ON THE ESTABLISHMENT OF HYPERBOLICITY OF SHALLOW WATER MOMENT EQUATIONS IN TWO DIMENSIONS

By

Matthew Bauerle

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics—Doctor of Philosophy

2023

ABSTRACT

In this thesis, we investigate the two-dimensional extension of a recently introduced set of shallow water models based on a regularized moment expansion of the incompressible Navier-Stokes equations [22, 21]. We show the rotational invariance of the proposed moment models with two different approaches. The first proof involves the split of the coefficient matrix into the conservative and non-conservative parts and prove the rotational invariance for each part, while the second one relies on the special block structure of the coefficient matrices. With the aid of rotational invariance, the analysis of the hyperbolicity for the moment model in 2D is reduced to the real diagonalizability of the coefficient matrix in 1D. Then we prove the real diagonalizability by deriving the analytical form of the characteristic polynomial. Furthermore, we extend the model to include a more general class of closure relations than the original model and establish that this set of general closure relations retain both rotational invariance and hyperbolicity.

ACKNOWLEDGMENTS

This work was impacted by many people. It has been said that we stand on the shoulders of giants so it only seems fair we mention them.

The main thanks for research goes to my advisor Professor Andrew Christlieb for organizing the research group that led to this thesis. I thank him for including me in the group and helping me through the struggles of graduate studies. The meetings and encouragement has kept me on track.

The Christlieb group AKA SPECTRE (it is from James Bond and we will tell the story if you wish) has been an enjoyable space to both work and spend free time in. I look forward to seeing what other work and tales will come out from its members.

My committee deserves thanks for reviewing this thesis and offering feedback.

Thanks goes to Professor Juntao Huang for formalizing the proofs and Dr. Mingchang Ding for the numerical analysis of the equations.

I would like to thank Michigan State University and Texas Tech University for the support of this work, through the ongoing support of the PIs and students. I would further like to thank Keith Promislow for his helpful discussions and insight during the development of this effort. The reading course we set up was also greatly appreciated as a way to finish out credits and review literature. I would like to thank the Office of Naval Research for the support of this work under grant number N00014-18-1-2552.

Last, I would like thank my family for helping me through the ups and downs. My mother was my first teacher and my father first got me interested in mathematics and engineering. I still appreciate their encouragement and motivation (sometimes).

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Shallow water equations are widely used in modeling meteorological and oceanographic phenomena. They are most useful in scenarios where the vertical dimension is much smaller than horizontal dimensions of the problem domain. This is often the case with simulations spanning thousands of kilometers horizontally but only a couple of kilometers vertically. An introduction to shallow water equations and proof of hyperbolicity in 2 dimensions is shown in [29]. The definition of hyperbolicity for 1st order systems of PDEs in multiple spacial dimensions is given in [25]. The Euler equations were shown to be rotationally invariant and therefore the system is hyperbolic if the system is hyperbolic in one direction. While the vertical dimension is often the smallest dimension, vertical dynamics play a critical role in many scenarios. Shallow water moment equations provide an intermediate level of resolution between shallow water and a full 3 dimensional model. The first expansion to the shallow water model is to replace the horizontal velocities with a polynomial expansion [22]. This creates a system of first order ODEs. For the 0th order model the system is the shallow water equations which form a conservative hyperbolic system. When one moment is added the system is still globally hyperbolic but the system loses the form of a conservation law. For two or more moments the system ceases to be hyperbolic. Subsequent work revealed this could lead to instabilities in solutions and the 1 dimensional equations were altered to make them hyperbolic [21]. In [26] the same process of regularization was extended to apply to the 2 dimensional problem. While it is conjectured in that work that the 2 dimensional regularized shallow water moment equations are globally hyperbolic it is not proven. Some extensions to shallow water equations are presented in [3]. Various boundary conditions and modifications are discussed. One of the most important modifications is the introduction of multilevel shallow water models. The examples given is of water dynamics in the strait of Gibraltar and submarine avalanches. Hyperbolicity loss is also a problem for multilevel shallow water equation. This is associated with a shear velocity differential between the

layers that physically creates Kelvin-Helmholtz instabilities which cannot be modeled with a piece-wise constant flow [6].

This work deals with that two-dimensional extension of shallow water models based on a regularized moment expansion of the incompressible Navier-Stokes (NS) equations [22, 21]. Moment models have a long history of constructing computationally efficient representations of complex dynamics arising in a higher dimensional model. Historically, this approach has been widely used in the reduction of kinetic equations to hierarchies of moment models, where the resulting evolution equation for the m^{th} moment equation depends on the $(m+1)^{st}$ moment [14]. At some point, one needs to restrict the expansion to m moments and introduce a model for the $(m+1)^{st}$ moment in terms of lower moments to obtain a solvable system of equations. This leads to the well-known moment closure problem, which is the study of what kind of model preserves the desired hyperbolic structure [23, 4, 5, 11, 8].

In this work, we are looking at a new class of models for describing systems traditionally modeled with shallow water equations [28] and multi-layer shallow water equations [2, 24, 10, 9]. In [22], the new class of models, called the shallow water moment equations (SWME), was derived by taking moments with respect to the Legendre polynomials of the vertical direction of the three-dimensional (3D) incompressible NS equations. The first two moments of the system yield the traditional shallow water equations. In principle, higher-order moments offer an approach to include vertical information without introducing a mesh in the vertical direction, as one would need for the 3D NS equations. Fundamentally, this is seeking to address a multi-scale problem by providing a path to increased fidelity of the flow dynamics without needing to introduce a mesh to resolve the vertical direction. The system as originally proposed is mathematically elegant, but does not preserve hyperbolicity for higher numbers of moments in 1D and 2D. The developers of the model realized this and introduced a class of regularizations in 1D called the hyperbolic shallow water moment equations (HSWME) and proceeded to show that this regularized system was provably hyperbolic [21]. In addition, they have extended the 1D system to a variety of settings and

compared depth averaged results of the incompressible 2D Navier Stokes to the regularized moment based models. In their work, they demonstrated increased fidelity when adding additional moments [22, 21]. More recently, there are some follow-up works, including the equilibrium stability analysis [19], well-balanced schemes [20], efficient time discretizations [1], axisymmetric model [27], multilayer-moment models [13], and extension to sediment transport [12].

In this work, our main result is that we establish that the 2D extension to the regularized moment expansion of the incompressible NS equations is rotational invariant and hyperbolic for arbitrary number of moments. We present two different proofs of the rotational invariance. The first proof involves the split of the coefficient matrix into the conservative and non-conservative parts and prove the rotational invariance for each part, while the second one relies on the special block structure of the coefficient matrix. With the aid of rotational invariance, the hyperbolicity in 2D reduces to the real diagonalizability of the matrix in 1D. To analyze the eigenvalues, we make use of the associated polynomial sequence and derive the characteristic polynomial of the coefficient matrix analytically. We find that the eigenvalues are related to the Gauss-Lobatto quadrature points (i.e. the zeros of derivative of Legendre polynomials) and Gauss-Legendre quadrature points (i.e. the zeros of Legendre polynomials). In particular, we show that the eigenvalues of the moment system are real and distinct for arbitrary number of moments. More importantly, we establish the general closure relations such that the rotational invariance and hyperbolicity are guaranteed. This opens the door to the development of data-driven closures that preserve hyperbolicity, as in our past work for the radiation transfer equations [15, 17, 16]. Data-driven closures are the subject of our future work.

The remaining parts of this paper are structured as follows: Section 2 introduces the models. In Section 3, we show two proofs to the rotational invariance of the models. Section 4 analyzes the eigen structure of the models and establishes hyperbolicity of the models. In Section 5, we give conclusions and talk about future directions.

CHAPTER 2

SHALLOW WATER MOMENT EQUATIONS

2.1 Equation Derivation

In this section, we review the main ideas and the results for the derivation of the shallow water moment model in [22]. We also show the 1D hyperbolic shallow water moment model proposed in [21].

We start by considering the 3D incompressible Navier-Stokes equations:

$$\begin{aligned} \nabla \cdot U &= 0, \\ \partial_t U + \nabla \cdot (UU) &= -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\nabla \cdot \sigma + g. \end{aligned} \tag{2.1.1}$$

Here, $U = (u, v, w)^T$ is the velocity vector, p is the pressure and σ is the stress tensor. The density ρ is constant and $g = (e_x, e_y, e_z)g$ with (e_x, e_y, e_z) a constant unit vector denotes the gravitational acceleration. The often used shallow water coordinate system is recovered by choosing $e_x = e_y = 0$ and $g = (0, 0, -1)^T g$.

Under the shallowness assumption, i.e., the horizontal scales of the flow are much larger than the vertical scale, the Navier-Stokes equations (2.1.1) can be reduced by an asymptotic analysis to:

$$\begin{aligned} \partial_x u + \partial_y v + \partial_z w &= 0, \\ \partial_t u + \partial_x(u^2) + \partial_y(uv) + \partial_z(uw) &= -\frac{1}{\rho}\partial_x p + \frac{1}{\rho}\partial_z \sigma_{xz} + g e_x, \\ \partial_t v + \partial_x(uv) + \partial_y(v^2) + \partial_z(vw) &= -\frac{1}{\rho}\partial_y p + \frac{1}{\rho}\partial_z \sigma_{yz} + g e_y, \end{aligned} \tag{2.1.2}$$

where the hydrostatic pressure is given by

$$p(x, y, z, t) = (h_s(x, y, t) - z)\rho g e_z, \tag{2.1.3}$$

with $h_s(x, y, t)$ being the profile of the upper free surface. A summary of this reduction is presented in the appendix of this thesis and details can be found in Appendix A of [22].

To derive the shallow water moment model from (2.1.2), the first idea in [22] is to introduce a scaled vertical variable $\zeta(x,y,t)$ given by

$$\zeta(x,y,t) := \frac{z - h_b(x,y,t)}{h_s(x,y,t) - h_b(x,y,t)} = \frac{z - h_b(x,y,t)}{h(x,y,t)}, \tag{2.1.4}$$

with $h(x,y,t) := h_s(x,y,t) - h_b(x,y,t)$ is the water height from the bottom $h_b(x,y,t)$ to the surface $h_s(x,y,t)$. This transforms the z-direction from a physical space $z \in [h_b, h_s]$ to a projected space $\zeta \in [0,1]$. For any function $\psi = \psi(x,y,z,t)$, the corresponding mapped function $\tilde{\psi} = \tilde{\psi}(x,y,\zeta,t)$ is given by

$$\tilde{\psi}(x,y,\zeta,t) := \psi(x,y,h(x,y,t)\zeta + h_b(x,y,t)). \tag{2.1.5}$$

The complete vertically resolved shallow flow system has the form [22]

$$\begin{aligned}
\partial_t h + \partial_x(hu_m) + \partial_y(hv_m) &= 0,\\
\partial_t(h\tilde{u}) + \partial_x(h\tilde{u}^2 + \frac{g}{2}e_z h^2) + \partial_y(h\tilde{u}\tilde{v}) + \partial_\zeta(h\tilde{u}\omega - \frac{1}{\rho}\tilde{\sigma}_{xz}) &= gh(e_x - e_z\partial_x h_b),\\
\partial_t(h\tilde{v}) + \partial_x(h\tilde{u}\tilde{v}) + \partial_y(h\tilde{v}^2 + \frac{g}{2}e_z h^2) + \partial_\zeta(h\tilde{v}\omega - \frac{1}{\rho}\tilde{\sigma}_{yz}) &= gh(e_y - e_z\partial_y h_b),
\end{aligned} \tag{2.1.6}$$

where $u_m(x,y,t) = \int_0^1 u(x,y,\zeta,t)d\zeta$ and $v_m(x,y,t) = \int_0^1 v(x,y,\zeta,t)d\zeta$ denote the mean velocities and ω is the vertical coupling

$$\omega = \frac{1}{h}\overline{\partial_x(h\tilde{u}) + \partial_y(h\tilde{v})}, \tag{2.1.7}$$

with the average for any function $\psi = \psi(\zeta)$ defined by

$$\bar{\psi}(\zeta) := \int_0^\zeta \left(\int_0^1 \psi(\check{\zeta})d\check{\zeta} - \psi(\hat{\zeta}) \right) d\hat{\zeta}. \tag{2.1.8}$$

Note that, for a constant flow profile in ζ, the vertical coupling coefficient ω vanishes. In that case, if in addition shear stresses are negligible $\sigma_{xz} = \sigma_{yz} = 0$, the system reduces to the shallow water equations.

Before deriving the moment equation, we introduce some assumptions. First, we use the Newtonian constitutive law:

$$\sigma_{xz} = \mu\partial_z u, \quad \sigma_{yz} = \mu\partial_z v,$$

where μ stands for the dynamic viscosity and $\nu = \mu/\rho$ the kinematic viscosity. In order to solve it, we need to specify dynamic boundary conditions in the form of a velocity boundary condition both at the free-surface, and at the bottom topography. At the free-surface, the stress-free conditions are assumed:

$$\partial_z u = \partial_z v = 0, \quad \text{at} \quad z = h_s(x, y, t).$$

At the basal surface, the slip boundary conditions are assumed:

$$u - \frac{\lambda}{\mu}\sigma_{xz} = v - \frac{\lambda}{\mu}\sigma_{yz} = 0, \quad \text{at} \quad z = h_b(x, y, t).$$

Here, λ stands for the slip length.

By assuming a polynomial expansion of the velocity components:

$$u(x, y, z, t) = u_m(x, y, t) + \sum_{j=1}^{N} \alpha_j(x, y, t)\phi_j(z),$$

$$v(x, y, z, t) = v_m(x, y, t) + \sum_{j=1}^{N} \beta_j(x, y, t)\phi_j(z),$$

with the scaled Legendre polynomials ϕ_j, orthogonal on the interval $[0, 1]$ and normalized by

$\phi_j(0) = 1$, the shallow water moment equations (SWME) can be derived [22]:

$$
\begin{aligned}
&\partial_t h + \partial_x(hu_m) + \partial_y(hv_m) = 0,\\
&\partial_t(hu_m) + \partial_x\left(h(u_m^2 + \sum_{j=1}^{N}\frac{\alpha_j^2}{2j+1}) + \frac{g}{2}e_z h^2\right) + \partial_y\left(h(u_m v_m + \sum_{j=1}^{N}\frac{\alpha_j\beta_j}{2j+1})\right)\\
&\qquad\qquad = -\frac{\nu}{\lambda}(u_m + \sum_{j=1}^{N}\alpha_j) + hg(e_x - e_z\partial_x h_b),\\
&\partial_t(hv_m) + \partial_x\left(h(u_m v_m + \sum_{j=1}^{N}\frac{\alpha_j\beta_j}{2j+1})\right) + \partial_y\left(h(v_m^2 + \sum_{j=1}^{N}\frac{\beta_j^2}{2j+1}) + \frac{g}{2}e_z h^2\right)\\
&\qquad\qquad = -\frac{\nu}{\lambda}(v_m + \sum_{j=1}^{N}\beta_j) + hg(e_y - e_z\partial_y h_b),\\
&\partial_t(h\alpha_i) + \partial_x\left(h(2u_m\alpha_i + \sum_{j,k=1}^{N}A_{ijk}\alpha_j\alpha_k)\right) + \partial_y\left(h(u_m\beta_i + v_m\alpha_i + \sum_{j,k=1}^{N}A_{ijk}\alpha_j\beta_k)\right)\\
&\qquad = u_m D_i - \sum_{j,k=1}^{N}B_{ijk}D_j\alpha_k - (2i+1)\frac{\nu}{\lambda}\left(u_m + \sum_{j=1}^{N}(1 + \frac{\lambda}{h}C_{ij})\alpha_j\right), \quad i = 1, 2, \cdots, N,\\
&\partial_t(h\beta_i) + \partial_x\left(h(u_m\beta_i + v_m\alpha_i + \sum_{j,k=1}^{N}A_{ijk}\alpha_j\beta_k)\right) + \partial_y\left(h(2v_m\beta_i + \sum_{j,k=1}^{N}A_{ijk}\beta_j\beta_k)\right)\\
&\qquad = v_m D_i - \sum_{j,k=1}^{N}B_{ijk}D_j\beta_k - (2i+1)\frac{\nu}{\lambda}\left(v_m + \sum_{j=1}^{N}(1 + \frac{\lambda}{h}C_{ij})\beta_j\right), \quad i = 1, 2, \cdots, N.
\end{aligned}
\tag{2.1.9}
$$

Here the right-hand-side (RHS) contains non-conservative terms involving the expression

$$D_i := \partial_x(h\alpha_i) + \partial_y(h\beta_i) \tag{2.1.10}$$

and the constants A_{ijk}, B_{ijk}, C_{ij} are related to the integrals of the Legendre polynomials:

$$
\begin{aligned}
A_{ijk} &= (2i+1)\int_0^1 \phi_i\phi_j\phi_k d\zeta, \quad i, j, k = 1, \cdots, N,\\
B_{ijk} &= (2i+1)\int_0^1 \phi_i'\left(\int_0^\zeta \phi_j d\hat{\zeta}\right)\phi_k d\zeta, \quad i, j, k = 1, \cdots, N,\\
C_{ij} &= \int_0^1 \phi_i'\phi_j' d\zeta, \quad i, j = 1, \cdots, N.
\end{aligned}
\tag{2.1.11}
$$

The above system (2.1.9) can be written as

$$\partial_t U + A(U)\partial_x U + B(U)\partial_y U = S(U), \tag{2.1.12}$$

with the unknown variables

$$U = (h,\, hu_m,\, hv_m,\, h\alpha_1,\, h\beta_1,\, h\alpha_2,\, h\beta_2,\, \cdots, h\alpha_N,\, h\beta_N)^T. \tag{2.1.13}$$

The coefficient matrices in (2.1.12) can be split into the conservative and non-conservative parts:

$$A(U) = \partial_U F(U) + P(U), \quad B(U) = \partial_U G(U) + Q(U). \tag{2.1.14}$$

Here the physical fluxes $F(U)$ and $G(U)$ for the conservative parts are

$$\begin{aligned} F(U) = (&hu, h(u^2 + \sum_{j=1}^{N} \frac{\alpha_j^2}{2j+1}) + \frac{1}{2}gh^2,\, h(uv + \sum_{j=1}^{N} \frac{\alpha_j\beta_j}{2j+1}), \\ &h(2u\alpha_1 + \sum_{j,k=1}^{N} A_{1jk}\alpha_j\alpha_k),\, h(u\beta_1 + v\alpha_1 + \sum_{j,k=1}^{N} A_{1jk}\alpha_j\beta_k), \\ &\cdots, \\ &h(2u\alpha_n + \sum_{j,k=1}^{N} A_{njk}\alpha_j\alpha_k), h(u\beta_N + v\alpha_n + \sum_{j,k=1}^{N} A_{njk}\alpha_j\beta_k))^T, \end{aligned} \tag{2.1.15}$$

$$\begin{aligned} G(U) = (&hv, h(uv + \sum_{j} \frac{\alpha_j\beta_j}{2j+1}),\, h(v^2 + \sum_{j} \frac{\beta_j^2}{2j+1}) + \frac{1}{2}gh^2, \\ &h(u\beta_1 + v\alpha_1 + \sum_{j,k} A_{1jk}\alpha_j\beta_k),\, h(2v\beta_1 + \sum_{j,k} A_{1jk}\beta_j\beta_k), \\ &\cdots, \\ &h(u\beta_n + v\alpha_n + \sum_{j,k} A_{njk}\alpha_j\beta_k),\, h(2v\beta_n + \sum_{j,k} A_{njk}\beta_j\beta_k))^T. \end{aligned} \tag{2.1.16}$$

The matrices for the non-conservative part can be further decomposed into two parts:

$$P(U) = P_1(U) + P_2(U), \quad Q(U) = Q_1(U) + Q_2(U). \tag{2.1.17}$$

Here $P_1(U)$ and $Q_1(U)$ describe the terms $u_m D_i$ and $v_m D_i$ on the RHS of (2.1.9):

$$P_1(U) = \mathrm{diag}(0_{3\times 3}, p(U), \cdots, p(U)) \in \mathbb{R}^{(2N+3)\times(2N+3)} \tag{2.1.18}$$

and

$$Q_1(U) = \mathrm{diag}(0_{3\times 3}, q(U), \cdots, q(U)) \in \mathbb{R}^{(2N+3)\times(2N+3)} \tag{2.1.19}$$

with

$$p(U) = \begin{pmatrix} -u & 0 \\ -v & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad q(U) = \begin{pmatrix} 0 & -u \\ 0 & -v \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{2.1.20}$$

The second part $P_2(U)$ and $Q_2(U)$ describe the terms $\sum_{j,k=1}^{N} B_{ijk} D_j \alpha_k$ and $\sum_{j,k=1}^{N} B_{ijk} D_j \beta_k$ on the RHS of (2.1.9):

$$P_2(U) = \operatorname{diag}(0_{3\times 3}, G(U)) \in \mathbb{R}^{(2N+3)\times(2N+3)} \tag{2.1.21}$$

and

$$Q_2(U) = \operatorname{diag}(0_{3\times 3}, H(U)) \in \mathbb{R}^{(2N+3)\times(2N+3)} \tag{2.1.22}$$

with

$$G(U) = (g_{ij})_{1\le i,j\le N} \in \mathbb{R}^{2N\times 2N}, \quad H(U) = (h_{ij})_{1\le i,j\le N} \in \mathbb{R}^{2N\times 2N} \tag{2.1.23}$$

and

$$g_{ij}(U) = \begin{pmatrix} \sum_k B_{ijk}\alpha_k & 0 \\ \sum_k B_{ijk}\beta_k & 0 \end{pmatrix} \in \mathbb{R}^{2\times 2}, \quad h_{ij}(U) = \begin{pmatrix} 0 & \sum_k B_{ijk}\alpha_k \\ 0 & \sum_k B_{ijk}\beta_k \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{2.1.24}$$

This system (2.1.12) is called the shallow water moment equations (SWME).

In the one-dimensional case, the SWME (2.1.12) with $N \geq 2$ is not globally hyperbolic. In [21], the author proposed to linearize the system matrix around linear deviations from equilibrium/constant velocity.

$$\partial_t U + A_H(U)\partial_x U + B_H(U)\partial_y U = S(U), \tag{2.1.25}$$

with

$$A_H(U) := A(h,\, hu_m,\, hv_m,\, h\alpha_1,\, h\beta_1,\, 0,\, 0,\, \cdots,\, 0,\, 0),$$

and

$$B_H(U) := B(h,\, hu_m,\, hv_m,\, h\alpha_1,\, h\beta_1,\, 0,\, 0,\, \cdots,\, 0,\, 0).$$

Keeping α_1 allows to capture a large part of the structure despite its simplicity. For example, there will still be a coupling between the different higher order equations. In 1D, it is proved to be hyperbolic in [21]. This system is called the hyperbolic shallow water moment equations (HSWME).

CHAPTER 3

ANALYSIS OF ROTATIONAL INVARIANCE

In this section, we present two approaches to prove the rotational invariance of the SWME (2.1.12) and the HSWME (2.1.25). The first approach is to first show that the SWME (2.1.12) is indeed rotational invariant by decomposing the convection term into conservative part and non-conservative part. The second approach is to exploit the systems block structure. Motivated by the second approach, we propose the general closure relation such that the rotational invariance is satisfied.

3.1 Rotational invariance of the SWME and the HSWME

In this part, we show that the SWME (2.1.12) is invariant under the rotation of the coordinate system. We first introduce the definition of rotational invariance, which guarantees that the form of the system remains unchanged under a new rotated coordinate system.

Definition 3.1.1 (rotational invariance)**.** *Consider the first-order system*

$$\partial_t U + A(U)\partial_x U + B(U)\partial_y U = S(U) \tag{3.1.1}$$

with $U = (h,\, hu,\, hv,\, h\alpha_1,\, h\beta_1,\, \cdots,\, h\alpha_N,\, h\beta_N)^T \in \mathbb{R}^{2N+3}$*. It is said to satisfy the rotational invariance property if the following relation holds:*

$$\cos\theta\, A(U) + \sin\theta\, B(U) = T^{-1} A(TU) T, \tag{3.1.2}$$

for any angle $0 \le \theta < 2\pi$ *and any vector* U*. Here* $T = T(\theta)$ *is the rotation matrix given by*

$$T(\theta) = diag(1, T_2(\theta), T_2(\theta), \cdots, T_2(\theta)) \in \mathbb{R}^{(2N+3)\times(2N+3)} \tag{3.1.3}$$

with

$$T_2(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{3.1.4}$$

Since the coefficient matrices in the SWME (2.1.12) can be split into the conservative part and non-conservative part in (2.1.14), we will prove the rotational invariance property in two steps. Before the proof, we prepare a set of equalities used in the proof.

Proposition 3.1.1. *For $u, v, \alpha, \beta, \theta \in \mathbb{R}$, we introduce the rotated variables in the new coordinate system:*

$$u_\theta := \cos\theta\, u + \sin\theta\, v, \quad v_\theta := -\sin\theta\, u + \cos\theta\, v, \tag{3.1.5}$$

and

$$\alpha_\theta := \cos\theta\, \alpha + \sin\theta\, \beta, \quad \beta_\theta := -\sin\theta\, \alpha + \cos\theta\, \beta. \tag{3.1.6}$$

Then the following equalities hold true:

$$\cos\theta\, u_\theta - \sin\theta\, v_\theta = u, \tag{3.1.7}$$

$$\sin\theta\, u_\theta + \cos\theta\, v_\theta = v, \tag{3.1.8}$$

$$\cos\theta\, (u_\theta)^2 - \sin\theta\, u_\theta v_\theta = u u_\theta, \tag{3.1.9}$$

$$\sin\theta\, (u_\theta)^2 + \cos\theta\, u_\theta v_\theta = v u_\theta, \tag{3.1.10}$$

$$2\cos\theta\, u_\theta \alpha_\theta - \sin\theta\, (u_\theta \beta_\theta + v_\theta \alpha_\theta) = 2\cos\theta\, u\alpha + \sin\theta\, (u\beta + v\alpha), \tag{3.1.11}$$

$$2\sin\theta\, u_\theta \alpha_\theta + \cos\theta\, (u_\theta \beta_\theta + v_\theta \alpha_\theta) = \cos\theta\, (u\beta + v\alpha) + 2\sin\theta\, v\beta. \tag{3.1.12}$$

Proof. See the proof in Appendix A.2. □

We first prove the rotational invariance for the conservative part:

Lemma 3.1.1 (rotational invariance for conservative part)**.** *The conservative part in* (2.1.12) *satisfies the rotational invariance:*

$$\cos\theta\, F(U) + \sin\theta\, G(U) = T^{-1} F(TU). \tag{3.1.13}$$

for any θ and U.

Proof. We first compute TU:

$$
TU = \begin{pmatrix}
1 & & & & & & & \\
& \cos\theta & \sin\theta & & & & & \\
& -\sin\theta & \cos\theta & & & & & \\
& & & \cos\theta & \sin\theta & & & \\
& & & -\sin\theta & \cos\theta & & & \\
& & & & & \ddots & & \\
& & & & & & \cos\theta & \sin\theta \\
& & & & & & -\sin\theta & \cos\theta
\end{pmatrix}
\begin{pmatrix} h \\ hu \\ hv \\ h\alpha_1 \\ h\beta_1 \\ \vdots \\ h\alpha_N \\ h\beta_N \end{pmatrix}
$$

$$
= \begin{pmatrix}
h \\
h(\cos\theta u + \sin\theta v) \\
h(-\sin\theta u + \cos\theta v) \\
h(\cos\theta\alpha_1 + \sin\theta\beta_1) \\
h(-\sin\theta\alpha_1 + \cos\theta\beta_1) \\
\vdots \\
h(\cos\theta\alpha_N + \sin\theta\beta_N) \\
h(-\sin\theta\alpha_N + \cos\theta\beta_N)
\end{pmatrix}
= \begin{pmatrix}
h \\
hu_\theta \\
hv_\theta \\
h(\alpha_1)_\theta \\
h(\beta_1)_\theta \\
\vdots \\
h(\alpha_N)_\theta \\
h(\beta_N)_\theta
\end{pmatrix}.
$$

Here, for convenience, we introduce the notation

$$
u_\theta := \cos\theta\, u + \sin\theta\, v, \quad v_\theta := -\sin\theta\, u + \cos\theta\, v, \tag{3.1.14}
$$

and

$$
(\alpha_i)_\theta := \cos\theta\, \alpha_i + \sin\theta\, \beta_i, \quad (\beta_i)_\theta := -\sin\theta\, \alpha_i + \cos\theta\, \beta_i, \quad i = 1, \cdots, N. \tag{3.1.15}
$$

Next, we compute $F(TU)$:

$$
\begin{aligned}
F(TU) = (&hu_\theta,\, h(u_\theta^2 + \sum_j \frac{(\alpha_1)_\theta^2}{2j+1}) + \frac{1}{2}gh^2,\, h(u_\theta v_\theta + \sum_j \frac{(\alpha_j)_\theta(\beta_j)_\theta}{2j+1}),\\
&h(2u_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\alpha_k)_\theta),\, h(u_\theta(\beta_1)_\theta + v_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\beta_k)_\theta),\\
&\cdots,\\
&h(2u_\theta(\alpha_n)_\theta + \sum_{j,k} A_{njk}(\alpha_j)_\theta(\alpha_k)_\theta),\, h(u_\theta(\beta_n)_\theta + v_\theta(\alpha_n)_\theta + \sum_{j,k} A_{njk}(\alpha_j)_\theta(\beta_k)_\theta))^T,
\end{aligned}
$$

Then we compute $T^{-1}F(TU)$ for each component and prove that it is equal to the LHS in (3.1.13). Notice that

$$
T^{-1} = \begin{pmatrix}
1 & & & & & & & \\
 & \cos\theta & -\sin\theta & & & & & \\
 & \sin\theta & \cos\theta & & & & & \\
 & & & \cos\theta & -\sin\theta & & & \\
 & & & \sin\theta & \cos\theta & & & \\
 & & & & & \ddots & & \\
 & & & & & & \cos\theta & -\sin\theta \\
 & & & & & & \sin\theta & \cos\theta
\end{pmatrix}. \tag{3.1.16}
$$

We start with the first component in (3.1.13):

$$
\text{RHS} = hu_\theta = h(\cos\theta\, u + \sin\theta\, v) = \cos\theta\, hu + \sin\theta\, hv = \text{LHS}.
$$

Next, we compute the second component in (3.1.13):

$$\begin{aligned}
\text{RHS} &= \cos\theta\left(h(u_\theta^2+\sum_j\frac{(\alpha_1)_\theta^2}{2j+1})+\frac{1}{2}gh^2\right)-\sin\theta h(u_\theta v_\theta+\sum_j\frac{(\alpha_j)_\theta(\beta_j)_\theta}{2j+1})\\
&= h\left(\cos\theta u_\theta^2-\sin\theta u_\theta v_\theta\right)+\cos\theta\frac{1}{2}gh^2+h\sum_j\frac{1}{2j+1}\left(\cos\theta(\alpha_j)_\theta^2-\sin\theta(\alpha_j)_\theta(\beta_j)_\theta\right)\\
&\overset{(3.1.9)}{=} huu_\theta+\cos\theta\frac{1}{2}gh^2+h\sum_j\frac{1}{2j+1}\alpha_j(\alpha_j)_\theta\\
&= hu(\cos\theta u+\sin\theta v)+\cos\theta\frac{1}{2}gh^2+h\sum_j\frac{1}{2j+1}\alpha_j(\cos\theta\alpha_j+\sin\theta\beta_j)\\
&= \cos\theta\left(h(u^2+\sum_j\frac{\alpha_j^2}{2j+1})+\frac{1}{2}gh^2\right)+\sin\theta\left(h(uv+\sum_j\frac{\alpha_j\beta_j}{2j+1})\right)\\
&= \text{LHS}.
\end{aligned}$$

The third component in the RHS of (3.1.13) is:

$$\begin{aligned}
\text{RHS} &= \sin\theta\left(h(u_\theta^2+\sum_j\frac{(\alpha_j)_\theta^2}{2j+1})+\frac{1}{2}gh^2\right)+\cos\theta h(u_\theta v_\theta+\sum_j\frac{(\alpha_j)_\theta(\beta_j)_\theta}{2j+1})\\
&= h\left(\sin\theta u_\theta^2+\cos\theta u_\theta v_\theta\right)+\sin\theta\frac{1}{2}gh^2+h\sum_j\frac{1}{2j+1}\left(\sin\theta(\alpha_j)_\theta^2+\cos\theta(\alpha_j)_\theta(\beta_j)_\theta\right)\\
&\overset{(3.1.10)}{=} hvu_\theta+\sin\theta\frac{1}{2}gh^2+h\sum_j\frac{1}{2j+1}\beta_j(\alpha_j)_\theta\\
&= hv(\cos\theta u+\sin\theta v)+h\sum_j\frac{1}{2j+1}\beta_j(\cos\theta\alpha_j+\sin\theta\beta_j)+\sin\theta\frac{1}{2}gh^2\\
&= \cos\theta\left(h(uv+\sum_j\frac{\alpha_j\beta_j}{2j+1})\right)+\sin\theta\left(h(v^2+\sum_j\frac{\beta_j^2}{2j+1})+\frac{1}{2}gh^2\right)\\
&= \text{LHS}.
\end{aligned}$$

The fourth component in the RHS of (3.1.13) is

$$\begin{aligned}
\text{RHS} &= \cos\theta h \left(2u_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\alpha_k)_\theta) \right) - \sin\theta h \left(u_\theta(\beta_1)_\theta + v_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\beta_k)_\theta) \right) \\
&= h\left(2\cos\theta u_\theta(\alpha_1)_\theta - \sin\theta(u_\theta(\beta_1)_\theta + v_\theta(\alpha_1)_\theta)\right) + h\sum_{j,k} A_{1jk}(\alpha_j)_\theta \left(\cos\theta(\alpha_k)_\theta - \sin\theta(\beta_k)_\theta\right) \\
&\overset{(3.1.11),(3.1.7)}{=} h\left(2\cos\theta u\alpha_1 + \sin\theta(u\beta_1 + v\alpha_1)\right) + h\sum_{j,k} A_{1jk}(\alpha_j)_\theta\alpha_k \\
&= h\left(2\cos\theta u\alpha_1 + \sin\theta(u\beta_1 + v\alpha_1)\right) + h\sum_{j,k} A_{1jk}(\cos\theta\alpha_j + \sin\theta\beta_j)\alpha_k \\
&= \cos\theta h(2u\alpha_1 + \sum_{j,k} A_{1jk}\alpha_j\alpha_k) + \sin\theta h((u\beta_1 + v\alpha_1) + \sum_{j,k} A_{1jk}\alpha_j\beta_k) \\
&= \text{LHS}.
\end{aligned}$$

Then we compute the fifth component:

$$\begin{aligned}
\text{RHS} &= \sin\theta h \left(2u_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\alpha_k)_\theta) \right) + \cos\theta h \left(u_\theta(\beta_1)_\theta + v_\theta(\alpha_1)_\theta + \sum_{j,k} A_{1jk}(\alpha_j)_\theta(\beta_k)_\theta) \right) \\
&= h\left(2\sin\theta u_\theta(\alpha_1)_\theta + \cos\theta(u_\theta(\beta_1)_\theta + v_\theta(\alpha_1)_\theta)\right) + h\sum_{j,k} A_{1jk}(\alpha_j)_\theta \left(\sin\theta(\alpha_k)_\theta + \cos\theta(\beta_k)_\theta\right) \\
&\overset{(3.1.12),(3.1.8)}{=} h\left(\cos\theta(u\beta_1 + v\alpha_1) + 2\sin\theta v\beta_1\right) + h\sum_{j,k} A_{1jk}(\alpha_j)_\theta\beta_k \\
&= h(\cos\theta(u\beta_1 + v\alpha_1) + 2\sin\theta v\beta_1) + h\sum_{j,k} A_{1jk}(\cos\theta\alpha_j + \sin\theta\beta_j)\beta_k \\
&= \cos\theta h((u\beta_1 + v\alpha_1) + \sum_{j,k} A_{1jk}\alpha_j\beta_k) + \sin\theta h(2v\beta_1 + \sum_{j,k} A_{1jk}\beta_j\beta_k) \\
&= \text{LHS}.
\end{aligned}$$

For the remaining components, the proof is similar to the fourth and fifth ones and the details are omitted here. □

Next, we prove the rotational invariance for the non-conservative part:

Lemma 3.1.2 (rotational invariance for non-conservative part)**.** *The non-conservative part in* (2.1.12) *satisfies the rotational invariance:*

$$\cos\theta\, P(U) + \sin\theta\, Q(U) = T^{-1}P(TU)T, \tag{3.1.17}$$

for any θ and U.

Proof. The matrix for the non-conservative part in (2.1.17) consists of two parts. We start the proof with the first part. Notice that $T(\theta)$ has the same block diagonal structure as $P_1(U)$ in (2.1.18) and $Q_1(U)$ in (2.1.19). Therefore, it suffices to check that

$$\cos\theta\, p(U) + \sin\theta\, q(U) = T_2^{-1} p(TU) T_2, \tag{3.1.18}$$

where the matrices $p(U)$ and $q(U)$ are defined in (2.1.20) and T_2 defined in (3.1.4). This equality can be easily verified by direct calculations.

The proof of the second part $P_2(U)$ in (2.1.21) and $Q_2(U)$ in (2.1.22) follows similarly by using the multiplication of the block matrices and the rotational invariance property of each sub-block matrices $g_{ij}(U)$ and $h_{ij}(U)$ defined in (2.1.24). □

Combining the above two lemmas, we have the following theorem:

Theorem 3.1.1 (rotational invariance of SWME)**.** *The SWME* (2.1.12) *satisfies the rotational invariance:*

$$\cos\theta\, A(U) + \sin\theta\, B(U) = T^{-1} A(TU) T. \tag{3.1.19}$$

Proof. Taking the derivative with respect to U on both sides of (3.1.13) in Lemma 3.1.1, we have

$$\cos\theta\, \partial_U F(U) + \sin\theta\, \partial_U G(U) = T^{-1} \partial_U F(TU) T. \tag{3.1.20}$$

Combining this with (3.1.17) in Lemma 3.1.2 , one immediately obtains

$$\cos\theta\, A(U) + \sin\,\theta B(U) = T^{-1} A(TU) T. \tag{3.1.21}$$

□

Since the HSWME (2.1.25) is obtained by evaluating the coefficient matrices in the SWME (2.1.12) at $\alpha_i = \beta_i = 0$ for $2 \le i \le N$, its rotational invariance follows immediately:

Theorem 3.1.2 (rotational invariance of HSWME)**.** *The HSWME* (2.1.25) *satisfies the rotational invariance:*

$$\cos\theta\, A_H(U) + \sin\theta\, B_H(U) = T^{-1} A_H(TU) T. \tag{3.1.22}$$

3.2 Other proof of rotational invariance of the SWME and the HSWME

In the previous part, we prove the rotational invariance of the HSWME (2.1.25) by first proving the rotational invariance of the SWME (2.1.12). In this part, we will show an alternative proof of the rotational invariance of the HSWME (2.1.25) with the aid of its block structure.

We first show the explicit form of the coefficient matrices of the HSWME (2.1.25). This is also given in Theorem 4.3.1 and Theorem 4.3.2 in [26] with another order of variables. For completeness, we include the result and the proof here.

Lemma 3.2.1 (coefficient matrices of HSWME)**.** *The coefficient matrices of the HSWME* (2.1.25) *are given by:*

$$A_H = \begin{pmatrix}
0 & 1 & 0 & & & & & & & \\
-u^2 - \frac{\alpha_1^2}{3} + gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-2u\alpha_1 & 2\alpha_1 & 0 & u & 0 & \frac{3}{5}\alpha_1 & 0 & \cdots & 0 & 0 \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & 0 & u & \frac{1}{5}\beta_1 & \frac{2}{5}\alpha_1 & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1^2 & 0 & 0 & \frac{1}{3}\alpha_1 & 0 & u & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & -\frac{1}{3}\beta_1 & \frac{2}{3}\alpha_1 & 0 & u & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N+1}{2N+1}\alpha_1 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{1}{2N+1}\beta_1 & \frac{N}{2N+1}\alpha_1 \\
 & & & & & & \frac{N-1}{2N-1}\alpha_1 & 0 & u & 0 \\
 & & & & & & -\frac{1}{2N-1}\beta_1 & \frac{N}{2N-1}\alpha_1 & 0 & u
\end{pmatrix} \tag{3.2.1}$$

in the x direction and

$$B_H = \begin{pmatrix}
0 & 0 & 1 & & & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-v^2 - \frac{\beta_1^2}{3} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & & & & & \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & v & 0 & \frac{2}{5}\beta_1 & \frac{1}{5}\alpha_1 & & & \\
-2v\beta_1 & 0 & 2\beta_1 & 0 & v & 0 & \frac{3}{5}\beta_1 & & & \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & \frac{2}{3}\beta_1 & -\frac{1}{3}\alpha_1 & v & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\beta_1^2 & 0 & 0 & 0 & \frac{1}{3}\beta_1 & 0 & v & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N}{2N+1}\beta_1 & \frac{1}{2N+1}\alpha_1 \\
 & & & & & \ddots & \ddots & \ddots & 0 & \frac{N+1}{2N+1}\beta_1 \\
 & & & & & & \frac{N}{2N-1}\beta_1 & -\frac{1}{2N-1}\alpha_1 & v & 0 \\
 & & & & & & 0 & \frac{N-1}{2N-1}\beta_1 & 0 & v
\end{pmatrix} \tag{3.2.2}$$

in the y direction.

Proof. See the proof in Appendix A.3. □

Now we present the alternative proof of Theorem 3.1.2 by using the coefficient matrices of the HSWME (2.1.25) given in Lemma 3.2.1.

Alternative proof of Theorem 3.1.2. The coefficient matrix A_H in (3.2.1) can be written as in the block form:

$$A_H = \begin{pmatrix}
0 & d_1 & & & & \\
d_2 & A_{11} & A_{12} & & & \\
d_3 & A_{21} & A_{22} & A_{23} & & \\
d_4 & & \ddots & \ddots & \ddots & \\
 & & & A_{N,N-1} & A_{N,N} & A_{N,N+1} \\
 & & & & A_{N+1,N} & A_{N+1,N+1}
\end{pmatrix} \tag{3.2.3}$$

with for $i > 1$

$$d_1 = (1,0), \quad d_2 = \begin{pmatrix} -u^2 - \frac{\alpha_1^2}{3} + gh \\ -uv - \frac{\alpha_1\beta_1}{3} \end{pmatrix}, \quad d_3 = \begin{pmatrix} -2u\alpha_1 \\ -(u\beta_1 + v\alpha_1) \end{pmatrix}, \quad d_4 = \begin{pmatrix} -\frac{2}{3}\alpha_1^2 \\ -\frac{2}{3}\alpha_1\beta_1 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} 2u & 0 \\ v & u \end{pmatrix}, \quad A_{21} = \begin{pmatrix} 2\alpha_1 & 0 \\ \beta_1 & \alpha_1 \end{pmatrix}$$

$$A_{ii} = \begin{pmatrix} u & 0 \\ 0 & u \end{pmatrix}, \quad A_{i,i+1} = \begin{pmatrix} \frac{i+1}{2i+1}\alpha_1 & 0 \\ \frac{1}{2i+1}\beta_1 & \frac{i}{2i+1}\alpha_1 \end{pmatrix}, \quad A_{i+1,i} = \begin{pmatrix} \frac{i-1}{2i-1}\alpha_1 & 0 \\ \frac{-1}{2i-1}\beta_1 & \frac{i}{2i-1}\alpha_1 \end{pmatrix}.$$

The coefficient matrix B_H in (3.2.2) can be written as in the block form similarly:

$$B_H = \begin{pmatrix} 0 & f_1 & & & & \\ f_2 & B_{11} & B_{12} & & & \\ f_3 & B_{21} & B_{22} & B_{23} & & \\ f_4 & & \ddots & \ddots & \ddots & \\ & & & B_{N,N-1} & B_{N,N} & B_{N,N+1} \\ & & & & B_{N+1,N} & B_{N+1,N+1} \end{pmatrix} \tag{3.2.4}$$

with

$$f_1 = (0,1), \quad f_2 = \begin{pmatrix} -uv - \frac{\alpha_1\beta_1}{3} \\ -v^2 - \frac{\beta_1^2}{3} + gh \end{pmatrix}, \quad f_3 = \begin{pmatrix} -(u\beta_1 + v\alpha_1) \\ -2v\beta_1 \end{pmatrix}, \quad f_4 = \begin{pmatrix} -\frac{2}{3}\alpha_1\beta_1 \\ -\frac{2}{3}\beta_1^2 \end{pmatrix},$$

$$B_{11} = \begin{pmatrix} v & u \\ 0 & 2v \end{pmatrix}, \quad B_{ii} = \begin{pmatrix} v & 0 \\ 0 & v \end{pmatrix}$$

$$B_{21} = \begin{pmatrix} \beta_1 & \alpha_1 \\ 0 & 2\beta_1 \end{pmatrix}, \quad B_{i,i+1} = \begin{pmatrix} \frac{i}{2i+1}\beta_1 & \frac{1}{2i+1}\alpha_1 \\ 0 & \frac{i+1}{2i+1}\beta_1 \end{pmatrix}, \quad B_{i+1,i} = \begin{pmatrix} \frac{i}{2i-1}\beta_1 & \frac{-1}{2i-1}\alpha_1 \\ 0 & \frac{i-1}{2i-1}\beta_1 \end{pmatrix}.$$

Next, we compute $T^{-1}A_H(TU)T$ and verify that it is equal to $\cos\theta A_H(U) + \sin\theta B_H(U)$.

$$T^{-1}A_H(TU)T = \mathrm{diag}(1, T_2^{-1}, \cdots, T_2^{-1}) \begin{pmatrix} 0 & d_1 & & & & & \\ d_2 & A_{11} & A_{12} & & & & \\ d_3 & A_{21} & A_{22} & A_{23} & & & \\ d_4 & & \ddots & \ddots & \ddots & & \\ & & & A_{N,N-1} & A_{N,N} & A_{N,N+1} \\ & & & & A_{N+1,N} & A_{N+1,N+1} \end{pmatrix} T$$

$$= \begin{pmatrix} 0 & d_1 & & & & & \\ T_2^{-1}d_2 & T_2^{-1}A_{11} & T_2^{-1}A_{12} & & & & \\ T_2^{-1}d_3 & T_2^{-1}A_{21} & T_2^{-1}A_{22} & T_2^{-1}A_{23} & & & \\ T_2^{-1}d_4 & & \ddots & \ddots & \ddots & & \\ & & & T_2^{-1}A_{N,N-1} & T_2^{-1}A_{N,N} & T_2^{-1}A_{N,N+1} \\ & & & & T_2^{-1}A_{N+1,N} & T_2^{-1}A_{N+1,N+1} \end{pmatrix} T$$

$$= \begin{pmatrix} 0 & d_1T_2 & & & & & \\ T_2^{-1}d_2 & T_2^{-1}A_{11}T_2 & T_2^{-1}A_{12}T_2 & & & & \\ T_2^{-1}d_3 & T_2^{-1}A_{21}T_2 & T_2^{-1}A_{22}T_2 & T_2^{-1}A_{23}T_2 & & & \\ T_2^{-1}d_4 & & \ddots & \ddots & \ddots & & \\ & & & T_2^{-1}A_{N,N-1}T_2 & T_2^{-1}A_{N,N}T_2 & T_2^{-1}A_{N,N+1}T_2 \\ & & & & T_2^{-1}A_{N+1,N}T_2 & T_2^{-1}A_{N+1,N+1}T_2 \end{pmatrix}$$

where the independent variable TU in d_i for $1 \le i \le 4$ and A_{ij} for $1 \le i, j \le N+1$ is omitted for simplicity.

Now we compute each block of the matrix:

$$d_1T_2 = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \end{pmatrix} = \cos\theta\, d_1 + \sin\theta\, f_1. \tag{3.2.5}$$

$$T_2^{-1}d_2(TU) = \cos\theta\, d_2(U) + \sin\theta\, f_2(U). \tag{3.2.6}$$

$$T_2^{-1}d_3(TU) = \cos\theta\, d_3(U) + \sin\theta\, f_3(U). \tag{3.2.7}$$

$$T_2^{-1} d_4(TU) = \cos\theta\, d_4(U) + \sin\theta\, f_4(U). \tag{3.2.8}$$

$$\begin{aligned} T_2^{-1} A_{ii} T_2 &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 2u & 0 \\ v & u \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\ &= \cos\theta \begin{pmatrix} 2u & 0 \\ v & u \end{pmatrix} + \sin\theta \begin{pmatrix} v & u \\ 0 & 2v \end{pmatrix} \\ &= \cos\theta A_{ii} + \sin\theta B_{ii}. \end{aligned} \tag{3.2.9}$$

$$T_2^{-1} A_{i,i+1} T_2 = \cos\theta A_{i,i+1} + \sin\theta B_{i,i+1}. \tag{3.2.10}$$

$$T_2^{-1} A_{i+1,i} T_2 = \cos\theta A_{i+1,i} + \sin\theta B_{i+1,i}. \tag{3.2.11}$$

Therefore, we have proved

$$\cos\theta\, A_H(U) + \sin\theta\, B_H(U) = T^{-1} A_H(TU) T. \tag{3.2.12}$$

□

Remark 3.2.1. *The rotational invariance of the SWME* (2.1.12) *can also be proved in the similar line as the above proof by using the block structure of the coefficient matrices* (A.3.24) *and* (A.3.44) *explicitly. We omit it here for space considerations.*

3.3 General closure relation with the rotational invariance

From the above alternative proof, we observe that the rotational invariance of the HSWME relies on the rotational invariance of each sub-block (of size 2×2) of the coefficient matrices. Motivated by this observation, we would like to analyze the rotational invariance of matrices of size 2×2 and find out some general relations which satisfy the rotational invariance. We note that this will be key in deriving our new model, which has a more general closure relation than that of HSWME.

Definition 3.3.1 (rotational invariance of 2×2 matrices)**.** *Consider two matrices $A(V)$ and $B(V)$ of size 2×2 given by*

$$A(V) = \begin{pmatrix} a_{11}(V) & a_{12}(V) \\ a_{21}(V) & a_{22}(V) \end{pmatrix} \in \mathbb{R}^{2\times 2}, \tag{3.3.1}$$

and

$$B(V) = \begin{pmatrix} b_{11}(V) & b_{12}(V) \\ b_{21}(V) & b_{22}(V) \end{pmatrix} \in \mathbb{R}^{2\times 2}, \tag{3.3.2}$$

with $V = (p,\, q)^T \in \mathbb{R}^2$. *We say that* $A(V)$ *and* $B(V)$ *satisfy the rotational invariance if*

$$T_2^{-1} A(T_2 V) T_2 = \cos\theta\, A(V) + \sin\theta\, B(V), \tag{3.3.3}$$

for any $0 \le \theta < 2\pi$ *and any* $V \in \mathbb{R}^2$ *with* T_2 *being the rotational matrix*

$$T_2 = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}. \tag{3.3.4}$$

Note that, in the definition, the dummy variables (p, q) could be (u, v) or (α_i, β_i) for $1 \le i \le N$ in the shallow water moment model. In the following part, we will derive some conditions for the rotational invariance of 2×2 matrices to be satisfied. We first present several necessary conditions:

Lemma 3.3.1. *If the matrices* $A(V)$ *and* $B(V)$ *given by*

$$A(V) = \begin{pmatrix} a_{11}(V) & a_{12}(V) \\ a_{21}(V) & a_{22}(V) \end{pmatrix}, \quad B(V) = \begin{pmatrix} b_{11}(V) & b_{12}(V) \\ b_{21}(V) & b_{22}(V) \end{pmatrix} \tag{3.3.5}$$

satisfy the rotational invariance of 2×2 *matrices, then the following relations hold:*

1. *All the entries in* $B(V)$ *are determined by* $A(V)$ *in the following way:*

$$\begin{aligned} b_{11}(p,q) &= a_{22}(q,-p), \\ b_{12}(p,q) &= -a_{21}(q,-p), \\ b_{21}(p,q) &= -a_{12}(q,-p), \\ b_{22}(p,q) &= a_{11}(q,-p). \end{aligned} \tag{3.3.6}$$

2. $A(V)$ *is an odd function in the sense that:*

$$A(-V) = -A(V). \tag{3.3.7}$$

Proof. 1. By taking $\theta = \frac{\pi}{2}$ in (3.3.3), we have

$$T_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and

$$T_2 V = \begin{pmatrix} q \\ -p \end{pmatrix}.$$

Therefore, we have

$$T_2^{-1} A(T_2 V) T_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_{11}(TV) & a_{12}(TV) \\ a_{21}(TV) & a_{22}(TV) \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} a_{22}(TV) & -a_{21}(TV) \\ -a_{12}(TV) & a_{11}(TV) \end{pmatrix}.$$

Then the relation (3.3.3) reduces to

$$\begin{pmatrix} b_{11}(V) & b_{12}(V) \\ b_{21}(V) & b_{22}(V) \end{pmatrix} = \begin{pmatrix} a_{22}(TV) & -a_{21}(TV) \\ -a_{12}(TV) & a_{11}(TV) \end{pmatrix},$$

which completes the proof.

2. Taking θ to be $(\theta + \pi)$ in (3.3.3), we have

$$\cos(\theta + \pi) A(V) + \sin(\theta + \pi) B(V) = (-T_2)^{-1} A(-T_2 V)(-T_2)$$

which implies

$$-\cos(\theta) A(V) - \sin(\theta) B(V) = T_2^{-1} A(-T_2 V) T_2.$$

Comparing the above equation with (3.3.3), we have

$$A(T_2 V) = -A(-T_2 V).$$

Since T_2 is invertible, we have that A is an odd function.

□

Next, we will restrict to the case of linear functions and find out the necessary and sufficient conditions for the matrices to be rotational invariant. The conditions will be presented in Thereom 3.3.1. To prove the theorem, we prepare the following lemma:

Lemma 3.3.2. *Assume that $A(V)$ and $B(V)$ satisfy the rotational invariance of 2×2 matrices and they are linear functions of V.*

1. *If $A(V)$ only has two non-zero entries in the first column:*

$$A(V) = \begin{pmatrix} a_{11}(V) & 0 \\ a_{21}(V) & 0 \end{pmatrix}, \tag{3.3.8}$$

then $A(V)$ and $B(V)$ have to be of the form:

$$A(V) = c_1 \begin{pmatrix} p & 0 \\ q & 0 \end{pmatrix} + c_2 \begin{pmatrix} q & 0 \\ -p & 0 \end{pmatrix}, \tag{3.3.9}$$

and

$$B(V) = c_1 \begin{pmatrix} 0 & p \\ 0 & q \end{pmatrix} + c_2 \begin{pmatrix} 0 & q \\ 0 & -p \end{pmatrix}, \tag{3.3.10}$$

where $c_1, c_2 \in \mathbb{R}$.

2. *If $A(V)$ only has two non-zero entries in the diagonal:*

$$A(V) = \begin{pmatrix} a_{11}(V) & 0 \\ 0 & a_{22}(V) \end{pmatrix}, \tag{3.3.11}$$

then $A(V)$ and $B(V)$ have to be of the form:

$$A(V) = c_1 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} + c_2 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix}, \tag{3.3.12}$$

and

$$B(V) = c_1 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} - c_2 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}, \tag{3.3.13}$$

where $c_1, c_2 \in \mathbb{R}$.

3. *If $A(V)$ only has two non-zero entries in the second column:*

$$A(V) = \begin{pmatrix} 0 & a_{21}(V) \\ 0 & a_{22}(V) \end{pmatrix}, \tag{3.3.14}$$

then $A(V)$ and $B(V)$ have to be of the form:

$$A(V) = c_1 \begin{pmatrix} 0 & p \\ 0 & q \end{pmatrix} + c_2 \begin{pmatrix} 0 & q \\ 0 & -p \end{pmatrix}, \tag{3.3.15}$$

and

$$B(V) = c_1 \begin{pmatrix} -p & 0 \\ -q & 0 \end{pmatrix} + c_2 \begin{pmatrix} -q & 0 \\ p & 0 \end{pmatrix}, \tag{3.3.16}$$

where $c_1, c_2 \in \mathbb{R}$.

Proof. See the proof in Appendix A.4. □

Theorem 3.3.1. *Assume that the matrices $A(V)$ and $B(V)$ satisfy the rotational invariance of 2×2 matrices and they are linear functions of $V = (p, q)^T$. Then they must be of the form*

$$A(V) = c_1 \begin{pmatrix} p & 0 \\ q & 0 \end{pmatrix} + c_2 \begin{pmatrix} q & 0 \\ -p & 0 \end{pmatrix} + c_3 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} + c_4 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} + c_5 \begin{pmatrix} 0 & p \\ 0 & q \end{pmatrix} + c_6 \begin{pmatrix} 0 & q \\ 0 & -p \end{pmatrix}, \tag{3.3.17}$$

and

$$B(V) = c_1 \begin{pmatrix} 0 & p \\ 0 & q \end{pmatrix} + c_2 \begin{pmatrix} 0 & q \\ 0 & -p \end{pmatrix} + c_3 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix} - c_4 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} - c_5 \begin{pmatrix} p & 0 \\ q & 0 \end{pmatrix} + c_6 \begin{pmatrix} -q & 0 \\ p & 0 \end{pmatrix}, \tag{3.3.18}$$

where $c_i \in \mathbb{R}$ for $1 \leq i \leq 6$.

Proof. Since the matrix $B(V)$ is determined by $A(V)$ by Lemma 3.3.1, we only need to consider the form of $A(V)$.

For any matrix $A(V)$ of the form

$$A(V) = \begin{pmatrix} a_{11}(V) & a_{12}(V) \\ a_{21}(V) & a_{22}(V) \end{pmatrix},$$

we can decompose it as

$$A(V) = \begin{pmatrix} \tilde{a}_{21}(V) & 0 \\ a_{21}(V) & 0 \end{pmatrix} + \begin{pmatrix} a_{11}(V) - \tilde{a}_{21}(V) & a_{12}(V) \\ 0 & a_{22}(V) \end{pmatrix},$$

where $\tilde{a}_{21}(V)$ is a linear function uniquely determined by $a_{21}(V)$ such that

$$P(V) := \begin{pmatrix} \tilde{a}_{21}(V) & 0 \\ a_{21}(V) & 0 \end{pmatrix}$$

satisfies the form of rotational invariance given in Case 1 in Lemma 3.3.2.

Since $A(V)$ and $P(V)$ both satisfy the rotational invariance, we have that the remaining part

$$Q(V) := \begin{pmatrix} a_{11}(V) - \tilde{a}_{21}(V) & a_{12}(V) \\ 0 & a_{22}(V) \end{pmatrix}$$

also satisfies the rotational invariance. Next, we decompose $Q(V)$ into two parts:

$$Q(V) = \begin{pmatrix} a_{11}(V) - \tilde{a}_{21}(V) & 0 \\ 0 & \tilde{a}_{22}(V) \end{pmatrix} + \begin{pmatrix} 0 & a_{12}(V) \\ 0 & a_{22}(V) - \tilde{a}_{22}(V) \end{pmatrix} \tag{3.3.19}$$

where $\tilde{a}_{22}(V)$ is a linear function uniquely determined by $(a_{11}(V) - \tilde{a}_{21}(V))$ such that

$$R(V) := \begin{pmatrix} a_{11}(V) - \tilde{a}_{21}(V) & 0 \\ 0 & \tilde{a}_{22}(V) \end{pmatrix} \tag{3.3.20}$$

satisfies the rotational invariance given in Case 2 in Lemma 3.3.2.

Lastly, the second part in (3.3.19)

$$S(V) := \begin{pmatrix} 0 & a_{12}(V) \\ 0 & a_{22}(V) - \tilde{a}_{22}(V) \end{pmatrix} \tag{3.3.21}$$

must satisfy the rotational invariance and falls into Case 3 in Lemma 3.3.2. The proof is completed.

□

Motivated by the constraint given in Theorem 3.3.1, we can modify the coefficient matrices in the HSWME (2.1.25) in the following way to make it satisfy the rotational invariance property.

Theorem 3.3.2 (general closure relation with rotational invariance)**.** *Suppose that the matrices* $A_{ij}(U) \in \mathbb{R}^{2\times 2}$ *and* $B_{ij}(U) \in \mathbb{R}^{2\times 2}$ *satisfy the rotational invariance for the* 2×2 *matrices for* $1 \le i, j \le N+1$*. Then the matrices* A *and* B *given by*

$$A = A_H + \begin{pmatrix} 0 & & & & \\ & A_{11} & A_{12} & \cdots & A_{1,N+1} \\ & A_{11} & A_{12} & \cdots & A_{1,N+1} \\ & \vdots & \vdots & & \vdots \\ & A_{N+1,1} & A_{N+1,2} & \cdots & A_{N+1,N+1} \end{pmatrix} \tag{3.3.22}$$

in the x *direction and*

$$B = B_H + \begin{pmatrix} 0 & & & & \\ & B_{11} & B_{12} & \cdots & B_{1,N+1} \\ & B_{11} & B_{12} & \cdots & B_{1,N+1} \\ & \vdots & \vdots & & \vdots \\ & B_{N+1,1} & B_{N+1,2} & \cdots & B_{N+1,N+1} \end{pmatrix} \tag{3.3.23}$$

in the y *direction, satisfy the rotational invariance. Here* A_H *and* B_H *are the coefficient matrices in the HSWME* (2.1.25)*.*

Proof. The proof is similar to the alternative proof of Theorem 3.1.2 in Section 3.2. □

Remark 3.3.1. *From Theorem 3.3.2, to preserve the rotational invariance of the moment model, we can modify any entries except the first row and the first column, as long as the sub-blocks satisfy the rotational invariance for the* 2×2 *matrices. However, in practice, we will only modify the last row blocks, i.e.,* $A_{N+1,j}$ *and* $B_{N+1,j}$ *for* $1 \le j \le N+1$*, to guarantee provable hyperbolicity. This will be illustrated in the next section in detail.*

CHAPTER 4

ANALYSIS OF THE HYPERBOLICITY

In this section, we analyze the hyperbolicity of the moment models. With the aid of the rotational invariance, the hyperbolicity in 2D is equivalent to the hyperbolicity in the x direction or y direction. Therefore, it suffices to only analyze the real diagonalizability of the coefficient matrix in x direction. We will first prove the hyperbolicity of the HSWME (2.1.25) in 2D. Then we generalize the β-HSMWE in 1D proposed in [21] to 2D and show its hyperbolicity. Lastly, we propose the general framework for constructing provable hyperbolic moment models with specified propogation speeds.

4.1 Hyperbolicity of the HSWME

In this part, we will prove the hyperbolicity of the HSWME (2.1.25) in 2D. This reduces to check the real diagonalizability of the coefficient matrix A_H in the x direction.

Note that the characteristic polynomial of A_H was analyzed in Theorem 4.3.3 in [26]. However, the proof in [26] only shows that the eigenvalues of A_H are real but not necessarily distinct. Therefore, the proof is incomplete since the real diagonalizability requires not only the real eigenvalues but also a complete set of eigenvectors. In this part, we will prove the real diagonalizability of A_H with the aid of the associated polynomial sequence and show that the eigenvalues are related to the Gauss-Lobatto and Gauss-Legendre quadrature points.

To analyze the hyperbolicity, we use another ordering of variables:

$$W = (h,\, hu,\, h\alpha_1,\, \cdots,\, h\alpha_N,\, hv,\, h\beta_1,\, \cdots,\, h\beta_N)^T. \tag{4.1.1}$$

Note that using different order of variables will not change the rotational invariance or the hyperbolicity of the model, but it does simplify the analysis.

Using this set of variables (4.1.1), the coefficient matrix (3.2.1) in the x direction in HSWME (2.1.25) can be written as

$$\tilde{A}_H(W) = \begin{pmatrix} \tilde{A}_{11}(W) & 0 \\ \tilde{A}_{21}(W) & \tilde{A}_{22}(W) \end{pmatrix}, \tag{4.1.2}$$

where the block matrices $\tilde{A}_{11}(W) \in \mathbb{R}^{(N+2)\times(N+2)}$, $\tilde{A}_{21}(W) \in \mathbb{R}^{(N+1)\times(N+2)}$ and $\tilde{A}_{22}(W) \in \mathbb{R}^{(N+1)\times(N+1)}$ are given by

$$\tilde{A}_{11}(W) = \begin{pmatrix} 0 & 1 & & & & & & \\ gh - u^2 - \frac{1}{3}\alpha_1^2 & 2u & \frac{2}{3}\alpha_1 & & & & & \\ -2u\alpha_1 & 2\alpha_1 & u & \frac{3}{5}\alpha_1 & & & & \\ -\frac{2}{3}\alpha_1^2 & 0 & \frac{1}{3}\alpha_1 & u & \frac{4}{7}\alpha_1 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \frac{N-2}{2N-3}\alpha_1 & u & \frac{N+1}{2N+1}\alpha_1 & \\ & & & & & \frac{N-1}{2N-1}\alpha_1 & u & \end{pmatrix}, \tag{4.1.3}$$

$$\tilde{A}_{21}(W) = \begin{pmatrix} -uv - \frac{\alpha_1\beta_1}{3} & v & \frac{\beta_1}{3} & & & & & \\ -(u\beta_1 + v\alpha_1) & \beta_1 & 0 & \frac{1}{5}\beta_1 & & & & \\ -\frac{2}{3}\alpha_1\beta_1 & 0 & -\frac{1}{3}\beta_1 & 0 & \frac{1}{7}\beta_1 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & -\frac{1}{2N-1}\beta_1 & 0 & \frac{1}{2N+1}\beta_1 & \\ & & & & & -\frac{1}{2N-1}\beta_1 & 0 & \end{pmatrix}, \tag{4.1.4}$$

and

$$\tilde{A}_{22}(W) = \begin{pmatrix} u & \frac{\alpha_1}{3} & & & & & \\ \alpha_1 & u & \frac{2\alpha_1}{5} & & & & \\ & \frac{2}{3}\alpha_1 & u & \frac{3\alpha_1}{7} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \frac{N-1}{2N-3}\alpha_1 & u & \frac{N}{2N+1}\alpha_1 & \\ & & & & \frac{N}{2N-1}\alpha_1 & u & \end{pmatrix}. \tag{4.1.5}$$

Therefore, the characteristic polynomial of $\tilde{A}_H(W)$ is

$$\det(\lambda I - \tilde{A}_H(W)) = \det\begin{pmatrix} \lambda I - \tilde{A}_{11}(W) & 0 \\ -\tilde{A}_{21}(W) & \lambda I - \tilde{A}_{22}(W) \end{pmatrix} = \det(\lambda I - \tilde{A}_{11}(W)) \det(\lambda I - \tilde{A}_{22}(W)).$$

Next, we focus on the analysis of the characteristic polynomial of $\tilde{A}_{11}(W)$ and $\tilde{A}_{22}(W)$.

Notice that both $\tilde{A}_{11}(W)$ and $\tilde{A}_{22}(W)$ are lower Hessenberg matrices. Before the discussion, we review important properties of the Hessenberg matrix. These properties facilitate directly relating the eigenvalues of a Hessenberg matrix to the roots of some associated polynomial. We start with the definitions of the (unreduced) lower Hessenberg matrix and the associated polynomial sequence [7]:

Definition 4.1.1 (lower Hessenberg matrix)**.** *The matrix $A = (a_{ij})_{n\times n}$ is called lower Hessenberg matrix if $a_{ij} = 0$ for $j > i + 1$. It is called unreduced lower Hessenberg matrix if further $a_{i,i+1} \neq 0$ for $i = 1, 2, \cdots, n-1$.*

Definition 4.1.2 (associated polynomial sequence [7])**.** *Let $A = (a_{ij})_{n\times n}$ be an unreduced lower Hessenberg matrix. The associated polynomial sequence $\{q_i\}_{0\leq i\leq n}$ is defined as follows:*

$$\begin{aligned} &q_0(x) = 1, \\ &q_i(x) = \frac{1}{a_{i,i+1}}\left(xq_{i-1}(x) - \sum_{j=1}^{i} a_{ij}q_{j-1}(x)\right), \quad 1 \leq i \leq n, \end{aligned} \tag{4.1.6}$$

with $a_{n,n+1} := 1$.

Theorem 4.1.1 ([7])**.** *Let $A = (a_{ij})_{n\times n}$ be an unreduced lower Hessenberg matrix and $\{q_i\}_{0\leq i\leq n}$ is the associated polynomial sequence with A. The following conclusions hold true:*

1. *If λ is a root of q_n, then λ is an eigenvalue of the matrix A and a corresponding eigenvector is $(q_0(\lambda), q_1(\lambda), \cdots, q_{n-1}(\lambda))^T$.*

2. *If all the roots of q_n are simple, then the characteristic polynomial of A is given by*

$$\det(xI - A) = \rho q_n(x),$$

with $\rho = \Pi_{i=1}^{n-1} a_{i,i+1}$.

With the aid of the associated polynomial sequence, we are able to obtain the analytical form of the characteristic polynomials of $\tilde{A}_{11}$ in (4.1.3) and $\tilde{A}_{22}$ in (4.1.5).

Lemma 4.1.1 (characteristic polynomial of $\tilde{A}_{11}$ in (4.1.3))**.** *The associated polynomial sequence of $\tilde{A}_{11}$ in* (4.1.3) *is given by*

$$\begin{aligned} q_0(x) &= 1, \\ q_1(x) &= x, \\ q_2(x) &= \frac{3(x-u)^2 - 3gh + \alpha_1^2}{2\alpha_1}, \end{aligned}$$

$$q_n(x) = \frac{2n-1}{n(n-1)\alpha_1} P'_{n-1}\left(\frac{x-u}{\alpha_1}\right)\left((x-u)^2 - gh - \alpha_1^2\right), \quad 3 \le n \le N+1, \tag{4.1.7}$$

and

$$q_{N+2}(x) = \frac{1}{N+1} P'_{N+1}(\xi)\left((x-u)^2 - gh - \alpha_1^2\right). \tag{4.1.8}$$

Here $P_n(\xi)$ is the Legendre polynomial on $[-1,1]$ with the standardization condition $P_n(1) = 1$.

Proof. For convenience, we first introduce the notations:

$$\xi := \frac{x-u}{\alpha_1},$$

and

$$p_g(x) := (x-u)^2 - gh - \alpha_1^2.$$

The first several associated polynomials can be obtained by direct computation:

$$q_0(x) = 1.$$

$$q_1(x) = \frac{1}{a_{12}}(xq_0(x) - a_{11}q_0(x)) = \frac{1}{1}(x-0) = x.$$

$$\begin{aligned} q_2(x) &= \frac{1}{a_{23}}(xq_1(x) - (a_{21}q_0(x) + a_{22}q_1(x))) \\ &= \frac{1}{\frac{2}{3}\alpha_1}\left(x^2 - (gh - u^2 - \frac{1}{3}\alpha_1^2 + 2ux)\right) \\ &= \frac{3(x-u)^2 - 3gh + \alpha_1^2}{2\alpha_1}. \end{aligned}$$

$$
\begin{aligned}
q_3(x) &= \frac{1}{a_{34}}\left(xq_2(x) - (a_{31}q_0(x) + a_{32}q_1(x) + a_{33}q_2(x))\right) \\
&= \frac{1}{\frac{3}{5}\alpha_1}\left(x\frac{3(x-u)^2 - 3gh + \alpha_1^2}{2\alpha_1} - (-2u\alpha_1 + 2\alpha_1 x + u\frac{3(x-u)^2 - 3gh + \alpha_1^2}{2\alpha_1})\right) \\
&= \frac{5(x-u)((x-u)^2 - gh - \alpha_1^2)}{2\alpha_1^2} \\
&= \frac{5\xi}{2\alpha_1}p_g(x) \\
&= \frac{5}{6\alpha_1}P_2'(\xi)p_g(x).
\end{aligned}
$$

$$
\begin{aligned}
q_4(x) &= \frac{1}{a_{45}}(xq_3(x) - (a_{41}q_0(x) + a_{42}q_1(x) + a_{43}q_2(x) + a_{44}q_3(x))) \\
&= \frac{1}{\frac{4}{7}\alpha_1}\left((x-u)\frac{5\xi}{2\alpha_1}p_g(x) - (-\frac{2}{3}\alpha_1^2 + 0 + \frac{1}{3}\alpha_1\frac{3(x-u)^2 - 3gh + \alpha_1^2}{2\alpha_1})\right) \\
&= \frac{7(5(x-u)^2 - \alpha_1^2)((x-u)^2 - gh - \alpha_1^2)}{8\alpha_1^3} \\
&= \frac{7(5\xi^2 - 1)}{8\alpha_1}p_g(x) \\
&= \frac{7}{12\alpha_1}P_3'(\xi)p_g(x).
\end{aligned}
$$

Now we prove by induction and assume that (4.1.7) holds for $4 \le n \le k$ with $k \le N$. We will prove it also holds for $n = k+1$.

$$
\begin{aligned}
q_{k+1}(x) &= \frac{1}{a_{k+1,k+2}}\left(xq_k(x) - \sum_{j=1}^{k+1} a_{k+1,j}q_{j-1}(x)\right) \\
&= \frac{1}{a_{k+1,k+2}}\left(xq_k(x) - (a_{k+1,k}q_{k-1}(x) + a_{k+1,k+1}q_k(x))\right) \\
&= \frac{1}{\frac{k+1}{2k+1}\alpha_1}\left((x-u)\frac{2k-1}{k(k-1)\alpha_1}P_{k-1}'(\xi)p_g(x) - \frac{k-2}{2k-3}\alpha_1\frac{2k-3}{(k-1)(k-2)\alpha_1}P_{k-2}'(\xi)p_g(x)\right) \\
&= \frac{2k+1}{(k+1)\alpha_1}\left(\frac{2k-1}{k(k-1)}\xi P_{k-1}'(\xi) - \frac{1}{k-1}P_{k-2}'(\xi)\right)p_g(x) \\
&= \frac{2k+1}{(k+1)\alpha_1}\left(\frac{1}{k(k-1)}((k-1)P_k'(\xi) + kP_{k-2}'(\xi)) - \frac{1}{k-1}P_{k-2}'(\xi)\right)p_g(x) \\
&= \frac{2k+1}{k(k+1)\alpha_1}P_k'(\xi)p_g(x).
\end{aligned}
$$

where we use the relation $(2k-1)xP_{k-1}'(x) = (k-1)P_k'(x) + kP_{k-2}'(x)$.

Lastly, we compute $q_{N+2}(x)$:

$$\begin{aligned} q_{N+2}(x) &= \frac{1}{a_{N+2,N+3}} \left(x q_{N+1}(x) - (a_{N+2,N+1} q_N(x) + a_{N+2,N+2} q_{N+1}(x)) \right) \\ &= x q_{N+1}(x) - (a_{N+2,N+1} q_N(x) + a_{N+2,N+2} q_{N+1}(x)) \\ &= \frac{1}{N+1} P'_{N+1}(\xi) p_g(x). \end{aligned} \tag{4.1.9}$$

The proof is completed. □

Corollary 4.1.1. *Since the roots of $P'_{N+1}(\xi)$ are real and distinct, the characteristic polynomial of $\tilde{A}_{11}$ is*

$$\det(xI - \tilde{A}_{11}) = \frac{N!}{(2N+1)!!} \alpha_1^N P'_{N+1}\left(\frac{x-u}{\alpha_1}\right) \left((x-u)^2 - gh - \alpha_1^2\right). \tag{4.1.10}$$

Threfore, the eigenvalues of $\tilde{A}_{11}$ are given by

$$\lambda_{1,2} = u \pm \sqrt{gh + \alpha_1^2},$$

and

$$\lambda_{i+2} = u + r_i \alpha_1, \quad i = 1, 2, \cdots, N.$$

where r_i with $i = 1, 2, \cdots, N$ are the roots of $P'_{N+1}(\xi)$, i.e. the Gauss-Lobatto quadrature points in $[-1, 1]$.

Lemma 4.1.2 (associated polynomial sequence of $\tilde{A}_{22}$)**.** *For the matrix $\tilde{A}_{22}$ given in* (4.1.5)*, the associated polynomial sequences satisfy:*

$$q_n(x) = (2n+1) P_n\left(\frac{x-u}{\alpha_1}\right), \quad 0 \le n \le N. \tag{4.1.11}$$

and

$$q_{N+1}(x) = (N+1)\alpha_1 P_{N+1}\left(\frac{x-u}{\alpha_1}\right). \tag{4.1.12}$$

Proof. We compute the associated polynomial sequence by recurrence relation:

$$q_0(x) = 1,$$

$$q_1(x) = \frac{1}{a_{12}}(xq_0(x) - a_{11}q_0(x)) = \frac{x-u}{\frac{\alpha_1}{3}} = \frac{3(x-u)}{\alpha_1} = 3\xi = 3P_1(\xi),$$

$$q_2(x) = \frac{1}{a_{23}}\left(xq_1(x) - (a_{21}q_0(x) + a_{22}q_1(x))\right) = \frac{1}{\frac{2}{5}\alpha_1}\left((x-u)\frac{3(x-u)}{\alpha_1} - \alpha_1\right) = \frac{5}{2}(3\xi^2 - 1),$$

$$= 5P_2(\xi) \text{ with } \xi := \tfrac{x-u}{\alpha_1}.$$

Now we prove by induction and assume that the formula (4.1.11) holds for $2 \le n \le k$. We will prove it also holds for $n = k+1$.

$$\begin{aligned}
q_{k+1}(x) &= \frac{1}{a_{k+1,k+2}}\left(xq_k(x) - \sum_{j=1}^{k+1} a_{k+1,j}q_{j-1}(x)\right) \\
&= \frac{1}{a_{k+1,k+2}}\left(xq_k(x) - (a_{k+1,k}q_{k-1}(x) + a_{k+1,k+1}q_k(x))\right) \\
&= \frac{1}{\frac{k+1}{2k+3}\alpha_1}\left((x-u)(2k+1)P_k(\xi) - \frac{k}{2k-1}\alpha_1(2k-1)P_{k-1}(\xi)\right) \\
&= \frac{2k+3}{k+1}\left((2k+1)\xi P_k(\xi) - kP_{k-1}(\xi)\right) \\
&= \frac{2k+3}{k+1}(k+1)P_{k+1}(\xi) \\
&= (2k+3)P_{k+1}(\xi),
\end{aligned}$$

Lastly,

$$q_{N+1}(x) = xq_N(x) - (a_{N+1,N}q_{k-1}(x) + a_{N+1,N+1}q_k(x)) = (N+1)\alpha_1 P_{N+1}\left(\frac{x-u}{\alpha_1}\right).$$

□

Corollary 4.1.2 (characteristic polynomial of $\tilde{A}_{22}$)**.** *The matrix $\tilde{A}_{22}$ is real diagonalizable and its characteristic polynomial is*

$$\det(\lambda I - \tilde{A}_{22}) = \frac{(N+1)!}{(2N+1)!!}\alpha_1^{N+1}P_{N+1}\left(\frac{x-u}{\alpha_1}\right). \tag{4.1.13}$$

Moreover, the eigenvalue of $\tilde{A}_{22}$ is given by

$$\lambda_i = s_i\alpha_1, \quad i = 1, 2, \cdots, N+1, \tag{4.1.14}$$

where s_i for $i = 1, 2, \cdots, N+1$ are the roots of Legendre polynomial $P_{N+1}(\xi)$, i.e. the Gauss-Legendre quadrature points in $[-1, 1]$.

Since the roots of $P_{N+1}(\xi)$ are all distinct, $P'_{N+1}(\xi)$ and $P_{N+1}(\xi)$ have no common roots, we immediately have that all the eigenvalues of A_H are real and distinct. The result is summarized as follows:

Theorem 4.1.2 (real diagonalizability of A_H)**.** *The matrix A_H is real diagonalizable. Its characteristic polynomial is given by:*

$$\det(\lambda I - A_H) = \frac{N!(N+1)!}{((2N+1)!!)^2}\alpha_1^{2N+1} P'_{N+1}\left(\frac{x-u}{\alpha_1}\right) P_{N+1}\left(\frac{x-u}{\alpha_1}\right)\left((x-u)^2 - gh - \alpha_1^2\right). \tag{4.1.15}$$

Moreover, the eigenvalues are given by

$$\lambda_{1,2} = u \pm \sqrt{gh + \alpha_1^2},$$

$$\lambda_{i+2} = u + r_i\alpha_1, \quad i = 1, 2, \cdots, N,$$

$$\lambda_{i+N+2} = u + s_i\alpha_1, \quad i = 1, 2, \cdots, N+1,$$

where r_i for $i = 1, 2, \cdots, N$ are the roots of the derivative of Legendre polynomial $P'_{N+1}(\xi)$ and s_i for $i = 1, 2, \cdots, N+1$ are the roots of Legendre polynomial $P_{N+1}(\xi)$.

Combining the real diagonalizability of A_H in Theorem 4.1.2 with the rotational invariance in Theorem 3.1.2, we have the hyperbolicity of the HSWME (2.1.25) in 2D:

Theorem 4.1.3 (hyperbolicity of the HSWME)**.** *The HSWME model* (2.1.25) *in 2D is hyperbolic.*

Remark 4.1.1. *The analytical form of the eigenvectors can be derived by Theorem 4.1.1. This will be useful in some Riemann solvers.*

4.2 Hyperbolicity of the β-HSWME

In [21], a new version of shallow water moment model, called the β-HSWME, is proposed by modifying the last row of the coefficient matrix, so that predefined propagation speeds can be obtained.

The coefficient matrix of the β-HSWME [21] reads as:

$$\tilde{A}_{\beta,11}(W) = \begin{pmatrix} 0 & 1 & & & & & & \\ gh-u^2-\frac{1}{3}\alpha_1^2 & 2u & \frac{2}{3}\alpha_1 & & & & & \\ -2u\alpha_1 & 2\alpha_1 & u & \frac{3}{5}\alpha_1 & & & & \\ -\frac{2}{3}\alpha_1^2 & 0 & \frac{1}{3}\alpha_1 & u & \frac{4}{7}\alpha_1 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \frac{N-2}{2N-3}\alpha_1 & u & \frac{N+1}{2N+1}\alpha_1 & \\ & & & & & \frac{(N-1)(2N+1)}{(N+1)(2N-1)}\alpha_1 & u \end{pmatrix}. \quad (4.2.1)$$

In Theorem 3.5 in [21], it was shown that the characteristic polynomial of $\tilde{A}_{\beta,11}$ in (4.2.1) is related to the Legendre polynomial by numerical computation up to order $N = 100$. Here, we will prove that this holds true for any $N \geq 1$:

Lemma 4.2.1 (characteristic polynomial of $\tilde{A}_{\beta,11}$)**.** *The characteristic polynomial of* $\tilde{A}_{\beta,11}$ *in* (4.2.1) *is*

$$\det(\lambda I - \tilde{A}_{\beta,11}) = \frac{N!}{(2N-1)!!}\alpha_1^N P_N\left(\frac{x-u}{\alpha_1}\right)\left((x-u)^2 - gh - \alpha_1^2\right). \quad (4.2.2)$$

Proof. Notice that the matrix $\tilde{A}_{\beta,11}$ in (4.2.1) only differs from $\tilde{A}_{11}$ in (4.1.3) in the last row. Thus, the associated polynomial sequences for two matrices are the same for q_i with $0 \leq i \leq N+1$. We only need to compute $q_{N+2}(x)$:

$$\begin{aligned} q_{N+2}(x) &= \frac{1}{a_{N+2,N+3}}(xq_{N+1}(x) - \sum_{j=1}^{N+2} a_{N+2,j}q_{j-1}(x)) \\ &= xq_{N+1}(x) - (a_{N+2,N+1}q_N(x) + a_{N+2,N+2}q_{N+1}(x)) \\ &= (x-u)q_{N+1}(x) - \frac{(N-1)(2N+1)}{(N+1)(2N-1)}q_N(x) \\ &= (x-u)\frac{2N+1}{N(N+1)\alpha_1}P'_N(\xi)p_g(x) - \frac{(N-1)(2N+1)}{(N+1)(2N-1)}\frac{2N-1}{N(N-1)\alpha_1}P'_{N-1}(\xi)p_g(x) \\ &= \frac{2N+1}{N(N+1)}\left(\xi P'_N(\xi) - P'_{N-1}(\xi)\right)p_g(x) \\ &= \frac{2N+1}{N(N+1)}NP_N(\xi)p_g(x) \\ &= \frac{2N+1}{N+1}P_N(\xi)p_g(x). \end{aligned}$$

Here we denote $\xi := \frac{x-u}{\alpha_1}$ and $p_g(x) := (x-u)^2 - gh - \alpha_1^2$ and use the relation $\xi P_n'(\xi) - P_{n-1}'(\xi) = nP_n(\xi)$.

Therefore, the characteristic polynomial of $\tilde{A}_{\beta,11}$ is

$$\det(\lambda I - \tilde{A}_{\beta,11}) = \frac{(N+1)!}{(2N+1)!!}\alpha_1^N \frac{2N+1}{N+1} P_N(\xi) p_g(x) = \frac{N!}{(2N-1)!!}\alpha_1^N P_N(\xi) p_g(x). \quad (4.2.3)$$

□

With the matrix $\tilde{A}_{\beta,11}$ at hand, there is still some degree of freedom to choose the matrix $\tilde{A}_{\beta,22}$ to make the matrix $\tilde{A}_\beta$ hyperbolic. One simple choice is to keep $\tilde{A}_{22}$ unchanged. In this case, the corresponding matrix in x direction is

$$\tilde{A}_\beta = \begin{pmatrix} \tilde{A}_{\beta,11} & 0 \\ \tilde{A}_{\beta,21} & \tilde{A}_{22} \end{pmatrix},$$

where $\tilde{A}_{\beta,21}$ is determined by the rotational invariance constraint given in Theorem 3.3.1:

$$\tilde{A}_{\beta,21}(W) = \begin{pmatrix} -uv - \frac{\alpha_1\beta_1}{3} & v & \frac{\beta_1}{3} & & & & \\ -(u\beta_1 + v\alpha_1) & \beta_1 & 0 & \frac{1}{5}\beta_1 & & & \\ -\frac{2}{3}\alpha_1\beta_1 & 0 & -\frac{1}{3}\beta_1 & 0 & \frac{1}{7}\beta_1 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -\frac{1}{2N-3}\beta_1 & 0 & -\frac{1}{2N+1}\beta_1 \\ & & & & & \frac{N^2}{(2N-1)(N+1)}\beta_1 & 0 \end{pmatrix}. \quad (4.2.4)$$

We can also write the coefficient matrices in the original order of variables:

$$A_\beta = \begin{pmatrix}
0 & 1 & 0 & & & & & & & \\
-u^2 - \frac{\alpha_1^2}{3} + gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-2u\alpha_1 & 2\alpha_1 & 0 & u & 0 & \frac{3}{5}\alpha_1 & 0 & \cdots & 0 & 0 \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & 0 & u & \frac{1}{5}\beta_1 & \frac{2}{5}\alpha_1 & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1^2 & 0 & 0 & \frac{1}{3}\alpha_1 & 0 & u & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & -\frac{1}{3}\beta_1 & \frac{2}{3}\alpha_1 & 0 & u & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N+1}{2N+1}\alpha_1 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{1}{2N+1}\beta_1 & \frac{N}{2N+1}\alpha_1 \\
 & & & & & & \frac{(N-1)(2N+1)}{(2N-1)(N+1)}\alpha_1 & 0 & u & 0 \\
 & & & & & & \frac{N^2}{(2N-1)(N+1)}\beta_1 & \frac{N}{2N-1}\alpha_1 & 0 & u
\end{pmatrix} \tag{4.2.5}$$

in the x direction and

$$B_\beta = \begin{pmatrix}
0 & 0 & 1 & & & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-v^2 - \frac{\beta_1^2}{3} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & & & & & \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & v & 0 & \frac{2}{5}\beta_1 & \frac{1}{5}\alpha_1 & & & \\
-2v\beta_1 & 0 & 2\beta_1 & 0 & v & 0 & \frac{3}{5}\beta_1 & & & \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & \frac{2}{3}\beta_1 & -\frac{1}{3}\alpha_1 & v & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\beta_1^2 & 0 & 0 & 0 & \frac{1}{3}\beta_1 & 0 & v & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N}{2N+1}\beta_1 & \frac{1}{2N+1}\alpha_1 \\
 & & & & & \ddots & \ddots & \ddots & 0 & \frac{N+1}{2N+1}\beta_1 \\
 & & & & & & \frac{N}{2N-1}\beta_1 & \frac{N^2}{(2N-1)(N+1)}\alpha_1 & v & 0 \\
 & & & & & & 0 & \frac{(N-1)(2N+1)}{(2N-1)(N+1)}\beta_1 & 0 & v
\end{pmatrix} \tag{4.2.6}$$

in the y direction.

Since $P_N(\xi)$ in Lemma 4.2.1 and $P_{N+1}(\xi)$ in Corollary 4.1.2 both have real and distinct roots and there are no common roots due to the interlacing property of zeros of orthogonal

polynomials, we have that the coefficient matrix (4.2.5) in the x direction has real and distinct roots and thus real diagonalizable. Combining with the rotational invariance in Theorem 3.3.2, we immediately have the hyperbolicity of the β-HSWME model in 2D:

Theorem 4.2.1 (hyperbolicity of the β-HSWME)**.** *The β-HSWME model in 2D with the coefficient matrices given by* (4.2.5) *in x direction and* (4.2.6) *in y direction is hyperbolic.*

4.3 A framework for constructing general closure relations with rotational invariance and hyperbolicity

Besides the previous hyperbolic shallow water moment models, we can also modify both $\tilde{A}_{11}$ and $\tilde{A}_{22}$ in (4.1.2), as long as each of them has real and distinct eigenvalues and they have no common eigenvalues. In this case, the real diagonalizability of the matrix $\tilde{A}$ in (4.1.2) is guaranteed. Here, we can borrow the idea from [21] to modify the entries in the last row of the matrix such that the modified matrix has predefined propagation speeds. Since only the last row is modified, the associated polynomial sequence remains unchanged except for the last one. Therefore, it is easy to derive the analytical form of the characteristic polynomial where the coefficients have a linear dependence on the entries in the last row. Next, by matching the coefficients using Vieta's formulas which relate the coefficients of a polynomial to sums and products of its roots, or using the appropriate recurrence relation, the entries in the last row can be solved analytically. Similar ideas are also applied in machine learning moment closures for radiative transfer equation [16] where the roots are represented by the neural networks.

After the modified $\tilde{A}_{11}$ and $\tilde{A}_{22}$ are obtained, the next step is to determine the form of $\tilde{A}_{21}$ by the rotational invariance constraint in Theorem 3.3.1. The coefficient matrix B in the y direction can be derived by this constraint as well. Then we have a moment model in 2D with provable rotational invariance and hyperbolicity.

4.4 An example of constructing a general closure

To illustrate the framework in the previous part, we show an example by modifying the entries in the last row of $\tilde{A}_{22}$ in (4.1.5) such that its characteristic polynomial is related to

the derivative of the Legendre polynomial. Since only the last row is modified, the associated polynomial sequence given in Lemma 4.1.2 will not be changed except the last one. Therefore, we compute the last associated polynomial:

$$\begin{aligned}
q_{N+1}(x) &= xq_N(x) - \sum_{j=1}^{N+1} a_{N+1,j} q_{j-1}(x) \\
&= (x - a_{N+1,N+1}) q_N(x) - \sum_{j=1}^{N} a_{N+1,j} q_{j-1}(x) \\
&= ((x-u) - (a_{N+1,N+1} - u))(2N+1)P_N(\xi) - \sum_{j=1}^{N} a_{N+1,j}(2j-1)P_{j-1}(\xi) \\
&= \xi\alpha_1(2N+1)P_N(\xi) - (a_{N+1,N+1} - u)(2N+1)P_N(\xi) - \sum_{j=1}^{N} a_{N+1,j}(2j-1)P_{j-1}(\xi) \\
&= \alpha_1(NP_{N-1}(\xi) + (N+1)P_{N+1}(\xi)) - (a_{N+1,N+1} - u)(2N+1)P_N(\xi) \\
&\quad - \sum_{j=1}^{N} a_{N+1,j}(2j-1)P_{j-1}(\xi) \\
&= \alpha_1(N+1)P_{N+1}(\xi) - (a_{N+1,N+1} - u)(2N+1)P_N(\xi) + (\alpha_1 N - a_{N+1,N}(2N-1))P_{N-1}(\xi) \\
&\quad - \sum_{j=1}^{N-1} a_{N+1,j}(2j-1)P_{j-1}(\xi),
\end{aligned}$$

where we use the recursion relation $(2n+1)\xi P_n(\xi) = nP_{n-1}(\xi) + (n+1)P_{n+1}(\xi)$.

Next, we would like to take appropriate values of $a_{N+1,j}$ for $1 \le j \le N+1$ such that $q_{N+1}(x)$ is proporional to $P'_{N+2}(\xi)$. We use the relation

$$P'_{N+2}(\xi) = (2N+3)P_{N+1}(\xi) + (2N-1)P_{N-1}(\xi) + (2N-5)P_{N-3}(\xi) + \cdots . \tag{4.4.1}$$

Matching the coefficient of $P_{N+1}(\xi)$ in $P'_{N+2}(\xi)$ and $q_{N+1}(x)$, we have

$$q_{N+1}(x) = \alpha_1 \frac{N+1}{2N+3} P'_{N+2}(\xi),$$

which can be expanded as

$$\begin{aligned}
&\alpha_1(N+1)P_{N+1}(\xi) - (a_{N+1,N+1} - u)(2N+1)P_N(\xi) + (\alpha_1 N - a_{N+1,N}(2N-1))P_{N-1}(\xi) \\
&\quad - \sum_{j=1}^{N-1} a_{N+1,j}(2j-1)P_{j-1}(\xi) \\
&= \alpha_1(N+1)P_{N+1}(\xi) + \alpha_1 \frac{N+1}{2N+3}(2N-1)P_{N-1}(\xi) + \alpha_1 \frac{N+1}{2N+3}(2N-5)P_{N-3}(\xi) + \cdots .
\end{aligned}$$

Now we match the remaining coefficients in the above equation. For the coefficient of $P_N(\xi)$, we have

$$a_{N+1,N+1} = u.$$

For the coefficient of $P_{N-1}(\xi)$, we have

$$\alpha_1 N - a_{N+1,N}(2N-1) = \alpha_1 \frac{N+1}{2N+3}(2N-1),$$

from which we solve for $a_{N+1,N}$ and obtain

$$a_{N+1,N} = \frac{2N+1}{(2N-1)(2N+3)}\alpha_1.$$

For the remaining entries, it is easy to obtain:

$$a_{N+1,j} = \begin{cases} -\dfrac{N+1}{2N+3}\alpha_1, & \text{if } j \equiv N \bmod 2, \\ 0, & \text{otherwise.} \end{cases} \tag{4.4.2}$$

for $1 \le j \le N-1$.

Therefore, the modified $\tilde{A}_{22}$ is

$$\tilde{A}_{22}(W) = \begin{pmatrix} u & \frac{\alpha_1}{3} & & & & & & \\ \alpha_1 & u & \frac{2\alpha_1}{5} & & & & & \\ & \frac{2}{3}\alpha_1 & u & \frac{3\alpha_1}{7} & & & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \frac{N-1}{2N-3}\alpha_1 & u & \frac{N}{2N+1}\alpha_1 \\ \cdots & 0 & -\frac{N+1}{2N+3}\alpha_1 & 0 & -\frac{N+1}{2N+3}\alpha_1 & 0 & \frac{2N+1}{(2N-1)(2N+3)}\alpha_1 & u \end{pmatrix}. \tag{4.4.3}$$

Next, we keep $\tilde{A}_{11}$ in (4.1.3) unchanged and write down $\tilde{A}_{21}$ based on the rotational invariance constraint given in Theorem 3.3.1. We can also write the coefficient matrices in

the original order of variables:

$$A = \begin{pmatrix}
0 & 1 & 0 & & & & & & & \\
-u^2 - \frac{\alpha_1^2}{3} + gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-2u\alpha_1 & 2\alpha_1 & 0 & u & 0 & \frac{3\alpha_1}{5} & 0 & \cdots & 0 & 0 \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & 0 & u & \frac{\beta_1}{5} & \frac{2\alpha_1}{5} & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1^2 & 0 & 0 & \frac{\alpha_1}{3} & 0 & u & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & -\frac{\beta_1}{3} & \frac{2\alpha_1}{3} & 0 & u & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{(N+1)\alpha_1}{2N+1} & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{\beta_1}{2N+1} & \frac{N\alpha_1}{2N+1} \\
\cdots & \cdots & 0 & 0 & 0 & 0 & \frac{(N-1)\alpha_1}{2N-1} & 0 & u & 0 \\
\cdots & \cdots & \frac{(N+1)\beta_1}{2N+3} & -\frac{(N+1)\alpha_1}{2N+3} & 0 & 0 & \frac{(2N^2-N-4)\beta_1}{(2N-1)(N+1)} & \frac{(2N+1)\alpha_1}{(2N-1)(2N+3)} & 0 & u
\end{pmatrix} \tag{4.4.4}$$

in the x direction and

$$B = \begin{pmatrix}
0 & 0 & 1 & & & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-v^2 - \frac{\beta_1^2}{3} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & & & & & \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & v & 0 & \frac{2\beta_1}{5} & \frac{\alpha_1}{5} & & & \\
-2v\beta_1 & 0 & 2\beta_1 & 0 & v & 0 & \frac{3\beta_1}{5} & & & \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & \frac{2\beta_1}{3} & -\frac{\alpha_1}{3} & v & 0 & \cdots & 0 & 0 \\
-\frac{2}{3}\beta_1^2 & 0 & 0 & 0 & \frac{\beta_1}{3} & 0 & v & \cdots & 0 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N}{2N+1}\beta_1 & \frac{1}{2N+1}\alpha_1 \\
 & & & & & \ddots & \ddots & \ddots & 0 & \frac{N+1}{2N+1}\beta_1 \\
\cdots & \cdots & -\frac{(N+1)\beta_1}{2N+3} & \frac{(N+1)\alpha_1}{2N+3} & 0 & 0 & \frac{(2N+1)\beta_1}{(2N-1)(2N+3)} & \frac{(2N^2-N-4)\alpha_1}{(2N-1)(N+1)} & v & 0 \\
\cdots & \cdots & 0 & 0 & 0 & 0 & 0 & \frac{(N-1)\beta_1}{2N-1} & 0 & v
\end{pmatrix} \tag{4.4.5}$$

in the y direction.

Since $P'_{N+1}(\xi)$ in Corollary 4.1.1 and $P'_{N+2}(\xi)$ both have real and distinct roots and there are no common roots, we have that the coefficient matrix (4.4.4) in the x direction has real and distinct roots and thus we have that the matrix is real diagonalizable. Combining with the rotational invariance in Theorem 3.3.2, we have established the hyperbolicity of the new model in 2D:

Theorem 4.4.1 (hyperbolicity of an examplore general moment closure)**.** *The model in 2D with the coefficient matrices given by* (4.4.4) *in x direction and* (4.4.5) *in y direction is hyperbolic.*

CHAPTER 5

CONCLUDING REMARKS

In this work, we investigate the rotational invariance and hyperbolicity of the shallow water moment equations in 2D. We establish a general closure relation such that the resulting shallow water moment equations are hyperbolic and rotationally invariant. Moreover, we propose a new class of shallow water moment equations, which is a generalization of the HSWME model.

In future work, it remains to be seen whether the derived shallow water moment models are good approximations of the original incompressible NS equations both numerically and theoretically. Another interesting direction is to apply data-driven methods to learn the closure relations while preserving the rotational invariance and hyperbolicity, as we have done in [18, 15, 17, 16]. These generalizations are the subject of our ongoing work.

BIBLIOGRAPHY

[1] A. Amrita and J. Koellermeier. Projective integration for hyperbolic shallow water moment equations. *Axioms*, 11(5):235, 2022.

[2] F. Bouchut and V. Zeitlin. A robust well-balanced scheme for multi-layer shallow water equations. *Discrete and Continuous Dynamical Systems-Series B*, 13(4):739–758, 2010.

[3] D. Bresch. Chapter 1 - shallow-water equations and related topics. In C. Dafermos and M. Pokorný, editors, *Handbook of Differential Equations*, volume 5 of *Handbook of Differential Equations: Evolutionary Equations*, pages 1–104. North-Holland, 2009.

[4] Z. Cai, Y. Fan, and R. Li. Globally hyperbolic regularization of Grad's moment system in one-dimensional space. *Communications in Mathematical Sciences*, 11(2):547–571, 2013.

[5] Z. Cai, Y. Fan, and R. Li. Globally hyperbolic regularization of Grad's moment system. *Communications on Pure and Applied Mathematics*, 67(3):464–518, 2014.

[6] Castro, Manuel, Macías, Jorge, and Parés, Carlos. A q-scheme for a class of systems of coupled conservation laws with source term. application to a two-layer 1-d shallow water system. *ESAIM: M2AN*, 35(1):107–127, 2001.

[7] M. Elouafi and A. D. A. Hadj. A recursion formula for the characteristic polynomial of hessenberg matrices. *Applied mathematics and computation*, 208(1):177–179, 2009.

[8] Y. Fan, J. Koellermeier, J. Li, R. Li, and M. Torrilhon. Model reduction of kinetic equations by operator projection. *Journal of Statistical Physics*, 162(2):457–486, 2016.

[9] E. D. Fernández-Nieto, J. Garres-Díaz, A. Mangeney, and G. Narbona-Reina. A multilayer shallow model for dry granular flows with the-rheology: application to granular collapse on erodible beds. *Journal of Fluid Mechanics*, 798:643–681, 2016.

[10] E. D. Fernández-Nieto, E. H. Koné, and T. Chacón Rebollo. A multilayer method for the hydrostatic navier-stokes equations: a particular weak solution. *Journal of Scientific Computing*, 60:408–437, 2014.

[11] R. Fox and F. Laurent. Hyperbolic quadrature method of moments for the one-dimensional kinetic equation. *arXiv preprint arXiv:2103.10138*, 2021.

[12] J. Garres-Diaz, M. J. Castro Diaz, J. Koellermeier, and T. Morales de Luna. Shallow water moment models for bedload transport problems. *Communications In Computational Physics*, 30(3):903–941, 2021.

[13] J. Garres-Díaz, C. Escalante, T. M. De Luna, and M. C. Díaz. A general vertical decomposition of euler equations: multilayer-moment models. *Applied Numerical Mathematics*, 183:236–262, 2023.

[14] H. Grad. On the kinetic theory of rarefied gases. *Communications on Pure and Applied Mathematics*, 2(4):331–407, 1949.

[15] J. Huang, Y. Cheng, A. J. Christlieb, and L. F. Roberts. Machine learning moment closure models for the radiative transfer equation I: directly learning a gradient based closure. *Journal of Computational Physics*, 453:110941, 2022.

[16] J. Huang, Y. Cheng, A. J. Christlieb, and L. F. Roberts. Machine learning moment closure models for the radiative transfer equation III: enforcing hyperbolicity and physical characteristic speeds. *Journal of Scientific Computing*, 94(1):7, 2023.

[17] J. Huang, Y. Cheng, A. J. Christlieb, L. F. Roberts, and W.-A. Yong. Machine learning moment closure models for the radiative transfer equation II: Enforcing global hyperbolicity in gradient-based closures. *Multiscale Modeling & Simulation*, 21(2):489–512, 2023.

[18] J. Huang, Z. Ma, Y. Zhou, and W.-A. Yong. Learning thermodynamically stable and Galilean invariant partial differential equations for non-equilibrium flows. *Journal of Non-Equilibrium Thermodynamics*, 2021.

[19] Q. Huang, J. Koellermeier, and W.-A. Yong. Equilibrium stability analysis of hyperbolic shallow water moment equations. *Mathematical Methods in the Applied Sciences*, 45(10):6459–6480, 2022.

[20] J. Koellermeier and E. Pimentel-García. Steady states and well-balanced schemes for shallow water moment equations with topography. *Applied Mathematics and Computation*, 427:127166, 2022.

[21] J. Koellermeier and M. Rominger. Analysis and numerical simulation of hyperbolic shallow water moment equations. *Communications in Computational Physics*, 28(3):1038–1084, 2020.

[22] J. Kowalski and M. Torrilhon. Moment approximations and model cascades for shallow flow. *Communications in Computational Physics*, 25(3):669–702, 2018.

[23] C. D. Levermore. Moment closure hierarchies for kinetic theories. *Journal of statistical Physics*, 83(5):1021–1065, 1996.

[24] A. L. Stewart and P. J. Dellar. Multilayer shallow water equations with complete coriolis force. part 1. derivation on a non-traditional beta-plane. *Journal of Fluid Mechanics*, 651:387–413, 2010.

[25] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer-Verlag, Berlin Heidelberg, 1997.

[26] R. Verbiest. Analysis and numerical simulation of two-dimensional shallow water moment equations, 2022.

[27] R. Verbiest and J. Koellermeier. Hyperbolic axisymmetric shallow water moment equations. *arXiv preprint arXiv:2302.07952*, 2023.

[28] C. B. Vreugdenhil. *Numerical methods for shallow-water flow*, volume 13. Springer Science & Business Media, 1994.

[29] P. õa Navarro, P. Brufau, J. Burguete, and J. Murillo. The shallow water equations: An example of hyperbolic system. *Monografías de la Real Academia de Ciencias de Zaragoza*, 31, 01 2008.

APPENDIX

A.1 Discussion of Dimensional Scaling

Under the shallow water assumptions certain terms are neglected under the assumption that the ratio of the characteristic vertical length scale H to the characteristic horizontal length scale L is small. i.e. $H/L = \epsilon \ll 1$ The starting point is the Navier Stokes equations

$$\begin{aligned} \nabla \cdot U &= 0, \\ \partial_t U + \nabla \cdot (UU) &= -\frac{1}{\rho}\nabla p + \frac{1}{\rho}\nabla \cdot \sigma + g. \end{aligned} \tag{A.1.1}$$

We will scale the spacial variables according to the characteristic length scales

$$x = L\hat{x}, \quad y = L\hat{y}, \quad z = H\hat{z} \tag{A.1.2}$$

The velocity variables are also scaled according to a characteristic horizontal velocity V. Because of the shallow water assumption the characteristic vertical velocity is much smaller, i.e. on the order of ϵV.

$$u = V\hat{u}, \quad v = V\hat{v}, \quad w = \epsilon V\hat{w} \tag{A.1.3}$$

The time is also scaled by a factor involving the characteristic horizontal length and velocity scales.

$$t = \frac{L}{V}\hat{t} \tag{A.1.4}$$

A characteristic stress scale S is introduced so the pressure and stress scaling are

$$p = \rho g H \hat{p}, \quad \sigma_{\text{basal}} = S\hat{\sigma}_{\text{basal}}, \quad \sigma_{\text{other}} = S\epsilon\hat{\sigma}_{\text{other}} \tag{A.1.5}$$

where basal stresses are σ_{xz} and σ_{yz} which will be larger than the other stress components due to the shallow water assumption. Substituting all this

$$\begin{aligned} \partial_{\hat{x}}\hat{u} + \partial_{\hat{y}}\hat{v} + \partial_{\hat{z}}\hat{w} &= 0 \\ F^2\epsilon(\partial_{\hat{t}}\hat{u} + \partial_{\hat{x}}\hat{u}^2 + \partial_{\hat{y}}(\hat{u}\hat{v}) + \partial_{\hat{z}}(\hat{u}\hat{w})) &= -\epsilon\partial_{\hat{x}}\hat{p} + \epsilon^2 G\partial_{\hat{x}}\hat{\sigma}_{xx} + \epsilon G\partial_{\hat{y}}\hat{\sigma}_{xy} + G\partial_{\hat{z}}\hat{\sigma}_{xz} + e_x \\ F^2\epsilon(\partial_{\hat{t}}\hat{v} + \partial_{\hat{x}}(\hat{u}\hat{v}) + \partial_{\hat{y}}\hat{v}^2 + \partial_{\hat{z}}(\hat{v}\hat{w})) &= -\epsilon\partial_{\hat{y}}\hat{p} + \epsilon G\partial_{\hat{x}}\hat{\sigma}_{xy} + \epsilon^2 G\partial_{\hat{y}}\hat{\sigma}_{yy} + G\partial_{\hat{z}}\hat{\sigma}_{yz} + e_y \\ F^2\epsilon^2(\partial_{\hat{t}}\hat{w} + \partial_{\hat{x}}(\hat{u}\hat{w}) + \partial_{\hat{y}}(\hat{v}\hat{w}) + \partial_{\hat{z}}\hat{w}^2) &= -\epsilon\partial_{\hat{z}}\hat{p} + \epsilon G\partial_{\hat{x}}\hat{\sigma}_{xz} + \epsilon G\partial_{\hat{y}}\hat{\sigma}_{yz} + \epsilon G\partial_{\hat{z}}\hat{\sigma}_{xz} + e_z \end{aligned} \tag{A.1.6}$$

where $F = V/\sqrt{gH}$ is the Froude number and $G = S/(\rho g H)$ is the ratio of characteristic stress to characteristic hydrostatic pressure. For shallow water both ϵ and G will be much smaller than 1 so terms with ϵ^2 and $G\epsilon$ will be ignored. This allows for pressure to be solved for directly and leads to the equations at the start of the shallow water moment equation derivation.

A.2 Proof of Proposition 3.1.1

Proof. We prove the equalities in Proposition 3.1.1 by direct calculations:

1.
$$\text{LHS} = \cos\theta\,(\cos\theta\,u + \sin\theta\,v) - \sin\theta\,(-\sin\theta\,u + \cos\theta\,v) = u = \text{RHS}.$$

2.
$$\text{LHS} = \sin\theta\,(\cos\theta\,u + \sin\theta\,v) + \cos\theta\,(-\sin\theta\,u + \cos\theta\,v) = v = \text{RHS}.$$

3.
$$\begin{aligned}\text{LHS} = {} & \cos\theta\,(\cos\theta\,u + \sin\theta\,v)^2 - \sin\theta\,(\cos\theta\,u + \sin\theta\,v)(-\sin\theta\,u + \cos\theta\,v) \\ & = (\cos\theta\,u + \sin\theta\,v)\,(\cos\theta\,(\cos\theta\,u + \sin\theta\,v) - \sin\theta\,(-\sin\theta\,u + \cos\theta\,v)) \\ & = u(\cos\theta\,u + \sin\theta\,v) = \text{RHS}.\end{aligned}$$

4.
$$\begin{aligned}\text{LHS} = {} & \cos\theta(\cos\theta\,u + \sin\theta\,v)^2 - \sin\theta(\cos\theta\,u + \sin\theta\,v)(-\sin\theta\,u + \cos\theta\,v) \\ & = (\cos\theta\,u + \sin\theta\,v)\,(\cos\theta(\cos\theta\,u + \sin\theta\,v) - \sin\theta(-\sin\theta\,u + \cos\theta\,v)) \\ & = u(\cos\theta\,u + \sin\theta\,v) = \text{RHS}.\end{aligned}$$

5.

$$\begin{aligned}
\text{LHS} &= 2\cos\theta\left(\cos^2\theta\, u\alpha + \cos\theta\,\sin\theta\,(u\beta + v\alpha) + \sin^2\theta\, v\beta\right) \\
&\quad - \sin\theta\,\left(\cos^2\theta\, u\beta + \cos\theta\,\sin\theta\,(v\beta - u\alpha) - \sin^2\theta\, v\alpha\right) \\
&\quad - \sin\theta\,\left(\cos^2\theta v\alpha + \cos\theta\,\sin\theta\,(-u\alpha + v\beta) - \sin^2\theta u\beta\right) \\
&= u\alpha\cos\theta\,(2\cos^2\theta + \sin^2\theta + \sin^2\theta) + v\alpha\sin\theta\,(2\cos^2\theta + \sin^2\theta - \cos^2\theta) \\
&\quad + u\beta\sin\theta\,(2\cos^2\theta - \cos^2\theta + \sin^2\theta) + v\beta\cos\theta(2\sin^2\theta - \sin^2\theta - \sin^2\theta) \\
&= 2u\alpha\cos\theta + (u\beta + v\alpha)\sin\theta \\
&= \text{RHS}
\end{aligned}$$

6.

$$\begin{aligned}
\text{LHS} &= 2\sin\theta\left(\cos^2\theta\, u\alpha + \cos\theta\,\sin\theta\,(u\beta + v\alpha) + \sin^2\theta v\beta\right) \\
&\quad + \cos\theta\,\left(\cos^2\theta\, u\beta + \cos\theta\,\sin\theta\,(v\beta - u\alpha) - \sin^2\theta\, v\alpha\right) \\
&\quad + \cos\theta\,\left(\cos^2\theta\, v\alpha + \cos\theta\,\sin\theta\,(-u\alpha + v\beta) - \sin^2\theta\, u\beta\right) \\
&= u\alpha\sin\theta\,(2\cos^2\theta\, - \cos^2\theta - \cos^2\theta) + v\alpha\cos\theta\,(2\sin^2\theta - \sin^2\theta + \cos^2\theta) \\
&\quad + u\beta\cos\theta(2\sin^2\theta + \cos^2\theta - \sin^2\theta) + v\beta\sin\theta(2\sin^2\theta + \cos^2\theta + \cos^2\theta) \\
&= (u\beta + v\alpha)\cos\theta + 2v\beta\sin\theta = \text{RHS}
\end{aligned}$$

□

A.3 Proof of Lemma 3.2.1

We split the proof into three parts: (1) the computation of the conservative part in the x-direction; (2) the computation of the conservative part in the y-direction; (3) the computation of the nonconservative part.

Since the proof relies on the properties of the coefficients A_{ijk} and B_{ijk}, we summarize in the following:

Lemma A.3.1 ([21]). *For the coefficient A_{ijk} given by*

$$A_{ijk} = (2i+1)\int_0^1 \phi_i(x)\phi_j(x)\phi_k(x)dx \tag{A.3.1}$$

we have the following properties:

1. $A_{ijk} = A_{ikj}$ *for any* i, j, k.

2. $$A_{ij1} = \begin{cases} \frac{i}{2i-1}, & \text{if} \quad j = i-1. \\ \frac{i+1}{2i+3}, & \text{if} \quad j = i+1. \\ 0, & \text{otherwise.} \end{cases} \tag{A.3.2}$$

Lemma A.3.2 ([21])**.** *For the coefficient* B_{ijk} *given by*

$$B_{ijk} = (2i+1)\int_0^1 \phi_i' \left(\int_0^\zeta \phi_j d\hat{\zeta} \right) \phi_k d\zeta, \tag{A.3.3}$$

it satisfies the properties:

$$B_{ij1} = \begin{cases} -\frac{i+1}{2i-1}, & \text{if} \quad j = i-1. \\ -\frac{i}{2i+3}, & \text{if} \quad j = i+1. \\ 0, & \text{otherwise.} \end{cases} \tag{A.3.4}$$

A.3.1 The conservative part in the x-direction

We first compute the Jacobian matrix in x-dimension $\frac{\partial F}{\partial U}$ where the physical flux F is given by (2.1.15):

1. For the first component

$$F_1(U) = hu, \tag{A.3.5}$$

we compute the gradient

$$\frac{\partial F_1(U)}{\partial U} = (0, 1, 0, 0, \cdots, 0)^T. \tag{A.3.6}$$

2. For the second component

$$F_2(U) = h(u^2 + \sum_j \frac{\alpha_j^2}{2j+1}) + \frac{1}{2}gh^2 = \frac{(hu)^2}{h} + \sum_j \frac{1}{2j+1}\frac{(h\alpha_j)^2}{h} + \frac{1}{2}gh^2, \tag{A.3.7}$$

the derivatives are

$$\frac{\partial F_2(U)}{\partial h} = -\frac{(hu)^2}{h^2} - \sum_j \frac{1}{2j+1}\frac{(h\alpha_j)^2}{h^2} + gh = -u^2 - \sum_j \frac{\alpha_j^2}{2j+1} + gh,$$

$$\begin{aligned}
\frac{\partial F_2(U)}{\partial (hu)} &= \frac{2(hu)}{h} = 2u \\
\frac{\partial F_2(U)}{\partial (h\alpha_j)} &= \frac{1}{2j+1}\frac{2(h\alpha_j)}{h} = \frac{2\alpha_j}{2j+1} \\
\frac{\partial F_2(U)}{\partial (hv)} &= \frac{\partial F_2(U)}{\partial (h\beta_j)} = 0.
\end{aligned}$$

Therefore, we have

$$\frac{\partial F_2(U)}{\partial U} = (-u^2 - \sum_j \frac{\alpha_j^2}{2j+1} + gh, 2u, 0, \frac{2\alpha_1}{3}, 0, \frac{2\alpha_2}{5}, \cdots, 0, \frac{2\alpha_N}{2N+1}, 0)^T. \quad \text{(A.3.8)}$$

3. For the third component

$$F_3(U) = h(uv + \sum_j \frac{\alpha_j\beta_j}{2j+1}) = \frac{(hu)(hv)}{h} + \sum_j \frac{1}{2j+1}\frac{(h\alpha_j)(h\beta_j)}{h} \quad \text{(A.3.9)}$$

the derivatives are

$$\frac{\partial F_3(U)}{\partial h} = -\frac{(hu)(hv)}{h^2} - \sum_j \frac{1}{2j+1}\frac{(h\alpha_j)(h\beta_j)}{h^2} = -uv - \sum_j \frac{\alpha_j\beta_j}{2j+1} \quad \text{(A.3.10)}$$

$$\begin{aligned}
\frac{\partial F_3(U)}{\partial (hu)} &= \frac{(hv)}{h} = v \\
\frac{\partial F_3(U)}{\partial (hv)} &= \frac{(hu)}{h} = u \\
\frac{\partial F_3(U)}{\partial (h\alpha_j)} &= \frac{1}{2j+1}\frac{(h\beta_j)}{h} = \frac{\beta_j}{2j+1} \\
\frac{\partial F_3(U)}{\partial (h\beta_j)} &= \frac{1}{2j+1}\frac{(h\alpha_j)}{h} = \frac{\alpha_j}{2j+1}
\end{aligned} \quad \text{(A.3.11)}$$

We have

$$\frac{\partial F_3(U)}{\partial U} = (-uv - \sum_j \frac{\alpha_j\beta_j}{2j+1}, v, u, \frac{\beta_1}{3}, \frac{\alpha_1}{3}, \frac{\beta_2}{5}, \frac{\alpha_2}{5}, \cdots, \frac{\beta_N}{2N+1}, \frac{\alpha_N}{2N+1})^T. \quad \text{(A.3.12)}$$

4. For the remaining component, we denote

$$g_i(U) := h(2u\alpha_i + \sum_{j,k} A_{ijk}\alpha_j\alpha_k) = \frac{2(hu)(h\alpha_i)}{h} + \sum_{j,k} A_{ijk}\frac{(h\alpha_j)(h\alpha_k)}{h} \quad \text{(A.3.13)}$$

for $i = 1, 2, \cdots, N$. Then the gradient is

$$\frac{\partial g_i(U)}{\partial h} = -\frac{2(hu)(h\alpha_i)}{h^2} - \sum_{j,k} A_{ijk}\frac{(h\alpha_j)(h\alpha_k)}{h^2} = -2u\alpha_i - \sum_{j,k} A_{ijk}\alpha_j\alpha_k \tag{A.3.14}$$

$$\frac{\partial g_i(U)}{\partial (hu)} = \frac{2(h\alpha_i)}{h} = 2\alpha_i \tag{A.3.15}$$

$$\begin{aligned}\frac{\partial g_i(U)}{\partial (h\alpha_l)} &= \frac{2(hu)}{h}\delta_{il} + \sum_{j,k} A_{ijk}(\delta_{jl}\frac{(h\alpha_k)}{h} + \delta_{kl}\frac{(h\alpha_j)}{h}) \\ &= 2u\delta_{ij} + \sum_k A_{ilk}\alpha_k + \sum_j A_{ijl}\alpha_j \\ &= 2u\delta_{ij} + \sum_j A_{ilj}\alpha_j + \sum_j A_{ijl}\alpha_j \\ &= 2u\delta_{ij} + 2\sum_j A_{ilj}\alpha_j\end{aligned} \tag{A.3.16}$$

where in the last step we use $A_{ijk} = A_{ikj}$.

$$\frac{\partial g_i(U)}{\partial (hv)} = \frac{\partial g_i(U)}{\partial (h\beta_j)} = 0 \tag{A.3.17}$$

5. For the component, we denote

$$h_i(U) := h(u\beta_i + v\alpha_i + \sum_{j,k} A_{ijk}\alpha_j\beta_k) = \frac{(hu)(h\beta_i)}{h} + \frac{(hv)(h\alpha_i)}{h} + \sum_{j,k} A_{ijk}\frac{(h\alpha_j)(h\beta_k)}{h} \tag{A.3.18}$$

for $i = 1, 2, \cdots, N$. Then the gradient is

$$\frac{\partial h_i(U)}{\partial h} = -\frac{(hu)(h\beta_i)}{h^2} - \frac{(hv)(h\alpha_i)}{h^2} - \sum_{j,k} A_{ijk}\frac{(h\alpha_j)(h\beta_k)}{h^2} = -u\beta_i - v\alpha_i - \sum_{j,k} A_{ijk}\alpha_j\beta_k \tag{A.3.19}$$

$$\frac{\partial h_i(U)}{\partial (hu)} = \frac{(h\beta_i)}{h} = \beta_i \tag{A.3.20}$$

$$\frac{\partial h_i(U)}{\partial (hv)} = \frac{(h\alpha_i)}{h} = \alpha_i \tag{A.3.21}$$

$$\begin{aligned}\frac{\partial h_i(U)}{\partial (h\alpha_l)} &= \frac{(hv)}{h}\delta_{il} + \sum_{j,k} A_{ijk}\delta_{jl}\frac{(h\beta_k)}{h} \\ &= v\delta_{il} + \sum_k A_{ilk}\beta_k\end{aligned} \tag{A.3.22}$$

$$\begin{aligned}\frac{\partial h_i(U)}{\partial(h\beta_l)} &= \frac{(hu)}{h}\delta_{il} + \sum_{j,k} A_{ijk}\delta_{kl}\frac{(h\alpha_j)}{h} \\ &= u\delta_{il} + \sum_j A_{ijl}\alpha_j \\ &= u\delta_{il} + \sum_j A_{ilj}\alpha_j\end{aligned} \tag{A.3.23}$$

Therefore, the Jacobian matrix $\frac{\partial F(U)}{\partial U}$ is

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
-u^2 - \frac{\alpha_j^2}{2j+1} + gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & \cdots & \frac{2\alpha_N}{2N+1} & 0 \\
-uv - \frac{\alpha_j\beta_j}{2j+1} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & \cdots & \frac{\beta_N}{2N+1} & \frac{\alpha_N}{2N+1} \\
-2u\alpha_1 - A_{1jk}\alpha_j\alpha_k & 2\alpha_1 & 0 & 2u + 2A_{11j}\alpha_j & 0 & \cdots & 2A_{1Nj}\alpha_j & 0 \\
-(u\beta_1 + v\alpha_1) - A_{1jk}\alpha_j\beta_k & \beta_1 & \alpha_1 & v + A_{11j}\beta_j & u + A_{11j}\alpha_j & \cdots & A_{1Nj}\beta_j & A_{1Nj}\alpha_j \\
-2u\alpha_2 - A_{2jk}\alpha_j\alpha_k & 2\alpha_2 & 0 & 2A_{21j}\alpha_j & 0 & \cdots & 2A_{2Nj}\alpha_j & 0 \\
-(u\beta_2 + v\alpha_2) - A_{2jk}\alpha_j\beta_k & \beta_2 & \alpha_2 & A_{21j}\beta_j & A_{21j}\alpha_j & \cdots & A_{2Nj}\beta_j & A_{2Nj}\alpha_j \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
-2u\alpha_N - A_{Njk}\alpha_j\alpha_k & 2\alpha_N & 0 & 2A_{N1j}\alpha_j & 0 & \cdots & 2u + 2A_{NNj}\alpha_j & 0 \\
-(u\beta_N + v\alpha_N) - A_{Njk}\alpha_j\beta_k & \beta_N & \alpha_N & A_{N1j}\beta_j & A_{N1j}\alpha_j & \cdots & v + A_{NNj}\beta_j & u + A_{NNj}\alpha_j
\end{pmatrix} \tag{A.3.24}$$

Here, for saving space, we use Einstein summation convention and the summation notation is omitted.

Evaluating the above matrix at $U_0 = (h, hu, hv, h\alpha_1, h\beta_1, 0, 0, \cdots, 0, 0)^T$, we have

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\
-u^2 - \frac{\alpha_1^2}{3} + gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & \cdots & 0 & 0 \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & \cdots & 0 & 0 \\
-2u\alpha_1 - A_{111}\alpha_1^2 & 2\alpha_1 & 0 & 2u + 2A_{111}\alpha_1 & 0 & \cdots & 2A_{1N1}\alpha_1 & 0 \\
-(u\beta_1 + v\alpha_1) - A_{111}\alpha_1\beta_1 & \beta_1 & \alpha_1 & v + A_{111}\beta_1 & u + A_{111}\alpha_1 & \cdots & A_{1N1}\beta_1 & A_{1N1}\alpha_1 \\
-A_{211}\alpha_1^2 & 0 & 0 & 2A_{211}\alpha_1 & 0 & \cdots & 2A_{2N1}\alpha_1 & 0 \\
-A_{211}\alpha_1\beta_1 & 0 & 0 & A_{211}\beta_1 & A_{211}\alpha_1 & \cdots & A_{2N1}\beta_1 & A_{2N1}\alpha_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
-A_{N11}\alpha_1^2 & 0 & 0 & 2A_{N11}\alpha_1 & 0 & \cdots & 2u + 2A_{NN1}\alpha_1 & 0 \\
-A_{N11}\alpha_1\beta_1 & 0 & 0 & A_{N11}\beta_1 & A_{N11}\alpha_1 & \cdots & v + A_{NN1}\beta_1 & u + A_{NN1}\alpha_1
\end{pmatrix} \tag{A.3.25}$$

Next, we use the property of A_{ijk} given in Lemma A.3.1 and further simplify it to

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & & & & & \\
-u^2-\frac{\alpha_1^2}{3}+gh & 2u & 0 & \frac{2\alpha_1}{3} & 0 & & & & & \\
-uv-\frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & & \\
-2u\alpha_1 & 2\alpha_1 & 0 & 2u & 0 & \frac{4}{5}\alpha_1 & 0 & & & \\
-(u\beta_1+v\alpha_1) & \beta_1 & \alpha_1 & v & u & \frac{2}{5}\beta_1 & \frac{2}{5}\alpha_1 & & & \\
-\frac{2}{3}\alpha_1^2 & 0 & 0 & \frac{4}{3}\alpha_1 & 0 & 2u & 0 & \ddots & & \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & \frac{2}{3}\beta_1 & \frac{2}{3}\alpha_1 & v & u & \ddots & & \\
 & & & & & \ddots & \ddots & \ddots & \frac{2N}{2N+1}\alpha_1 & 0 \\
 & & & & & \ddots & \ddots & \ddots & \frac{N}{2N+1}\beta_1 & \frac{N}{2N+1}\alpha_1 \\
 & & & & & & \frac{2N}{2N-1}\alpha_1 & 0 & 2u & 0 \\
 & & & & & & \frac{N}{2N-1}\beta_1 & \frac{N}{2N-1}\alpha_1 & v & u
\end{pmatrix} \tag{A.3.26}$$

A.3.2 The conservative part in the y direction

Then we compute the Jacobian matrix in the y direction.

1. For the first component

$$G_1(U) = hv \tag{A.3.27}$$

we have

$$\frac{\partial G_1(U)}{\partial U} = (0,0,1,0,0,\cdots,0)^T \tag{A.3.28}$$

2. For the second component

$$G_2(U) = h(uv + \sum_j \frac{\alpha_j\beta_j}{2j+1}) = \frac{(hu)(hv)}{h} + \sum_j \frac{1}{2j+1}\frac{(h\alpha_j)(h\beta_j)}{h} \tag{A.3.29}$$

it is the same as the third component $F_3(U)$ in x direction.

3. For the third component

$$G_3(U) = h(v^2 + \sum_j \frac{\beta_j^2}{2j+1}) + \frac{1}{2}gh^2 = \frac{(hv)^2}{h} + \sum_j \frac{1}{2j+1}\frac{(h\beta_j)^2}{h} + \frac{1}{2}gh^2 \tag{A.3.30}$$

the gradient is

$$\frac{\partial G_3(U)}{\partial h} = -\frac{(hv)^2}{h^2} - \sum_j \frac{1}{2j+1}\frac{(h\beta_j)^2}{h^2} + gh = -v^2 - \sum_j \frac{\beta_j^2}{2j+1} + gh \tag{A.3.31}$$

$$\frac{\partial G_3(U)}{\partial (hu)} = 0 \tag{A.3.32}$$

$$\frac{\partial G_3(U)}{\partial (hv)} = \frac{2(hv)}{h} = 2v \tag{A.3.33}$$

$$\frac{\partial G_3(U)}{\partial (h\alpha_j)} = 0 \tag{A.3.34}$$

$$\frac{\partial G_3(U)}{\partial (h\beta_j)} = \frac{1}{2j+1}\frac{2(h\beta_j)}{h} = \frac{2\beta_j}{2j+1} \tag{A.3.35}$$

Therefore, we have

$$\frac{\partial G_3(U)}{\partial U} = (-v^2 - \sum_j \frac{\beta_j^2}{2j+1} + gh, 0, 2v, 0, \frac{2\beta_1}{3}, 0, \frac{2\beta_2}{5}, \cdots, 0, \frac{2\beta_N}{2N+1})^T \tag{A.3.36}$$

4. For the component

$$h_i(U) := h(u\beta_i + v\alpha_i + \sum_{j,k} A_{ijk}\alpha_j\beta_k) \tag{A.3.37}$$

it is the same as the component in x direction.

5. For the component

$$g_i(U) := h(2v\beta_i + \sum_{j,k} A_{ijk}\beta_j\beta_k) = \frac{2(hv)(h\beta_i)}{h} + \sum_{j,k} A_{ijk}\frac{(h\beta_j)(h\beta_k)}{h} \tag{A.3.38}$$

the gradient is

$$\frac{\partial g_i(U)}{\partial h} = -\frac{2(hv)(h\beta_i)}{h^2} - \sum_{j,k} A_{ijk}\frac{(h\beta_j)(h\beta_k)}{h^2} = -2v\beta_i - \sum_{j,k} A_{ijk}\beta_j\beta_k \tag{A.3.39}$$

$$\frac{\partial g_i(U)}{\partial (hu)} = 0 \tag{A.3.40}$$

$$\frac{\partial g_i(U)}{\partial (hv)} = \frac{2(h\beta_i)}{h} = 2\beta_i \tag{A.3.41}$$

$$\frac{\partial g_i(U)}{\partial (h\alpha_j)} = 0 \tag{A.3.42}$$

$$
\begin{aligned}
\frac{\partial g_i(U)}{\partial(h\beta_l)} &= \frac{2(hv)}{h}\delta_{il} + \sum_{j,k} A_{ijk}(\delta_{jl}\frac{(h\beta_k)}{h} + \delta_{kl}\frac{(h\beta_j)}{h}) \\
&= 2v\delta_{ij} + \sum_k A_{ilk}\beta_k + \sum_j A_{ijl}\beta_j \\
&= 2v\delta_{ij} + \sum_j A_{ilj}\beta_j + \sum_j A_{ijl}\beta_j \\
&= 2v\delta_{ij} + 2\sum_j A_{ilj}\beta_j
\end{aligned}
\tag{A.3.43}
$$

where in the last step we use $A_{ijk} = A_{ikj}$.

Therefore, the Jacobian matrix $\frac{\partial G(U)}{\partial U}$ is

$$
\begin{pmatrix}
0 & 0 & 1 & 0 & & & & \\
-uv - \frac{\alpha_j\beta_j}{2j+1} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & \cdots & \frac{\beta_N}{2N+1} & \frac{\alpha_N}{2N+1} \\
-v^2 - \sum_j \frac{\beta_j^2}{2j+1} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & \cdots & 0 & \frac{2\beta_N}{2N+1} \\
-(u\beta_1 + v\alpha_1) - A_{1jk}\alpha_j\beta_k & \beta_1 & \alpha_1 & v + A_{11j}\beta_j & u + A_{11j}\alpha_j & \cdots & A_{1Nj}\beta_j & A_{1Nj}\alpha_j \\
-2v\beta_1 - A_{1jk}\beta_j\beta_k & 0 & 2\beta_1 & 0 & 2v + 2A_{11j}\beta_j & \cdots & 0 & 2A_{1Nj}\beta_j \\
-(u\beta_2 + v\alpha_2) - A_{2jk}\alpha_j\beta_k & \beta_2 & \alpha_2 & A_{21j}\beta_j & A_{21j}\alpha_j & \cdots & A_{2Nj}\beta_j & A_{2Nj}\alpha_j \\
-2v\beta_2 - A_{2jk}\beta_j\beta_k & 0 & 2\beta_2 & 0 & 2A_{21j}\beta_j & \cdots & 0 & 2A_{2Nj}\beta_j \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
-(u\beta_N + v\alpha_N) - A_{Njk}\alpha_j\beta_k & \beta_N & \alpha_N & A_{N1j}\beta_j & A_{N1j}\alpha_j & \cdots & v + A_{NNj}\beta_j & u + A_{NNj}\alpha_j \\
-2v\beta_N - A_{Njk}\beta_j\beta_k & 0 & 2\beta_N & 0 & 2A_{N1j}\beta_j & \cdots & 0 & 2v + 2A_{NNj}\beta_j
\end{pmatrix}
\tag{A.3.44}
$$

Here, for saving space, we use Einstein summation convention and the summation notation is omitted.

Evaluating the above matrix at $U_0 = (h, hu, hv, h\alpha_1, h\beta_1, 0, 0, \cdots, 0, 0)^T$, we have

$$\begin{pmatrix}
0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & \cdots & 0 & 0 \\
-v^2 - \frac{\beta_1^2}{3} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & \cdots & 0 & 0 \\
-(u\beta_1 + v\alpha_1) - A_{111}\alpha_1\beta_1 & \beta_1 & \alpha_1 & v + A_{111}\beta_1 & u + A_{111}\alpha_1 & \cdots & A_{1N1}\beta_1 & A_{1N1}\alpha_1 \\
-2v\beta_1 - A_{111}\beta^2 & 0 & 2\beta_1 & 0 & 2v + 2A_{111}\beta_1 & \cdots & 0 & 2A_{1N1}\beta_1 \\
-A_{211}\alpha_1\beta_1 & 0 & 0 & A_{211}\beta_1 & A_{211}\alpha_1 & \cdots & A_{2N1}\beta_1 & A_{2N1}\alpha_1 \\
-A_{211}\beta_1^2 & 0 & 0 & 0 & 2A_{211}\beta_1 & \cdots & 0 & 2A_{2N1}\beta_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
-A_{N11}\alpha_1\beta_1 & 0 & 0 & A_{N11}\beta_1 & A_{N11}\alpha_1 & \cdots & v + A_{NN1}\beta_1 & u + A_{NN1}\alpha_1 \\
-A_{N11}\beta_1^2 & 0 & 0 & 0 & 2A_{N11}\beta_1 & \cdots & 0 & 2v + 2A_{NN1}\beta_1
\end{pmatrix} \tag{A.3.45}$$

and further simplify it to

$$\begin{pmatrix}
0 & 0 & 1 & & & & & & \\
-uv - \frac{\alpha_1\beta_1}{3} & v & u & \frac{\beta_1}{3} & \frac{\alpha_1}{3} & & & & \\
-v^2 - \frac{\beta_1^2}{3} + gh & 0 & 2v & 0 & \frac{2\beta_1}{3} & & & & \\
-(u\beta_1 + v\alpha_1) & \beta_1 & \alpha_1 & v & u & \frac{2}{5}\beta_1 & \frac{2}{5}\alpha_1 & & \\
-2v\beta_1 & 0 & 2\beta_1 & 0 & 2v & 0 & \frac{4}{5}\beta_1 & & \\
-\frac{2}{3}\alpha_1\beta_1 & 0 & 0 & \frac{2}{3}\beta_1 & \frac{2}{3}\alpha_1 & v & u & \ddots & \\
-\frac{2}{3}\beta_1^2 & 0 & 0 & 0 & \frac{4}{3}\beta_1 & 0 & 2v & \ddots & \\
 & & & & & \ddots & \ddots & \ddots & \frac{N}{2N+1}\beta_1 & \frac{N}{2N+1}\alpha_1 \\
 & & & & & \ddots & \ddots & \ddots & 0 & \frac{2N}{2N+1}\beta_1 \\
 & & & & & & \frac{N}{2N-1}\beta_1 & \frac{N}{2N-1}\alpha_1 & v & u \\
 & & & & & & 0 & \frac{2N}{2N-1}\beta_1 & 0 & 2v
\end{pmatrix} \tag{A.3.46}$$

A.3.3 The nonconservative part in x and y directions

Evaluating the nonconservative part in x direction at the state $(h, hu, hv, h\alpha_1, h\beta_1, \cdots, 0, 0)$, we have

$$P(U) = \begin{pmatrix} 0_{3\times 3} & & & & & & & \\ & -u & 0 & -\frac{1}{5}\alpha_1 & 0 & & & \\ & -v & 0 & -\frac{1}{5}\beta_1 & 0 & & & \\ & -\frac{3}{3}\alpha_1 & 0 & -u & 0 & & & \\ & -\frac{3}{3}\beta_1 & 0 & -v & 0 & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & -u & 0 & -\frac{N-1}{2N+1}\alpha_1 & 0 \\ & & & & & -v & 0 & -\frac{N-1}{2N+1}\beta_1 & 0 \\ & & & & & -\frac{N+1}{2N-1}\alpha_1 & 0 & -u & 0 \\ & & & & & -\frac{N+1}{2N-1}\beta_1 & 0 & -v & 0 \end{pmatrix} \tag{A.3.47}$$

Here, we use the property of the coefficient B_{ijk}.

The nonconservative part in y direction is similar to the one in x direction, we have

$$Q(U) = \begin{pmatrix} 0_{3\times 3} & & & & & & & \\ & 0 & -u & 0 & -\frac{1}{5}\alpha_1 & & & \\ & 0 & -v & 0 & -\frac{1}{5}\beta_1 & & & \\ & 0 & -\frac{3}{3}\alpha_1 & 0 & -u & & & \\ & 0 & -\frac{3}{3}\beta_1 & 0 & -v & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & & 0 & -u & 0 & -\frac{N-1}{2N+1}\alpha_1 \\ & & & & & 0 & -v & 0 & -\frac{N-1}{2N+1}\beta_1 \\ & & & & & 0 & -\frac{N+1}{2N-1}\alpha_1 & 0 & -u \\ & & & & & 0 & -\frac{N+1}{2N-1}\beta_1 & 0 & -v \end{pmatrix} \tag{A.3.48}$$

Combining the previous results, we obtain the explicit form given in (3.2.1).

A.4 Proof of Lemma 3.3.2

A.4.1 Proof of the first case

For the first case, $A(V)$ only has two non-zero entries in the first column:

$$A(V) = \begin{pmatrix} a_{11}(V) & 0 \\ a_{21}(V) & 0 \end{pmatrix}. \tag{A.4.1}$$

From the necessary condition given in Lemma 3.3.1, we have that $B(V)$ only has two non-zero entries in the second column:

$$B(V) = \begin{pmatrix} 0 & b_{12}(V) \\ 0 & b_{22}(V) \end{pmatrix}. \tag{A.4.2}$$

Now we compute

$$\begin{aligned} T_2^{-1}A(T_2V)T_2 &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a_{11}(T_2V) & 0 \\ a_{21}(T_2V) & 0 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\ &= \begin{pmatrix} \cos\theta(\cos\theta a_{11}(T_2V) - \sin\theta a_{21}(T_2V)) & \sin\theta(\cos\theta a_{11}(T_2V) - \sin\theta a_{21}(T_2V)) \\ \cos\theta(\sin\theta a_{11}(T_2V) + \cos\theta a_{21}(T_2V)) & \sin\theta(\sin\theta a_{11}(T_2V) + \cos\theta a_{21}(T_2V)) \end{pmatrix} \end{aligned}$$

which should be equal to

$$\cos\theta A(V) + \sin\theta B(V) = \begin{pmatrix} \cos\theta a_{11}(V) & \sin\theta b_{12}(V) \\ \cos\theta a_{21}(V) & \sin\theta b_{22}(V) \end{pmatrix}. \tag{A.4.3}$$

Then we have

$$\begin{aligned} \cos\theta(\cos\theta\, a_{11}(T_2V) - \sin\theta\, a_{21}(T_2V)) &= \cos\theta\, a_{11}(V), \\ \sin\theta(\cos\theta\, a_{11}(T_2V) - \sin\theta\, a_{21}(T_2V)) &= \sin\theta\, b_{12}(V), \\ \cos\theta(\sin\theta\, a_{11}(T_2V) + \cos\theta\, a_{21}(T_2V)) &= \cos\theta\, a_{21}(V), \\ \sin\theta(\sin\theta\, a_{11}(T_2V) + \cos\theta\, a_{21}(T_2V)) &= \sin\theta\, b_{22}(V). \end{aligned} \tag{A.4.4}$$

The above equations can be further simplified as

$$\begin{aligned} \cos\theta a_{11}(T_2V) - \sin\theta a_{21}(T_2V) &= a_{11}(V), \\ \sin\theta a_{11}(T_2V) + \cos\theta a_{21}(T_2V) &= a_{21}(V), \end{aligned} \tag{A.4.5}$$

and

$$\begin{aligned} a_{11}(V) &= b_{12}(V), \\ a_{21}(V) &= b_{22}(V). \end{aligned} \tag{A.4.6}$$

Since $a_{11}(V)$ and $a_{21}(V)$ are linear functions of V, we have

$$\begin{aligned} a_{11}(V) &= c_1 p + c_2 q, \\ a_{21}(V) &= c_3 p + c_4 q, \end{aligned} \tag{A.4.7}$$

where c_1, c_2, c_3, c_4 are constants. Plugging (A.4.7) into (A.4.5), we have

$$\begin{aligned} &\cos\theta(c_1(\cos\theta p + \sin\theta q) + c_2(-\sin\theta p + \cos\theta q)) - \sin\theta(c_3(\cos\theta p + \sin\theta q) + c_4(-\sin\theta p + \cos\theta q)) \\ &\qquad = c_1 p + c_2 q \\ &\sin\theta(c_1(\cos\theta p + \sin\theta q) + c_2(-\sin\theta p + \cos\theta q)) + \cos\theta(c_3(\cos\theta p + \sin\theta q) + c_4(-\sin\theta p + \cos\theta q) \\ &\qquad = c_3 p + c_4 q \end{aligned}$$

which can be simplified as

$$\begin{aligned} &(c_1\cos^2\theta - (c_2 + c_3)\cos\theta\sin\theta + c_4\sin^2\theta)p + (c_2\cos^2\theta + (c_1 - c_4)\cos\theta\sin\theta - c_3\sin^2\theta)q \\ &\qquad = c_1 p + c_2 q \\ &(c_3\cos^2\theta + (c_1 - c_4)\cos\theta\sin\theta - c_2\sin^2\theta)p + (c_4\cos^2\theta - (c_2 + c_3)\cos\theta\sin\theta + c_1\sin^2\theta)q \\ &\qquad = c_3 p + c_4 q \end{aligned}$$

This implies

$$\begin{aligned} c_1\cos^2\theta - (c_2 + c_3)\cos\theta\sin\theta + c_4\sin^2\theta &= c_1 \\ c_2\cos^2\theta + (c_1 - c_4)\cos\theta\sin\theta - c_3\sin^2\theta &= c_2 \end{aligned} \tag{A.4.8}$$

Then we derive the following relations:

$$c_1 = c_4, \quad c_2 = -c_3 \tag{A.4.9}$$

Therefore, we have

$$\begin{aligned} a_{11}(V) &= c_1 p - c_2 q \\ a_{21}(V) &= c_2 p + c_1 q \end{aligned} \tag{A.4.10}$$

Thus, the matrices $A(V)$ and $B(V)$ are

$$A(V) = \begin{pmatrix} c_1p + c_2q & 0 \\ -c_2p + c_1q & 0 \end{pmatrix} = c_1 \begin{pmatrix} p & 0 \\ q & 0 \end{pmatrix} + c_2 \begin{pmatrix} q & 0 \\ -p & 0 \end{pmatrix}, \tag{A.4.11}$$

and

$$B(V) = \begin{pmatrix} 0 & c_1p + c_2q \\ 0 & -c_2p + c_1q \end{pmatrix} = c_1 \begin{pmatrix} 0 & p \\ 0 & q \end{pmatrix} + c_2 \begin{pmatrix} 0 & q \\ 0 & -p \end{pmatrix}. \tag{A.4.12}$$

A.4.2 Proof of the second case

In this case, we assume that $A(V)$ is a diagonal matrix

$$A(V) = \begin{pmatrix} a_{11}(V) & 0 \\ 0 & a_{22}(V) \end{pmatrix}.$$

From Lemma 3.3.1, we have $B(V)$ is also a diagonal matrix

$$B(V) = \begin{pmatrix} b_{11}(V) & 0 \\ 0 & b_{22}(V) \end{pmatrix}.$$

Then we compute

$$\begin{aligned} T_2^{-1}A(T_2V)T_2 &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a_{11}(T_2V) & 0 \\ 0 & a_{22}(T_2V) \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\ &= \begin{pmatrix} \cos^2\theta a_{11}(T_2V) + \sin^2\theta a_{22}(T_2V) & \cos\theta\sin\theta(a_{11}(T_2V) - a_{22}(T_2V)) \\ \cos\theta\sin\theta(a_{11}(T_2V) - a_{22}(T_2V)) & \sin^2\theta a_{11}(T_2V) + \cos^2\theta a_{22}(T_2V) \end{pmatrix}, \end{aligned}$$

which should be equal to

$$\cos\theta A(V) + \sin\theta B(V) = \begin{pmatrix} \cos\theta a_{11}(V) + \sin\theta b_{11}(V) & 0 \\ 0 & \cos\theta a_{22}(V) + \sin\theta b_{22}(V) \end{pmatrix}.$$

Then we have

$$\begin{aligned} a_{11}(T_2V) &= a_{22}(T_2V), \\ \cos^2\theta a_{11}(T_2V) + \sin^2\theta a_{22}(T_2V) &= \cos\theta a_{11}(V) + \sin\theta b_{11}(V), \\ \sin^2\theta a_{11}(T_2V) + \cos^2\theta a_{22}(T_2V) &= \cos\theta a_{22}(V) + \sin\theta b_{22}(V). \end{aligned} \tag{A.4.13}$$

This is reduced to

$$\begin{aligned} a_{11}(V) &= a_{22}(V), \\ b_{11}(V) &= b_{22}(V), \\ a_{11}(T_2V) &= \cos\theta a_{11}(V) + \sin\theta b_{11}(V). \end{aligned} \tag{A.4.14}$$

Next, we assume the linear functions

$$\begin{aligned} a_{11}(V) = a_{22}(V) &= c_1 p + c_2 q, \\ b_{11}(V) = b_{22}(V) &= c_3 p + c_4 q, \end{aligned} \tag{A.4.15}$$

where c_1, c_2, c_3, c_4 are constants. Then we have

$$c_1(\cos\theta p + \sin\theta q) + c_2(-\sin\theta p + \cos\theta q) = \cos\theta(c_1 p + c_2 q) + \sin\theta(c_3 p + c_4 q), \tag{A.4.16}$$

which means that

$$c_1 = c_4, \quad c_2 = -c_3. \tag{A.4.17}$$

Therefore, we have

$$A(V) = \begin{pmatrix} c_1 p + c_2 q & 0 \\ 0 & c_1 p + c_2 q \end{pmatrix} = c_1 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} + c_2 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix}, \tag{A.4.18}$$

and

$$B(V) = \begin{pmatrix} c_3 p + c_4 q & 0 \\ 0 & c_3 p + c_4 q \end{pmatrix} = -c_2 \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} + c_1 \begin{pmatrix} q & 0 \\ 0 & q \end{pmatrix}. \tag{A.4.19}$$

A.4.3 Proof of the third case

In the last case, we assume that $A(V)$ only has non-zero entries in the second column:

$$A(V) = \begin{pmatrix} 0 & a_{12}(V) \\ 0 & a_{22}(V) \end{pmatrix}, \tag{A.4.20}$$

and from Lemma 3.3.1, we have

$$B(V) = \begin{pmatrix} b_{11}(V) & 0 \\ b_{21}(V) & 0 \end{pmatrix}. \tag{A.4.21}$$

We compute

$$
\begin{aligned}
T_2^{-1}A(T_2V)T_2 &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 0 & a_{12}(T_2V) \\ 0 & a_{22}(T_2V) \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \begin{pmatrix} 0 & \cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V) \\ 0 & \sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V) \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \\
&= \begin{pmatrix} -\sin\theta(\cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V)) & \cos\theta(\cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V)) \\ -\sin\theta(\sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V)) & \cos\theta(\sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V)) \end{pmatrix}
\end{aligned}
$$

which should be equal to

$$
\cos\theta A(V) + \sin\theta B(V) = \begin{pmatrix} \sin\theta b_{11}(V) & \cos\theta a_{12}(V) \\ \sin\theta b_{21}(V) & \cos\theta a_{22}(V) \end{pmatrix}. \tag{A.4.22}
$$

Therefore, we have

$$
\begin{aligned}
-(\cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V)) &= b_{11}(V) \\
\cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V) &= a_{12}(V) \\
-(\sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V)) &= b_{21}(V) \\
\sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V) &= a_{22}(V)
\end{aligned} \tag{A.4.23}
$$

which is reduced to

$$
\begin{aligned}
b_{11}(V) &= -a_{12}(V) \\
b_{21}(V) &= -a_{22}(V) \\
\cos\theta a_{12}(T_2V) - \sin\theta a_{22}(T_2V) &= a_{12}(V) \\
\sin\theta a_{12}(T_2V) + \cos\theta a_{22}(T_2V) &= a_{22}(V)
\end{aligned} \tag{A.4.24}
$$

Next, we assume the linear function

$$
\begin{aligned}
a_{12}(V) &= c_1 p + c_2 q, \\
a_{22}(V) &= c_3 p + c_4 q,
\end{aligned} \tag{A.4.25}
$$

where c_1, c_2, c_3, c_4 are constants. Then we have

$$\begin{aligned}\cos\theta(c_1(\cos\theta p+\sin\theta q)+c_2(-\sin\theta p+\cos\theta q))-\sin\theta(c_3(\cos\theta p+\sin\theta q)\\ +c_4(-\sin\theta p+\cos\theta q))=c_1p+c_2q,\\ \sin\theta(c_1(\cos\theta p+\sin\theta q)+c_2(-\sin\theta p+\cos\theta q))+\cos\theta(c_3(\cos\theta p+\sin\theta q)\\ +c_4(-\sin\theta p+\cos\theta q))=c_3p+c_4q,\end{aligned}$$

from which we solve out

$$c_1=c_4,\quad c_2=-c_3. \tag{A.4.26}$$

Therefore, we have

$$A(V)=\begin{pmatrix}0 & c_1p+c_2q\\ 0 & -c_2p+c_1q\end{pmatrix}=c_1\begin{pmatrix}0 & p\\ 0 & q\end{pmatrix}+c_2\begin{pmatrix}0 & q\\ 0 & -p\end{pmatrix}, \tag{A.4.27}$$

and

$$B(V)=\begin{pmatrix}-c_1p-c_2q & 0\\ c_2p-c_1q & 0\end{pmatrix}=c_1\begin{pmatrix}-p & 0\\ -q & 0\end{pmatrix}+c_2\begin{pmatrix}-q & 0\\ p & 0\end{pmatrix}. \tag{A.4.28}$$