ESAIM: M2AN 49 (2015) 1019–1046
DOI: 10.1051/m2an/2014065

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

NUMERICAL SOLUTION OF THE VISCOUS SURFACE WAVE WITH DISCONTINUOUS GALERKIN METHOD *

Lei Wu[1] and Chi-Wang Shu[1]

Abstract. We consider an incompressible viscous flow without surface tension in a finite-depth domain of two dimensions, with free top boundary and fixed bottom boundary. This system is governed by the Navier–Stokes equations in this moving domain and the transport equation on the moving boundary. In this paper, we construct a stable numerical scheme to simulate the evolution of this system by discontinuous Galerkin method, and discuss the error analysis of the fluid under certain assumptions. Our formulation is mainly based on the geometric structure introduced in [Y. Guo and Ian Tice, *Anal. PDE* **6** (2013) 287–369; Y. Guo and Ian Tice, *Arch. Ration. Mech. Anal.* **207** (2013) 459–531; L. Wu, *SIAM J. Math. Anal.* **46** (2014) 2084–2135], and the natural energy estimate, which is rarely used in the numerical study of this system before.

Mathematics Subject Classification. 35Q30, 35R35, 74S05.
Received June 1, 2014.
Published online 25 June 2015.

1. Introduction

We consider an incompressible viscous flow in the moving domain

$$\Lambda(t) = \{y = (y_1, y_2) \in \mathbb{T} \times \mathbb{R} \ : \ -1 < y_2 < \eta(y_1, t)\}, \tag{1.1}$$

where $\mathbb{T}$ denotes the 1-torus. We denote the initial domain $\Lambda(0) = \Lambda_0$. For each t, the flow is described by its velocity and pressure $(u, p) : \Lambda(t) \mapsto \mathbb{R}^2 \times \mathbb{R}$ which satisfies the incompressible Navier–Stokes equations

$$\begin{cases} \partial_t u + u \cdot \nabla u + \nabla p = \nu \Delta u + S & \text{in} \quad \Lambda(t), \\ \nabla \cdot u = 0 & \text{in} \quad \Lambda(t), \\ (pI - \nu \mathbb{D}(u))\mu = g\eta\mu & \text{on} \quad \{y_2 = \eta(y_1, t)\}, \\ u = 0 & \text{on} \quad \{y_2 = -1\}, \\ \partial_t \eta = u_2 - u_1 \partial_{y_1} \eta & \text{on} \quad \{y_2 = \eta(y_1, t)\}, \\ \\ u(t = 0) = u_0 & \text{in} \quad \Lambda_0, \\ \eta(t = 0) = \eta_0 & \text{on} \quad \mathbb{T}, \end{cases} \tag{1.2}$$

Keywords and phrases. Stability, free boundary, Navier–Stokes equation.

* *Research supported by DOE grant DE-FG02-08ER25863 and NSF grants DMS-1112700 and DMS-1418750.*

[1] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Lei_Wu@brown.edu; shu@dam.brown.edu

Article published by EDP Sciences

for μ the outward-pointing unit normal vector on $\{y_2 = \eta\}$, I the 2×2 identity matrix, $(\mathbb{D}u)_{ij} = \partial_i u_j + \partial_j u_i$ the symmetric gradient of u, g the gravitational constant, $\nu > 0$ the viscosity and $S = S(t, y_1, y_2)$ an external source term. The fifth equation in (1.2) implies the free surface is convected with the fluid. Note in (1.2), we have shifted the actual pressure $\bar{p}$ by the constant atmosphere pressure p_{atm} according to $p = \bar{p} + gy_2 - p_{atm}$.

We always assume the natural condition that there exists a positive number δ such that $\eta_0 + 1 \geq \delta > 0$ on Σ, which means the initial free surface is strictly separated from the bottom.

Traditionally, based on the handling of the free surface, this type of problems can be solved via moving-grid technique as in [17], marker-and-cell method as in [13], volume-of-fluid method as in [14] and level-set method as in [9]. In each case, finite difference method, finite volume method and finite element method can be applied to solve the Navier–Stokes equation in $\Lambda(t)$. However, to the best of authors' knowledge, there is very little in the literature on solving this problem with discontinuous Galerkin method other than [9]. On the other hand, in spite of many computational tests presented in the literature, there is very little discussion on the stability and convergence of the numerical scheme. In this paper, we employ the idea from [10, 11, 20], to construct a stable numerical scheme and give detailed analysis.

Our central idea is to flatten the free surface *via* a coordinates transform. First, we define a fixed domain

$$\Omega = \{x = (x_1, x_2) \in \Sigma \times \mathbb{R} \mid -1 < x_2 < 0\}, \tag{1.3}$$

for which we write the coordinates $x \in \Omega$. In this slab, we take $\Sigma : \{x_2 = 0\}$ as the upper boundary and $\Sigma_b : \{x_2 = -1\}$ as the lower boundary.

Consider the geometric transform from Ω to $\Lambda(t)$, which is first introduced in [2] and further extended in [10, 20]:

$$\Phi : (x_1, x_2) \mapsto (x_1, x_2 + \eta\,(1 + x_2)) = (y_1, y_2)\,. \tag{1.4}$$

We may directly verify this transform maps Ω into $\Lambda(t)$ with the Jacobi matrix

$$\nabla\Phi = \begin{pmatrix} 1 & 0 \\ A & J \end{pmatrix}, \tag{1.5}$$

and the transform matrix

$$\mathscr{A} = \left((\nabla\Phi)^{-1}\right)^T = \begin{pmatrix} 1 & -AK \\ 0 & K \end{pmatrix}, \tag{1.6}$$

where

$$\begin{aligned} \tilde{b} &= 1 + x_2, \qquad & A &= \partial_1\eta\tilde{b}, \\ J &= 1 + \eta, \qquad & K &= 1/J. \end{aligned} \tag{1.7}$$

Here we denote the derivative with respect to x_1 as ∂_1 and with respect to x_2 as ∂_2. Define the transformed operators as follows:

$$\begin{aligned} (\nabla_{\mathscr{A}} f)_i &= \mathscr{A}_{ij}\partial_j f, \\ \nabla_{\mathscr{A}} \cdot \boldsymbol{g} &= \mathscr{A}_{ij}\partial_j g_i, \\ \Delta_{\mathscr{A}} f &= \nabla_{\mathscr{A}} \cdot \nabla_{\mathscr{A}} f, \\ \mathscr{N} &= (-\partial_1\eta, 1)\,, \\ \chi &= \partial_1\eta, \\ (\mathbb{D}_{\mathscr{A}} u)_{ij} &= \mathscr{A}_{ik}\partial_k u_j + \mathscr{A}_{jk}\partial_k u_i, \\ S_{\mathscr{A}}(p, u) &= pI - \mathbb{D}_{\mathscr{A}} u, \end{aligned} \tag{1.8}$$

where the summation index should be understood in the Einstein convention. If we extend the divergence $\nabla_{\mathscr{A}}\cdot$ to act on symmetric tensors in the natural way, then a straightforward computation reveals $\nabla_{\mathscr{A}}\cdot S_{\mathscr{A}}(p,u) = \nabla_{\mathscr{A}} p - \Delta_{\mathscr{A}} u$ for vector fields satisfying $\nabla_{\mathscr{A}}\cdot u = 0$.

In our new coordinates, the original system (1.2) becomes

$$\begin{cases} \partial_t u - \partial_t \eta \tilde{b} K \partial_2 u + u\cdot\nabla_{\mathscr{A}} u - \nu\Delta_{\mathscr{A}} u + \nabla_{\mathscr{A}} p = S & \text{in} \quad \Omega, \\ \nabla_{\mathscr{A}}\cdot u = 0 & \text{in} \quad \Omega, \\ S_{\mathscr{A}}(p,u)\mathscr{N} = g\eta\mathscr{N} & \text{on} \quad \Sigma, \\ u = 0 & \text{on} \quad \Sigma_b, \\ u(x,0) = u_0(x) & \text{in} \quad \Omega, \\ \\ \partial_t \eta = u\cdot\mathscr{N} & \text{on} \quad \Sigma, \\ \eta(x',0) = \eta_0(x') & \text{on} \quad \Sigma. \end{cases} \tag{1.9}$$

Since $\mathscr{A}$ depends on η through the transform, most of the operators in the Navier–Stokes equations are related to the free surface η. Hence, the Navier–Stokes equations and the transport equation are essentially coupled.

Based on [10], equation (1.9) with $S = 0$ possesses a natural energy equality as follows:

$$\int_\Omega J(t)\,|u(t)|^2 + g\int_\Sigma |\eta(t)|^2 + \nu\int_0^t\int_\Omega J(s)\,|\mathbb{D}_{\mathscr{A}} u(s)|^2\,\mathrm{d}s = \int_\Omega J(0)\,|u(0)|^2 + g\int_\Sigma |\eta(0)|^2\,. \tag{1.10}$$

Hence, we try to construct a numerical scheme to recover this energy stability for the numerical solution.

In the following, we refer to the term "continuous case" when we consider the exact solution triple (u,p,η) which is sufficiently smooth. On the other hand, we refer to the term "discrete case" when we consider the numerical solution triple (u_h,p_h,η_h).

Throughout this paper, $C>0$ denotes a positive constant that only depends on the parameters Ω, g and ν of the problem, and the exact solution (u,p,η). It is referred as universal and can change from one inequality to another. When we write $C(z)$, it means a positive constant depending on the quantity z. $a\lesssim b$ denotes $a\le Cb$, where C is a universal constant as defined above.

The method we discuss in this paper belongs to the class of discontinuous Galerkin (DG) methods. DG methods are finite element methods which use completely discontinuous piecewise polynomial solution and test spaces. They are particularly useful for convection dominated wave problems, and have the advantage of flexibility in adaptivity and efficient parallel implementation. We refer to the references [3–5,19] for more details.

2. Numerical scheme

2.1. Fundamental settings

In our construction and analysis of the numerical scheme, we focus on the semi-discrete form of the system. The time discretization is based on the Runge–Kutta method for differential-algebraic equations of index 2, as presented in [12]. We do not discuss the fully-discrete scheme in this paper.

We choose the bulk domain as $\Omega : (x_1,x_2)\in[0,1]\times[-1,0]$ and the surface domain as $\Sigma : x_1\in[0,1]$, which are 1-periodic in the x_1 direction. The surface mesh is defined by dividing Σ into N uniform elements with length $h = 1/N$. The bulk mesh construction can be divided into two steps: first we divide Ω into $N\times N$ uniform squares, where each element is sized $h\times h$. Then, we divide each square into two right triangles by cutting along the diagonal from the left-up corner to the right-down corner. The bulk mesh is shown in Figure 1.

Remark 2.1. Since the discretization of the weighted differential operators $\nabla_{\mathscr{A}}$, $\nabla_{\mathscr{A}}\cdot$ and $\Delta_{\mathscr{A}}$ depends on η, *i.e.* the surface variable, we need the bulk discretization to match the surface discretization in the vertical direction. Hence, we need to first choose the strip-shape mesh as in Figure 2, whose projection on the upper

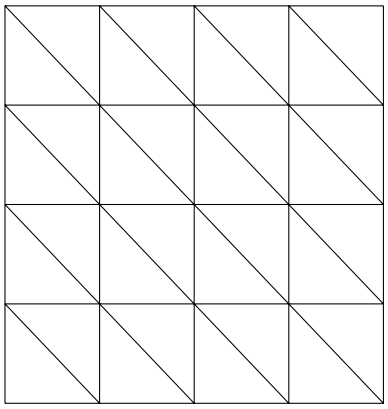

FIGURE 1. Mesh distribution in $N = 4$.

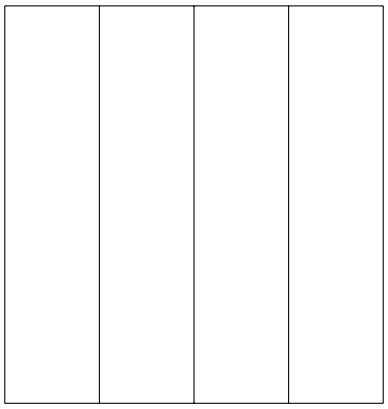

FIGURE 2. Strip–shape mesh distribution in $N = 4$.

boundary is exactly the surface mesh. Then in each strip, any triangulation is permitted. For convenience, we made the choice as in Figure 1.

Let $\mathcal{E}_h$ be the set consisting of all the triangular elements in the bulk mesh and $\partial\mathcal{E}_h$ be the set of all the sides of the triangular cells. Let $\mathcal{F}_h$ be the set consisting of all the interval elements in the surface mesh and $\partial\mathcal{F}_h$ be the set of all the boundary points of intervals.

Define the usual Sobolev space in Ω as $H^k(\Omega)$ and in Σ as $H^k(\Sigma)$. We define the space $P_h^k(\Omega)$ where $f \in P_h^k(\Omega)$ if and only if $f|_E \in P^k(E)$ for all $E \in \mathcal{E}_h$ ($P^k(E)$ denotes the set of polynomials of degree at most k defined on E). Similarly, we can define the space $P_h^k(\Sigma)$.

We define the solution spaces as

$$X_h^k = \left\{ u_h \in \left(L^2(\Omega)\right)^2 : \ u_h \in \left(P_h^k(\Omega)\right)^2 \right\}, \tag{2.1}$$

$$M_h^k = \left\{ p_h \in L^2(\Omega) : \ p_h \in P_h^k(\Omega) \right\}, \tag{2.2}$$

$$S_h^k = \left\{ \eta_h \in L^2(\Sigma) : \ \eta_h \in P_h^k(\Sigma) \right\}, \tag{2.3}$$

for $k \geq 1$.

For discrete functions $f_h \in X_h^k$ and $g_h \in S_h^k$, we may define the H_h norms as follows:

$$\|f_h\|_{H_h}^2 = \sum_{E \in \mathcal{E}_h} \|f_h\|_{H^1(E)}^2, \tag{2.4}$$

$$\|g_h\|_{H_h}^2 = \sum_{F \in \mathcal{F}_h} \|g_h\|_{H^1(F)}^2. \tag{2.5}$$

Define the X_h norm for $v_h \in X_h^k$ as

$$\|v_h\|_{X_h}^2 = \sum_{E \in \mathcal{E}_h} \left(\|\nabla v_h\|_{L^2(E)}^2 + \int_{\partial E} \frac{1}{h} [v_h]^2 \right). \tag{2.6}$$

Before stating the scheme, we announce the notations used below. For a given boundary belonging to a reference cell $E \in \mathcal{E}_h$ and a function f defined in Ω, $f^{\mathrm{ext}(E)}$ denotes the value of f read from the exterior direction on the

boundary and $f^{\mathrm{int}(E)}$ denotes that read from the interior direction. Also for the boundary shared by two cells, we define

$$\{f\} = \frac{1}{2}\left(f^{\mathrm{ext}(E)} + f^{\mathrm{int}(E)}\right), \tag{2.7}$$

$$[f] = f^{\mathrm{int}(E)} - f^{\mathrm{ext}(E)}. \tag{2.8}$$

For the side on the boundary of Ω, we make a special definition that on Σ_b, $f^{\mathrm{int}(E)} = \{f\} = [f]$ and $f^{\mathrm{ext}(E)} = 0$, while on Σ, $f^{\mathrm{int}(E)} = f^{\mathrm{ext}(E)} = \{f\}$ and $[f] = 0$. The definitions are reasonable, since the system (1.9) for u gives Dirichlet-type condition on Σ_b and Neumann-type condition on Σ.

The numerical scheme is based on the weak formulation of the system (1.9). Taking test functions $v_h \in X_h^k$ and $q_h \in M_h^{k-1}$. We multiply Jv_h and Jq_h respectively on both sides of the Navier–Stokes equations and integrate over $E \in \mathcal{E}_h$ to obtain

$$\begin{aligned}&\int_E J\partial_t u \cdot v_h - \int_E J\left(\partial_t \eta \tilde{b} K \partial_2 u\right)\cdot v_h \\ &+ \int_E J(u\cdot\nabla_{\mathscr{A}} u)\cdot v_h - \nu\int_E J\Delta_{\mathscr{A}} u\cdot v_h + \int_E J\nabla_{\mathscr{A}} p\cdot v_h = \int_E J\left(S\cdot v_h\right),\end{aligned} \tag{2.9}$$

$$\int_E J(\nabla_{\mathscr{A}}\cdot u) q_h = 0. \tag{2.10}$$

Considering the transport equation in (1.9), we make an addition and subtraction to modify two terms in above formulation as

$$\begin{aligned}&\left(\int_E J\partial_t u\cdot v_h - \int_E J\left(\partial_t\eta\tilde{b}K\partial_2 u\right)\cdot v_h + \frac{1}{2}\int_{\partial E\cap\Sigma}\partial_t\eta\left(u\cdot v_h\right)\right) \\ &+\left(\int_E J(u\cdot\nabla_{\mathscr{A}}u)\cdot v_h - \frac{1}{2}\int_{\partial E\cap\Sigma}(u\cdot\mathscr{N})\left(u\cdot v_h\right)\right) - \nu\int_E J\Delta_{\mathscr{A}}u\cdot v_h + \int_E J\nabla_{\mathscr{A}}p\cdot v_h = \int_E J\left(S\cdot v_h\right),\end{aligned} \tag{2.11}$$

$$\int_E J(\nabla_{\mathscr{A}}\cdot u) q_h = 0, \tag{2.12}$$

where ∂E denotes the boundary of the triangular element E. This trick is to enforce the transport relation and shows its power in the stability proof.

Multiplying a test functions $\phi_h \in S_h^k$ on both sides of the transport equation and integrating over $F \in \mathcal{F}_h$, we obtain

$$\int_F \partial_t\eta\phi_h + \int_F \bar{u}_1\partial_1\eta\phi_h - \int_F \bar{u}_2\phi_h = 0, \tag{2.13}$$

where $\bar{u}_1$ and $\bar{u}_2$ denote the traces of u_1 and u_2 on Σ.

The weak formulations of the system (1.9) are based on the integration by parts of (2.11), (2.12) and (2.13). In the following, we define and analyze the numerical scheme term by term.

2.2. Discretization

We define the semi-discrete scheme as follows:

$$\begin{cases} \kappa_F\left(\eta_h, u_h, \phi_h\right) = 0, \\ \lambda_F\left(\chi_h, \psi_h\right) = 0, \\ \zeta_E\left(\eta_h, u_h, v_h\right) + \gamma_E\left(u_h, \eta_h, u_h, v_h\right) + \nu\alpha_E\left(\eta_h, u_h, \eta_h, v_h\right) \\ \qquad + \beta_E\left(p_h, \eta_h, v_h\right) + g\mu_E\left(\eta_h, \eta_h, v_h\right) = \omega_E\left(\eta_h, v_h\right), \\ \rho_E\left(u_h, \eta_h, q_h\right) = 0, \end{cases} \tag{2.14}$$

for any test functions $\phi_h \in S_h^k$, $\psi_h \in S_h^k$, $v_h \in X_h^k$ and $q_h \in M_h^{k-1}$, where the first two equations denote the transport discretization and the last two equations denote the fluid discretization. This forms a complete differential-algebraic system for $\eta_h(t) \in S_h^k$, $\chi_h(t) \in S_h^k$, $u_h(t) \in X_h^k$ and $p_h(t) \in M_h^{k-1}$. The detailed definitions of above multi-linear terms κ, λ, ζ, γ, α, β, μ, ω, and ρ are in the following.

2.2.1. Discretization of transport terms as κ_F and λ_F

We define

$$\kappa_F(\eta_h, u_h, \phi_h) = \int_F \partial_t \eta_h \phi_h - \int_F (\bar{u}_2)_h \phi_h - \int_F \partial_1 (\bar{u}_1)_h \eta_h \phi_h - \int_F (\bar{u}_1)_h \eta_h \partial_1 \phi_h$$
$$+ (\hat{u}_1)_h \hat{\eta}_h \phi_h^-|_{F+1/2} - (\hat{u}_1)_h \hat{\eta}_h \phi_h^+|_{F-1/2} + \frac{1}{2}\left(\eta_h^- \phi_h^- [u_h]|_{F+1/2} - \eta_h^+ \phi_h^+ [u_h]|_{F-1/2}\right) \quad (2.15)$$

$$\lambda_F(\chi_h, \psi_h) = \int_F \chi_h \psi_h + \int_F \eta_h \partial_1 \psi_h - \eta_h^+ \psi_h^-|_{F+1/2} + \eta_h^+ \psi_h^+|_{F-1/2}, \quad (2.16)$$

where $F-1/2$ and $F+1/2$ denote the left and right boundary points of F. Here, for each boundary point, "$-$" means the value read from the left and "$+$" means the value read from the right. Also, $(\hat{u}_1)_h$ $\hat{\eta}_h$ denotes the numerical flux on the boundary. We always take

$$(\hat{u}_1)_h = \{(\bar{u}_1)_h\}, \quad (2.17)$$

and $\hat{\eta}_h$ should be determined from the sign of $\{(\bar{u}_1)_h\}$ following the upwinding rule as

$$\hat{\eta}_h = \begin{cases} \eta_h^- \text{ if } \{(\bar{u}_1)_h\} \geq 0, \\ \eta_h^+ \text{ if } \{(\bar{u}_1)_h\} < 0, \end{cases} \quad (2.18)$$

where $(\bar{u}_1)_h$ is the trace of $(u_1)_h$ on Σ. Note $\frac{1}{2}\left(\eta_h^- \phi_h^- [u_h]|_{F+1/2} - \eta_h^+ \phi_h^+ [u_h]|_{F-1/2}\right)$ are extra penalty terms, which helps to show the stability. Since u_h is also discontinuous across the boundary point, these penalty terms are used to transform u_h into $\{u_h\}$ on the boundary after integrating by parts in (2.15).

In all the applications below, we use χ_h to discretize $\partial_1 \eta$, and η_h to discretize η. Hence, we denote $\mathscr{N}_h$ for the vector $(-\chi_h, 1)$. Also, A_h, J_h, K_h and $\mathscr{A}_h$ can be defined in the same convention.

2.2.2. Spacial discretization of temporal terms as ζ_E

We define

$$\zeta_E(\eta_h, u_h, v_h) = \int_E \partial_t (J_h u_h) \cdot v_h + \int_E \partial_t \eta_h \tilde{b} u_h \cdot \partial_2 v_h - \int_{\partial E} \partial_t \eta_h \tilde{b} \left(\hat{u}_h \cdot v_h^{\text{int}(E)}\right) n_2$$
$$+ \frac{1}{2} \int_{\partial E \cap \Sigma} \partial_t \eta_h (u_h \cdot v_h), \quad (2.19)$$

where $n_E = (n_1, n_2)$ denotes the outward normal vector on ∂E. The last term in (2.19) is the newly-added penalty term in the formulation (2.11). The argument η_h in $\zeta_E(\eta_h, u_h, v_h)$ denotes η_h and its derivatives. We utilize the upwinding flux in this scheme, *i.e.*

$$\hat{u}_h = \begin{cases} u_h^+ \text{ if } \partial_t \eta_h \tilde{b} \geq 0, \\ u_h^- \text{ if } \partial_t \eta_h \tilde{b} < 0, \end{cases} \quad (2.20)$$

where "$+$" denotes the value read from the up direction and "$-$" denote that from the down direction. If a side is vertical, then its contribution is zero.

2.2.3. Discretization of convection terms as γ_E

We define

$$\begin{aligned}\gamma_E\left(u_h,\eta_h,u_h,v_h\right)=&-\frac{1}{2}\int_{\partial E\cap\Sigma}\left(u_h\cdot\mathscr{N}_h\right)\left(u_h\cdot v_h\right)+\int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h}u_h\right)\cdot v_h\\&+\frac{1}{2}\int_E J_h\left(\nabla_{\mathscr{A}_h}\cdot u_h\right)\left(u_h\cdot v_h\right)-\frac{1}{2}\int_{\partial E\setminus\partial\Omega}\left[u_h\cdot J_h\mathscr{A}_h\right]\cdot n_E\frac{u_h^{\mathrm{int}(E)}\cdot v_h^{\mathrm{int}(E)}}{2}\\&+\int_{\partial E^-}\left|\left\{u_h\cdot J_h\mathscr{A}_h\right\}\cdot n_E\right|\left(u_h^{\mathrm{int}(E)}-u_h^{\mathrm{ext}(E)}\right)\cdot v_h^{\mathrm{int}(E)},\end{aligned}\tag{2.21}$$

where $\nabla_{\mathscr{A}_h}$ is understood as in (1.8) with $\mathscr{A}$ replaced by $\mathscr{A}_h$ and

$$\partial E^-=\left\{x\in\partial E:\left\{u_h\cdot J_h\mathscr{A}_h\right\}\cdot n_E<0\right\}.\tag{2.22}$$

The first term in (2.21) is the newly added penalty term in the formulation (2.11). Since

$$J_h\mathscr{A}_h=\begin{pmatrix}J_h & -A_h\\0 & 1\end{pmatrix},\tag{2.23}$$

we only need one η_h argument in $\gamma_E(u_h,\eta_h,u_h,v_h)$.

2.2.4. Discretization of diffusion terms as α_E

We utilize the Symmetric Internal Penalty Galerkin Method (SIPG) or the Nonsymmetric Internal Penalty Galerkin Method (NIPG) to define

$$\begin{aligned}\alpha_E\left(\eta_h,u_h,\eta_h,v_h\right)=&\frac{1}{2}\int_E J_h\mathbb{D}_{\mathscr{A}_h}u_h:\mathbb{D}_{\mathscr{A}_h}v_h-\int_{\partial E\setminus\Sigma}\left\{\mathbb{D}_{\mathscr{A}_h}u_hJ_h\mathscr{A}_h\right\}\cdot n_E\cdot v_h^{\mathrm{int}(E)}\\&\pm\frac{1}{2}\int_{\partial E\setminus\Sigma}\mathbb{D}_{\mathscr{A}_h^{\mathrm{int}(E)}}v_h^{\mathrm{int}(E)}J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}\cdot n_E\cdot[u_h]+\frac{\sigma}{h}\int_{\partial E}[u_h]\cdot v_h^{\mathrm{int}(E)}.\end{aligned}\tag{2.24}$$

For the third term, if we take $+$, it is NIPG and if we take $-$, it is SIPG. The penalty constant σ can always be taken as 1 for NIPG and a sufficiently large number for SIPG. Note that we need two η_h arguments in $\alpha_E(\eta_h,u_h,\eta_h,v_h)$ to denote the dependence of $\mathscr{A}_h$.

2.2.5. Discretization of pressure terms as β_E

We define

$$\beta_E\left(p_h,\eta_h,v_h\right)=-\int_E J_hp_h\nabla_{\mathscr{A}_h}\cdot v_h+\int_{\partial E\setminus\Sigma}v_h^{\mathrm{int}(E)}\cdot\left\{p_hJ_h\mathscr{A}_h\right\}\cdot n_E.\tag{2.25}$$

2.2.6. Discretization of forcing terms as μ_E

We define

$$\mu_E\left(\eta_h,\eta_h,v_h\right)=\int_{\partial E\cap\Sigma}\eta_h\left(v_2\right)_h-\int_{\partial E\cap\Sigma}\eta_h\partial_1\eta_h\left(\bar{v}_1\right)_h+\frac{1}{4}\left[\eta_h^2\right]\bar{v}_h^-|_{\partial E\cap\Sigma+1/2}-\frac{1}{4}\left[\eta_h^2\right]\bar{v}_h^+|_{\partial E\cap\Sigma-1/2},\tag{2.26}$$

where $\partial E\cap\Sigma+1/2$ and $\partial E\cap\Sigma-1/2$ denote the left and right boundary points of $\partial E\cap\Sigma$ respectively, *i.e.* similar to $F-1/2$ and $F+1/2$. Note that the forcing term is nontrivial only for the top cells and for all other cells it is zero, $\mu_E(\eta_h,\eta_h,v_h)=0$. $\frac{1}{4}[\eta_h^2]\bar{v}_h^-|_{\partial E\cap\Sigma+1/2}-\frac{1}{4}[\eta_h^2]\bar{v}_h^+|_{\partial E\cap\Sigma-1/2}$ are extra penalty terms, which help to show the stability. Since η_h is discontinuous across the boundary, these penalty terms are used to transform η_h^2 into $\{\eta_h^2\}$ on the boundary after integrating by parts in (2.26).

2.2.7. Discretization of source terms as ω_E

We define

$$\omega_E\left(\eta_h, v_h\right) = \int_E J_h S \cdot v_h. \tag{2.27}$$

2.2.8. Discretization of divergence terms as ρ_E

We define

$$\rho_E\left(u_h, \eta_h, q_h\right) = -\int_E J_h q_h \nabla_{\mathscr{A}_h} \cdot u_h + \frac{1}{2}\int_{\partial E \setminus \Sigma} [u_h] \cdot q_h^{\mathrm{int}(E)} J_h^{\mathrm{int}(E)} \mathscr{A}_h^{\mathrm{int}(E)} \cdot n_E. \tag{2.28}$$

2.3. Properties and estimates of discretization

2.3.1. Estimate of temporal terms

We consider the temporal terms, *i.e.*

$$\left(\int_E J\partial_t u \cdot v_h - \int_E J\left(\partial_t \eta \tilde{b} K \partial_2 u\right) \cdot v_h + \frac{1}{2}\int_{\partial E \cap \Sigma} \partial_t \eta \left(u \cdot v_h\right)\right). \tag{2.29}$$

In the continuous case, a direct integration by parts reveals

$$\int_E J\partial_t u \cdot v_h = \int_E \partial_t (Ju) \cdot v_h - \int_E \partial_t J \left(u \cdot v_h\right), \tag{2.30}$$

$$\begin{aligned} -\int_E J\left(\partial_t \eta \tilde{b} K \partial_2 u\right) \cdot v_h &= -\int_E \partial_t \eta \tilde{b} \partial_2 u \cdot v_h \\ &= \int_E \partial_t \eta \tilde{b} u \cdot \partial_2 v_h + \int_E \partial_2\left(\partial_t \eta \tilde{b}\right)\left(u \cdot v_h\right) - \int_{\partial E} \partial_t \eta \tilde{b}\left(u \cdot v_h\right) n_2. \end{aligned} \tag{2.31}$$

Since $\partial_t J = \partial_2(\partial_t \eta \tilde{b})$, we can simplify above terms into

$$\int_E J\partial_t u \cdot v_h - \int_E J\left(\partial_t \eta \tilde{b} K \partial_2 u\right) \cdot v_h = \int_E \partial_t (Ju) \cdot v_h + \int_E \partial_t \eta \tilde{b} u \cdot \partial_2 v_h - \int_{\partial E} \partial_t \eta \tilde{b}\left(u \cdot v_h\right) n_2. \tag{2.32}$$

Hence, we may define the discretization of the temporal term as (2.19). Since in our bulk mesh, there are only two effective boundary integrals for each triangle, we may call them the upper boundary $\partial E + 1/2$ and lower boundary $\partial E - 1/2$ without ambiguity. If we denote $f(u_h) = -\partial_t \eta_h \tilde{b} u_h$, then (2.19) is actually

$$\begin{aligned} &\zeta_E\left(\eta_h, u_h, v_h\right) - \frac{1}{2}\int_{\partial E \cap \Sigma} \partial_t \eta_h \left(u_h \cdot v_h\right) \\ &\qquad = \int_E \partial_t\left(J_h u_h\right) \cdot v_h - \int_E f\left(u_h\right) \cdot \partial_2 v_h + \int_{\partial E + 1/2} \hat{f}\left(u_h\right) \cdot v_h^- - \int_{\partial E - 1/2} \hat{f}\left(u_h\right) \cdot v_h^+, \end{aligned} \tag{2.33}$$

where the flux reduces to

$$\hat{f}\left(u_h\right) = f\left(\hat{u}_h\right) = \begin{cases} f\left(u_h^+\right) \text{ if } \partial_t \eta_h \tilde{b} \geq 0, \\ f\left(u_h^-\right) \text{ if } \partial_t \eta_h \tilde{b} < 0. \end{cases} \tag{2.34}$$

Lemma 2.2. *If we take $v_h = u_h$, the discretized temporal term satisfies*

$$\sum_{E \in \mathcal{E}_h} \zeta_E\left(\eta_h, u_h, u_h\right) \geq \frac{1}{2}\partial_t \int_\Omega J_h \left|u_h\right|^2. \tag{2.35}$$

Proof. Note the fact that $\partial_t J_h = \partial_2(\partial_t \eta_h \tilde{b})$. We can directly compute

$$\begin{aligned}
&\zeta_E(\eta_h, u_h, u_h) - \frac{1}{2}\int_{\partial E\cap\Sigma} \partial_t\eta_h |u_h|^2 \\
&= \int_E \partial_t(J_h u_h)\cdot u_h - \int_E f(u_h)\cdot\partial_2 u_h + \int_{\partial E+1/2} \hat{f}(u_h)\cdot u_h^- - \int_{\partial E-1/2} \hat{f}(u_h)\cdot u_h^+ \\
&= \frac{1}{2}\partial_t \int_E J_h |u_h|^2 + \frac{1}{2}\int_E \partial_t J_h |u_h|^2 - \frac{1}{2}\int_E \partial_2\left(\partial_t\eta_h\tilde{b}\right)|u_h|^2 \\
&\quad -\frac{1}{2}\int_{\partial E+1/2} f\left(u_h^-\right)\cdot u_h^- + \frac{1}{2}\int_{\partial E-1/2} f\left(u_h^+\right)\cdot u_h^+ + \int_{\partial E+1/2}\hat{f}(u_h)\cdot u_h^- - \int_{\partial E-1/2}\hat{f}(u_h)\cdot u_h^+ \\
&= \frac{1}{2}\partial_t \int_E J_h |u_h|^2 \\
&\quad -\frac{1}{2}\int_{\partial E+1/2} f\left(u_h^-\right)\cdot u_h^- + \frac{1}{2}\int_{\partial E-1/2} f\left(u_h^+\right)\cdot u_h^+ + \int_{\partial E+1/2}\hat{f}(u_h)\cdot u_h^- - \int_{\partial E-1/2}\hat{f}(u_h)\cdot u_h^+ \\
&= \frac{1}{2}\partial_t \int_E J_h |u_h|^2 + \hat{F}_{\partial E+1/2} - \hat{F}_{\partial E-1/2} + \Theta_E,
\end{aligned} \tag{2.36}$$

where

$$\hat{F}_{\partial E+1/2} = -\frac{1}{2}\int_{\partial E+1/2} f\left(u_h^-\right)\cdot u_h^- + \int_{\partial E+1/2}\hat{f}(u_h)\cdot u_h^-, \tag{2.37}$$

$$\hat{F}_{\partial E-1/2} = -\frac{1}{2}\int_{\partial E-1/2} f\left(u_h^-\right)\cdot u_h^- + \int_{\partial E-1/2}\hat{f}(u_h)\cdot u_h^-, \tag{2.38}$$

and

$$\Theta_E = \frac{1}{2}\int_{\partial E-1/2}\left(f\left(u_h^+\right)\cdot u_h^+ - f\left(u_h^-\right)\cdot u_h^-\right) + \int_{\partial E-1/2}\left(\hat{f}(u_h)\cdot u_h^- - \hat{f}(u_h)\cdot u_h^+\right). \tag{2.39}$$

Based on the flux definition (2.34), we have the estimate

$$\Theta_E = \int_{\partial E-1/2}\left(\partial_t\eta_h\tilde{b}\left(u_h^+ - u_h^-\right)\cdot\left(\hat{u}_h - \frac{u_h^+ + u_h^-}{2}\right)\right) \geq 0. \tag{2.40}$$

Hence, combining with (2.36), we obtain

$$\zeta_E(\eta_h, u_h, u_h) - \frac{1}{2}\int_{\partial E\cap\Sigma}\partial_t\eta_h|u_h|^2 \geq \frac{1}{2}\partial_t\int_E J_h|u_h|^2 + \hat{F}_{\partial E+1/2} - \hat{F}_{\partial E-1/2}. \tag{2.41}$$

When summing up over all $E\in\mathcal{E}_h$, we can easily see when $\partial E\subset\Omega$, all the terms involving $\hat{F}$ are canceled out. Therefore, only the terms on $\partial\Omega$ remain, *i.e.*

$$\sum_{E\in\mathcal{E}_h}\zeta_E(\eta_h,u_h,u_h) - \frac{1}{2}\int_\Sigma\partial_t\eta_h|u_h|^2 \geq \sum_{E\in\mathcal{E}_h}\frac{1}{2}\partial_t\int_E J_h|u_h|^2 + \sum_{\partial E\cap\Sigma\neq\emptyset}\hat{F}_{\partial E+1/2} - \sum_{\partial E\cap\Sigma_b\neq\emptyset}\hat{F}_{\partial E-1/2}. \tag{2.42}$$

On Σ_b we have $u_h^- = 0$, so $-\hat{F}_{\partial E-1/2} = 0$, *i.e.*

$$\sum_{\partial E\cap\Sigma_b\neq\emptyset}\hat{F}_{\partial E-1/2} = 0. \tag{2.43}$$

On Σ we have $u_h^+ = u_h^- = u_h|_\Sigma$ and we always take $\hat{u}_h = u_h$, so it implies

$$\hat{F}_{\partial E+1/2} = -\frac{1}{2}\int_{\partial E+1/2} f\left(u_h^-\right) u_h^- + \int_{\partial E+1/2} \hat{f}\left(u_h\right) u_h^- = -\frac{1}{2}\int_{\partial E\cap\Sigma} \partial_t \eta_h \left|u_h\right|^2. \tag{2.44}$$

Then summing up over $E \in \mathcal{E}_h$ gives a full integration over Σ, *i.e.*

$$\sum_{\partial E\cap\Sigma\neq\emptyset} \hat{F}_{\partial E+1/2} = -\frac{1}{2}\int_{\Sigma} \partial_t \eta_h \left|u_h\right|^2. \tag{2.45}$$

Therefore, combining (2.42), (2.43) and (2.45), we deduce

$$\sum_{E\in\mathcal{E}_h} \zeta_E\left(\eta_h, u_h, u_h\right) - \frac{1}{2}\int_{\Sigma} \partial_t \eta_h \left|u_h\right|^2 \geq \frac{1}{2}\partial_t \int_E J_h \left|u_h\right|^2 - \frac{1}{2}\int_{\Sigma} \partial_t \eta_h \left|u_h\right|^2. \tag{2.46}$$

Hence, our result easily follows. □

Remark 2.3. This proof is based on the mesh as in Figure 1. For general triangulation, it is possible to have three effective boundaries for each element. However, it is easy to see Lemma 2.2 still holds.

2.3.2. Estimate of convection terms

We consider the convection term

$$\int_E J(u\cdot\nabla_{\mathscr{A}} u)\cdot v_h - \frac{1}{2}\int_F (u\cdot\mathscr{N})\,(u\cdot v_h). \tag{2.47}$$

This is the key nonlinear term in the Navier–Stokes equations. Our discretization is inspired by the idea in [7]. In the continuous case, when summing up over all $E \in \mathcal{E}_h$, we can see

$$\sum_{E\in\mathcal{E}_h} \gamma_E\left(u,\eta,u,v_h\right) = \int_\Omega J\left(u\cdot\nabla_{\mathscr{A}} u\right)\cdot v_h + \frac{1}{2}\int_\Omega J\left(\nabla_{\mathscr{A}}\cdot u\right)(u\cdot v_h) - \frac{1}{2}\int_\Sigma (u\cdot\mathscr{N})\,(u\cdot v_h), \tag{2.48}$$

where all the other boundary terms vanish. Hence, considering the continuous $\mathscr{A}$-divergence-free condition for u, this discretization is consistent.

Lemma 2.4. *If we take $v_h = u_h$, the discretized convection term satisfies*

$$\sum_{E\in\mathcal{E}_h} \gamma_E\left(u_h,\eta_h,u_h,u_h\right) \geq 0. \tag{2.49}$$

Proof. We divide the proof into several steps:

Step 1: Direct integration by parts.
We plug the test function $v_h = u_h$ into (2.21) to obtain

$$\gamma_E\left(u_h,\eta_h,u_h,u_h\right) + \frac{1}{2}\int_F (u_h\cdot\mathscr{N}_h)\left|u_h\right|^2 = \int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h} u_h\right)\cdot u_h + \frac{1}{2}\int_E J_h\left(\nabla_{\mathscr{A}_h}\cdot u_h\right)(u_h\cdot u_h)$$
$$-\frac{1}{2}\int_{\partial E\setminus\partial\Omega} [u_h\cdot J_h\mathscr{A}_h]\cdot n_E \frac{u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}}{2} + \int_{\partial E^-} \left|\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\right|\left(u_h^{\mathrm{int}(E)} - u_h^{\mathrm{ext}(E)}\right)\cdot u_h^{\mathrm{int}(E)}. \tag{2.50}$$

A direct integration by parts yields

$$\int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h}u_h\right)\cdot u_h = -\int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h}u_h\right)\cdot u_h - \int_E J_h\left(\nabla_{\mathscr{A}_h}\cdot u_h\right)\left(u_h\cdot u_h\right) + \int_{\partial E}\left(u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}\cdot n_E\right)\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right). \tag{2.51}$$

Therefore, (2.50) can be simplified as

$$\begin{aligned}
&\gamma_E\left(u_h,\eta_h,u_h,u_h\right)+\frac{1}{2}\int_{\partial E\cap\Sigma}\left(u_h\cdot\mathscr{N}_h\right)|u_h|^2\\
&=-\left(\int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h}u_h\right)\cdot u_h+\frac{1}{2}\int_E J_h\left(\nabla_{\mathscr{A}_h}\cdot u_h\right)\left(u_h\cdot u_h\right)\right)\\
&\quad-\frac{1}{2}\int_{\partial E\backslash\partial\Omega}\left[u_h\cdot J_h\mathscr{A}_h\right]\cdot n_E\frac{u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}}{2}\\
&\quad+\int_{\partial E^-}\left|\left\{u_h\cdot J_h\mathscr{A}_h\right\}\cdot n_E\right|\left(u_h^{\mathrm{int}(E)}-u_h^{\mathrm{ext}(E)}\right)\cdot u_h^{\mathrm{int}(E)}\\
&\quad+\int_{\partial E}\left(u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}\cdot n_E\right)\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&=I+II+III+IV,
\end{aligned}\tag{2.52}$$

where

$$I=-\left(\int_E J_h\left(u_h\cdot\nabla_{\mathscr{A}_h}u_h\right)\cdot u_h+\frac{1}{2}\int_E J_h\left(\nabla_{\mathscr{A}_h}\cdot u_h\right)\left(u_h\cdot u_h\right)\right), \tag{2.53}$$

and II, III and IV can be understood respectively.

Step 2: Estimates of II and IV.
In IV, for $e\in\partial E\backslash\partial\Omega$, we have the decomposition

$$u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}=\left\{u_h\cdot J_h\mathscr{A}_h\right\}+\frac{1}{2}\left(u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}-u_h^{\mathrm{ext}(E)}\cdot J_h^{\mathrm{ext}(E)}\mathscr{A}_h^{\mathrm{ext}(E)}\right). \tag{2.54}$$

For $e\subset\partial E\backslash\partial\Omega$, we can sum up the second term on the right-hand side of (2.54) over E and its neighboring cells, which have the opposition outward normal vectors, to show

$$\begin{aligned}
&\int_e\frac{1}{2}\left(u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}-u_h^{\mathrm{ext}(E)}\cdot J_h^{\mathrm{ext}(E)}\mathscr{A}_h^{\mathrm{ext}(E)}\right)\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&+\int_e\frac{1}{2}\left(u_h^{\mathrm{ext}(E)}\cdot J_h^{\mathrm{ext}(E)}\mathscr{A}_h^{\mathrm{ext}(E)}-u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}\right)\cdot(-n_E)\left(u_h^{\mathrm{ext}(E)}\cdot u_h^{\mathrm{ext}(E)}\right)\\
&=\int_e\left[u_h\cdot J_h\mathscr{A}_h\right]\cdot n_e\left\{u_h\cdot u_h\right\},
\end{aligned}\tag{2.55}$$

where n_e denotes the normal vector read from the same reference cell as $[u_h\cdot J_h\mathscr{A}_h]$. Note for $e\subset\partial E\cap\partial\Omega$, $u_h^{\mathrm{int}(E)}\cdot J_h^{\mathrm{int}(E)}\mathscr{A}_h^{\mathrm{int}(E)}=\{u_h\cdot J_h\mathscr{A}_h\}$. Therefore, we obtain

$$\sum_{E\in\mathcal{E}_h}IV=\sum_{e\in\partial\mathcal{E}_h}\int_e\left\{u_h\cdot J_h\mathscr{A}_h\right\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)+\sum_{e\in\partial\mathcal{E}_h\backslash\partial\Omega}\int_e\left[u_h\cdot J_h\mathscr{A}_h\right]\cdot n_e\left\{u_h\cdot u_h\right\}. \tag{2.56}$$

Also, in II, for $e \subset \partial E \backslash \partial\Omega$, we can sum up over E and its neighboring cells to achieve

$$-\frac{1}{2}\int_e [u_h \cdot J_h \mathscr{A}_h] \cdot n_E \frac{u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}}{2} - \frac{1}{2}\int_e (-[u_h \cdot J_h \mathscr{A}_h]) \cdot (-n_E) \frac{u_h^{\mathrm{ext}(E)} \cdot u_h^{\mathrm{ext}(E)}}{2}$$
$$= -\frac{1}{2}\int_e [u_h \cdot J_h \mathscr{A}_h] \cdot n_e \{u_h \cdot u_h\}, \quad (2.57)$$

which further leads to

$$\sum_{E\in\mathcal{E}_h} II = -\sum_{e\in\partial\mathcal{E}_h\backslash\partial\Omega} \frac{1}{2}\int_e [u_h \cdot J_h \mathscr{A}_h] \cdot n_e \{u_h \cdot u_h\}. \quad (2.58)$$

Step 3: Further estimates in (2.52).
Hence, summing over all $E \in \mathcal{E}_h$ in (2.52) and combining (2.56) and (2.58), we deduce

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h} \gamma_E (u_h, \eta_h, u_h, u_h) + \frac{1}{2}\int_\Sigma (u_h \cdot \mathscr{N}_h) |u_h|^2 \\
&= -\left(\int_\Omega J_h (u_h \cdot \nabla_{\mathscr{A}_h} u_h) \cdot u_h + \frac{1}{2}\int_\Omega J_h (\nabla_{\mathscr{A}_h} \cdot u_h)(u_h \cdot u_h) - \sum_{e\in\partial\mathcal{E}_h\backslash\partial\Omega} \frac{1}{2}\int_e [u_h \cdot J_h \mathscr{A}_h] \cdot n_E \{u_h \cdot u_h\} \right) \\
&\quad + \sum_{E\in\mathcal{E}_h} \left(\int_{\partial E^-} |\{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E| \left(u_h^{\mathrm{int}(E)} - u_h^{\mathrm{ext}(E)}\right) \cdot u_h^{\mathrm{int}(E)} + \int_{\partial E} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right) \right). \quad (2.59)
\end{aligned}$$

We can further estimate the last two terms in (2.59) as follows:

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h} \left(\int_{\partial E^-} |\{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E| \left(u_h^{\mathrm{int}(E)} - u_h^{\mathrm{ext}(E)}\right) \cdot u_h^{\mathrm{int}(E)} + \int_{\partial E} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right) \right) \\
&= \sum_{E\in\mathcal{E}_h} \int_{\partial E^-} |\{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E| \, u_h^{\mathrm{ext}(E)} \cdot \left(u_h^{\mathrm{ext}(E)} - u_h^{\mathrm{int}(E)}\right) \\
&\quad + \sum_{E\in\mathcal{E}_h} \int_{\partial E^+\cap\partial\Omega} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right) + Z \sum_{E\in\mathcal{E}_h} \int_{\partial E^-\cap\Sigma} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right). \quad (2.60)
\end{aligned}$$

Then combining (2.59) and (2.60), we have the complete form

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h} \gamma_E (u_h, \eta_h, u_h, u_h) + \frac{1}{2}\int_\Sigma (u_h \cdot \mathscr{N}_h) |u_h|^2 \\
&= -\left(\int_\Omega J_h (u_h \cdot \nabla_{\mathscr{A}_h} u_h) \cdot u_h + \frac{1}{2}\int_\Omega J_h (\nabla_{\mathscr{A}_h} \cdot u_h)(u_h \cdot u_h) \right. \\
&\quad \left. - \sum_{e\in\partial\mathcal{E}_h} \frac{1}{2}\int_{e\not\subseteq\partial\Omega} [u_h \cdot J_h \mathscr{A}_h] \cdot n_E \{u_h \cdot u_h\} \right) + \sum_{E\in\mathcal{E}_h} \int_{\partial E^-} |\{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E| \, u_h^{\mathrm{ext}(E)} \cdot \left(u_h^{\mathrm{ext}(E)} - u_h^{\mathrm{int}(E)}\right) \\
&\quad + \sum_{E\in\mathcal{E}_h} \int_{\partial E^+\cap\partial\Omega} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right) + \sum_{E\in\mathcal{E}_h} \int_{\partial E^-\cap\Sigma} \{u_h \cdot J_h \mathscr{A}_h\} \cdot n_E \left(u_h^{\mathrm{int}(E)} \cdot u_h^{\mathrm{int}(E)}\right). \quad (2.61)
\end{aligned}$$

Step 4: Synthesis.
Summing up (2.50) over $E \in \mathcal{E}_h$ and adding it to (2.61) imply

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h} \gamma_E\left(u_h,\eta_h,u_h,u_h\right)+\frac{1}{2}\int_{\Sigma}\left(u_h\cdot\mathscr{N}_h\right)|u_h|^2\\
&=\frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_{\partial E^-}|\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E|\left|u_h^{\mathrm{ext}(E)}-u_h^{\mathrm{int}(E)}\right|^2+\sum_{E\in\mathcal{E}_h}\frac{1}{2}\int_{\partial E^+\cap\partial\Omega}\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&\quad+\sum_{E\in\mathcal{E}_h}\frac{1}{2}\int_{\partial E^-\cap\Sigma}\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&\geq\sum_{E\in\mathcal{E}_h}\frac{1}{2}\int_{\partial E^+\cap\Sigma}\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)+\sum_{E\in\mathcal{E}_h}\frac{1}{2}\int_{\partial E^-\cap\Sigma}\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&=\frac{1}{2}\int_{\Sigma}\{u_h\cdot J_h\mathscr{A}_h\}\cdot n_E\left(u_h^{\mathrm{int}(E)}\cdot u_h^{\mathrm{int}(E)}\right)\\
&=\frac{1}{2}\int_{\Sigma}\left(u_h\cdot\mathscr{N}_h\right)|u_h|^2,
\end{aligned}\tag{2.62}$$

where the last equality can be directly verified by the definitions of J_h, $\mathscr{A}_h$ and $\mathscr{N}_h$. Then our result naturally follows. □

2.3.3. Estimate of diffusion terms

In order to show the coercivity of the diffusion term, we need the discrete form of Korn's inequality. The proof here is based on [18].

Lemma 2.5. *Assume $f_h \in X_h^k$ satisfies for $e \subset \partial\mathcal{E}_h \cap \Sigma_b$, it holds that $f_h^{\mathrm{ext}}|_e = 0$. Then for sufficiently large $\sigma_0 > 0$, we have*

$$\sum_{E\in\mathcal{E}_h}\int_E|\mathbb{D}f_h|^2+\sum_{e\in\partial\mathcal{E}_h}\frac{\sigma_0}{h}\int_e[f_h]^2\gtrsim\|f_h\|_{X_h}^2\gtrsim\|f_h\|_{H_h}^2.\tag{2.63}$$

Proof. The second inequality has been shown in [7], so we turn to the first one. It is easy to see the key part is to show the derivatives of f_h can be controlled. Hence, we only need to show

$$\sum_{E\in\mathcal{E}_h}\int_E|\mathbb{D}f_h|^2+\sum_{e\in\partial\mathcal{E}_h}\frac{\sigma_0}{h}\int_e[f_h]^2\gtrsim\|f_h\|_{H_h}^2.\tag{2.64}$$

If this is not true, then we can construct a sequence $f_h^n \in X_h^k$ satisfying

$$\|f_h^n\|_{H_h}=1,\tag{2.65}$$

and

$$\sum_{E\in\mathcal{E}_h}\|\mathbb{D}f_h^n\|_{L^2(E)}+\sum_{e\in\partial\mathcal{E}_h}\frac{\sigma_0}{h}\int_e[f_h^n]^2\leq\frac{1}{n}.\tag{2.66}$$

To abuse the notations, we can extract the weakly convergent subsequence

$$f_h^n\rightharpoonup f_h\ \text{ in }\ H^1(E)\ \text{ for }\ \forall E\in\mathcal{E}_h.\tag{2.67}$$

By the compact embedding theorem in each cell E and the weak lower semi-continuity of $H^1(E)$ norm, we have the strongly convergent subsequence

$$f_h^n \to f_h \ \text{ in } \ L^2(\Omega). \tag{2.68}$$

By (2.66), we also have

$$\sum_{E\in\mathcal{E}_h} \|\mathbb{D}f_h^n\|_{L^2(\Omega)} \to 0. \tag{2.69}$$

Notice the fact that in each cell E, we still have the continuous Korn's inequality

$$\|f_h^m - f_h^n\|_{H^1(E)} \lesssim \|\mathbb{D}f_h^m - \mathbb{D}f_h^n)\|_{L^2(E)} + \|f_h^m - f_h^n\|_{L^2(E)} \quad \text{for } \ \forall m,n\in\mathbb{N}. \tag{2.70}$$

Hence, combining all above, we know f_h^n is a Cauchy's sequence under the norm $H^1(E)$, which means it is a Cauchy's sequence under the norm H_h. Then

$$f_h^n \to f_h \ \text{ in } \ H_h. \tag{2.71}$$

Thus this means

$$\|f_h\|_{H_h} = 1, \tag{2.72}$$

which further implies

$$\mathbb{D}f_h = 0 \ \text{ in } \ \forall E\in\mathcal{E}_h, \tag{2.73}$$

and

$$\sum_{e\in\partial\mathcal{E}_h} \frac{\sigma_0}{h}\int_e [f_h]^2 = 0. \tag{2.74}$$

This means

$$f_h = (a,b) + c\,(-x_1, x_2)\,, \tag{2.75}$$

for some constant a, b, c and f_h is continuous in $\bar{\Omega}$. Certainly, the zero bottom implies $f_h = 0$, which contradicts $\|f_h\|_{H_h} = 1$. Therefore, our hypothesis is invalid and (2.64) holds. □

Lemma 2.6. *Assume f_h satisfies for $e\subset\partial\mathcal{E}_h\cap\Sigma_b$, it holds that $f_h^{\text{ext}}|_e = 0$. Also, η_h satisfies $\eta_h + 1 \geq \delta > 0$ and*

$$Q = \sup_{F\in\mathcal{F}_h} \|\eta_h\|_{W^{1,\infty}(F)} < \infty. \tag{2.76}$$

Then for sufficiently large $\sigma_0 > 0$, we have

$$\sum_{E\in\mathcal{E}_h}\int_E J_h\left|\mathbb{D}_{\mathscr{A}_h} f_h\right|^2 + \sum_{e\in\partial\mathcal{E}_h}\frac{\sigma_0}{h}\int_e [f_h]^2 \gtrsim C(Q)\,\|f_h\|_{X_h}^2 \gtrsim C(Q)\,\|f_h\|_{H_h}^2\,. \tag{2.77}$$

Proof. Note that in each cell E, the free surface η_h is smooth. Hence, we can change it back by the transform Φ_h^{-1} to a curved cell $\Phi_h^{-1}E$ which satisfies

$$\int_E J_h(x)\left|\mathbb{D}_{\mathscr{A}_h(x)} f_h(x)\right|^2 \mathrm{d}x = \int_{\Phi_h^{-1}E} |\mathbb{D}f_h(y)|^2\,\mathrm{d}y. \tag{2.78}$$

Hence, in this curved cell, we have the Korn's inequality

$$\|f_h\|_{H^1\left(\Phi_h^{-1}E\right)} \lesssim \|\mathbb{D} f_h\|_{L^2\left(\Phi_h^{-1}E\right)} + \|f_h\|_{L^2\left(\Phi_h^{-1}E\right)}. \tag{2.79}$$

Since the transform Φ_h^{-1} is a diffeomorphism between E and $\Phi_h^{-1}(E)$, then the H^1 norms in these two spaces are comparable. Hence, we have

$$\|f_h\|_{H^1(E)} \lesssim C(Q)\left(\int_E J_h \left|\mathbb{D}_{\mathscr{A}_h} f_h\right|^2 + \|f_h\|_{L^2(E)}\right). \tag{2.80}$$

Then a similar proof as that of Lemma 2.5 naturally yields the desired result. □

Lemma 2.7. *Assume f_h satisfies for $e \subset \partial\mathcal{E}_h \cap \Sigma_b$, it holds that $f_h^{\text{ext}}|_e = 0$. Also, η_h satisfies $\eta_h + 1 \geq \delta > 0$ and*

$$Q = \sup_{F\in\mathcal{F}_h} \|\eta_h\|_{W^{1,\infty}(F)} < \infty. \tag{2.81}$$

Then there exists a sufficiently small constant $\delta' > 0$ such that for η' satisfying

$$\sup_{F\in\mathcal{F}_h} \|\eta_h' - \eta_h\|_{W^{1,\infty}(F)} \leq \delta', \tag{2.82}$$

and for sufficiently large $\sigma_0' > 0$, we have

$$\sum_{E\in\mathcal{E}_h}\int_E J_h' \left|\mathbb{D}_{\mathscr{A}_h'} f_h\right|^2 + \sum_{e\in\partial\mathcal{E}_h} \frac{\sigma_0'}{h}\int_e [f_h]^2 \gtrsim C(Q)\|f_h\|_{X_h}^2 \gtrsim C(Q)\|f_h\|_{H_h}^2. \tag{2.83}$$

Proof. By Lemma 2.6, we know

$$\sum_{E\in\mathcal{E}_h}\int_E J_h \left|\mathbb{D}_{\mathscr{A}_h} f_h\right|^2 + \sum_{e\in\partial\mathcal{E}_h} \frac{\sigma_0}{h}\int_e [f_h]^2 \gtrsim C(Q)\|f_h\|_{X_h}^2 \gtrsim C(Q)\|f_h\|_{H_h}^2. \tag{2.84}$$

We can rewrite our formula in a perturbed form as

$$\begin{aligned}
&\sum_{E\in\mathcal{E}_h}\int_E J_h' \left|\mathbb{D}_{\mathscr{A}_h'} f_h\right|^2 - \sum_{E\in\mathcal{E}_h}\int_E J_h \left|\mathbb{D}_{\mathscr{A}_h} f_h\right|^2 \\
&= \sum_{E\in\mathcal{E}_h}\int_E J_h' \left(\mathbb{D}_{\mathscr{A}_h'-\mathscr{A}_h} f_h\right)\left(2\mathbb{D}_{\mathscr{A}_h} f_h + \mathbb{D}_{\mathscr{A}_h'-\mathscr{A}_h} f_h\right) + \sum_{E\in\mathcal{E}_h}\int_E (J_h' - J_h)\left|\mathbb{D}_{\mathscr{A}_h} f_h\right|^2 \\
&\lesssim \delta' \|f_h\|_{H_h}^2.
\end{aligned} \tag{2.85}$$

Naturally, when taking δ' sufficiently small, we can absorb the perturbation into the principle part. Then our result easily follows. □

For the diffusion term

$$-\int_E J\Delta_{\mathscr{A}} u \cdot v_h, \tag{2.86}$$

a direct integration by parts implies

$$-\sum_{E\in\mathcal{E}_h}\int_E J\Delta_{\mathscr{A}} u\cdot v_h = \frac{1}{2}\int_\Omega J\mathbb{D}_{\mathscr{A}} u : \mathbb{D}_{\mathscr{A}} v_h - \sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e \{\mathbb{D}_{\mathscr{A}} u J\mathscr{A}\cdot n_E\}\cdot[v_h] - \int_\Sigma (\mathbb{D}_{\mathscr{A}} u)\,\mathscr{N}\cdot v_h^{\text{int}(E)}. \tag{2.87}$$

Also in the continuous case, our discretization yields

$$\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta,u,\eta,v_h\right)=\frac{1}{2}\int_\Omega J\mathbb{D}_{\mathscr{A}}u:\mathbb{D}_{\mathscr{A}}v_h-\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e\left\{\mathbb{D}_{\mathscr{A}}uJ\mathscr{A}\cdot n_E\right\}\cdot\left[v_h\right]. \tag{2.88}$$

We can notice the difference is the extra physical boundary term $-\int_\Sigma(\mathbb{D}_{\mathscr{A}}u)\mathscr{N}\cdot v_h^{\mathrm{int}(E)}$, which contributes to the boundary condition on Σ. This is separately added to the scheme as a forcing term later combined with the pressure contribution. Hence, our scheme is consistent.

Lemma 2.8. *Assume η_h satisfies $\eta_h+1\geq\delta>0$. If we take $v_h=u_h$ and choose the penalty constant σ properly, the discretized diffusion term with NIPG satisfies*

$$\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,u_h\right)\geq\int_\Omega J_h\left|\mathbb{D}_{\mathscr{A}_h}u_h\right|^2. \tag{2.89}$$

If we further assume η_h satisfies

$$Q=\sup_{F\in\mathcal{F}_h}\left\|\eta_h\right\|_{W^{1,\infty}(F)}<\infty, \tag{2.90}$$

then the discretized diffusion term with SIPG or NIPG satisfies

$$\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,u_h\right)\gtrsim C(Q)\left\|u_h\right\|_{X_h}^2. \tag{2.91}$$

Proof. We can directly compute

$$\begin{aligned}\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,v_h\right)=&\frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E J_h\mathbb{D}_{\mathscr{A}_h}u_h:\mathbb{D}_{\mathscr{A}_h}v_h-\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e\left\{\mathbb{D}_{\mathscr{A}_h}u_hJ_h\mathscr{A}_h\cdot n_E\right\}\cdot\left[v_h\right]\\&\pm\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e\left\{\mathbb{D}_{\mathscr{A}_h}v_hJ_h\mathscr{A}_h\cdot n_E\right\}\cdot\left[u_h\right]+\frac{\sigma}{h}\sum_{e\in\partial\mathcal{E}_h}\left[u_h\right]\left[v_h\right].\end{aligned} \tag{2.92}$$

For NIPG, when $v_h=u_h$, (2.92) reduces to

$$\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,u_h\right)=\frac{1}{2}\int_\Omega J_h\left|\mathbb{D}_{\mathscr{A}_h}u_h\right|^2+\frac{\sigma}{h}\sum_e\left[u_h\right]^2. \tag{2.93}$$

Hence, we may simply take the penalty $\sigma=1$ and by the discrete Korn's inequality in Lemma 2.6, our result naturally follows.

For SIPG, when $v_h=u_h$, (2.92) reduces to

$$\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,u_h\right)=\frac{1}{2}\int_\Omega J_h\left|\mathbb{D}_{\mathscr{A}_h}u_h\right|^2-2\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e\left\{\mathbb{D}_{\mathscr{A}_h}u_hJ_h\mathscr{A}_h\cdot n_E\right\}\cdot\left[u_h\right]+\frac{\sigma}{h}\sum_{e\in\partial\mathcal{E}_h}\left[u_h\right]^2. \tag{2.94}$$

By the discrete Korn's inequality in Lemma 2.6, we have

$$\sum_{E\in\mathcal{E}_h}\int_E J_h\left|\mathbb{D}_{\mathscr{A}_h}u_h\right|^2+\sum_{e\in\partial\mathcal{E}_h}\frac{\sigma_0}{h}\int_e\left[u_h\right]^2\gtrsim C(Q)\left\|u_h\right\|_{X_h}^2\gtrsim C(Q)\left\|u_h\right\|_{H_h}^2. \tag{2.95}$$

Then we utilize Hölder's inequality, the trace theorem and Cauchy's inequality to estimate

$$\begin{aligned}\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\int_e\{\mathbb{D}_{\mathscr{A}_h}u_hJ_h\mathscr{A}_h\cdot n_E\}\cdot[u_h] &\le C(Q)\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\|\nabla u_h\|_{L^2(e)}\|[u_h]\|_{L^2(e)}\\ &\le C(Q)\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\|u_h\|_{H^1(e)}\|[u_h]\|_{L^2(e)}\\ &\le CC(Q)\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}h\|u_h\|^2_{H^1(e)}+\frac{C(Q)}{4C}\frac{1}{h}\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\|[u_h]\|^2_{L^2(e)}\\ &\le CC(Q)\sum_{E\in\mathcal{E}_h}h\|u_h\|^2_{H^{3/2}(E)}+\frac{C(Q)}{4C}\frac{1}{h}\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\|[u_h]\|^2_{L^2(e)}\\ &\le CC(Q)\|u_h\|^2_{H_h}+\frac{C(Q)}{4C}\frac{1}{h}\sum_{e\in\partial\mathcal{E}_h\setminus\partial\Omega}\|[u_h]\|^2_{L^2(e)}.\end{aligned}\tag{2.96}$$

The last inequality is valid since $u_h\in P_h^k$. Then we take C sufficiently small, and $\sigma\ge\sigma+C(Q)/(4C)$ to absorb (2.96) into (2.95). Hence, our result easily follows. □

2.3.4. Estimate of pressure terms

For the pressure term

$$\int_E J\nabla_{\mathscr{A}}p\cdot v_h,\tag{2.97}$$

a direct integration by parts reveals the following equality in the continuous case

$$\sum_{E\in\mathcal{E}_h}\int_E J\nabla_{\mathscr{A}}p\cdot v_h=-\int_\Omega Jp\nabla_{\mathscr{A}}\cdot v_h+\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e[v_h]\cdot(pJ\mathscr{A})\cdot n_e)+\int_\Sigma p\mathscr{N}\cdot v_h^{\mathrm{int}(E)}.\tag{2.98}$$

It is easy to see in the continuous case, our discretization reduces to

$$\sum_{E\in\mathcal{E}_h}\beta_E(p,\eta,v_h)=-\int_\Omega Jp\nabla_{\mathscr{A}}\cdot v_h+\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e[v_h]\cdot(pJ\mathscr{A})\cdot n_e.\tag{2.99}$$

We can notice the difference is the extra physical boundary term $\int_\Sigma p\mathscr{N}\cdot v^{\mathrm{int}(E)}$, which contributes to the boundary condition on Σ. This is separately added to the scheme as a forcing term later combined with the diffusion contribution. Hence, our scheme is consistent.

Lemma 2.9. *When we take $v_h=u_h$, the discretized pressure term satisfies*

$$\sum_{E\in\mathcal{E}_h}\beta_E(p_h,\eta_h,u_h)=-\int_\Omega J_hp_h\nabla_{\mathscr{A}_h}\cdot u_h+\sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e[u_h]\cdot\{p_hJ_h\mathscr{A}_h\}\cdot n_e,\tag{2.100}$$

where n_e denotes the outward normal vector read from the same direction as $[u_h]$.

Proof. This is a direct corollary of the definition, so we omit the proof here. □

2.3.5. Estimate of forcing terms

Considering the physical boundary condition on Σ of the system (1.9), we need to take the upper boundary condition as a forcing term, *i.e.*

$$(pI - \mathbb{D}_{\mathscr{A}} u)\,\mathscr{N} = \eta \mathscr{N} \quad \text{on} \quad \Sigma. \tag{2.101}$$

Multiplying the test function v_h and integrating over F yield

$$\int_F \eta \mathscr{N} v_h = \int_F \eta \left(\bar{v}_2\right)_h - \int_F \eta \partial_1 \eta \left(\bar{v}_1\right)_h. \tag{2.102}$$

In the continuous case, our discretization reduces to

$$\sum_{E \in \mathcal{E}_h} \mu_E \left(\eta, \eta, v_h\right) = \int_\Sigma \eta \left(\bar{v}_2\right)_h - \int_\Sigma \eta \partial_1 \eta \left(\bar{v}_1\right)_h, \tag{2.103}$$

and the penalty terms vanish. Hence, this discretization is consistent.

Lemma 2.10. *When we take $v_h = u_h$, the discretized forcing term satisfies*

$$\sum_{E \in \mathcal{E}_h} \mu_E \left(\eta_h, \eta_h, u_h\right) \geq \frac{1}{2} \int_\Sigma \partial_t \left|\eta_h\right|^2. \tag{2.104}$$

Proof. We may sum up over $E \in \mathcal{E}_h$ to obtain

$$\begin{aligned} &\sum_{E \in \mathcal{E}_h} \mu_E \left(\eta_h, \eta_h, v_h\right) \\ &= \int_\Sigma \eta_h \left(\bar{v}_2\right)_h + \int_\Sigma \eta_h \partial_1 \eta_h \left(\bar{v}_1\right)_h + \frac{1}{2} \sum_{F \in \mathcal{F}_h} \left(\left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^-\right)^2 |_{F+1/2} - \left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^+\right)^2 |_{F-1/2} \right). \end{aligned} \tag{2.105}$$

Taking $\phi_h = \eta_h$ in the transport discretization (2.15) and integrating by parts imply

$$\begin{aligned} &\frac{1}{2} \int_F \partial_t \left|\eta_h\right|^2 - \int_F \left(\bar{u}_2\right)_h \eta_h + \int_F \left(\bar{u}_1\right)_h \eta_h \partial_1 \eta_h - \left(\bar{u}_1\right)_h^- \left(\eta_h^-\right)^2 |_{F+1/2} + \left(\bar{u}_1\right)_h^+ \left(\eta_h^+\right)^2 |_{F-1/2} \\ &+ \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^- |_{F+1/2} - \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^+ |_{F-1/2} + \frac{1}{2} \left(\left(\eta_h^-\right)^2 \left[u_h\right] |_{F+1/2} - \left(\eta_h^+\right)^2 \left[u_h\right] |_{F-1/2} \right) = 0, \end{aligned} \tag{2.106}$$

which further leads to

$$\begin{aligned} &\frac{1}{2} \int_F \partial_t \left|\eta_h\right|^2 - \int_F \left(\bar{u}_2\right)_h \eta_h + \int_F \left(\bar{u}_1\right)_h \eta_h \partial_1 \eta_h \\ &- \left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^-\right)^2 |_{F+1/2} + \left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^+\right)^2 |_{F-1/2} + \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^- |_{F+1/2} - \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^+ |_{F-1/2} = 0. \end{aligned} \tag{2.107}$$

Summing over $F \in \mathcal{F}_h$ implies

$$\begin{aligned} &\frac{1}{2} \int_\Sigma \partial_t \left|\eta_h\right|^2 - \int_\Sigma \left(\bar{u}_2\right)_h \eta_h + \int_\Sigma \left(\bar{u}_1\right)_h \eta_h \partial_1 \eta_h - \sum_{F \in \mathcal{F}_h} \left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^-\right)^2 |_{F+1/2} \\ &+ \sum_{F \in \mathcal{F}_h} \left\{\left(\bar{u}_1\right)_h\right\} \left(\eta_h^+\right)^2 |_{F-1/2} - \sum_{F \in \mathcal{F}_h} \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^- |_{F+1/2} + \sum_{F \in \mathcal{F}_h} \left(\hat{u}_1\right)_h \hat{\eta}_h \eta_h^+ |_{F-1/2} = 0. \end{aligned} \tag{2.108}$$

Since we always take $\hat{u}_1 = \{(\bar{u}_1)_h\}$, combining (2.105) and (2.108) yields

$$\sum_{E\in\mathcal{E}_h} \mu_E(\eta_h, \eta_h, v_h) = \frac{1}{2}\int_\Sigma \partial_t |\eta_h|^2 + \sum_{F\in\mathcal{F}_h} \Bigg(-\frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F+1/2} + \frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^+\right)^2|_{F-1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F+1/2} - \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^+|_{F-1/2}\Bigg). \tag{2.109}$$

We can decompose the boundary terms in (2.109) to obtain

$$-\frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F+1/2} + \frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^+\right)^2|_{F-1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F+1/2} - \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^+|_{F-1/2} = \Psi_{F+1/2} - \Psi_{F-1/2} + \Theta_F, \tag{2.110}$$

where

$$\Psi_{F+1/2} = -\frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F+1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F+1/2}, \tag{2.111}$$

$$\Psi_{F-1/2} = -\frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F-1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F-1/2}, \tag{2.112}$$

and

$$\Theta_F = \frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^+\right)^2|_{F-1/2} - \frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F-1/2} - \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^+|_{F-1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F-1/2}. \tag{2.113}$$

By the flux definition (2.18), we can derive

$$\Theta_F = \{(\bar{u}_1)_h\}\left(\eta_h^+ - \eta_h^-\right)\left(\frac{\eta_h^+ + \eta_h^-}{2} - \hat{\eta}_h\right) \geq 0. \tag{2.114}$$

When we sum up (2.110) over $F \in \mathcal{F}_h$, all $\Psi_{F+1/2}$ are canceled out due to the periodicity. Then we have

$$\sum_{F\in\mathcal{F}_h} \Bigg(-\frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^-\right)^2|_{F+1/2} + \frac{1}{2}\{(\bar{u}_1)_h\}\left(\eta_h^+\right)^2|_{F-1/2} + \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^-|_{F+1/2} - \{(\bar{u}_1)_h\}\hat{\eta}_h\eta_h^+|_{F-1/2}\Bigg) \geq 0. \tag{2.115}$$

Hence, combining (2.109) and (2.115), we can obtain the desired result. □

2.3.6. Estimate of divergence terms

In the continuous case, our discretization reduces to

$$\sum_{E\in\mathcal{E}_h} \rho_E(u, \eta, q_h) = -\int_\Omega J\left(\nabla_{\mathscr{A}}\cdot u\right) q_h. \tag{2.116}$$

Hence, this discretization is consistent.

Lemma 2.11. *When we take $q_h = p_h$, the discretized divergence term satisfies*

$$\sum_{E\in\mathcal{E}_h} \rho_E(u_h, \eta_h, p_h) = -\int_\Omega J_h p_h \nabla_{\mathscr{A}_h}\cdot u_h + \sum_{e\in\partial\mathcal{E}_h\setminus\Sigma}\int_e [u_h]\cdot\{p_h J_h \mathscr{A}_h\}\cdot n_e. \tag{2.117}$$

Proof. This is a natural corollary of the definition, so we omit the proof here. □

3. Stability analysis

Condition 3.1. The free surface $\eta_h(t)$ satisfies $1+\eta_h(t)\geq\delta>0$ for some $\delta>0$ independent of h and t.

Condition 3.2. The free surface $\eta_h(t)$ satisfies $\sup_{F\in\mathcal{F}_h}\|\eta_h(t)\|_{W^{1,\infty}(F)}\leq Q$ for some $Q>0$ independent of h and t.

Theorem 3.3. *If $S=0$, suppose Condition 3.1 is valid in $t\in[0,T]$ for some $T>0$. Then there exists a unique numerical solution triple $(u_h,p_h,\eta_h)\in X_h^k\times M_h^{k-1}\times S_h^k$ to the scheme (2.14) with NIPG, which satisfies the estimate*

$$\left\|\sqrt{J_h(t)}u_h(t)\right\|_{L^2(\Omega)}^2+g\left\|\eta_h(t)\right\|_{L^2(\Sigma)}^2+\nu\int_0^t\int_\Omega J_h(s)\left|\mathbb{D}_{\mathscr{A}_h(s)}u_h(s)\right|^2\mathrm{d}s \leq\left\|\sqrt{J_h(0)}u_h(0)\right\|_{L^2(\Omega)}^2+g\left\|\eta_h(0)\right\|_{L^2(\Sigma)}^2, \tag{3.1}$$

for any $t\in[0,T]$.

For general S, suppose Conditions 3.1 and 3.2 are valid in $t\in[0,T]$ for some $T>0$. Then there exists a unique numerical solution triple $(u_h,p_h,\eta_h)\in X_h^k\times M_h^{k-1}\times S_h^k$ to the scheme (2.14) with SIPG or NIPG, which satisfies the estimate

$$\left\|\sqrt{J_h(t)}u_h(t)\right\|_{L^2(\Omega)}^2+\left\|\eta_h(t)\right\|_{L^2(\Sigma)}^2+\int_0^t\left\|u_h(s)\right\|_{X_h}^2\mathrm{d}s \lesssim C(Q)\left(\left\|\sqrt{J_h(0)}u_h(0)\right\|_{L^2(\Omega)}^2+\left\|\eta_h(0)\right\|_{L^2(\Sigma)}^2+\int_0^t\int_\Omega\left|S(r)\right|^2\mathrm{d}r\right), \tag{3.2}$$

for any $t\in[0,T]$.

Proof. The existence and uniqueness follow from a standard argument for the differential-algebraic equations, so we omit it here and focus on the energy estimate. In the system (2.14), we take the test function $v_h=u_h$ and $q_h=p_h$, and sum up over $E\in\mathcal{E}_h$. Then it yields

$$\sum_{E\in\mathcal{E}_h}\zeta_E\left(\eta_h,u_h,u_h\right)+\sum_{E\in\mathcal{E}_h}\gamma_E\left(u_h,\eta_h,u_h,u_h\right)$$

$$+\nu\sum_{E\in\mathcal{E}_h}\alpha_E\left(\eta_h,u_h,\eta_h,u_h\right)+\sum_{E\in\mathcal{E}_h}\beta_E\left(p_h,\eta_h,u_h\right)+g\sum_{E\in\mathcal{E}_h}\mu_E\left(\eta_h,\eta_h,u_h\right)=\sum_{E\in\mathcal{E}_h}\omega_E\left(\eta_h,u_h\right), \tag{3.3}$$

$$\sum_{E\in\mathcal{E}_h}\rho_E\left(u_h,\eta_h,p_h\right)=0. \tag{3.4}$$

By Lemmas 2.9 and 2.11, we have

$$\sum_{E\in\mathcal{E}_h}\beta_E\left(p_h,\eta_h,u_h\right)=\sum_{E\in\mathcal{E}_h}\rho_E\left(u_h,\eta_h,p_h\right)=0. \tag{3.5}$$

Hence, we can simplify (3.3) into

$$\sum_{E\in\mathcal{E}_h}\zeta_E(\eta_h,u_h,u_h)+\sum_{E\in\mathcal{E}_h}\gamma_E(u_h,\eta_h,u_h,u_h)+\nu\sum_{E\in\mathcal{E}_h}\alpha_E(\eta_h,u_h,\eta_h,u_h)+g\sum_{E\in\mathcal{E}_h}\mu_E(\eta_h,\eta_h,u_h)=\sum_{E\in\mathcal{E}_h}\omega_E(\eta_h,u_h). \tag{3.6}$$

Then by Lemmas 2.2, 2.4, 2.8 and 2.10, we have

$$\frac{1}{2}\partial_t \int_\Omega J_h |u_h|^2 + \frac{g}{2}\partial_t \int_\Sigma |\eta_h|^2 + CC(Q)\|u_h\|_{X_h}^2 \le \sum_{E\in\mathcal{E}_h} \omega_E(u_h). \tag{3.7}$$

Note Conditions 3.1 and 3.2 guarantee the existence of $C(Q) > 0$ which is independent of t. An application of Cauchy's inequality implies

$$\sum_{E\in\mathcal{E}_h} \omega_E(u_h) \le C' \int_\Omega J_h^2 |u_h|^2 + \frac{1}{4C'}\int_\Omega |S|^2 \le Q^2 C' \|u_h\|_{X_h}^2 + \frac{1}{4C'}\int_\Omega |S|^2. \tag{3.8}$$

In (3.8), when taking C' sufficiently small, we can always absorb it into $CC(Q)\|u_h\|_{X_h}^2$ in (3.7). Hence, we have

$$\frac{1}{2}\partial_t \int_\Omega J_h |u_h|^2 + \frac{g}{2}\partial_t \int_\Sigma |\eta_h|^2 + \frac{CC(Q)}{2}\|u_h\|_{X_h}^2 \le C(Q)\int_\Omega |S|^2. \tag{3.9}$$

Then integrating over $[0,t]$ leads to the desired result. The $S = 0$ case is easily derived from the general case without discussion of the source term. □

Remark 3.4. Conditions 3.1 and 3.2 are not always satisfied *a priori*. In [1, 15, 16], this type of assumptions were also introduced in the numerical analysis.

4. Discussion on the error analysis

For the continuous solution (u,η), the analysis in [20] reveals in the Navier–Stokes equations of (1.9), we need H^1 norm of $\eta(t)$ to bound L^2 norm of $u(t)$. However, the result in [6] implies in the transport equation of (1.9), we need H^2 norm of $u(t)$ to control H^1 norm of $\eta(t)$. This type of inconsistent coupling cannot be improved even if we go to higher order derivatives. Hence, this implies the coupled system in (1.9) is not closed in the usual Sobolev norms, which means we cannot expect to obtain the error estimates for the whole system (1.9).

In the following, we mainly analyze the error in the Navier–Stokes equations provided we have the error estimates of the free surface.

Condition 4.1. The free surface η_h satisfies the error estimates

$$\|\eta_h - \eta\|_{L^2} \lesssim h^{k+1}, \tag{4.1}$$

$$\|\eta_h - \eta\|_{L^\infty} \lesssim h^{k}, \tag{4.2}$$

$$\|\chi_h - \chi\|_{L^2} \lesssim h^{k}, \tag{4.3}$$

$$\|\chi_h - \chi\|_{L^\infty} \lesssim h^{k-1}, \tag{4.4}$$

$$\|\partial_t(\eta_h - \eta)\|_{L^2} \lesssim h^{k}, \tag{4.5}$$

$$\|\partial_t(\eta_h - \eta)\|_{L^\infty} \lesssim h^{k-1}. \tag{4.6}$$

4.1. Velocity error analysis

We decompose the velocity error and the pressure error as follows:

$$u_h - u = (u_h - \mathcal{P}u) + (\mathcal{P}u - u) = \epsilon_u + \delta_u, \tag{4.7}$$

$$p_h - p = (p_h - \mathcal{P}p) + (\mathcal{P}p - p) = \epsilon_p + \delta_p, \tag{4.8}$$

where $\mathcal{P}$ is some projection such that $\mathcal{P}u \in X_h^k$ and $\mathcal{P}p \in M_h^{k-1}$ achieving the optimal accuracy, *i.e.*

$$\|\delta_u\|_{L^2} \lesssim h^{k+1}, \tag{4.9}$$

$$\|\delta_p\|_{L^2} \lesssim h^{k}. \tag{4.10}$$

Hence, the key part of the error estimates is ϵ_u and ϵ_p.

Lemma 4.2. *Suppose Conditions 3.1, 3.2 and 4.1 are valid. Assume the exact solution satisfies $u \in C^2(\Omega)$, $p \in C^1(\Omega)$ and $\eta \in C^1(\Sigma)$. Then for h sufficiently small, the numerical solution to the scheme (2.14) satisfies the estimate*

$$\partial_t \int_\Omega J_h |\epsilon_u|^2 + \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)\left(h^{k-1-r}\|\epsilon_u\|_{H_h} + h^{k-r}\|\epsilon_p\|_{L^2} + \|\epsilon_u\|_{L^2}^2\right) \quad \text{for} \quad 1/2 < r < 1. \tag{4.11}$$

Proof. The discretization of the Navier–Stokes equations is

$$\zeta_E(\eta_h, u_h, v_h) + \gamma_E(u_h, \eta_h, u_h, v_h) + \nu\alpha_E(\eta_h, u_h, \eta_h, v_h) + \beta_E(p_h, \eta_h, v_h) + g\mu_E(\eta_h, \eta_h, v_h) = \omega_E(\eta_h, v_h), \tag{4.12}$$

$$\rho_E(u_h, \eta_h, q_h) = 0. \tag{4.13}$$

The consistency of the scheme (2.14) implies the exact solution (u, p, η) also satisfies this scheme, *i.e.*

$$\zeta_E(\eta, u, v_h) + \gamma_E(u, \eta, u, v_h) + \nu\alpha_E(\eta, u, \eta, v_h) + \beta_E(p, \eta, v_h) + g\mu_E(\eta, \eta, v_h) = \omega_E(\eta, v_h), \tag{4.14}$$

$$\rho_E(u, \eta, q_h) = 0. \tag{4.15}$$

Therefore, taking the difference of above two sets of equations, we obtain the error equations

$$\begin{aligned}
&(\zeta_E(\eta_h, u_h, v_h) - \zeta_E(\eta, u, v_h)) + (\gamma_E(u_h, \eta_h, u_h, v_h) - \gamma_E(u, \eta, u, v_h)) \\
&\qquad + \nu(\alpha_E(\eta_h, u_h, \eta_h, v_h) - \alpha_E(\eta, u, \eta, v_h)) \\
&+ (\beta_E(p_h, \eta_h, v_h) - \beta_E(p, \eta, v_h)) + g(\mu_E(\eta_h, \eta_h, v_h) - \mu_E(\eta, \eta, v_h)) = (\omega_E(\eta_h, v_h) - \omega_E(\eta, v_h)),
\end{aligned} \tag{4.16}$$

$$(\rho_E(u_h, \eta_h, q_h) - \rho_E(u, \eta, q_h)) = 0. \tag{4.17}$$

Now we need to analyze each term in the error equations (4.16) and (4.17). We always decompose the difference of the bilinear forms into two parts: the energy part W_* and the remaining part R_*, such that

$$\text{Difference} = W_* + R_*, \tag{4.18}$$

where W_* builds the main body of the error equations and is put in the left-hand side (LHS), and R_* is moved to the right-hand side (RHS) and can be taken as the perturbed source term. The $*$ can be ζ, γ or α, *etc.* For example, we can decompose the convection difference $\gamma(u_h, \eta_h, u_h, v_h) = \sum_{E\in\mathcal{E}_h}\gamma_E(u_h, \eta_h, u_h, v_h)$ as follows:

$$\begin{aligned}
&\gamma(u_h, \eta_h, u_h, v_h) - \gamma(u, \eta, u, v_h) \\
&= \gamma(u_h, \eta_h, u_h - u, v_h) + \gamma(u_h, \eta_h - \eta, u, v_h) + \gamma(u_h - u, \eta, u, v_h) \\
&= \gamma(u_h, \eta_h, \epsilon_u, v_h) + \gamma(u_h, \eta_h, \delta_u, v_h) + \gamma(u_h, \eta_h - \eta, u, v_h) + \gamma(u_h - u, \eta, u, v_h) \\
&= \gamma(u_h, \eta_h, \epsilon_u, \epsilon_u) + (\gamma(u_h, \eta_h, \delta_u, \epsilon_u) + \gamma(u_h, \eta_h - \eta, u, \epsilon_u) + \gamma(u_h - u, \eta, u, \epsilon_u)) \\
&= W_\gamma + R_\gamma,
\end{aligned} \tag{4.19}$$

where

$$W_\gamma = \gamma\left(u_h, \eta_h, \epsilon_u, \epsilon_u\right), \tag{4.20}$$

$$R_\gamma = \gamma\left(u_h, \eta_h, \delta_u, \epsilon_u\right) + \gamma\left(u_h, \eta_h - \eta, u, \epsilon_u\right) + \gamma\left(u_h - u, \eta, u, \epsilon_u\right). \tag{4.21}$$

W_* can be bounded as in Theorem 3.3 and R_* can be estimated based on the following rules:

Rule 1:
For the estimates of the product terms as

$$\int F_1 F_2 \dots F_k, \tag{4.22}$$

we apply Hölder's inequality to obtain two terms in the L^2 norm and other terms in the L^∞ norm. The priority of taking the L^∞ norm is as follows:

(1) F_i only containing the exact solution (u, p, η) has the highest priority.
(2) F_i containing the free surface error $\eta_h - \eta$, $\chi_h - \chi$ or $\partial_t(\eta_h - \eta)$, or the numerical solution η_h, χ_h or $\partial_t \eta_h$ has the second priority.
(3) F_i containing the projection error $u - \mathcal{P}u$ or $p - \mathcal{P}p$ for some $\mathcal{P}$ has the third priority.
(4) F_i only containing the velocity error $u_h - \mathcal{P}u$ or the pressure error $p_h - \mathcal{P}p$ for some projection $\mathcal{P}$, or the numerical solution u_h or p_h, has the lowest priority.

In this fashion, we can make the best use of the regularity of the exact solutions and the known error estimates in the free surface to avoid the estimates of (u_h, p_h) in undesired norms due to the embedding theorem.

Rule 2:
For the boundary term, we apply the trace theorem

$$\|F\|_{L^2(\partial E)} \lesssim \|F\|_{H^r(E)}, \tag{4.23}$$

for $r > 1/2$. This introduces more regularity in the estimates, so we should always try to avoid to use it directly and apply certain projection property to eliminate the boundary terms.

Rule 3:
Note the simple fact that for $r > 0$, we have

$$h^r \|f_h\|_{H^r(E)} \lesssim \|f_h\|_{L^2(E)}, \tag{4.24}$$

where h is the mesh size and $f_h \in P_h^k$. This can bound the higher order Sobolev norm by the lower norm at the price of some order of h.

In the error equations (4.16) and (4.17), we always take the test functions $v_h = \epsilon_u$ and $q_h = \epsilon_p$. Since the estimates are standard, we omit the details here and just present the results. We can simplify the error equations (4.16) and (4.17) as:

$$W_\zeta + W_\gamma + W_\alpha + W_\beta = -R_\zeta - R_\gamma - R_\alpha - R_\beta - R_\mu - R_\omega \tag{4.25}$$

$$W_\rho = -R_\rho. \tag{4.26}$$

Then we can utilize a similar argument as in the proof of Theorem 3.3 to bound the energy part as

$$\frac{1}{2}\partial_t \int_\Omega J_h |\epsilon_u|^2 + C(Q)\|\epsilon_u\|_{H_h}^2 \lesssim R_\rho - R_\zeta - R_\gamma - R_\alpha - R_\beta - R_\mu - R_\omega. \tag{4.27}$$

Utilizing the estimates about the remaining part, we can further obtain

$$\frac{1}{2}\partial_t \int_\Omega J_h \left|\epsilon_u\right|^2 + C(Q)\left\|\epsilon_u\right\|_{H_h}^2 \lesssim$$

$$h^{k-1-r}\left\|\epsilon_u\right\|_{H_h} + h^{k-r}\left\|\epsilon_p\right\|_{L^2} + h^{k-1}(1+\left\|\epsilon_u\right\|_{L^2})\left\|\epsilon_u\right\|_{H_h} + \left\|\epsilon_u\right\|_{L^2}\left\|\epsilon_u\right\|_{H_h} \quad \text{for } 1/2 < r < 1. \tag{4.28}$$

The stability shows both u_h and u are bounded in L^2, so ϵ_u is also bounded in L^2. Then we get

$$\frac{1}{2}\partial_t \int_\Omega J_h \left|\epsilon_u\right|^2 + C(Q)\left\|\epsilon_u\right\|_{H_h}^2 \lesssim h^{k-1-r}\left\|\epsilon_u\right\|_{H_h} + h^{k-r}\left\|\epsilon_p\right\|_{L^2} + \left\|\epsilon_u\right\|_{L^2}\left\|\epsilon_u\right\|_{H_h} \quad \text{for } 1/2 < r < 1. \tag{4.29}$$

Applying Cauchy's inequality to the last term of (4.29) implies

$$\frac{1}{2}\partial_t \int_\Omega J_h \left|\epsilon_u\right|^2 + C(Q)\left\|\epsilon_u\right\|_{H_h}^2 \lesssim h^{k-1-r}\left\|\epsilon_u\right\|_{H_h} + h^{k-r}\left\|\epsilon_p\right\|_{L^2} + \frac{1}{4C'}\left\|\epsilon_u\right\|_{L^2}^2 + C'\left\|\epsilon_u\right\|_{H_h}^2$$

$$\text{for } 1/2 < r < 1\ . \tag{4.30}$$

Taking C' sufficiently small, we can absorb it into the left-hand side to achieve

$$\frac{1}{2}\partial_t \int_\Omega J_h \left|\epsilon_u\right|^2 + \frac{C(Q)}{2}\left\|\epsilon_u\right\|_{H_h}^2 \lesssim h^{k-1-r}\left\|\epsilon_u\right\|_{H_h} + h^{k-r}\left\|\epsilon_p\right\|_{L^2} + C(Q)\left\|\epsilon_u\right\|_{L^2}^2 \quad \text{for } 1/2 < r < 1. \tag{4.31}$$

This is the desired result. □

4.2. Pressure error analysis

In order for further analysis, we first need to show the inf-sup condition of our pressure discretization.

Lemma 4.3. *In the pressure discretization $\beta(p_h, \eta_h, v_h) = \sum_{E\in\mathcal{E}_h}\beta_E(p_h, \eta_h, v_h)$, suppose Conditions 3.1, 3.2 and 4.1 are valid. Assume the exact solution satisfies $\eta(t) \in C^1(\Sigma)$. Then for sufficiently small time T which is independent of h, there exists a constant $\Xi > 0$ such that*

$$\inf_{p_h\in M_h^{k-1}}\sup_{v_h\in X_h^k}\frac{\beta\left(p_h, \eta_h, v_h\right)}{\left\|p_h\right\|_{L^2}\left\|v_h\right\|_{H_h}} \geq \Xi, \tag{4.32}$$

for $t \in [0, T]$, where Ξ is independent of h.

Proof. This result is similar to ([8], Thm. 4.5), so we omit the details here. Note the fact that for $J_0\mathscr{A}_0$ which is the value of $J\mathscr{A}$ at $t = 0$, and $\bar{J}_0\bar{\mathscr{A}}_0$ the average of $J_0\mathscr{A}_0$ in each cell E which is piecewise constants, we naturally have

$$\left\|J_0\mathscr{A}_0 - \bar{J}_0\bar{\mathscr{A}}_0\right\|_{L^\infty} \lesssim h. \tag{4.33}$$

Hence, the weight functions in the differential operators is only a small perturbation, which can be absorbed into the main part. □

Lemma 4.4. *Suppose Conditions 3.1, 3.2 and 4.1 are valid. Assume the exact solution satisfies $u \in C^2(\Omega)$, $p \in C^1(\Omega)$ and $\eta \in C^1(\Sigma)$. Then for sufficiently small T and h, the numerical solution to the scheme (2.14) satisfies the estimate*

$$\int_0^t \left\|\epsilon_p\right\|_{L^2} \lesssim C(Q)\left(h^{k-1-r} + \int_0^t \left\|\epsilon_u\right\|_{L^2} + \left\|\epsilon_u(t)\right\|_{L^2} + \frac{1}{h}\int_0^t \left\|\epsilon_u\right\|_{H_h}\right) \quad \textit{for } 1/2 < r < 1, \tag{4.34}$$

within $t \in [0, T]$.

Proof. We can take any test function $v_h \in X_h^k$ in the error equation (4.16) and integrate over time $[0,t]$. Here we do not distinguish between the energy part and remaining part, but only concentrate on

$$\int_0^t \beta\left(p_h, \eta_h, v_h\right). \tag{4.35}$$

All the other terms can be moved into the right-hand side and estimated in terms of $\|v_h\|_{H_h}$. Basically, the estimates are similar to the velocity error estimates in the proof of Lemma 4.2, so we omit the details here. Based on the inf-sup condition in Lemma 4.3, we can deduce the desired result. □

4.3. Error analysis of the fluid

Theorem 4.5. *Suppose Conditions 3.1, 3.2 and 4.1 are valid. Assume the exact solution satisfies $u \in C^2(\Omega)$, $p \in C^1(\Omega)$ and $\eta \in C^1(\Sigma)$. Then for sufficiently small T and h, the numerical solution to the scheme (2.14) satisfies the estimates*

$$\|u_h - u\|_{L^2} \lesssim C(Q) h^{k-1-r}, \tag{4.36}$$

$$\left(\int_0^t \|u_h - u\|_{H_h}^2\right)^{1/2} \lesssim C(Q) h^{k-1-r}, \tag{4.37}$$

$$\int_0^t \|p_h - p\|_{L^2} \lesssim C(Q) h^{k-1-r} \quad for \quad 1/2 < r < 1, \tag{4.38}$$

in $t \in [0,T]$.

Proof. In Lemmas 4.2 and 4.4, we have shown

$$\partial_t \int_\Omega J_h |\epsilon_u|^2 + \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)\left(h^{k-1-r}\|\epsilon_u\|_{H_h} + h^{k-r}\|\epsilon_p\|_{L^2} + \|\epsilon_u\|_{L^2}^2\right) \quad \text{for} \quad 1/2 < r < 1, \tag{4.39}$$

and

$$\int_0^t \|\epsilon_p\|_{L^2} \lesssim C(Q)\left(h^{k-1-r} + \int_0^t \|\epsilon_u\|_{L^2} + \|\epsilon_u(t)\|_{L^2} + \frac{1}{h}\int_0^t \|\epsilon_u\|_{H_h}\right) \quad \text{for} \quad 1/2 < r < 1. \tag{4.40}$$

Integrating over $[0,t]$ for any $t \in [0,T]$ in (4.39), we obtain

$$\int_\Omega J_h(t)|\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)\left(h^{k+2} + h^{k-1-r}\int_0^t \|\epsilon_u\|_{H_h} + h^{k-r}\int_0^t \|\epsilon_p\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2}^2\right). \tag{4.41}$$

Then we plug (4.40) into (4.41) to eliminate ϵ_p and obtain

$$\begin{aligned}
&\int_\Omega J_h(t)|\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \\
&\lesssim C(Q)\left(h^{2k-1-2r} + h^{k-1-r}\int_0^t \|\epsilon_u\|_{H_h} + h^{k-r}\|\epsilon_u(t)\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2}^2\right) \\
&\lesssim C(Q)\left(h^{2k-1-2r} + h^{2k-2-2r} + \left(\int_0^t \|\epsilon_u\|_{H_h}\right)^2 + h^{k-r}\|\epsilon_u(t)\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2}^2\right) \\
&\lesssim C(Q)\left(h^{2k-2-2r} + \sqrt{t}\int_0^t \|\epsilon_u\|_{H_h}^2 + h^{k-r}\|\epsilon_u(t)\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2}^2\right).
\end{aligned} \tag{4.42}$$

When T is sufficiently small, we can absorb $C(Q)\sqrt{t}\int_0^t \|\epsilon_u\|_{H_h}^2$ into the left-hand side of (4.42) to achieve

$$\int_\Omega J_h(t)|\epsilon_u(t)|^2 + \int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)\left(h^{2k-2-2r} + h^{k-r}\|\epsilon_u(t)\|_{L^2} + \int_0^t \|\epsilon_u\|_{L^2}^2\right). \tag{4.43}$$

Since (4.43) holds for any $t \in [0,T]$, we can take the maximum to get

$$\max_{t\in[0,T]} \|\epsilon_u(t)\|_{L^2}^2 \lesssim C(Q)\left(h^{2k-2-2r} + h^{k-r}\max_{t\in[0,T]} \|\epsilon_u(t)\|_{L^2} + T\max_{t\in[0,T]} \|\epsilon_u(t)\|_{L^2}^2\right). \tag{4.44}$$

When T and h are sufficiently small, we can absorb $C(Q)(h^{k-r}+T)\max_{t\in[0,T]} \|\epsilon_u\|_{L^2}^2$ into the left-hand side of (4.44) to obtain

$$\max_{t\in[0,T]} \|\epsilon_u(t)\|_{L^2}^2 \lesssim C(Q)h^{2k-2-2r}. \tag{4.45}$$

Then this leads to

$$\max_{t\in[0,T]} \|\epsilon_u(t)\|_{L^2} \lesssim C(Q)h^{k-1-r}. \tag{4.46}$$

Hence, we plug (4.46) into (4.43) and (4.40) to achieve

$$\int_0^t \|\epsilon_u\|_{H_h}^2 \lesssim C(Q)h^{2k-2-2r}, \tag{4.47}$$

$$\int_0^t \|\epsilon_p\|_{L^2} \lesssim C(Q)h^{k-1-r}. \tag{4.48}$$

Combining with the projection errors δ_u and δ_p, we show the desired result. □

Remark 4.6. The convergent rate $k-1-r$ is not optimal. If we can obtain a nicer error estimate of the temporal term, then this rate can be further improved.

5. Numerical tests

We perform the accuracy tests for our numerical scheme. Our tests are based on a set of exact solutions

$$\eta = Z\sin\left(2\pi\left(x_1 - t\right)\right), \tag{5.1}$$

$$u_1 = Z\left(6x_2^5 + 15x_2^4 + 12x_2^3 + 3x_2^2\right)\sin\left(2\pi\left(x_1-t\right)\right) + \left(6x_2^5 + 5x_2^4 + 1\right), \tag{5.2}$$

$$\begin{aligned} u_2 = {} & \pi Z^2\left(4x_2^6 + 15x_2^5 + 21x_2^4 + 13x_2^3 + 3x_2^2\right)\sin\left(4\pi\left(x_1-t\right)\right) \\ & + 2\pi Z\left(4x_2^6 + 7x_2^5 + 2x_2^4 - x_2^3\right)\cos\left(2\pi\left(x_1-t\right)\right), \end{aligned} \tag{5.3}$$

$$p = Z\sin\left(2\pi\left(x_1-t\right)\right), \tag{5.4}$$

with the source term

$$\begin{aligned} S_1 = {} & \left(u_1\right)_t - \eta_t\left(1+x_2\right)K\partial_2 u_1 + \left(u_1\left(\partial_1 u_1 - AK\partial_2 u_1\right) + Ku_2\partial_2 u_1\right) \\ & -\nu\left(\partial_{11}u_1 + \left(1+A^2\right)K^2\partial_{22}u_1 - 2AK\partial_{12}u_1 + \left(AK^2\partial_2 A - A\partial_1 K - K\partial_1 A\right)\partial_2 u_1\right) \\ & + \left(\partial_1 p - AK\partial_2 p\right), \end{aligned} \tag{5.5}$$

$$\begin{aligned} S_2 = {} & \left(u_2\right)_t - \eta_t\left(1+x_2\right)K\partial_2 u_2 + \left(u_1\left(\partial_1 u_2 - AK\partial_2 u_2\right) + Ku_2\partial_2 u_2\right) \\ & -\nu\left(\partial_{11}u_2 + \left(1+A^2\right)K^2\partial_{22}u_2 - 2AK\partial_{12}u_2 + \left(AK^2\partial_2 A - A\partial_1 K - K\partial_1 A\right)\partial_2 u_2\right) \\ & + K\partial_2 p, \end{aligned} \tag{5.6}$$

where $0 < Z < 1$ can be any fixed constant, A and K are defined as in (1.7), and differential operators ∂_t, ∂_i and ∂_{ij} are defined in the usual sense. In the following test, we always take $Z = 0.1$, $\nu = 0.05$ and $T = 1/8$.

5.1. Accuracy tests with SIPG

The following are the error tables for our numerical scheme with SIPG:

TABLE 1. L^2 error table for $X_h^1 - M_h^0 - S_h^1$ formulation with SIPG.

(a) Free Surface Error $\eta - \eta_h$

N	Error	Order
4	9.4028E-4	–
8	2.5493E-4	1.8830
16	6.6612E-5	1.9362
32	1.7409E-5	1.9359
64	4.9397E-6	1.8173

(b) Velocity Error $u - u_h$

N	Error	Order
4	4.8233E-2	–
8	1.3863E-2	1.7988
16	3.6182E-3	1.9378
32	9.4871E-4	1.9312
64	2.4593E-4	1.9477

(c) Pressure Error $p - p_h$

N	Error	Order
4	3.1028E-3	–
8	1.6255E-3	0.9327
16	8.2709E-4	0.9748
32	4.1570E-4	0.9925
64	2.0818E-4	0.9977

TABLE 2. L^2 error table for $X_h^2 - M_h^1 - S_h^2$ formulation with SIPG.

(a) Free Surface Error $\eta - \eta_h$

N	Error	Order
4	1.4435E-4	–
8	1.6283E-5	3.1481
16	2.1495E-6	2.9213
32	2.9961E-7	2.8428

(b) Velocity Error $u - u_h$

N	Error	Order
4	5.8590E-3	–
8	7.7408E-4	2.9201
16	9.8674E-5	2.9717
32	1.2586E-5	2.9709

(c) Pressure Error $p - p_h$

N	Error	Order
4	9.6076E-4	–
8	2.3182E-4	2.0512
16	5.4527E-5	2.0880
32	1.2942E-5	2.0749

5.2. Accuracy tests with NIPG

The following are the error tables for our numerical scheme with NIPG:

TABLE 3. L^2 error table for $X_h^1 - M_h^0 - S_h^1$ formulation with NIPG.

(a) Free Surface Error $\eta - \eta_h$

N	Error	Order
4	9.3809E-4	–
8	2.5179E-4	1.8975
16	6.5861E-5	1.9347
32	1.7258E-5	1.9322
64	4.9108E-6	1.8132

(b) Velocity Error $u - u_h$

N	Error	Order
4	4.5428E-2	–
8	1.8218E-2	1.3182
16	6.7565E-3	1.4310
32	2.1758E-3	1.6347
64	7.5074E-4	1.5352

(c) Pressure Error $p - p_h$

N	Error	Order
4	3.2035E-3	–
8	1.6481E-3	0.9588
16	8.4487E-4	0.9640
32	4.2453E-4	0.9929
64	2.1239E-4	0.9992

TABLE 4. L^2 error table for $X_h^2 - M_h^1 - S_h^2$ formulation with NIPG.

(a) Free Surface Error $\eta - \eta_h$

N	Error	Order
4	1.4810E-4	–
8	1.7895E-5	3.0489
16	3.3218E-6	2.4295
32	8.5392E-7	1.9598

(b) Velocity Error $u - u_h$

N	Error	Order
4	6.0580E-3	–
8	1.2693E-3	2.2548
16	1.9305E-4	2.7170
32	2.7534E-5	2.8097

(c) Pressure Error $p - p_h$

N	Error	Order
4	9.2305E-4	–
8	2.8213E-4	1.7100
16	7.5701E-5	1.8980
32	2.0531E-5	1.8825

5.3. Discussion on the numerical tests

In above accuracy tests, we can obtain the optimal order of convergence when applying scheme (2.14) with SIPG, and the sub-optimal order with NIPG. However, in both cases, our numerical results are much better than the analytical result obtained in Theorem 4.5.

6. Conclusions and remarks

In this paper, we construct a stable numerical scheme to solve the system (1.9) with discontinuous Galerkin method, and discuss the error analysis in the fluid with certain assumptions.

Although we focus on the 2-D fluid throughout this paper, it is easy to see this scheme can be naturally extended to the 3-D case with periodic settings for two horizontal directions. The main restriction to our scheme is that we require the free surface to be a single-valued function of the horizontal variables, which is not always true in practice, especially when the topological structure of the free surface varies during the evolution. Our scheme is non-conservative. Hence, it might not give the qualitatively correct simulation if the exact solution possesses singularities. However, for smooth exact solutions, our scheme can give a quite good approximation. Here we use the piecewise polynomial P^k due to the connection between the bulk discretization and surface discretization. Note that Q^k, which denotes the piecewise polynomials of degree at most k for each variables, is acceptable in the scheme, but not suitable in analysis since it cannot satisfy the desired inf-sup condition for pressure term in the error analysis.

Acknowledgements. The author thanks the editor and referees for their constructive comments and suggestions.

References

[1] M.J. Ahn, H.Y. Lee and M.R. Ohm, Error estimates for fully discrete approximation to a free boundary problem in polymer technology. *Appl. Math. Comput.* **138** (2003) 227–238.

[2] J. Thomas Beale, The initial value problem for the Navier–Stokes equations with a free surface. *Commun. Pure Appl. Math.* **34** (1981) 359–392.

[3] B. Cockburn and C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comput.* **52** (1989) 411–435.

[4] B. Cockburn and C.-W. Shu, Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *J. Sci. Comput.* **16** (2001) 173–261.

[5] B. Cockburn, S. Hou and C.-W. Shu, The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case. *Math. Comput.* **54** (1990) 545–581.

[6] R. Danchin, Estimates in Besov spaces for transport and transport-diffusion equations with almost Lipschitz coefficients. *Rev. Mat. Iberoamericana* **21** (2005) 863–888.

[7] V. Girault, B. Riviere and M.F. Wheeler, A splitting method using discontinuous Galerkin for the transient incompressible Navier-Stokes equations. *ESAIM: M2AN* **39** (2005) 1115–1147.

[8] V. Girault, B. Riviere and M.F. Wheeler, A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier-Stokes problems. *Math. Comput.* **74** (2005) 249 53–84.

[9] J. Grooss and J.S. Hesthaven, A level set discontinuous Galerkin method for free surface flows. *Comput. Methods Appl. Mech. Engrg.* **195** (2006) 3406–3429.

[10] Y. Guo and I. Tice, Local well-posedness of the viscous surface wave problem without surface tension. *Anal. PDE* **6** (2013) 287–369.

[11] Y. Guo and I. Tice, Almost exponential decay of periodic viscous surface waves without surface tension. *Arch. Ration. Mech. Anal.* **207** (2013) 459–531.

[12] E. Hairer, C. Lubich and M. Roche, The numerical solution of differential-algebraic systems by Runge-Kutta methods. In vol. 1409 of *Lect. Notes Math.* Springer-Verlag, Berlin (1989).

[13] F.H. Harlow and J. Eddie Welch, Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Phys. Fluids* **8** (1965) 2182–2189.

[14] C.W. Hirt and B.D. Nichols, Volume of fluid (VOF) method for the dynamics of free boundaries. *J. Comput. Phys.* **39** (1981) 201–225.

[15] H.Y. Lee, Error analysis of finite element approximation of a Stefan problem with nonlinear free boundary condition. *J. Appl. Math. Comput.* **22** (2006) 223–235.

[16] R.H. Nochetto and C. Verdi, An efficient linear scheme to approximate parabolic free boundary problems: error estimates and implementation. *Math. Comput.* **51** (1988) 27–53.

[17] M. Sussman and M.Y. Hussaini, A discontinuous spectral element method for the level set equation. *J. Sci. Comput.* **19** (2003) 479–500.

[18] L.-H. Wang, On Korn's inequality. *J. Comput. Math.* **21** (2003) 321–324.

[19] M.F. Wheeler, An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.* **15** (1978) 152–161.

[20] L. Wu, Well-posedness and decay of the viscous surface wave. *SIAM J. Math. Anal.* **46** (2014) 2084–2135.