{1}{h}\int_{C_i} u_0(x)dx + \mathcal{O}(h), \qquad i = 1,\dots,N.$$

Further, by assumption we have the inequality

$$\|\overline{\mathbf{u}}^n\|_h^2 \leq Ke^{\theta n \Delta t}\left(\left\|\overline{\mathbf{u}}^0\right\|_h^2 + a\Delta t \sum_{j=1}^n \sum_{k=0}^q w_k |g(t^{j-1} + \xi_k \Delta t)|^2\right), \qquad \forall n \geq 1.$$

Recalling $\|U^n\|_{L^2(\Omega)}^2 = \|\overline{\mathbf{u}}^n\|_h^2$ and using Lemma 1 this gives

$$\|U^n\|_{L^2(\Omega)}^2 \leq Ke^{\theta n \Delta t}\left(M\|u_0\|_{L^2(\Omega)}^2 + a\Delta t \sum_{j=1}^n \sum_{k=0}^q w_k |g(t^{j-1} + \xi_k \Delta t)|^2\right), \qquad \forall n \geq 1.$$

Now,

$$\begin{aligned} M &\to 1 &&\text{as} && h \to 0, \\ \sum_{k=0}^q w_k \Delta t |g(t^{j+\xi_k})|^2 &\to \int_{t^j}^{t^{j+1}} |g(\tau)|^2 d\tau &&\text{as} && \Delta t \to 0, \end{aligned}$$

imply

$$\lim_{\Delta t,\, h \to 0} \|U^n\|_{L^2(\Omega)}^2 \leq Ce^{\eta n \Delta t}\left(\|u_0\|_{L^2(\Omega)}^2 + a\int_0^{t^n} |g(\tau)|^2 d\tau\right), \qquad \forall n \geq 1,$$

where $C = Ke^{(\theta-\eta)n\Delta t}$. □

3. Active Flux scheme for the linear advection IBVP

We apply the Active Flux method to the problem (1.1) using structured, regular control volumes with measure h. We obtain the finite volume scheme

$$\overline{u}_i^{n+1} = \overline{u}_i^n - a\frac{\Delta t}{h}\sum_{k=0}^{2} w_k \left(u_{i+1/2}^{n+k/2} - u_{i-1/2}^{n+k/2}\right), \tag{3.1}$$

where $w_0 = 1/6$, $w_1 = 4/6$ and $w_2 = 1/6$. Recall that the values $u_{i\pm 1/2}^{n+k/2}$ are obtained by applying the solution operator of the continuous problem to $R(x_{i\pm1/2})$. For the linear advection PDE (with $a > 0$ and $0 < a\Delta t \leq h$) this gives

$$\begin{aligned} u_{i-1/2}^{n+k/2} &= R_{i-1}(x_{i-1/2} - ak\Delta t/2), \qquad i \neq 1, \\ u_{i+1/2}^{n+k/2} &= R_i(x_{i+1/2} - ak\Delta t/2), \end{aligned}$$

and the boundary node $u_{1-1/2}^n$ is updated by the boundary condition,

$$u_{1-1/2}^{n+k/2} = g(t^n + k\Delta t/2).$$

Inserting this into (3.1) gives

$$\begin{aligned} \overline{u}_1^{n+1} &= \overline{u}_1^n - a\frac{\Delta t}{h}\sum_{k=0}^{2} w_k \Big(R_1(x_{1+1/2} - ak\Delta t/2) - g(t^n + k\Delta t/2)\Big) \\ \overline{u}_i^{n+1} &= \overline{u}_i^n - a\frac{\Delta t}{h}\sum_{k=0}^{2} w_k \Big(R_i(x_{i+1/2} - ak\Delta t/2) - R_{i-1}(x_{i-1/2} - ak\Delta t/2)\Big), \qquad (i \neq 1) \end{aligned}$$

and the vector $\mathbf{F}$ becomes

$$\mathbf{F} = a\sum_{k=0}^{2} w_k \begin{bmatrix} R_1(x_{1+1/2} - ak\Delta t/2) - g(t^n + k\Delta t/2) \\ R_2(x_{2+1/2} - ak\Delta t/2) - R_1(x_{1+1/2} - ak\Delta t/2) \\ \vdots \\ R_N(x_{N+1/2} - ak\Delta t/2) - R_{N-1}(x_{N-1/2} - ak\Delta t/2) \end{bmatrix}. \tag{3.2}$$

In order to analyze the scheme, we want to express $\mathbf{F}$ as an affine map of the degrees of freedom $\overline{\mathbf{u}}^n$ and $\mathbf{u}^n$. We have the following result:

Lemma 2. *Let $\mathbf{F}$ be given by (3.2) and $\{R_i\}_{i=1}^N$ by (2.2). Define α, β, and $\gamma \in \mathbb{R}$ by*

$$\alpha = -2\left(\frac{a\Delta t}{h}\right)^2 + 3\frac{a\Delta t}{h}, \qquad \beta = \left(\frac{a\Delta t}{h}\right)^2 - \frac{a\Delta t}{h}, \qquad \gamma = \left(\frac{a\Delta t}{h}\right)^2 - 2\frac{a\Delta t}{h} + 1. \tag{3.3}$$

Then, $QL_1 \in \mathbb{R}^{N\times N}$ and $QL_2 \in \mathbb{R}^{N\times(N+1)}$ given by

$$QL_1 = \begin{bmatrix} \alpha & & & \\ -\alpha & \alpha & & \\ & \ddots & \ddots & \\ & & -\alpha & \alpha \end{bmatrix}, \qquad QL_2 = \begin{bmatrix} \beta & \gamma & & & \\ -\beta & \beta-\gamma & \gamma & & \\ & \ddots & \ddots & \ddots & \\ & & -\beta & \beta-\gamma & \gamma \end{bmatrix}, \tag{3.4}$$

satisfy

$$\mathbf{F} = aQL_1\overline{\mathbf{u}}^n + aQL_2\mathbf{u}^n - a\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)e_1.$$

Proof. Note that the local reconstructions $R_i(x)$ can be written as $R_i(x) = \langle \psi_i(x), L_i(\overline{\mathbf{u}}^n, \mathbf{u}^n)\rangle$ where

$$\psi_i(x) = \begin{bmatrix} 1 \\ (x-x_i) \\ (x-x_i)^2 \end{bmatrix}, \qquad \text{and} \qquad L_i(\overline{\mathbf{u}}^n, \mathbf{u}^n) = \begin{bmatrix} \frac{1}{4}(6\overline{u}_i^n - u_{i-1/2}^n - u_{i+1/2}^n) \\ \frac{1}{h}(u_{i+1/2}^n - u_{i-1/2}^n) \\ \frac{3}{h^2}(-2\overline{u}_i^n + u_{i-1/2}^n + u_{i+1/2}^n) \end{bmatrix}. \tag{3.5}$$

We introduce the vector $\mathbf{L}\mathbf{u}^n = [L_1(\overline{\mathbf{u}}^n, \mathbf{u}^n), \dots, L_N(\overline{\mathbf{u}}^n, \mathbf{u}^n)]^T \in (\mathbb{R}^3)^N$ and matrix $Q \in \mathbb{R}^{N\times 3N}$ defined by

$$Q_{ij:j+2} = \begin{cases} \sum_{k=0}^2 w_k \psi_i(x_{i+1/2} - ak\Delta t/2), & j = 3(i-1)+1 \\ -\sum_{k=0}^2 w_k \psi_{i-1}(x_{i-1/2} - ak\Delta t/2), & j = 3(i-2)+1 \\ 0, & \text{otherwise.} \end{cases}$$

Then it follows by construction that we have

$$\mathbf{F} = aQ\mathbf{L}\mathbf{u}^n - a\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)e_1,$$

where $e_1 = [1, 0, 0, \dots]^T \in \mathbb{R}^N$. Moreover, (3.5) indicates that $L_i(\overline{\mathbf{u}}^n, \mathbf{u}^n) = L_{i,1}\overline{\mathbf{u}}^n + L_{i,2}\mathbf{u}^n$ with

$$L_{i,1}\overline{\mathbf{u}}^n = \begin{bmatrix} \frac{6}{4}\overline{u}_i^n \\ 0 \\ -\frac{6}{h^2}\overline{u}_i^n \end{bmatrix}, \qquad \text{and} \qquad L_{i,2}\mathbf{u}^n = \begin{bmatrix} \frac{1}{4}(-u_{i+1/2}^n - u_{i-1/2}^n) \\ \frac{1}{h}(u_{i+1/2}^n - u_{i-1/2}^n) \\ \frac{3}{h^2}(u_{i+1/2}^n + u_{i-1/2}^n) \end{bmatrix}.$$

Hence, $\mathbf{L}\mathbf{u}^n$ can be decomposed into $\mathbf{L}\mathbf{u}^n = L_1\overline{\mathbf{u}}^n + L_2\mathbf{u}^n$, where L_1 and L_2 are given by

$$L_1 \in \mathbb{R}^{3N\times N} = \begin{bmatrix} 6/4 & & & \\ 0 & & & \\ -6/(h^2) & & & \\ & \ddots & & \\ & \ddots & & \\ & \ddots & & \\ & & & 6/4 \\ & & & 0 \\ & & & -6/(h^2) \end{bmatrix}, \quad L_2 \in \mathbb{R}^{3N\times(N+1)} = \begin{bmatrix} -1/4 & -1/4 & & & \\ -1/h & 1/h & & & \\ 3/(h^2) & 3/(h^2) & & & \\ & \ddots & \ddots & & \\ & \ddots & \ddots & & \\ & \ddots & \ddots & & \\ & & & -1/4 & -1/4 \\ & & & -1/h & 1/h \\ & & & 3/(h^2) & 3/(h^2) \end{bmatrix},$$

and it follows that

$$\mathbf{F} = aQL_1\overline{\mathbf{u}}^n + aQL_2\mathbf{u}^n - a\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)e_1. \tag{3.6}$$

To see that QL_1 and QL_2 satisfy (3.4), note that $x_{i+1/2} - x_i = h/2$ for all i implies

$$\sum_{k=0}^{2} w_k\psi_i(x_{i+1/2} - ak\Delta t/2) = \begin{bmatrix} \sum_{k=0}^2 w_k \\ \sum_{k=0}^2 w_k(x_{i+1/2} - ak\Delta t/2 - x_i) \\ \sum_{k=0}^2 w_k(x_{i+1/2} - ak\Delta t/2 - x_i)^2 \end{bmatrix} = \begin{bmatrix} \sum_{k=0}^2 w_k \\ \sum_{k=0}^2 w_k(h/2 - ak\Delta t/2) \\ \sum_{k=0}^2 w_k(h/2 - ak\Delta t/2)^2 \end{bmatrix}.$$

Now,

$$(QL_1)_{ij} = \begin{cases} \sum_{k=0}^2 w_k \left(\frac{6}{4} - \frac{6}{h^2}(h/2 - ak\Delta t/2)^2\right), & j = i \\ -\sum_{k=0}^2 w_k \left(\frac{6}{4} - \frac{6}{h^2}(h/2 - ak\Delta t/2)^2\right), & j = i-1 \\ 0, & \text{otherwise,} \end{cases}$$

$$(QL_2)_{ij} = \begin{cases} \sum_{k=0}^2 w_k \left(-\frac{1}{4} - \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right), & j = i = 1 \\ \Bigg(\sum_{k=0}^2 w_k \left(-\frac{1}{4} - \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right) & \\ \quad - \sum_{k=0}^2 w_k \left(-\frac{1}{4} + \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right) \Bigg), & j = i \neq 1 \\ \sum_{k=0}^2 w_k \left(-\frac{1}{4} + \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right), & j = i+1 \\ -\sum_{k=0}^2 w_k \left(-\frac{1}{4} - \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right), & j = i-1 \\ 0, & \text{otherwise.} \end{cases}$$

Finally, simple calculation reveals

$$\begin{aligned} \sum_{k=0}^2 w_k \left(\frac{6}{4} - \frac{6}{h^2}(h/2 - ak\Delta t/2)^2\right) &= \alpha, \\ \sum_{k=0}^2 w_k \left(-\frac{1}{4} - \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right) &= \beta, \\ \sum_{k=0}^2 w_k \left(-\frac{1}{4} + \frac{1}{h}(h/2 - ak\Delta t/2) + \frac{3}{h^2}(h/2 - ak\Delta t/2)^2\right) &= \gamma. \end{aligned}$$

(Recall $w_0 = 1/6$, $w_1 = 4/6$ and $w_2 = 1/6$). Hence, we obtain

$$(QL_1)_{ij} = \begin{cases} \alpha, & j = i \\ -\alpha, & j = i-1 \\ 0, & \text{otherwise} \end{cases}, \qquad (QL_2)_{ij} = \begin{cases} \beta, & j = i = 1 \\ \beta - \gamma, & j = i \neq 1 \\ \gamma, & j = i+1 \\ -\beta, & j = i-1 \\ 0, & \text{otherwise} \end{cases}.$$

□

To summarize, we have found that the Active Flux scheme approximating (1.1) can be written

$$\overline{\mathbf{u}}^{n+1} = \overline{\mathbf{u}}^n - \Delta t H^{-1} \left(aQL_1 \overline{\mathbf{u}}^n + aQL_2 \mathbf{u}^n - a \sum_{k=0}^2 w_k g(t^n + k\Delta t/2) e_1 \right). \tag{3.7}$$

Remark. We remind the reader that the point values $\mathbf{u}^n$ must be updated as described in section 2.1. However, using the equation above it is not necessary to calculate and store the intermediate values $\mathbf{u}^{n+1/2}$.

4. Proof of main result

To simplify the analysis of the scheme (3.7), we introduce the CFL number $\lambda = a\Delta t/h$ and the matrix

$$D \in \mathbb{R}^{N \times N} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

Further, we introduce $\mathbf{v}^n$, $\mathbf{w}^n \in \mathbb{R}^N$ defined by

$$v_i^n = u_{i-1/2}^n, \qquad w_i^n = u_{i+1/2}^n, \qquad i = 1, \dots, N.$$

Clearly,

$$QL_1\overline{\mathbf{u}}^n = \alpha D\overline{\mathbf{u}}^n, \qquad QL_2\mathbf{u}^n = \beta D\mathbf{v}^n + \gamma D\mathbf{w}^n,$$

and by (3.6),

$$\mathbf{F} = a\Big(\alpha D\overline{\mathbf{u}}^n + \beta D\mathbf{v}^n + \gamma D\mathbf{w}^n - \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)e_1\Big). \tag{4.1}$$

Now we are ready to prove the following preliminary result.

Proposition 4.1. *The numerical solution of the scheme (3.7) satisfies*

$$\begin{aligned}
\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \left\|\overline{\mathbf{u}}^{n}\right\|_h^2 = {} & h\lambda\alpha(\lambda\alpha - 1)\left\|D\overline{\mathbf{u}}^n\right\|_{\ell^2}^2 + h\lambda^2\beta^2\left\|D\mathbf{v}^n\right\|_{\ell^2}^2 + h\lambda^2\gamma^2\left\|D\mathbf{w}^n\right\|_{\ell^2}^2 \\
& + h\lambda^2\left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)\right)^2 + 2h\lambda\beta(\lambda\alpha - 1)\langle\overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle \\
& + 2h\lambda^2\alpha\beta\langle\overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle + 2h\lambda\gamma(\lambda\alpha - 1)\langle\overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle \\
& + 2h\lambda^2\alpha\gamma\langle\overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle + 2h\lambda^2\beta\gamma\langle\mathbf{v}^n, D^TD\mathbf{w}^n\rangle \\
& - 2h\lambda(\lambda\alpha - 1)\overline{u}_1^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) - 2h\lambda^2\beta u_{1-1/2}^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \\
& - 2h\lambda^2\gamma u_{1+1/2}^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2).
\end{aligned}$$

for all $n \geq 0$.

Proof. Using (2.3),

$$\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \left\|\overline{\mathbf{u}}^{n}\right\|_h^2 = -2\Delta t\langle\overline{\mathbf{u}}^n, \mathbf{F}\rangle + \frac{(\Delta t)^2}{h}\langle\mathbf{F}, \mathbf{F}\rangle.$$

From (4.1) and the equality $h\lambda = a\Delta t$ we find

$$-2\Delta t\langle\overline{\mathbf{u}}^n, \mathbf{F}\rangle = -2h\lambda\alpha\langle\overline{\mathbf{u}}^n, D\overline{\mathbf{u}}^n\rangle - 2h\lambda\beta\langle\overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle - 2h\lambda\gamma\langle\overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle + 2h\lambda\overline{u}_1^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2).$$

Further, using (4.1) and writing $h\lambda^2 = a^2(\Delta t)^2/h$,

$$\begin{aligned}
\frac{(\Delta t)^2}{h}\langle\mathbf{F}, \mathbf{F}\rangle = {} & h\lambda^2\alpha^2\left\|D\overline{\mathbf{u}}^n\right\|_{\ell^2}^2 + h\lambda^2\beta^2\left\|D\mathbf{v}^n\right\|_{\ell^2}^2 + h\lambda^2\gamma^2\left\|D\mathbf{w}^n\right\|_{\ell^2}^2 + h\lambda^2\left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2)\right)^2 \\
& + 2h\lambda^2\alpha\beta\langle\overline{\mathbf{u}}^n, D^TD\mathbf{v}^n\rangle + 2h\lambda^2\alpha\gamma\langle\overline{\mathbf{u}}^n, D^TD\mathbf{w}^n\rangle + 2h\lambda^2\beta\gamma\langle\mathbf{v}^n, D^TD\mathbf{w}^n\rangle \\
& - 2h\lambda^2\alpha\overline{u}_1^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) - 2h\lambda^2\beta u_{1-1/2}^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \\
& - 2h\lambda^2\gamma u_{1+1/2}^n\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2).
\end{aligned}$$

Now, because

$$D^TD = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} = D + D^T,$$

it is not too hard to see that

$$\begin{aligned}\|D\overline{\mathbf{u}}^n\|_{\ell^2}^2 &= \langle D\overline{\mathbf{u}}^n, D\overline{\mathbf{u}}^n\rangle = \langle \overline{\mathbf{u}}^n, D^T D\overline{\mathbf{u}}^n\rangle = 2\langle \overline{\mathbf{u}}^n, D\overline{\mathbf{u}}^n\rangle,\\ \langle \overline{\mathbf{u}}^n, D^T D\mathbf{v}^n\rangle &= \langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle + \langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle,\\ \langle \overline{\mathbf{u}}^n, D^T D\mathbf{w}^n\rangle &= \langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle + \langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle.\end{aligned}$$

Adding everything together yields

$$\begin{aligned}\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \|\overline{\mathbf{u}}^n\|_h^2 = {}& h\lambda\alpha(\lambda\alpha-1)\,\|D\overline{\mathbf{u}}^n\|_{\ell^2}^2 + h\lambda^2\beta^2\,\|D\mathbf{v}^n\|_{\ell^2}^2 + h\lambda^2\gamma^2\,\|D\mathbf{w}^n\|_{\ell^2}^2\\ &+ h\lambda^2\left(\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right)^2 + 2h\lambda\beta(\lambda\alpha-1)\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle\\ &+ 2h\lambda^2\alpha\beta\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle + 2h\lambda\gamma(\lambda\alpha-1)\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle\\ &+ 2h\lambda^2\alpha\gamma\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle + 2h\lambda^2\beta\gamma\langle \mathbf{v}^n, D^T D\mathbf{w}^n\rangle\\ &- 2h\lambda(\lambda\alpha-1)\overline{u}_1^n\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2) - 2h\lambda^2\beta u_{1-1/2}^n\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\\ &- 2h\lambda^2\gamma u_{1+1/2}^n\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2).\end{aligned}$$

□

From proposition 4.1 we see immediately that the scheme is energy stable for $\lambda = 1$, as $(\lambda\alpha-1)$, β and γ have a root at $\lambda = 1$, giving

$$\left(\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \|\overline{\mathbf{u}}^n\|_h^2\right)\Bigg|_{\lambda=1} = a\Delta t\left(\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right)^2 \le a\Delta t\sum_{k=0}^{2} w_k|g(t^n+k\Delta t/2)|^2.$$

It is also trivially stable when $\lambda = 0$. Note that the sign of

$$\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle,\qquad \langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle,\qquad \langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle,\qquad \langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle,\qquad \langle \mathbf{v}^n, D^T D\mathbf{w}^n\rangle,$$

cannot be determined unless we assume the problem data is monotone. Further, we see that $h\lambda\alpha(\lambda\alpha-1)\,\|D\overline{\mathbf{u}}^n\|_{\ell^2}^2 \le 0$ for $\lambda\in[0,1]$. Thus, we obtain the rough estimate

$$\begin{aligned}\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \|\overline{\mathbf{u}}^n\|_h^2 \le {}& h\lambda^2\beta^2\,\|D\mathbf{v}^n\|_{\ell^2}^2 + h\lambda^2\gamma^2\,\|D\mathbf{w}^n\|_{\ell^2}^2 + h\lambda^2\left(\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right)^2\\ &+ |2h\lambda\beta(\lambda\alpha-1)||\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle| + |2h\lambda^2\alpha\beta||\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle|\\ &+ |2h\lambda\gamma(\lambda\alpha-1)||\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle| + |2h\lambda^2\alpha\gamma||\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle|\\ &+ |2h\lambda^2\beta\gamma||\langle \mathbf{v}^n, D^T D\mathbf{w}^n\rangle| + |2h\lambda(\lambda\alpha-1)\overline{u}_1^n|\left|\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right|\\ &+ |2h\lambda^2\beta u_{1-1/2}^n|\left|\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right| + |2h\lambda^2\gamma u_{1+1/2}^n|\left|\sum_{k=0}^{2} w_k g(t^n+k\Delta t/2)\right|.\end{aligned}$$

Next, we want to rewrite the above to the form

$$\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 \le (1+p_1(\lambda))\,\|\overline{\mathbf{u}}^n\|_h^2 + p_2(\lambda)\sum_{k=0}^{2} w_k|g(t^n+k\Delta t/2)|^2.$$

To achieve this we need the following proposition. From here on, we write $\|\cdot\|$ for the induced ℓ^2 matrix norm.

Proposition 4.2. *There exists $c \in \mathbb{R}^+$, independent of Δt, such that*

$$\begin{aligned} \|D\mathbf{v}^n\|_{\ell^2}^2 &\le 4c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,, & \|D\mathbf{w}^n\|_{\ell^2}^2 &\le 4c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,,\\ |\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle| &\le 2c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,, & |\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle| &\le 2c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,,\\ |\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle| &\le 4c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,, & |\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle| &\le 4c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,,\\ |\langle \mathbf{v}^n, D^TD\mathbf{w}^n\rangle| &\le 4c\,\|\overline{\mathbf{u}}^n\|_{\ell^2}^2\,, \end{aligned}$$

and $c \to 1$ as $h \to 0$.

Proof. Using the Cauchy-Schwarz inequality and the definition of $\|\cdot\|$ gives

$$\begin{aligned} \|D\mathbf{v}^n\|_{\ell^2}^2 &\le \|D\|^2\,\|\mathbf{v}^n\|_{\ell^2}^2\,, & \|D\mathbf{w}^n\|_{\ell^2}^2 &\le \|D\|^2\,\|\mathbf{w}^n\|_{\ell^2}^2\,,\\ |\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle| &\le \|D\|\,\|\overline{\mathbf{u}}^n\|_{\ell^2}\,\|\mathbf{v}^n\|_{\ell^2}\,, & |\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle| &\le \|D\|\,\|\overline{\mathbf{u}}^n\|_{\ell^2}\,\|\mathbf{w}^n\|_{\ell^2}\,,\\ |\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle| &\le \left\|D^T\right\|\,\|\overline{\mathbf{u}}^n\|_{\ell^2}\,\|\mathbf{v}^n\|_{\ell^2}\,, & |\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle| &\le \left\|D^T\right\|\,\|\overline{\mathbf{u}}^n\|_{\ell^2}\,\|\mathbf{w}^n\|_{\ell^2}\,,\\ |\langle \mathbf{v}^n, D^TD\mathbf{w}^n\rangle| &\le \left\|D^TD\right\|\,\|\mathbf{v}^n\|_{\ell^2}\,\|\mathbf{w}^n\|_{\ell^2}\,. \end{aligned}$$

Using Gershgorin's theorem and the fact that $\|L\| = \sqrt{\rho(L^*L)}$, where ρ denotes the spectral radius, we observe $\|D\| = \left\|D^T\right\| \le 2$. Moreover,

$$(D^TD)^T(D^TD) = \begin{bmatrix} 5 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ & & & & 1 & -4 & 5 \end{bmatrix}$$

implies $\left\|D^TD\right\| \le 4$. Next, we will show that there exists $C \in \mathbb{R}^+$ such that $\|\mathbf{v}^n\|_{\ell^2} \le C\,\|\overline{\mathbf{u}}^n\|_{\ell^2}$ and $\|\mathbf{w}^n\|_{\ell^2} \le C\,\|\overline{\mathbf{u}}^n\|_{\ell^2}$. Recall from section 2.1 that we constructed a function $R \in \mathcal{C}^0(\overline{\Omega}) \cap \mathcal{C}^\infty(\cup_{i=1}^N C_i)$ satisfying

$$R(x_{i\pm1/2}) = u_{i\pm1/2}^n, \qquad \frac{1}{h}\int_{C_i} R(x)dx = \overline{u}_i^n, \qquad i = 1,\dots,N.$$

By the definition of the supremum, we obtain $|u_{i\pm1/2}^n|^2 \le \sup_{x\in\overline{C_i}} |R(x)|^2$ for $i = 1,\dots,N$, and it follows that

$$\sum_{i=1}^N h|u_{i-1/2}^n|^2 \le U(|R|^2, \{C_i\}_{i=1}^N), \qquad \text{and} \qquad \sum_{i=1}^N h|u_{i+1/2}^n|^2 \le U(|R|^2, \{C_i\}_{i=1}^N),$$

where U is the upper Riemann sum. Since $|R|^2$ is Riemann integrable we obtain $\|\mathbf{v}^n\|_{\ell^2} \le (\epsilon_h/h)^{1/2}\,\|R\|_{L^2(\Omega)}$ and $\|\mathbf{w}^n\|_{\ell^2} \le (\epsilon_h/h)^{1/2}\,\|R\|_{L^2(\Omega)}$, where $1 < \epsilon_h$ and $\epsilon_h \to 1$ as $h \to 0$. Next, we have

$$L(|R|^2, \{C_i\}_{i=1}^N) \le \sum_{i=1}^N h|\overline{u}_i^n|^2 \implies (\delta_h/h)^{1/2}\,\|R\|_{L^2(\Omega)} \le \|\overline{\mathbf{u}}^n\|_{\ell^2}\,,$$

where L is the lower Riemann sum, $0 < \delta_h < 1$ and $\delta_h \to 1$ as $h \to 0$. Choosing $C = (\epsilon_h/\delta_h)^{1/2}$ gives $\|\mathbf{v}^n\|_{\ell^2} \le C\,\|\overline{\mathbf{u}}^n\|_{\ell^2}$ and $\|\mathbf{w}^n\|_{\ell^2} \le C\,\|\overline{\mathbf{u}}^n\|_{\ell^2}$. Combining all our arguments and noting that

$C < C^2$ yields

$$
\begin{aligned}
\|D\mathbf{v}^n\|_{l^2}^2 &\le 4C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , & \|D\mathbf{w}^n\|_{\ell^2}^2 &\le 4C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , \\
|\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n\rangle| &\le 2C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , & |\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n\rangle| &\le 2C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , \\
|\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n\rangle| &\le 4C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , & |\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n\rangle| &\le 4C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 , \\
|\langle \mathbf{v}^n, D^T D\mathbf{w}^n\rangle| &\le 4C^2 \|\overline{\mathbf{u}}^n\|_{\ell^2}^2 .
\end{aligned}
$$

Let $c = C^2$ and observe that $C^2 \to 1$ as $h \to 0$ to complete the proof. □

Next we apply Young's inequality for products to obtain

$$
\begin{aligned}
|2h\lambda(\lambda\alpha - 1)\overline{u}_1^n| \left| \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| &= |2\sqrt{h}(\lambda\alpha - 1)\overline{u}_1^n| \left| \sqrt{h}\lambda \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| \\
&\le \frac{\epsilon_1}{2} 4(\lambda\alpha - 1)^2 c \|\overline{\mathbf{u}}^n\|_h^2 + \frac{h\lambda^2}{2\epsilon_1} \left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right)^2 , \\
|2h\lambda^2 \beta u_{1-1/2}^n| \left| \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| &= |2\sqrt{h}\lambda\beta u_{1-1/2}^n| \left| \sqrt{h}\lambda \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| \\
&\le \frac{\epsilon_2}{2} 4\lambda^2\beta^2 c \|\overline{\mathbf{u}}^n\|_h^2 + \frac{h\lambda^2}{2\epsilon_2} \left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right)^2 , \\
|2h\lambda^2 \gamma u_{1+1/2}^n| \left| \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| &= |2\sqrt{h}\lambda\gamma u_{1+1/2}^n| \left| \sqrt{h}\lambda \sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right| \\
&\le \frac{\epsilon_3}{2} 4\lambda^2\gamma^2 c \|\overline{\mathbf{u}}^n\|_h^2 + \frac{h\lambda^2}{2\epsilon_3} \left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right)^2 .
\end{aligned}
$$

for all ϵ_1, ϵ_2 and $\epsilon_3 \in \mathbb{R}^+$. Using the above, and writing $h \|\overline{\mathbf{u}}^n\|_{\ell^2} = \|\overline{\mathbf{u}}^n\|_h$, we obtain

$$
\begin{aligned}
\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 - \|\overline{\mathbf{u}}^n\|_h^2 \le{}& 4c\lambda^2\beta^2 \|\overline{\mathbf{u}}^n\|_h^2 + 4c\lambda^2\gamma^2 \|\overline{\mathbf{u}}^n\|_h^2 + h\lambda^2 \left(1 + \frac{1}{2\epsilon_1} + \frac{1}{2\epsilon_2} + \frac{1}{2\epsilon_3}\right) \left(\sum_{k=0}^{2} w_k g(t^n + k\Delta t/2) \right)^2 \\
&+ 2c|2\lambda\beta(\lambda\alpha - 1)| \|\overline{\mathbf{u}}^n\|_h^2 + 4c|2\lambda^2\alpha\beta| \|\overline{\mathbf{u}}^n\|_h^2 \\
&+ 2c|2\lambda\gamma(\lambda\alpha - 1)| \|\overline{\mathbf{u}}^n\|_h^2 + 4c|2\lambda^2\alpha\gamma| \|\overline{\mathbf{u}}^n\|_h^2 \\
&+ 4c|2\lambda^2\beta\gamma| \|\overline{\mathbf{u}}^n\|_h^2 + \frac{\epsilon_1}{2} 4(\lambda\alpha - 1)^2 c \|\overline{\mathbf{u}}^n\|_h^2 \\
&+ \frac{\epsilon_2}{2} 4\lambda^2\beta^2 c \|\overline{\mathbf{u}}^n\|_h^2 + \frac{\epsilon_3}{2} 4\lambda^2\gamma^2 c \|\overline{\mathbf{u}}^n\|_h^2 .
\end{aligned}
$$

If we select $\epsilon_1 = \epsilon_2 = \epsilon_3 = \frac{3}{2}\lambda$, and add $\|\overline{\mathbf{u}}^n\|_h^2$, then this becomes

$$
\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 \le (1 + cp_1(\lambda)) \|\overline{\mathbf{u}}^n\|_h^2 + p_2(\lambda) \sum_{k=0}^{2} w_k |g(t^n + k\Delta t/2)|^2,
$$

where

$$
\begin{aligned}
p_1(\lambda) &= 4\lambda^2\beta^2 + 4\lambda^2\gamma^2 + 2|2\lambda\beta(\lambda\alpha - 1)| + 4|2\lambda^2\alpha\beta| + 2|2\lambda\gamma(\lambda\alpha - 1)| \\
&\quad + 4|2\lambda^2\alpha\gamma| + 4|2\lambda^2\beta\gamma| + 3\lambda(\lambda\alpha - 1)^2 + 3\lambda^3\beta^2 + 3\lambda^3\gamma^2 \\
p_2(\lambda) &= h\lambda(1 + \lambda).
\end{aligned}
$$

We have plotted $p_1(\lambda)$ in Figure 1, and we see that it is bounded by 8λ for $\lambda \in [0, 1]$. Noting that

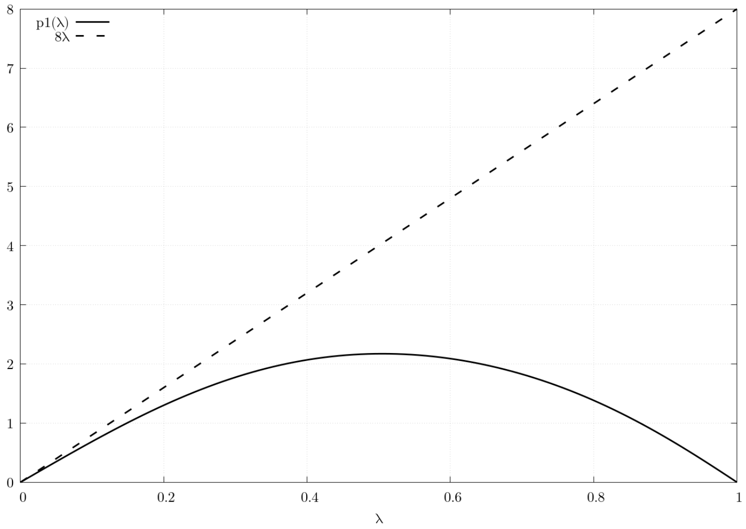

Figure 1. Plot of $p_1(\lambda)$ and $f(\lambda) = 8\lambda$ for $0 \leq \lambda \leq 1$.

$h\lambda = a\Delta t$, and $1 + \lambda \leq 2$ for $\lambda \in [0, 1]$, we obtain

$$\begin{aligned}
\left\|\overline{\mathbf{u}}^{n+1}\right\|_h^2 &\leq (1 + 8c\lambda) \left\|\overline{\mathbf{u}}^{n}\right\|_h^2 + 2a\Delta t \sum_{k=0}^{2} w_k |g(t^n + k\Delta t/2)|^2 \\
&\leq (1 + 8c\lambda)^2 \left\|\overline{\mathbf{u}}^{n-1}\right\|_h^2 + (1 + 8c\lambda) 2a\Delta t \sum_{k=0}^{2} w_k |g(t^{n-1} + k\Delta t/2)|^2 + 2a\Delta t \sum_{k=0}^{2} w_k |g(t^n + k\Delta t/2)|^2 \\
&\leq (1 + 8c\lambda)^{n+1} \left(\left\|\overline{\mathbf{u}}^{0}\right\|_h^2 + 2a\Delta t \sum_{j=1}^{n+1} (1 + 8c\lambda)^{-(j+1)} \sum_{k=0}^{2} w_k |g(t^{j-1} + k\Delta t/2)|^2 \right) \\
&\leq 2e^{8c\lambda(n+1)} \left(\left\|\overline{\mathbf{u}}^{0}\right\|_h^2 + a\Delta t \sum_{j=1}^{n+1} \sum_{k=0}^{2} w_k |g(t^{j-1} + k\Delta t/2)|^2 \right).
\end{aligned}$$

The deduction above is sometimes referred to as a discrete Grönwall lemma ([15]). This proves our main result:

Proposition 4.3 (Main result)**.** *If the CFL number λ satisfies $\lambda \in [0, 1]$, then the numerical solution of the scheme (3.7) satisfies*

$$\left\|\overline{\mathbf{u}}^{n}\right\|_h^2 \leq 2e^{8c\lambda n} \left(\left\|\overline{\mathbf{u}}^{0}\right\|_h^2 + a\Delta t \sum_{j=1}^{n} \sum_{k=0}^{2} w_k |g(t^{j-1} + k\Delta t/2)|^2 \right),$$

for all $n \geq 1$. Here $c \to 2$ as $h \to 0$.

5. Conclusions

We have obtained an energy estimate for the Active Flux scheme approximating the 1D scalar linear advection initial-boundary value problem. Due to the indeterminate sign of the inner products $\langle \overline{\mathbf{u}}^n, D\mathbf{v}^n \rangle$, $\langle \overline{\mathbf{u}}^n, D\mathbf{w}^n \rangle$, $\langle \overline{\mathbf{u}}^n, D^T\mathbf{v}^n \rangle$, $\langle \overline{\mathbf{u}}^n, D^T\mathbf{w}^n \rangle$ and $\langle \mathbf{v}^n, D^T D\mathbf{w}^n \rangle$, we could only give a rough energy estimate. Moreover, we used Gershgorin's theorem to approximate the spectral radius of several linear operators. A manual computation might yield a sharper estimate.

It should be noted that our estimate is only valid for Cartesian partitions of the spatial domain. We suspect a similar result can be obtained for more general partitions.

References

[1] T. Eymann and P. Roe, "Active flux schemes for systems," 06 2011.

[2] ——, "Multidimensional active flux schemes," 06 2013.

[3] W. Barsukow and R. Abgrall, "Extensions of active flux to arbitrary order of accuracy," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 57, 01 2023.

[4] 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

[5] W. Barsukow, J. Hohm, C. Klingenberg, and P. L. Roe, "The active flux scheme on cartesian grids and its low mach number limit," 2019.

[6] D. Fan and P. L. Roe, *Investigations of a New Scheme for Wave Propagation.* [Online]. Available: https://arc.aiaa.org/doi/abs/10.2514/6.2015-2449

[7] W. Barsukow and J. Berberich, "A well-balanced active flux method for the shallow water equations with wetting and drying," *Communications on Applied Mathematics and Computation*, 04 2023.

[8] E. Chudzik, C. Helzel, and D. Kerkmann, "The cartesian grid active flux method: Linear stability and bound preserving limiting," *Applied Mathematics and Computation*, vol. 393, p. 125501, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0096300320304598

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

[10] 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

[11] E. Tadmor, "From semidiscrete to fully discrete: stability of runge-kutta schemes by the energy method. ii," *Collected lectures on the preservation of stability under discretization*, vol. 109, pp. 25–49, 2002.

[12] G. J. Gassner, M. Svärd, and F. J. Hindenlang, "Stability issues of entropy-stable and/or split-form high-order schemes: Analysis of linear stability," *Journal of Scientific Computing*, vol. 90, pp. 1–36, 2022.

[13] W. Barsukow, "The active flux scheme for nonlinear problems," *Journal of Scientific Computing*, vol. 86, 01 2021.

[14] R. Abgrall, W. Barsukow, and C. Klingenberg, "The active flux method for the euler equations on cartesian grids," 2023.

[15] E. Emmrich, "Discrete versions of gronwall's lemma and their application to the numerical analysis of parabolic problems," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:221213518

Trondheim, Norway
Email address: `thomas.bjarne.hestvik@gmail.com`