Analysis of the SBP-SAT Stabilization with Finite Volume: Advection Equation

name

October 12, 2022

1 SBP-SAT Approximation

Consider the following scalar advection equation in 1D in the domain $x \in [0,1]$.

$$\frac{\partial q(x,t)}{\partial t} + c(x)\frac{\partial q(x,t)}{\partial x} = 0 \tag{1}$$

Here, $q(x,t)$ be the quantity of the water advected along the domain and $c(x)$ be the velocity which it advected. Let the initial condition in that domain be $q(x,0) = f(x)$. In this case, we let the water goes in from the left following the equation $q(0,t) = g(t)$ and advected throughout the domain with speed $c(x) = 1$ and let out at $q(1,t)$. Such a problem called Initial Boundary Value Problem (IBVP). We want to solve the equation numerically with semidiscrete method which energy stable.

1.1 Stability of The Model

Introducing the energy method by define the inner product on L_2 as well as the norm. For any smooth function p and q, define the inner product as

$$(p,q) = \int_0^1 p(x)q(x) \ \mathrm{d}x \tag{2}$$

and the L_2-norm as

$$\|q(\cdot,t)\|^2 = (q,q) \tag{3}$$

Later on, we want to mimic this energy method and redefine it in the semidiscrete method to prove the stability.

Before that, we will see how the boundary condition affects stability of our advection equation.

Now consider the advection equation (1) with initial condition $q(x,0) = f(x)$ and boundary condition $q(0,t) = g(t)$.

Now, multiplied by the solution q. We have $(q, q_t + q_x) = 0$ or equivalently

$$(q, q_t) + (q, q_x) = 0 \tag{4}$$

$$\int_0^1 q(x,t) \frac{\partial q(x,t)}{\partial t} \, dx + \int_0^1 q(x,t) \frac{\partial q(x,t)}{\partial x} \, dx = 0 \tag{5}$$

With the same argument, we will have

$$(q_t, q) + (q_x, q) = \int_0^1 \frac{\partial q(x,t)}{\partial t} q(x,t) \, dx + \int_0^1 \frac{\partial q(x,t)}{\partial x} q(x,t) \, dx = 0 \tag{6}$$

Adding the two, we will have

$$\int_0^1 \left(q(x,t) \frac{\partial q(x,t)}{\partial t} + \frac{\partial q(x,t)}{\partial t} q(x,t) \right) \, dx + 2 \int_0^1 \frac{\partial q(x,t)}{\partial x} q(x,t) \, dx = 0 \tag{7}$$

Integrated by part on the first integral gives us

$$\int_0^1 \left(q(x,t) \frac{\partial q(x,t)}{\partial t} + \frac{\partial q(x,t)}{\partial t} q(x,t) \right) \, dx = \int_0^1 \frac{\partial q(x,t)^2}{\partial t} \, dx = \frac{\partial}{\partial t} \int_0^1 q(x,t)^2 \, dx = \frac{\partial}{\partial t} \|q(x,t)\|. \tag{8}$$

Meanwhile, we also have

$$\int_0^1 2\frac{\partial q(x,t)}{\partial x} q(x,t) \, dx = \int_0^1 2 \, q(x,t)\partial q(x,t) = q(x,t)^2 \mid_0^1 = q(1,t)^2 - q(0,t)^2 \tag{9}$$

Now we have

$$\frac{\partial}{\partial t} \int_0^1 q(x,t)^2 \, dx = -q(1,t)^2 + q(0,t)^2 = -q(1,t)^2 + g(t)^2 \tag{10}$$

Thus, $\frac{\partial}{\partial t} \|u(\cdot, t)\| = -q(1,t)^2 + g(t)^2$. Since $q(1,t)^2$ always positive, then we could see that the energy bounded above by $g(t)^2$, so it is stable.

1.2 Semidiscrete Model

Introduce integration by part property with the product notation. Noting that

$$((\partial p/\partial x), q) = \int_0^1 \frac{\partial p}{\partial x} q \, dx = -\int_0^1 \frac{\partial q}{\partial x} p \, dx + p(0)q(0) - p(1)q(1) \tag{11}$$

$$= -(p, (\partial q/\partial x)) + p(1)q(1) - p(0)q(0) \tag{12}$$

2

or equivalently

$$((\partial p / \partial x), q) + (p, (\partial q / \partial x) = p(1)q(1) - p(0)q(0) \tag{13}$$

Definition 1.1. Let N be a natural number and $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ be a discretization of the interval that p and q are defined. Let $\mathbf{p}$ and $\mathbf{q}$ are the discretization of p and q on $\mathbf{x}$ respectively. A discrete operator D said to be have summation by part properties if there exist discrete inner product $< \cdot, \cdot >_H$ along with the symmetric matrix H such that

$$< D\mathbf{p}, \mathbf{q} > + < \mathbf{p}, D\mathbf{q} > = \mathbf{p}(x_N)\mathbf{q}(x_N) - \mathbf{p}(x_0)\mathbf{q}(x_0) \tag{14}$$

with H be positive semi definite matrices.

In order to approximate the time derivative numerically, we will use derivative operator in the form $D = H^{-1}Q$ with D being operator that derived from Finite Volume Method.

Consider more general equation than (1)

$$\frac{\partial q}{\partial t} + \frac{\partial f}{\partial x} = 0 \tag{15}$$

defined on the interval $[0, 1]$ for some operator f with initial and boundary condition $q(x, 0) = k(x)$ and $q(0, t) = g(t)$. for some operator f.

Consider Finite Volume Method on identical grid on the interval $[0, 1]$. Before continuing, we'll discretize the interval into some grids. Let N be a positive integer and I_i be the i-th control volume for all $i = 1, 2, ..., N$.. Define $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ be the length of the i-th interval and $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ are the left boundary and right boundary of the interval respectively.

From (15), we have

$$\int_{I_i} \frac{\partial q}{\partial t}(x, t) \, \mathrm{d}x + \int_{I_i} \frac{\partial f}{\partial x}(x, t) \, \mathrm{d}x = 0 \tag{16}$$

$$\frac{d}{dt} \int_{I_i} q(x, t) \, \mathrm{d}x + \left(f(x_{i+\frac{1}{2}}, t) - f(x_{i-\frac{1}{2}}, t) \right) = 0 \tag{17}$$

Now, introducing the following notation.

$$\bar{q}_i = \frac{1}{\Delta x_i} \int_{I_i} q(x, t) \, \mathrm{d}x \tag{18}$$

$$f_{i\pm\frac{1}{2}} = f(x_{i\pm\frac{1}{2}}, t) \tag{19}$$

3

Using the new notation and rearrange terms, the last equation becomes

$$\Delta x_i \frac{d\bar{q}_i}{dt} + \left(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}} \right) = 0 \tag{20}$$

Now what's left is to approximate the fluxes $f_{i\pm\frac{1}{2}}$ with numerical flux $F_{i\pm\frac{1}{2}}$

$$\Delta x_i \frac{d\bar{q}_i}{dt} + \left(F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} \right) = 0 \tag{21}$$

The local flux by the inflow and outflow by the average flux on the neighbouring cell, and adding the dissipative energy, we would get

$$F_{i+\frac{1}{2}} = \frac{f(\bar{q}_{i+1}) + f(\bar{q}_i)}{2} - \frac{\alpha}{2} (\bar{q}_{i+1} - \bar{q}_i) \tag{22}$$

with some constant $\alpha > 0$. Calculating $\left(F_{i-\frac{1}{2}} - F_{i+\frac{1}{2}} \right)$ we get

$$\left(F_{i-\frac{1}{2}} - F_{i+\frac{1}{2}} \right) = \left(\frac{f(\bar{q}_{i+1}) + f(\bar{q}_i)}{2} - \frac{\alpha}{2} (\bar{q}_{i+1} - \bar{q}_i) \right) - \left(\frac{f(\bar{q}_i) + f(\bar{q}_{i-1})}{2} - \frac{\alpha}{2} (\bar{q}_i - \bar{q}_{i-1}) \right) \tag{23}$$

$$= \frac{f(\bar{q}_{i+1}) - f(\bar{q}_{i-1})}{2} - \frac{\alpha (\bar{q}_{i+1} - 2\bar{q}_i + \bar{q}_{i-1})}{2} \tag{24}$$

Put this approximation into (21), we got

$$\Delta x \frac{d\bar{q}_i}{dt} + \frac{f(\bar{q}_{i+1}) - f(\bar{q}_{i-1})}{2} - \frac{\alpha}{2} (\bar{q}_{i+1} - 2\bar{q}_i + \bar{q}_{i-1}) = 0 \tag{25}$$

4

Meanwhile, in the left and right boundary, we don't have any cell on it's left and right, respectively. So we could say that the length of the leftmost (rightmost) cell is cut in a half. That means, instead of divide by Δx, we divide by $\Delta x/2$. That is

$$\frac{\Delta x}{2}\frac{d\bar{q}_1}{dt} + \left(F_{\frac{1}{2}} - F_0\right) = 0 \tag{26}$$

$$\frac{\Delta x}{2}\frac{d\bar{q}_N}{dt} + \left(F_N - F_{N-\frac{1}{2}}\right) = 0 \tag{27}$$

The fact that we don't have whole cell on the left and the right force us to set $F_0 = f(\bar{q}_0)$ and $F_N = f(\bar{q}_N)$ and (22) to (26) we have

$$\frac{\Delta x}{2}\frac{d\bar{q}_1}{dt} + \left(F_{\frac{1}{2}} - F_0\right) = \frac{\Delta x}{2}\frac{d\bar{q}_1}{dt} + \left(\frac{\bar{q}_1 + \bar{q}_0}{2} - \frac{\alpha(\bar{q}_1 - \bar{q}_0))}{2}\right) - \bar{q}_0 \tag{28}$$

$$= \frac{\Delta x}{2}\frac{d\bar{q}_1}{dt} + \frac{\bar{q}_1 - \bar{q}_0}{2} - \frac{\alpha(\bar{q}_1 - \bar{q}_0))}{2} \tag{29}$$

and with the same manner we also have the following.

$$\frac{\Delta x}{2}\frac{d\bar{q}_N}{dt} + \left(F_N - F_{N-\frac{1}{2}}\right) = \frac{\Delta x}{2}\frac{d\bar{q}_N}{dt} + \bar{q}_N - \left(\frac{\bar{q}_N + \bar{q}_{N-1}}{2} - \frac{\alpha(\bar{q}_N - \bar{q}_{N-1}))}{2}\right) \tag{30}$$

$$= \frac{\Delta x}{2}\frac{d\bar{q}_N}{dt} + \frac{\bar{q}_N - \bar{q}_{N-1}}{2} + \frac{\alpha(\bar{q}_N - \bar{q}_{N-1}))}{2} \tag{31}$$

Hence, in the boundary we'll have

$$\frac{\Delta x}{2}\frac{d\bar{q}_0}{dt} + \frac{f(\bar{q}_1) - f(\bar{q}_0)}{2} - \frac{\alpha}{2}(\bar{q}_0 - \bar{q}_1) = 0 \tag{32}$$

$$\frac{\Delta x}{2}\frac{d\bar{q}_N}{dt} + \frac{f(\bar{q}_N) - f(\bar{q}_{N-1})}{2} + \frac{\alpha}{2}(\bar{q}_N - \bar{q}_{N-1}) = 0 \tag{33}$$

Putting together equation (32),(25), and (33) respectively, we'll have the following system of equation.

$$\begin{aligned}
\frac{\Delta x_0}{2}\frac{d\bar{q}_0}{dt} + \frac{f(\bar{q}_1) - f(\bar{q}_0)}{2} + \frac{\alpha}{2}(\bar{q}_0 - \bar{q}_1) &= 0 \\
\Delta x_i \frac{d\bar{q}_i}{dt} + \frac{f(\bar{q}_{i+1}) - f(\bar{q}_{i-1})}{2} + \frac{\alpha}{2} - (\bar{q}_{i+1} + 2\bar{q}_i - \bar{q}_{i-1}) &= 0, \text{ for } i = 1, 2, \ldots, N-1 \\
\frac{\Delta x_N}{2}\frac{d\bar{q}_N}{dt} + \frac{f(\bar{q}_N) - f(\bar{q}_{N-1})}{2} + \frac{\alpha}{2}(\bar{q}_N - \bar{q}_{N-1}) &= 0
\end{aligned} \tag{34}$$

Introducing some notation $\mathbf{Q}, \mathbf{A}$, and $\mathbf{H}$ as follows:

$$\mathbf{Q} = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \cdots & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & \frac{1}{2} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \cdots & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{35}$$

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \tag{36}$$

and $\mathbf{H} = \mathrm{diag}(\frac{1}{2}\Delta x_0, \Delta x_1, \Delta x_2, \cdots, \Delta x_{N-1}, \frac{1}{2}\Delta x_N)$. Also notating the vector $\mathbf{q} = (q_0, q_1, \ldots, q_N)^\top$. We could write our equation as

$$\mathbf{H}\frac{d\bar{\mathbf{q}}}{dt} + \mathbf{Q}\mathbf{f}(\bar{\mathbf{q}}) + \alpha\mathbf{A}\bar{\mathbf{q}} = 0 \tag{37}$$

or if we write $\mathbf{D} = \mathbf{H}^{-1}\mathbf{Q}$

$$\frac{d\bar{\mathbf{q}}}{dt} + \mathbf{D}\mathbf{f}(\bar{\mathbf{q}}) + \alpha\mathbf{H}^{-1}\mathbf{A}\bar{\mathbf{q}} = 0 \tag{38}$$

We could see easily see that both $\mathbf{A}$ and $\mathbf{H}^{-1}\mathbf{A}$ are symmetric. Let the differential operator in the discrete level estimated by our operator $\mathbf{D}$.

$$\frac{d\mathbf{q}}{dx} \approx \mathbf{D}_x\mathbf{q} \tag{39}$$

38 In the next section, we will prove that our operator $\mathbf{D}$ satisfy SBP property from Definition 1.1.
39 Now, we're closing this section with implementing the boundary condition with penalty parameter τ
40 that goes as follows.

41
$$\frac{d\bar{\mathbf{q}}}{dt} + \mathbf{D}\mathbf{f}(\bar{\mathbf{q}}) + \alpha\mathbf{H}^{-1}\mathbf{A}\bar{\mathbf{q}} + \tau\mathbf{H}^{-1}\mathbf{e}_0(\bar{q}_0 - g(t)) = 0 \tag{40}$$

42 The last terms is used since the boundary condition imposed at the left boundary. that's why we

43 multiply the condition $(\bar{q}_0 - g(t))$ with $H^{-1}\mathbf{e}_0$.

44 1.3 Stability

45 To ensure the stability, we will check it with energy methods and adjust the penalty parameter τ. Now,

46 we well introduce a discrete inner product that related with the operator from previous subsection.

47 **Definition 1.2.** The inner product $\langle \cdot, \cdot \rangle_H$ defined by

$$48 \qquad \langle \mathbf{p}, \mathbf{q} \rangle_H = \mathbf{q}^\top H \mathbf{p}$$

49 with $H = \mathrm{diag}(\frac{1}{2}, 1, 1, \ldots, 1, \frac{1}{2})$. The norm based in the inner product defined by

$$50 \qquad \|\mathbf{q}\|^2 = \langle \mathbf{q}, \mathbf{q} \rangle_H.$$

We can easily prove that this is actually an inner product. Moreover, with this inner product, our operator D becomes an SBP operator since

$$\langle \mathbf{Dp}, \mathbf{q} \rangle + \langle \mathbf{p}, \mathbf{Dq} \rangle = \mathbf{q}^\top \mathbf{H}(\mathbf{Dp}) + (\mathbf{Dq})^\top \mathbf{Hp} \tag{41}$$

$$= \mathbf{q}^\top \mathbf{H}(\mathbf{H}^{-1}\mathbf{Q})\mathbf{p} + \mathbf{q}^\top (\mathbf{H}^{-1}\mathbf{Q})^\top \mathbf{Hp} \tag{42}$$

$$= \mathbf{q}^\top \mathbf{Q}\mathbf{p} + \mathbf{q}^\top \mathbf{Q}^\top \mathbf{p} \tag{43}$$

$$= \mathbf{q}^\top (\mathbf{Q} + \mathbf{Q}^\top)\mathbf{p} \tag{44}$$

$$= \mathbf{q}^\top (\mathbf{e}_N \mathbf{e}_N^\top - \mathbf{e}_0 \mathbf{e}_0^\top)\mathbf{p} \tag{45}$$

$$= \mathbf{q}_N \mathbf{p}_N - \mathbf{q}_0 \mathbf{p}_0 \tag{46}$$

51 Hence, by Definition (1.1), D is an SBP operator.

Now, moving to our expression (40), we will make sure that our numerical solution is stable. With abuse of notation, multiply expression (40) with $\bar{\mathbf{q}}$ from the left, we get

$$\left\langle \bar{\mathbf{q}}, \frac{d\bar{\mathbf{q}}}{dt} + \mathbf{D}\bar{\mathbf{q}} + \alpha \mathbf{H}^{-1} A\bar{\mathbf{q}} + \tau \mathbf{H}^{-1}\mathbf{e}_0(A\bar{\mathbf{q}} - g(t)) \right\rangle_{\mathbf{H}} = 0 \tag{47}$$

$$\left\langle \bar{\mathbf{q}}, \frac{d\bar{\mathbf{q}}}{dt} \right\rangle_{\mathbf{H}} + \langle \bar{\mathbf{q}}, \mathbf{D}\bar{\mathbf{q}} \rangle_{\mathbf{H}} + \langle \bar{\mathbf{q}}, \alpha \mathbf{H}^{-1}\mathbf{A}\bar{\mathbf{q}} \rangle_{\mathbf{H}} + \langle \bar{\mathbf{q}}, \tau \mathbf{H}^{-1}\mathbf{e}_0(\bar{q}_0 - g(t)) \rangle_{\mathbf{H}} = 0 \tag{48}$$

$$\frac{1}{2}\frac{d}{dt}\langle \bar{\mathbf{q}}, \bar{\mathbf{q}} \rangle_{\mathbf{H}} + \bar{\mathbf{q}}^\top \mathbf{D}^\top \mathbf{H}\bar{\mathbf{q}} + \alpha(\mathbf{H}^{-1}\mathbf{A}\bar{\mathbf{q}})^\top \mathbf{H}\bar{\mathbf{q}} + \tau(\mathbf{H}^{-1}\mathbf{e}_0(\bar{q}_0 - g(t)))^\top \mathbf{H}\bar{\mathbf{q}} = 0 \tag{49}$$

$$\frac{1}{2}\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + \bar{\mathbf{q}}^\top \mathbf{Q}^\top \bar{\mathbf{q}} + \alpha \bar{\mathbf{q}}^\top \mathbf{A}^\top \bar{\mathbf{q}} + \tau \bar{q}_0(\bar{q}_0 - g(t)) = 0 \tag{50}$$

52 The third row to fourth row use the following identity, $\mathbf{D} = \mathbf{H}^{-1}\mathbf{Q}$. If we multiply the equation from

7

the right, with the same way we will get

$$\frac{1}{2}\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + \bar{\mathbf{q}}^\top \mathbf{Q}\bar{\mathbf{q}} + \alpha\bar{\mathbf{q}}^\top \mathbf{A}\bar{\mathbf{q}} + \tau\bar{q}_0(\bar{q}_0 - g(t)) = 0$$

Adding both of them and remember that $\mathbf{A} = \mathbf{A}^\top$, we could get

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + \bar{\mathbf{q}}^\top(\mathbf{Q}^\top + \mathbf{Q})\bar{\mathbf{q}} + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + 2\tau\bar{q}_0(\bar{q}_0 - g(t)) = 0 \qquad (51)$$

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + (\bar{q}_N^2 - \bar{q}_0^2) + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + 2\tau\bar{q}_0(\bar{q}_0 - g(t)) = 0 \qquad (52)$$

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + \bar{q}_N^2 + (2\tau - 1)\bar{q}_0^2 - 2\tau\bar{q}_0 g(t) = 0 \qquad (53)$$

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + \bar{q}_N^2 + (2\tau - 1)\left(\bar{q}_0^2 - 2\frac{\tau}{2\tau - 1}\bar{q}_0 g(t)\right) = 0 \qquad (54)$$

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + \bar{q}_N^2 + (2\tau - 1)\left(\bar{q}_0^2 - 2\bar{q}_0\frac{\tau g(t)}{2\tau - 1} + \left(\frac{\tau g(t)}{2\tau - 1}\right)^2 - \left(\frac{\tau g(t)}{2\tau - 1}\right)^2\right) = 0 \qquad (55)$$

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 + 2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} + \bar{q}_N^2 + (2\tau - 1)\left(\bar{q}_0^2 - \left(\frac{\tau g(t)}{2\tau - 1}\right)\right)^2 + \left(-\frac{\tau^2 g(t)^2}{2\tau - 1}\right) = 0 \qquad (56)$$

The equation (52) came from the fact that $\mathbf{Q} + \mathbf{Q}^\top$ entries all zeros except the first entry of the diagonal and the last entry of the diagonal. Remember that A is a positive semi-definite matrix. If we could bound $2\tau - 1 > 0$, then $-\frac{1}{2\tau - 1} < 0$ and both last two terms in (56) are positive. That means

$$\frac{d}{dt}\|\bar{\mathbf{q}}\|^2 \leq -2\alpha\bar{\mathbf{q}}^\top \mathbf{A}^\top\bar{\mathbf{q}} - \bar{q}_N^2 - (2\tau - 1)\left(\bar{q}_0^2 - \left(\frac{\tau g(t)}{2\tau - 1}\right)\right)^2 + \frac{\tau^2 g(t)^2}{2\tau - 1} \leq \frac{\tau^2 g(t)^2}{2\tau - 1}. \qquad (57)$$

or if we had $\tau > \frac{1}{2}$ we could safely say that our numerical solution $\bar{\mathbf{q}}$ is stable.

2 Numerical Experiment

8

3 Shallow Water Equation

In this section we want to use SBP-SAT with finite volume to model the linearized one dimensional shallow water equation using the same method as the advection equation in the previous section.

3.1 Linearization

Let h be depth of the water and u be x-directional velocity. The one dimensional shallow water equation is the following system of equation.

$$\frac{\partial h}{\partial t} + \frac{\partial uh}{\partial x} = 0 \tag{58}$$

$$\frac{\partial (uh)}{\partial t} + \frac{\partial (u^2 h + \frac{1}{2}gh^2)}{\partial x} = 0 \tag{59}$$

We'll linearized this equation by letting $h = H + h'$ and $u = U + u'$ for some constant H and U and small perturbation h' and u'.

Now, putting it into the system that we have, we got the following,

$$\frac{\partial}{\partial t}(H + h') + \frac{\partial}{\partial x}\left((U + u')(H + h')\right) = 0 \tag{60}$$

$$\frac{\partial}{\partial t}\left((U + u')(H + h')\right) + \frac{\partial}{\partial x}\left((U + u')^2(H + h') + \frac{1}{2}g(H + h')\right) = 0 \tag{61}$$

Note that H and U are constants, we got

$$\frac{\partial h'}{\partial t} + H\frac{\partial u'}{\partial x} + U\frac{\partial h'}{\partial x} + \frac{\partial u'h'}{\partial x} = 0 \tag{62}$$

$$H\frac{\partial u'}{\partial t} + U\frac{\partial h'}{\partial t} + \frac{\partial u'h'}{\partial t} + 2UH\frac{\partial u'}{\partial x} + H\frac{\partial u'^2}{\partial x} + U^2\frac{\partial h'}{\partial x} + 2U\frac{\partial u'h'}{\partial x} + \frac{\partial u'^2 h}{\partial x} +$$

$$\frac{\partial \left(\frac{1}{2}gH^2\right)}{\partial x} + gH\frac{\partial h'}{\partial x} + \frac{1}{2}g\frac{\partial h'^2}{\partial x} = 0 \tag{63}$$

Ignoring the higher terms $\frac{\partial u'h'}{\partial t}, \frac{\partial u'h'}{\partial x}, \frac{\partial u'^2}{\partial x}, \frac{\partial u'^2 h'}{\partial x}$, we get

$$\frac{\partial h'}{\partial t} + H\frac{\partial u'}{\partial x} + U\frac{\partial h'}{\partial x} = 0 \tag{64}$$

$$H\frac{\partial u'}{\partial t} + U\frac{\partial h'}{\partial t} + 2UH\frac{\partial u'}{\partial x} + U^2\frac{\partial h'}{\partial x} + gH\frac{\partial h'}{\partial x} = 0 \tag{65}$$

Substituting (64) into (65) to eliminate $\frac{\partial h'}{\partial t}$, it becomes

$$H\frac{\partial u'}{\partial t} + U\left(-H\frac{\partial u'}{\partial x} - U\frac{\partial h'}{\partial x}\right) + 2UH\frac{\partial u'}{\partial x} + U^2\frac{\partial h'}{\partial x} + gH\frac{\partial h'}{\partial x} = 0 \tag{66}$$

9

or equivalently

$$H\frac{\partial u'}{\partial t} + UH\frac{\partial u'}{\partial x} + gH\frac{\partial h'}{\partial x} = 0 \tag{67}$$

Canceling out the constant H from the equation, we got the following Shallow Water equation linearized system.

$$\frac{\partial h'}{\partial t} + H\frac{\partial u'}{\partial x} + U\frac{\partial h'}{\partial x} = 0 \tag{68}$$

$$\frac{\partial u'}{\partial t} + U\frac{\partial u'}{\partial x} + g\frac{\partial h'}{\partial x} = 0 \tag{69}$$

or with change of notation, we could rewrite it to the following systems.

$$\frac{\partial h}{\partial t} + U\frac{\partial h}{\partial x} + H\frac{\partial u}{\partial x} = 0 \tag{70}$$

$$\frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} + U\frac{\partial u}{\partial x} = 0 \tag{71}$$

3.2   Analysis

In this section we would like to analyse the stability with energy method to get our boundary condition.

Multiply equation (70) with gh, and integrate with respect to x in the inteval $[0,1]$, we got

$$gh\frac{\partial h}{\partial t} + gUh\frac{\partial h}{\partial x} + gHh\frac{\partial u}{\partial x} = 0 \tag{72}$$

$$\int_0^1 gh\frac{\partial h}{\partial t} + gUh\frac{\partial h}{\partial x} + gHh\frac{\partial u}{\partial x} + gUh\frac{\partial h}{\partial x} = 0 \tag{73}$$

$$\int_0^1 gh\frac{\partial h}{\partial t}\ \mathrm{d}x + \int_0^1 gUh\frac{\partial h}{\partial x}\ \mathrm{d}x + \int_0^1 gHh\frac{\partial u}{\partial x}\ \mathrm{d}x = 0 \tag{74}$$

$$\frac{d}{dt}\int_0^1 \frac{1}{2}gh^2\ \mathrm{d}x + gU\left(\frac{1}{2}h^2\big|_0^1\right) + gH\left(uh\big|_0^1 - \int_0^1 u\frac{\partial h}{\partial x}\ \mathrm{d}x\right) = 0 \tag{75}$$

Multiply equation (71) with uH and integrate with respect to x in in the interval $[0,1]$, we got

$$uH\frac{\partial u}{\partial t} + uHg\frac{\partial h}{\partial x} + uHU\frac{\partial u}{\partial x} = 0 \tag{76}$$

$$\int_0^1 uH\frac{\partial u}{\partial t} + uHg\frac{\partial h}{\partial x} + uHU\frac{\partial u}{\partial x} \; dx = 0 \tag{77}$$

$$\int_0^1 uH\frac{\partial u}{\partial t} \; dx + \int_0^1 uHg\frac{\partial h}{\partial x} \; dx + \int_0^1 uHU\frac{\partial u}{\partial x} \; dx = 0 \tag{78}$$

$$\frac{d}{dt}\int_0^1 \frac{1}{2}Hu^2 \; dx + gH\int_0^1 u\frac{\partial h}{\partial x} \; dx + HU\int_0^1 u\frac{\partial u}{\partial x} \; dx \tag{79}$$

$$\frac{d}{dt}\int_0^1 \frac{1}{2}Hu^2 \; dx + gH\int_0^1 u\frac{\partial h}{\partial x} \; dx + HU\left(\frac{1}{2}u^2\big|_0^1\right) = 0 \tag{80}$$

Adding (75) with (80) will cancel out terms $gH\int_0^1 u\frac{\partial h}{\partial x} \; dx$ and we got

$$\frac{d}{dt}\int_0^1 \frac{1}{2}\left(gh^2 + Hu^2\right)dx + gU\left(\frac{1}{2}h^2\big|_0^1\right) + gH\left(uh\big|_0^1\right) + HU\left(\frac{1}{2}u^2\big|_0^1\right) = 0 \tag{81}$$

Here, we could see with energy method, that this system of equation is stable.

Now, we want to know how many boundary condition needed in this linearized shallow water equation.

Let's rewrite equation (81) by letting $E = \int_0^1 \frac{1}{2}\left(gh^2 + Hu^2\right)dx$ be the energy and $\mathbf{q} = [h \quad u]^\top$. The equation becomes

$$\frac{dE}{dt} + gU\left(\frac{1}{2}h^2\big|_0^1\right) + gH\left(uh\big|_0^1\right) + HU\left(\frac{1}{2}u^2\big|_0^1\right) = 0 \tag{82}$$

With some algebraic manipulation, we could get the equation equivalent to

$$\frac{dE}{dt} + \begin{bmatrix} h & u \end{bmatrix}\begin{bmatrix} \frac{1}{2}gU & \frac{1}{2}gH \\ \frac{1}{2}gH & \frac{1}{2}Hu \end{bmatrix}\begin{bmatrix} h \\ u \end{bmatrix}\Bigg|_0^1 = 0 \tag{83}$$

or

$$\frac{dE}{dt} + \frac{1}{2}\mathbf{q}^\top\begin{bmatrix} gU & gH \\ gH & HU \end{bmatrix}\mathbf{q}\Bigg|_0^1 = 0 \tag{84}$$

We want to see the eigenvector of the symmetric matrix $M = \begin{bmatrix} gU & gH \\ gH & HU \end{bmatrix}$ to get the boundary

11

condition right. Calculating $\det(M - \lambda I) = 0$ we get

$$\det(M - \lambda I) = \begin{vmatrix} gU - \lambda & gH \\ gH & HU - \lambda \end{vmatrix} = 0$$

$$(gU - \lambda)(HU - \lambda) - g^2 H^2 = 0$$

$$\lambda^2 - \lambda(gU + HU) + gHU^2 - g^2 H^2$$

$$\lambda_{1,2} = \frac{1}{2}(gU + HU) \pm \frac{1}{2}\sqrt{(gU + HU)^2 - 4gH(U^2 - gH)} \tag{85}$$

We could rewrite the expression inside the root square as we need. Two that useful would be

$$(gU + HU)^2 - 4gH(U^2 - gH) = (2gH)^2 + U^2(g - H)^2$$

to make sure that the eigenvalue are reals, and the other one is

$$(gU + HU)^2 - 4gH(U^2 - gH) = U^2(g + H)^2 + 4gH(gH - U^2)$$

to see the sign of one of the eigenvalue. More on the second one, if we let $U > 0$ and $gH - U^2 > 0$, then we have the following:

$$\lambda_2 = \frac{1}{2}U(g + H) - \frac{1}{2}\sqrt{U^2(g + H)^2 + 4gH(gH - U^2)} < \frac{1}{2}U(g + H) - \frac{1}{2}\sqrt{U^2(g + H)^2} < 0 \tag{86}$$

that means $\lambda_2 < 0$. Conversely, if we have $gH - U^2 < 0$, we will have $\lambda_2 > 0$. But this terms doesn't affect the first eigenvalue since both of the terms in the expression are positive so it would be always positive.

$$\frac{1}{2}U(g + H) + \frac{1}{2}\sqrt{U^2(g + H)^2 + 4gH(gH - U^2)}$$

Of course, this would be change if we consider the other case, that is when $U < 0$. Our λ_1 would be positive or negative depends on the sign of $gh - U^2$ since $\frac{1}{2}U(g + H)$ is negative.

$$\frac{1}{2}U(g + H) + \frac{1}{2}\sqrt{U^2(g + H)^2 + 4gH(gH - U^2)}$$

If $gh - U^2 > 0$ then the expression is positive, otherwise it's negative. Meanwhile, our λ_2 would always negative.

Now we have the following tables of sign

Now, lets back to the equation (84). For simplicity, we will take the factor gH to get

U	$gH - U^2$	λ_1	λ_2
+	+	+	-
+	-	+	+
-	+	+	-
-	-	-	-

Table 1: Eigenvalues sign

$$\frac{dE}{dt} + \frac{1}{2}gH\,\mathbf{q}^\top \begin{bmatrix} U/H & 1 \\ 1 & U/g \end{bmatrix} \mathbf{q}\,\Bigg|_0^1 = 0 \tag{87}$$

Addressing the factor, using the previous eigen analysis, we could see that this new matrices

$$A' = \begin{bmatrix} U/H & 1 \\ 1 & U/g \end{bmatrix} \tag{88}$$

has eigenvalues $\frac{1}{2gH}\left(U(g+H) \pm \sqrt{U^2(g+H)^2 + 4gH(gH - U^2)}\right)$ and would still have the same table of sign as Table 1.

Now, we would like to diagonalize matrix A' from (88) to get the boundary treatments. With the identity that we get from polynomial characteristic of matrix A', that is $\lambda_1 + \lambda_2 = \frac{U}{g} + \frac{U}{H}$, one could find that the vector

$$\begin{pmatrix} \lambda_1 - \frac{U}{g} \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} \lambda_2 - \frac{U}{g} \\ 1 \end{pmatrix} \tag{89}$$

are the eigenvectors of λ_1 and λ_2 respectively. Since the matrix A' are symmetric, this eigenvectors should be orthogonal to each other. But it also could came from the fact that

$$\left(\lambda_1 - \frac{U}{g}\right)\left(\lambda_2 - \frac{U}{g}\right) + 1 = \lambda_1\lambda_2 - (\lambda_1 + \lambda_2)\frac{U}{g} + \frac{U^2}{g^2} + 1 \tag{90}$$

$$= \left(\frac{U}{g}\frac{U}{H} - 1\right) - \left(\frac{U}{g} + \frac{U}{H}\right)\frac{U}{g} + \frac{U^2}{g^2} + 1 \tag{91}$$

$$= \left(\frac{U^2}{gH} - 1\right) - \left(\frac{U^2}{g^2} + \frac{U^2}{gH}\right) + \frac{U^2}{g^2} + 1 \tag{92}$$

$$= 0. \tag{93}$$

The fact that $\lambda_1\lambda_2 = \left(\frac{U}{g}\frac{U}{H} - 1\right)$ also came from the characteristic polynomial of A'.

Now, since the the eigenvectors are orthogonal, hence we could construct an orhtonormal matrix that diagonalize our matrix A' by normalizing the eigenvectors. Let S be the matrix we mentioned,

13

we have

$$S = \begin{bmatrix} \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right) & \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right) \\ \frac{1}{c} & \frac{1}{d} \end{bmatrix} \tag{94}$$

with $c = \sqrt{\left(\lambda_1 - \frac{U}{g}\right)^2 + 1}$ and $d = \sqrt{\left(\lambda_2 - \frac{U}{g}\right)^2 + 1}$, that is the length of our eigenvectors.

Note that S is orhtonormal, we have $S^{-1} = S^\top$. Therefore, our equation (87) becomes

$$\frac{dE}{dt} + \frac{1}{2}gH\,\mathbf{q}^\top S \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} S^\top \mathbf{q}\,\Bigg|_0^1 = 0 \tag{95}$$

or equivalently

$$\frac{dE}{dt} + \frac{1}{2}gH\,(S^\top \mathbf{q})^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} S^\top \mathbf{q}\,\Bigg|_0^1 = 0 \tag{96}$$

Notate $\mathbf{w} = S^\top \mathbf{q}$ and let $\mathbf{w} = [w_1\ w_2]^\top$, we have

$$w_1 = \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)h + \frac{1}{c}u, \quad w_2 = \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)h + \frac{1}{d}u \tag{97}$$

$$h = \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)w_1 + \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)w_2, \quad u = \frac{1}{c}w_1 + \frac{1}{d}w_2 \tag{98}$$

so that

$$\frac{dE}{dt} + \frac{1}{2}gH\,\mathbf{w}^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{w}\,\Bigg|_0^1 = 0 \tag{99}$$

Introducing

$$\mathbf{BT} = -\,\mathbf{w}^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{w}\,\Bigg|_0^1 \tag{100}$$

and moving terms to the right hand side we got

$$\frac{dE}{dt} = \frac{1}{2}gH\,\mathbf{BT} \tag{101}$$

We could only focusing our attention to $\mathbf{BT}$.

$$\mathbf{BT} = -\,\mathbf{w}^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{w}\,\Bigg|_0^1 = \mathbf{w}^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{w}\,\Bigg|_0 - \mathbf{w}^\top \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \mathbf{w}\,\Bigg|_1 \tag{102}$$

14

Using $\mathbf{w} = [w_1 \ w_2]^\top$, we could get

$$\mathbf{BT} = \lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 - (\lambda_1 w_1^2 + \lambda_2 w_2^2)\big|_1 . \tag{103}$$

3.2.1 First case: $\lambda_1 > 0$, $\lambda_2 < 0$

In this case we have $U > 0$ and $gh - U^2 > 0$. On the right boundary, $x = 1$, we have need to get these expression to be negative, that is

$$- \left(\lambda_1 w_1^2 + \lambda_2 w_2^2\right)\big|_1 < 0 \tag{104}$$

or equivalently

$$\lambda_1 w_1^2 + \lambda_2 w_2^2\big|_1 > 0 \tag{105}$$

Let $w_2 = \gamma_N w_1$ so that the expression becomes

$$\lambda_1 w_1^2 + \lambda_2 w_2^2\big|_1 = \lambda_1 w_1^2 + \lambda_2 (\gamma_N w_1)^2\big|_1 \tag{106}$$

$$= w_1^2 \left(\lambda_1 + \lambda_2 \gamma_N^2\right)\big|_1 \tag{107}$$

That means, we need to get $\lambda_1 + \lambda_2 \gamma_N^2 > 0$ when $x = 1$. Equivalently, we need to choose γ_N^2 such that

$$\gamma_N^2 < -\frac{\lambda_1}{\lambda_2} \tag{108}$$

That would be our right boundary condition.

Meanwhile, on the left boundary, $x = 0$, we need to have

$$\lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 < 0 \tag{109}$$

But we know that λ_1 is positive. Let $w_1 = \gamma_0 w_2$.

$$\lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 = \lambda_1 (\gamma_0 w_2)^2 + \lambda_2 w_2^2\big|_0 \tag{110}$$

$$= w_2^2 \left(\lambda_1 \gamma_0^2 + \lambda_2\right)\big|_0 \tag{111}$$

15

so we need to have $\left(\lambda_1 \gamma_0^2 + \lambda_2\right) < 0$ when $x = 0$. Hence, our choosing of should fulfill the following.

$$\gamma_0^2 < -\frac{\lambda_2}{\lambda_1} \tag{112}$$

3.2.2   Second case: $\lambda_1 > 0, \lambda_2 > 0$

The second case, when $U > 0$ and $gh - U^2 < 0$, that is when both λ_1 and λ_2 are positive, from (103), we could rest assure on the right side, since the terms $(-\lambda_1 w_1^2 - \lambda_2 w_2^2)\big|_1$ is always non-positive. On the other side, $x = 1$, is pretty harsh since we need the expression $(\lambda_1 w_1^2 + \lambda_2 w_2^2)\big|_0$ to be non-positive. That would only possible if we force $w_1 = 0$ and $w_2 = 0$, and that would be two boundary conditions on the right.

3.2.3   Third case: $\lambda_1 < 0, \lambda_2 > 0$

From the boundary terms (103), we need **BT** to be non positive.

$$\mathbf{BT} = \lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 - (\lambda_1 w_1^2 + \lambda_2 w_2^2)\big|_1 . \tag{113}$$

On the left boundary $x = 0$, we have $\lambda_1 < 0$ so we need to control the terms $\lambda_2 w_2^2\big|_0$. Setting $w_2 = \gamma_0 w_1$ we could get

$$\lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 = \lambda_1 w_1^2 + \lambda_2 (\gamma_0 w_1)^2\big|_0 \tag{114}$$

$$= w_1^2 (\lambda_1 + \lambda_2 \gamma_0^2)\big|_0 \tag{115}$$

We want this terms to be non negative so our choice of γ_0^2 must satisfy

$$\lambda_1 + \lambda_2 \gamma_0^2 \leq 0 \tag{116}$$

$$\gamma_0^2 \leq -\frac{\lambda_1}{\lambda_2}. \tag{117}$$

On the right boundary $x = 1$, we have $\gamma_2 > 0$ so we need to control $\lambda_1 w_1^2$. Setting $w_1 = \gamma_N w_2$, we got

$$(\lambda_1 w_1^2 + \lambda_2 w_2^2)\big|_1 = (\lambda_1 (\gamma_N w_2)^2 + \lambda_2 w_2^2)\big|_1 \tag{118}$$

$$= w_2^2 (\lambda_1 \gamma_N^2 + \lambda_2)\big|_1 \tag{119}$$

16

We want this terms to be non positive, so we need to have

$$\lambda_1 \gamma_N^2 + \lambda_2 \geq 0 \tag{120}$$

$$\gamma_N^2 \leq -\frac{\lambda_2}{\lambda_1} \tag{121}$$

126 The sign changed because $\lambda_1 < 0$. Now we need to have our choice of γ_N fulfill the condition (121).

127 **3.2.4 Fourth case: $\lambda_1 < 0, \lambda_2 < 0$**

128 In this case our **BT** that need to be non positive,

129
$$\mathbf{BT} = \lambda_1 w_1^2 + \lambda_2 w_2^2\big|_0 - (\lambda_1 w_1^2 + \lambda_2 w_2^2)\big|_1 . \tag{122}$$

130 would give us not so much room to relax. The condition $\lambda_1 < 0, \lambda_2 < 0$ will ensure the non-positivity
131 on the left boundary, and forcing us to have $w_1 = 0$ and $w_0 = 0$ on the right boundary.

132 ## 3.3 Discretization

133 We would like to do discretization of the interval $[0,1]$ as the previous section. Consider Finite Volume
134 Method on identical grid on the interval $[0,1]$. Let N be a positive integer and I_i be the i-th control
135 volume for all $i = 1, 2, ..., N..$ Define $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ be the length of the i-th interval and $x_{i-\frac{1}{2}}$
136 and $x_{i+\frac{1}{2}}$ are the left boundary and right boundary of the interval respectively.

137 Let's our differential operator estimated by SBP operator

138
$$\frac{\partial \mathbf{q}}{\partial x} \approx \mathbf{D}_x \mathbf{q} \tag{123}$$

139 $\mathbf{D}_x = \mathbf{P}^{-1}\mathbf{Q}$. and $\mathbf{Q} + \mathbf{Q}^\top = \text{diag}[-1, 0, 0, \ldots, 0, 1]$

140 discretize norm be

141
$$\langle u, v \rangle_{\mathbf{P}} = v^\top \mathbf{P} u. \tag{124}$$

Our shallow water equation (70) and (71) could be rewrite as follows.

$$P\frac{dh}{dt} + U\mathbf{Q}h + H\mathbf{Q}u = 0 \tag{125}$$

$$P\frac{du}{dt} + g\mathbf{Q}h + U\mathbf{Q}u = 0 \tag{126}$$

142 For now, we would not adding the dissipative terms. Consider the first case when $\lambda_1 > 0$ and $\lambda_2 < 0$
143 that we have one boundary condition on the left and one boundary condition on the right.

Add the SAT terms, we could write the discretize equation as

$$\mathbf{P}\frac{dh}{dt} + U\mathbf{Q}h + H\mathbf{Q}u + \tau_{01}e_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N1}e_N\left(w_1 - \gamma_N w_2\right) = 0 \tag{127}$$

$$\mathbf{P}\frac{du}{dt} + g\mathbf{Q}h + U\mathbf{Q}u + \tau_{02}e_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N2}e_N\left(w_1 - \gamma_N w_2\right) = 0 \tag{128}$$

Multiply the two equation with $gh^\top$ and $Hu^\top$ respectively we got

$$gh^\top\mathbf{P}\frac{dh}{dt} + gUh^\top\mathbf{Q}h + gHh^\top\mathbf{Q}u + \tau_{01}gh_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N1}gh_N\left(w_1 - \gamma_N w_2\right) = 0 \tag{129}$$

$$Hu^\top\mathbf{P}\frac{du}{dt} + gHu^\top\mathbf{Q}h + UHu^\top\mathbf{Q}u + \tau_{02}Hu_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N2}Hu_N\left(w_1 - \gamma_N w_2\right) = 0 \tag{130}$$

Transpose both of the two, we will get

$$g\left(\frac{dh}{dt}\right)^\top\mathbf{P}^\top h + gUh^\top\mathbf{Q}^\top h + gHu^\top\mathbf{Q}^\top h + \tau_{01}gh_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N1}gh_N\left(w_1 - \gamma_N w_2\right) = 0 \tag{131}$$

$$H\left(\frac{du}{dt}\right)^\top\mathbf{P}^\top u + gHh^\top\mathbf{Q}^\top u + UHu^\top\mathbf{Q}^\top u + \tau_{02}Hu_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N2}Hu_N\left(w_1 - \gamma_N w_2\right) = 0. \tag{132}$$

Adding (129), (130), (131), and (132) together and using the identity $\mathbf{Q} + \mathbf{Q}^\top = \mathrm{diag}[-1, 0, 0, \ldots, 0, 1]$ gives us

$$\frac{d}{dt}\left(gh^\top\mathbf{P}h + Hu^\top\mathbf{P}u\right) + gUh^\top\mathbf{B}h + UHu^\top\mathbf{B}u + gH\left(h^\top\mathbf{B}u + u^\top\mathbf{B}h\right)$$
$$+ \tau_{01}gh_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N1}gh_N\left(w_1 - \gamma_N w_2\right)$$
$$+ \tau_{02}Hu_0\left(w_2 - \gamma_0 w_1\right) + \tau_{N2}Hu_N\left(w_1 - \gamma_N w_2\right)$$
$$= 0 \tag{133}$$

We will treat the terms $\left(gh^\top\mathbf{P}h + Hu^\top\mathbf{P}u\right)$ as our energy. Move all the terms except the energy, and

evaluate some of the terms involving $u^\top \mathbf{Q} u$ and $h^\top \mathbf{Q} h$, we got

$$\frac{d}{dt}\left(gh^\top \mathbf{P} h + H u^\top \mathbf{P} u\right) = -gU h^\top \mathbf{B} h - U H u^\top \mathbf{B} u - gH\left(h^\top \mathbf{B} u + u^\top \mathbf{B} h\right)$$

$$- \tau_{01} g h_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N1} g h_N \left(w_1 - \gamma_N w_2\right)$$

$$- \tau_{02} H u_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N2} H u_N \left(w_1 - \gamma_N w_2\right) \tag{134}$$

$$= -gU(h_N^2 - h_0^2) - U H(u_N^2 - u_0^2) - 2gH(h_N u_N - h_0 u_0)$$

$$- \tau_{01} g h_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N1} g h_N \left(w_1 - \gamma_N w_2\right)$$

$$- \tau_{02} H u_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N2} H u_N \left(w_1 - \gamma_N w_2\right) \tag{135}$$

We could write the last expression (135) in the matrix form, treating left boundary and right boundary separately. Before that, for convenience, we would our boundary conditoin implemented in the form of h and u, namely, $w_2 - \gamma_0 w_1 = \alpha_0 h_0 + \beta_0 u_0$ on the left boundary; $w_1 - \gamma_N w_2 = \alpha_N h_N + \beta_N u_N$ on the right boundary.

Now, expression (135) becomes

$$- gU(h_N^2 - h_0^2) - U H(u_N^2 - u_0^2) - 2gH(h_N u_N - h_0 u_0)$$

$$- \tau_{01} g h_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N1} g h_N \left(w_1 - \gamma_N w_2\right)$$

$$- \tau_{02} H u_0 \left(w_2 - \gamma_0 w_1\right) - \tau_{N2} H u_N \left(w_1 - \gamma_N w_2\right) \tag{136}$$

$$= gU h_0^2 + U H u_0^2 + 2gH h_0 u_0 - \tau_{01} g h_0 (\alpha_0 h_0 + \beta_0 u_0) - \tau_{02} H u_0 (\alpha_0 h_0 + \beta_0 u_0) \tag{137}$$

$$- gU h_N^2 - U H u_N^2 - 2gH h_N u_N - \tau_{N1} g h_N (\alpha_N h_N + \beta_N u) - \tau_{N2} H u_N (\alpha_N h_N + \beta_N u_N) \tag{138}$$

From (137) we could express it as

$$gH \begin{pmatrix} h_0 \\ u_0 \end{pmatrix}^\top \begin{pmatrix} \frac{U}{H} - \frac{2\tau_{01}\alpha_0}{H} & 1 - \frac{\tau_{01}\beta_0}{H} - \frac{\tau_{02}\alpha_0}{g} \\ 1 - \frac{\tau_{01}\beta}{H} - \frac{\tau_{02}\alpha_0}{g} & \frac{U}{g} - \frac{2\tau_{02}\beta_0}{g} \end{pmatrix} \begin{pmatrix} h_0 \\ u_0 \end{pmatrix} \tag{139}$$

and from (138) we could express it as

$$-gH \begin{pmatrix} h_N \\ u_N \end{pmatrix}^\top \begin{pmatrix} \frac{U}{H} + \frac{2\tau_{N1}\alpha_N}{H} & 1 + \frac{\tau_{N1}\beta_N}{H} + \frac{\tau_{N2}\alpha_N}{g} \\ 1 + \frac{\tau_{N1}\beta_N}{H} + \frac{\tau_{N2}\alpha_N}{g} & \frac{U}{g} + \frac{2\tau_{N2}\beta_N}{g} \end{pmatrix} \begin{pmatrix} h_N \\ u_N \end{pmatrix} \tag{140}$$

19

The whole equation now becomes

$$\frac{d\mathbf{E}}{dt} = gH \begin{pmatrix} h_0 \\ u_0 \end{pmatrix}^{\top} \begin{pmatrix} \frac{U}{H} - \frac{2\alpha_0 \tau_{01}}{H} & 1 - \frac{\tau_{01}\beta_0}{H} - \frac{\alpha_0}{g} \\ 1 - \frac{\tau_{01}\beta}{H} - \frac{\alpha_0}{g} & \frac{U}{g} - \frac{2\beta_0 \tau_{02}}{g} \end{pmatrix} \begin{pmatrix} h_0 \\ u_0 \end{pmatrix}$$
$$- gH \begin{pmatrix} h_N \\ u_N \end{pmatrix}^{\top} \begin{pmatrix} \frac{U}{H} + \frac{2\alpha_N \tau_{N1}}{H} & 1 + \frac{\tau_{N1}\beta_N}{H} + \frac{\alpha_N}{g} \\ 1 + \frac{\tau_{N1}\beta_N}{H} + \frac{\alpha_N}{g} & \frac{U}{g} + \frac{2\beta_N \tau_{N2}}{g} \end{pmatrix} \begin{pmatrix} h_N \\ u_N \end{pmatrix} \tag{141}$$

To make it stable, we need to choose our penalty parameter so that the right hand side of the equation becomes non positive. We could make this easier by choose the parameter so that the matrix involved becomes diagonal. That is, for left boundary, $1 - \frac{\tau_{01}\beta_0}{H} - \frac{\tau_{02}\alpha_0}{g} = 0$ and for the right boundary $1 + \frac{\tau_{N1}\beta_0}{H} + \frac{\tau_{N2}\alpha_0}{g} = 0$. One particular choice would be:

$$\frac{\tau_{01}\beta_0}{H} = \frac{\tau_{02}\alpha_0}{g} = \frac{1}{2} \quad \text{and} \quad \frac{\tau_{N1}\beta_0}{H} = \frac{\tau_{N2}\alpha_0}{g} = -\frac{1}{2} \tag{142}$$

or equivalently

$$\tau_{01} = \frac{H}{2\beta_0}, \qquad \tau_{02} = \frac{g}{2\alpha_0}, \qquad \tau_{N1} = -\frac{H}{2\beta_0}, \qquad \tau_{N1} = -\frac{g}{2\alpha_0} \tag{143}$$

By those choice, equation (141) could be rewrite as

$$\frac{d\mathbf{E}}{dt} = gH \begin{pmatrix} h_0 \\ u_0 \end{pmatrix}^{\top} \begin{pmatrix} \frac{U}{H} - \frac{\alpha_0}{\beta_0} & 0 \\ 0 & \frac{U}{g} - \frac{\beta_0}{\alpha_0} \end{pmatrix} \begin{pmatrix} h_0 \\ u_0 \end{pmatrix}$$
$$- gH \begin{pmatrix} h_N \\ u_N \end{pmatrix}^{\top} \begin{pmatrix} \frac{U}{H} + \frac{\alpha_N}{\beta_N} & 0 \\ 0 & \frac{U}{g} + \frac{\alpha_N}{\beta_N} \end{pmatrix} \begin{pmatrix} h_N \\ u_N \end{pmatrix} \tag{144}$$

As we want to have the energy non increasing, we want the right hand side non positive. It is sufficient to have the eigenvalues $\frac{U}{H} - \frac{\alpha_0}{\beta_0}$ and $\frac{U}{g} - \frac{\beta_0}{\alpha_0}$ are non positive and the eigenvalues $\frac{U}{H} + \frac{\alpha_N}{\beta_N}$ and $\frac{U}{g} + \frac{\alpha_N}{\beta_N}$ are non negative. Equivalently, the following inequality must be fulfilled.

$$\frac{U}{H} \leq \frac{\alpha_0}{\beta_0} \qquad \text{and} \qquad \frac{U}{g} \leq \frac{\beta_0}{\alpha_0} \tag{145}$$
$$\frac{U}{H} + \frac{\alpha_N}{\beta_N} \geq 0 \qquad \text{and} \qquad \frac{U}{g} + \frac{\alpha_N}{\beta_N} \geq 0 \tag{146}$$

3.4   Boundary Conditions

The boundary condition that we will try to implementing should fulfill (145) and (146) to satisfy the stability in the semi discrete level with penalty parameter chosen as well as some condtion (112), (108), (117), and (121) depending on the case. But we have to back and fort from using the pair w_1 and w_2, to using the pair u and h. That's why we recall equation (97) and (98).

$$w_1 = \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)h + \frac{1}{c}u, \qquad\qquad w_2 = \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)h + \frac{1}{d}u \qquad (147)$$

$$h = \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)w_1 + \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)w_2, \qquad u = \frac{1}{c}w_1 + \frac{1}{d}w_2 \qquad (148)$$

Before we're going into details. We would like to show that we couldn't set the left boundary condition with $u = 0$ or $h = 0$. It's because if we look into equation (82), and let one of the $u = 0$ or $h = 0$ we would end up with either

$$\cdots \qquad (149)$$

which both of them are not energy stable.

3.4.1   First case: $\lambda_1 > 0, \lambda_2 < 0$.

In this case $\lambda_1 > 0, \lambda_2 < 0$ would equivalent to $U > 0$ and $gH > U^2$. This case require one boundary
condition in each side.

Right Boudary

Let our right boundary $h = 0$. Then, to have the stability in the semidiscrete level, we have to prove that this satisfy (112), (108). On the left, the condition $h = 0$ give us

$$h = 0 \qquad (150)$$

$$\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)w_1 + \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)w_2 = 0 \qquad (151)$$

$$w_2 - \left(-\frac{\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)}{\frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)}\right)w_1 = 0 \qquad (152)$$

So, we need our $\gamma_N^2 = \left(-\frac{\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)}{\frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)}\right)^2$ less than $-\frac{\lambda_1}{\lambda_2}$.
(Proof)

Or we could also set the right boundary to be $u = 0$. It will lead us to condition

$$u = 0 \tag{153}$$

$$\frac{1}{c}w_1 + \frac{1}{d}w_2 = 0 \tag{154}$$

$$w_2 - \left(-\frac{d}{c}\right)w_1 = 0 \tag{155}$$

and we need to prove that $\gamma_N^2 = \left(-\frac{d}{c}\right)^2 < -\frac{\lambda_1}{\lambda_2}$.

(Proof)

Left Boundary

In the left boundary, we need the linear combination of u and h. There's two sub-cases that interesting to be implemented. The first one is when $\sqrt{\frac{g}{H}}h + u = 0$ and the second one is $Hu + Uh = 0$ which is the linearized version of $q = 0$. Remember that $q = hu$. In linearized version, $q = (H + h)(U + u) = HU + Hu + Uh + uh$. Since we ignore the higher order terms, uh, and also HU is only constant, we could see that $q = 0$ could be interpreted as $Hu + Uh = 0$.

In the first sub-case, $\sqrt{\frac{g}{H}}h + u = 0$, we could interpreted this as

$$\sqrt{\frac{g}{H}}h + u = 0 \tag{156}$$

$$\sqrt{\frac{g}{H}}\left(\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)w_1 + \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)w_2\right) + \left(\frac{1}{c}w_1 + \frac{1}{d}w_2\right) = 0 \tag{157}$$

$$\frac{1}{c}\left(\sqrt{\frac{g}{H}}\left(\lambda_1 - \frac{U}{g}\right)\right)w_1 + \frac{1}{d}\left(\sqrt{\frac{g}{H}}\left(\lambda_2 - \frac{U}{g}\right)\right)w_2 = 0 \tag{158}$$

$$w_1 - \left(-\frac{\frac{1}{d}\left(\sqrt{\frac{g}{H}}\left(\lambda_2 - \frac{U}{g}\right)\right)}{\frac{1}{c}\left(\sqrt{\frac{g}{H}}\left(\lambda_1 - \frac{U}{g}\right)\right)}\right)w_2 = 0 \tag{159}$$

Now, we need to prove that

$$\gamma_0^2 = \left(-\frac{\frac{1}{d}\left(\sqrt{\frac{g}{H}}\left(\lambda_2 - \frac{U}{g}\right)\right)}{\frac{1}{c}\left(\sqrt{\frac{g}{H}}\left(\lambda_1 - \frac{U}{g}\right)\right)}\right)^2 < -\frac{\lambda_2}{\lambda_1}$$

The other sub-case is $Hu + Uh = 0$. We could change this into

22

$$Hu + Uh = 0 \tag{160}$$

$$H\left(\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)w_1 + \frac{1}{d}\left(\lambda_2 - \frac{U}{g}\right)w_2\right) + U\left(\frac{1}{c}w_1 + \frac{1}{d}w_2\right) = 0 \tag{161}$$

$$\left(\frac{H}{c}\left(\lambda_1 - \frac{U}{g}\right) + \frac{U}{c}\right)w_1 + \left(\frac{H}{d}\left(\lambda_2 - \frac{U}{g}\right) + \frac{U}{d}\right)w_2 = 0 \tag{162}$$

$$w_1 - \left(-\frac{\frac{H}{c}\left(\lambda_1 - \frac{U}{g}\right) + \frac{U}{c}}{\frac{H}{d}\left(\lambda_2 - \frac{U}{g}\right) + \frac{U}{d}}\right) = 0 \tag{163}$$

and as before, we need to prove that

$$\gamma_0^2 = \left(-\frac{\frac{H}{c}\left(\lambda_1 - \frac{U}{g}\right) + \frac{U}{c}}{\frac{H}{d}\left(\lambda_2 - \frac{U}{g}\right) + \frac{U}{d}}\right)^2 < -\frac{\lambda_2}{\lambda_1}.$$

$$\dots$$

In our implementation of SAT, we need to satisfy condition (145) and (146). On the left boundary, it's pretty straight forward. In the first sub-case, $\sqrt{\frac{g}{H}}h + u = 0$ give us $\alpha_0 = \sqrt{\frac{g}{H}}$ and $\beta_0 = 1$. We only need to show that

$$\frac{U}{H} \leq \frac{\alpha_0}{\beta_0} \qquad \text{and} \qquad \frac{U}{g} \leq \frac{\beta_0}{\alpha_0} \tag{164}$$

$$\frac{U}{H} \leq \sqrt{\frac{g}{H}} \qquad \text{and} \qquad \frac{U}{g} \leq \frac{1}{\sqrt{\frac{g}{H}}} \tag{165}$$

168 and both of the conditions equivalent in our first case $U^2 - gH < 0$.

The second sub-case, with $Hu + Uh = 0$, we have $\alpha_0 = U$ and $\beta_0 = H$. And we need to prove that

$$\frac{U}{H} \leq \frac{\alpha_0}{\beta_0} \qquad \text{and} \qquad \frac{U}{g} \leq \frac{\beta_0}{\alpha_0} \tag{166}$$

$$\frac{U}{H} \leq \frac{U}{H} \qquad \text{and} \qquad \frac{U}{g} \leq \frac{H}{U} \tag{167}$$

but thats equivalent to our condition $U^2 < gH$.

$$\dots$$

23

3.4.2   Second case: $\lambda_1 > 0, \lambda_2 > 0.$

In this case, we have $U > 0$ and $gH < U^2$. The two boundary conditions needed in this case are $w_1 = 0$ and $w_2 = 0$. We don't need to choose γ_0 nor γ_N in this case. The only conditions we need to met are (145) and (146). The boundary condition $w_1 = 0$ would equivalent to

$$w_1 = \frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)h_0 + \frac{1}{c}u_0 = \alpha_0 h_0 + \beta_0 u_0 \tag{168}$$

so that

$$\frac{\alpha_0}{\beta_0} = \frac{\frac{1}{c}\left(\lambda_1 - \frac{U}{g}\right)}{\frac{1}{c}} = \left(\lambda_1 - \frac{U}{g}\right). \tag{169}$$

Hence, we need to prove that

$$\left(\lambda_1 - \frac{U}{g}\right) \geq \frac{U}{H} \qquad \text{and} \qquad \frac{1}{\left(\lambda_1 - \frac{U}{g}\right)} \geq \frac{U}{g} \tag{170}$$

$$\lambda_1 \geq \frac{U^2}{gH} \qquad \text{and} \qquad \left(\lambda_1 - \frac{U}{g}\right) \leq \frac{g}{U} \tag{171}$$

We could see that this is true by looking this for any value of U and H.

<p align="center">DIAGRAM</p>

The second condition could be analyse with the same argument and get the following inequalities.

$$\lambda_2 \geq \frac{U^2}{gH} \qquad \text{and} \qquad \left(\lambda_2 - \frac{U}{g}\right) \leq \frac{g}{U} \tag{172}$$

<p align="center">DIAGRAM</p>

3.4.3   Third case: $\lambda_1 < 0, \lambda_2 > 0$

In this case, we have $U < 0$ and $gH > U^2$, one boundary condition on the left and one boundary condition on the right. The condition (145) and (146) do not depend on the cases of the eigenvalues, then we could reuse the result from previous cases. So in this case we only need to check our choice of γ_0 and γ_N.

Left Boundary

On the left boundary, we could only set the linear combination of u and h. We would look into two cases, $Uh + Hu = 0$ and $\sqrt{\frac{g}{H}}h + u$.

Right Boundary

3.4.4 Fourth case: $\lambda_1 < 0, \lambda_2 < 0$

In this case, we have $U < 0$ and $gH < U^2$, that is two boundary condition on the right. Now we have plenty of choices: $u = 0, h = 0$, and linear combination of u and h. Since the condition (145) and (146) do not depend on the cases of the eigenvalues, then we could reuse the result from previous cases. Also because this case has $w_1 = 0$ and $w_2 = 0$ we don't need to bother with γ_0 and γ_N.

4 Conclusion