Journal of the KSIAM

Chief Editor:
Kim, Chongam (Seoul National Univ., Seoul)

Managing Editor:
Kim, Do Wan (Inha Univ., Incheon)
Lee, June-Yub (Ewha Univ., Seoul)
Shin, Sang Joon (Seoul National Univ., Seoul)

EDITORIAL BOARD
Agarwal, P. Ravi (Florida Institute of Technology, USA)
Bae, Hyeong-ohk (Ajou Univ., Suwon)
Byeon, Jaeyoung (POSTECH, Pohang)
Chen, Zhangxin (Univ. of Calgary, Canada)
Chen, Falai (Univ. of Science and Technology, China)
Cheon, Jung Hee (Seoul National Univ., Seoul)
Cho, Jin-Yeon (Inha Univ., Suwon)
Chou, So Hsiang (Bowling Green State Univ., USA)
Ha, Seung Yeal (Seoul National Univ., Seoul)
Hwang, Gang Uk (KAIST, Daejeon)
Jung, Il Hyo (Pusan National Univ., Pusan)
Kim, Hong Chul (Gangneung-Wonju National Univ., Gangneung)
Kim, Hyo Youngseok (Dong-Eui Univ., Pusan)
Kim, Kyu Hong (Seoul National Univ., Seoul)
Kim, Seokchan (Changwon National Univ., Changwon)
Koolen, Jack (POSTECH, Pohang)
Kwak, Minkyu (Chonnam National Univ., Gwangju)
Kwon, Yong Hoon (POSTECH, Pohang)
Lee, Gue Myung (Pukyong National Univ., Pusan)
Lee, Jongwoo (Kwangwoon Univ., Seoul)
Park, Eun-Jae (Yonsei Univ., Seoul)
Shim, Gyoocheol (Ajou Univ., Suwon)
Shin, Byeong-Chun (Chonnam National Univ., Gwangju)
Shin, Eui-Sup (Chonbuk National Univ., Jeonju)
Stephan, Ernst P. (Univ. Hannover, Germany)
Wang, Hong (Univ. South Carolina, USA)
Wu, Zong Min (Fudan Univ., China)

Aims and Scope

Journal of the KSIAM is devoted to theory, experimentation, algorithms, numerical simulation, or applications in the fields of applied mathematics, engineering, economics, computer science, and physics as long as the work is creative and sound.

Copyright

It is a fundamental condition that submitted manuscripts have not been published and will not be simultaneously submitted or published elsewhere. By submitting a manuscript, the authors agree that the copyright for their article is transferred to the publisher if and when the article is accepted for publication. The copyright covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microform or any other reproductions of similar nature, and translation. Photographic reproduction, microform, or any other reproduction of text, figures or tables from this journal is prohibited without permission obtained from the publisher.

Subscription

Subscription rate for individuals is $50/year; Domestic university library rate is $150/year; Foreign or industrial institute rate is $300/year.

Contact info.

The KOREAN SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS
KAIST, Yusong-gu, Daejeon 305-701, Republic of KOREA
Lim, Chulho (e-mail: jksiam@ksiams.org)
Tel.: +82-42-350-5708, Fax: +82-42-350-5710

Journal of KSIAM was launched in 1997, is published four times a year by the Korean Society for Industrial and Applied Mathematics. Total or part of the articles in this journal are abstracted in KSCI, Mathematical Reviews, CrossRef, Korean Science, and NDSL. Full text is available at http://www.ksiam.org/archive.

Copyright © 2013, the Korean Society for Industrial and Applied Mathematics.
contents

A P-HIERARCHICAL ERROR ESTIMATOR FOR A FEM-BEM COUPLING OF AN EDDY CURRENT PROBLEM IN $\mathbb{R}^3$
   FLORIAN LEYDECKER, MATTHIAS MAISCHAK, ERNST P. STEPHAN,
   AND MATTHIAS TELTSCHER ................................................................. 139

CIRCLE APPROXIMATION BY QUARTIC $G^2$ SPLINE USING ALTERNATION OF ERROR FUNCTION
   SOO WON KIM AND YOUNG JOON AHN ..................................................... 171

ANALYTIC APPROACH FOR THE STUDY OF AIR AND/OR LIQUID FILLED GEOMEMBRANE TUBE SECTIONS ON A HORIZONTAL
   YOON-RAK CHOI ....................................................................................... 181

COMPARISON OF DIFFERENT NUMERICAL SCHEMES FOR THE CAHN-HILLIARD EQUATION
   SEUNGGYU LEE, CHAEYOUNG LEE, HYUN GEUN LEE, AND JUNSEOK KIM ........... 197

EXTENSION OF AUSMPW+ SCHEME FOR TWO-FLUID MODEL
   JIN SEOK PARK AND CHONGAM KIM ..................................................... 209
A P-HIERARCHICAL ERROR ESTIMATOR FOR A FEM-BEM COUPLING OF AN EDDY CURRENT PROBLEM IN $\mathbb{R}^3$

DEDICATED TO PROFESSOR WOLFGANG L. WENDLAND ON THE OCCASION OF HIS 75TH BIRTHDAY

FLORIAN LEYDECKER†, MATTHIAS MAISCHAK2, ERNST P. STEPHAN1, AND MATTHIAS TELTSCHER1

1INSTITUTE FOR APPLIED MATHEMATICS, LEIBNIZ UNIVERSITÄT HANNOVER, GERMANY
E-mail address: leydecker@ifam.uni-hannover.de, stephan@ifam.uni-hannover.de

2BRUNEL UNIVERSITY, UXBRIDGE, UK

ABSTRACT. We extend a $p$-hierarchical decomposition of the second degree finite element space of Nédélec for tetrahedral meshes in three dimensions given in [1] to meshes with hexahedral elements, and derive $p$-hierarchical decompositions of the second degree finite element space of Raviart-Thomas in two dimensions for triangular and quadrilateral meshes. After having proved stability of these subspace decompositions and requiring certain saturation assumptions to hold, we construct a local a posteriori error estimator for fem and bem coupling of a time-harmonic electromagnetic eddy current problem in $\mathbb{R}^3$. We perform some numerical tests to underline reliability and efficiency of the estimator and test its usefulness in an adaptive refinement scheme.

1. INTRODUCTION

This paper is concerned with the construction of a reliable and efficient $p$-hierarchical based local a posteriori error estimator for a fem-bem coupling of a time-harmonic electromagnetic problem in $\mathbb{R}^3$.

The use of boundary elements for exterior problems in electromagnetics is not new, we mention the early work of MacCamy & Stephan [2, 3, 4, 5] and Nédélec [6, 7]. The coupling of fem and bem in electromagnetics has been pursued most notably by Bossavit [8], Costabel & Stephan [9], Nédélec et al [10, 11, 12, 13] and Hiptmair [14, 15]. In this paper we will be considering a field-based coupling formulation for an eddy current problem taken from [14]. The problem is discretized by edge elements inside the conductor and the exterior region is taken into account by means of a suitable boundary integral coupling.

Given a conductor and a monochromatic exciting current, the task in eddy current computations is to compute the resulting magnetic and electric fields, in the conductor as well as in

† Corresponding author.

2000 Mathematics Subject Classification. 65N15, 65N30, 65N38, 65N50.

Key words and phrases. hierarchical based a posteriori error estimation, Maxwell’s equations, eddy currents, fem-bem coupling, Nédélec’s elements, Raviart-Thomas elements.
the exterior domain. To this end, let \( \Omega \subset \mathbb{R}^3 \) be a bounded, simply connected open Lipschitz polyhedron with boundary \( \Gamma = \partial \Omega \), and further set \( \Omega_E = \mathbb{R}^3 \setminus \bar{\Omega} \). The domain \( \Omega \) then represents the conductor with conductivity \( \sigma \in L^\infty(\mathbb{R}^3) \), \( \sigma_1 \geq \sigma(x) \geq \sigma_0 > 0 \) and magnetic permeability \( \mu \in L^\infty(\mathbb{R}^3) \), \( \mu_1 \geq \mu(x) \geq \mu_0 > 0 \) with positive constants \( \sigma_0, \sigma_1, \mu_0, \mu_1 \). In the exterior region \( \Omega_E \), which represents air, we set \( \sigma \equiv 0 \) and by scaling \( \mu \equiv 1 \). The elementwise regularity of the material parameters reflects the fact that \( \Omega \) can consist of different conducting materials, i.e. the conductivity and permeability can jump from one material to another. We further assume a source current \( \mathbf{J}_0 \in H(\text{div}, \mathbb{R}^3) \) with \( \text{supp}(\mathbf{J}_0) \subset \bar{\Omega} \). It follows that \( \mathbf{J}_0 \cdot \mathbf{n} = 0 \) on \( \Gamma \) (there is no flow of \( \mathbf{J}_0 \) through \( \Gamma \)), where \( \mathbf{n} \) denotes the unit normal vector field on \( \Gamma \), defined almost everywhere and pointing from \( \Omega \) into \( \Omega_E \).

A mathematical model of the resulting time-harmonic eddy current problem for low frequencies (cf. Ammari, Buffa & Nédélec[16], MacCamy & Stephan[5]) consists of Maxwell’s equations

\[
\text{curl} \mathbf{E} = -i \omega \mu \mathbf{H}, \quad \text{curl} \mathbf{H} = \sigma \mathbf{E} + \mathbf{J}_0 \quad \text{in } \mathbb{R}^3, \tag{1.1}
\]

the Coulomb gauge \( \text{div} \mathbf{E} = 0 \) in \( \Omega_E \) together with the transmission conditions

\[
[\mathbf{E} \times \mathbf{n}]_\Gamma = 0, \quad [\mathbf{H} \times \mathbf{n}]_\Gamma = 0, \tag{1.2}
\]

and the Silver-Müller radiation conditions

\[
\mathbf{E}(x) = \mathcal{O}\left( \frac{1}{|x|} \right), \quad \mathbf{H}(x) = \mathcal{O}\left( \frac{1}{|x|} \right) \quad \text{uniformly for } |x| \to \infty. \tag{1.3}
\]

The equations in (1.1) are just the time-harmonic Maxwell equations with neglected displacement currents (formally setting \( \omega \epsilon = 0 \), where \( \epsilon \) denotes the electric permittivity). This approximation is justified in view of low frequencies \( \omega \). Note that the second equation in (1.1) reduces to \( \text{curl} \mathbf{H} = 0 \) in the exterior domain \( \Omega_E \). Therefore \( \mathbf{E} \) cannot be uniquely determined in \( \Omega_E \) and requires the Coulomb gauge condition. The transmission conditions (1.2) result from requiring \( \text{curl} \mathbf{E} \) and \( \text{curl} \mathbf{H} \) to be in \( L^2_{\text{loc}}(\mathbb{R}^3) \).

It must be stated that in spite of the Coulomb gauge, \( \mathbf{E} \) is unique only up to harmonic Dirichlet vector fields in \( \Omega_E \) (cf. [16]). But \( \mathbf{H} := \frac{1}{i \omega \mu} \text{curl} \mathbf{E} \) (by (1.1)) remains unique, so a scheme for determining the magnetic field, which is in fact the interesting quantity in most applications, can consist in first computing a solution \( \mathbf{E} \) and then deriving \( \mathbf{H} \) from \( \mathbf{E} \). If in addition we require \( \int_\Gamma \mathbf{E} \cdot \mathbf{n} = 0 \), then the solution \( \mathbf{E} \) is unique.

In [14], Hiptmair derives an \( \mathbf{E} \)-based coupling method for solving the problem (1.1)–(1.3) which is based on Costabel’s symmetric coupling method [17] (see also [18]). It is this variational formulation that we will be working with. The unknowns of the coupled formulation in this paper are \( \mathbf{u} \), the electrical field \( \mathbf{E} \) in \( \Omega \), and \( \mathbf{\lambda} \), the twisted tangential trace of the magnetic field on the transmission surface \( \Gamma \). The natural Sobolev space for \( \mathbf{u} \) is \( H(\text{curl}, \Omega) \), the space of \( L^2 \)-fields in \( \Omega \) with rotation in \( L^2(\Omega) \), and the space for \( \mathbf{\lambda} \) turns out to be a trace space of \( H(\text{curl}, \Omega) \). The discretization of \( \mathbf{u} \) uses the lowest order \( H(\text{curl}, \Omega) \)-conforming finite element space of Nédélec [19]. We use the corresponding trace space for discretizing \( \mathbf{\lambda} \), which is just a generalization of the lowest order finite element space of Raviart-Thomas on \( \Gamma \).
spaces belong to the class of edge element spaces, as their degrees of freedom correspond to edges of the grid.

Let \((E, \lambda)\) be the solution of the continuous problem for the above mentioned fem-bem coupling formulations and let \((E_h, \lambda_h)\) be the solution of the discrete problem. Then we are interested in finding a reliable and efficient \(p\)-hierarchical error estimator for the Galerkin error \((E - E_h, \lambda - \lambda_h)\) in the energy norm. One of the main reasons such local a posteriori estimators are so valuable is their usefulness in adaptive mesh refinement schemes. For a residual type error estimator of the fem-bem coupling solution of (1.1)–(1.3) see [20].

Though hierarchical error estimators have long been in use for elliptic problems (cf. [21, 22] for \(h\)-hierarchical estimation) and even for fem-bem coupling problems [23], the investigation of their usefulness in electromagnetics has only begun recently. A \(p\)-hierarchical error estimator for an eddy current problem in three dimensions using tetrahedral Nédélec elements can be found in [1]. As usual, this estimator depends on a stable subspace decomposition of the higher order finite element space and requires a saturation assumption. We extend the results of that paper in two directions – first, we also consider hexahedral meshes, but more importantly we now deal with coupling formulations, i.e. we have additional boundary element terms. Thus, stable subspace decompositions of higher order Raviart-Thomas elements are needed as well. We derive these from the decompositions of the Nédélec elements by virtue of the twisted tangential trace mapping. In [24] we have applied the results of this paper to the fem-bem coupling of a time-harmonic scattering problem. This work can also be seen as an extension of [25, 26, 27, 28, 23], which deal with a posteriori error estimates for (non-electromagnetic) coupling problems. Concerning electromagnetics, other recent articles dealing with a posteriori error estimators for edge elements are [29, 30, 31, 32].

The paper is organized as follows: In Section 2 we present the coupling formulation for the eddy current problem and in Section 3 we discuss the Galerkin method. Section 4 defines the finite element space \(ND_k(T_h)\) of first kind Nédélec elements on a mesh \(T_h\) in \(\Omega\) and investigates \(RT_k(K_h)\), the twisted tangential trace of \(ND_k(T_h)\), defined on the trace mesh \(K_h\). Starting from the decomposition of \(ND_2(T_h)\) for tetrahedra given in [1], we then construct stable \(p\)-hierarchical decompositions of \(ND_2(T_h)\) for hexahedra and of \(RT_2(K_h)\) for triangles and quadrilaterals in Section 5. In Section 6, we then apply the theory of the last section to find a local a posteriori error estimator for an eddy current fem-bem coupling formulation. Finally, the last section is devoted to numerically underlining the efficiency and reliability of a simplified form of the error estimator for the eddy current problem. We also test its usefulness in an adaptive mesh refinement scheme.

2. COUPLING FORMULATION

In this paper, we assume \(\Omega\) to be a simply connected polyhedron. Let us then denote the planar boundary faces by \(\Gamma_i, i = 1, \ldots, N_\Gamma\) such that \(\partial\Omega = \Gamma = \bigcup_{i=1}^{N_\Gamma} \Gamma_i\).

The complex duality pairings in \(\Omega\) and on \(\Gamma\) will be denoted by \(\langle \cdot, \cdot \rangle_\Omega\) and \(\langle \cdot, \cdot \rangle_\Gamma\). We use the usual Sobolev spaces \(H^s(\Omega)\) for scalar functions and \(H^s(\Omega)\) for vector fields of order \(s \in \mathbb{R}\)
(cf. Grisvard [33]). Furthermore we use the spaces
\[ H(\text{curl}, \Omega) := \{ v \in L^2(\Omega) : \text{curl} v \in L^2(\Omega) \}, \]
\[ H(\text{div}, \Omega) := \{ v \in L^2(\Omega) : \text{div} v \in L^2(\Omega) \}, \]
and the spaces of distributional tangential fields \( H_{\|}^{-1/2}(\Gamma) \) and \( H_{\perp}^{-1/2}(\Gamma) \) as introduced in [34] together with the trace spaces
\[ H_{\|}^{-1/2}(\text{div}_\Gamma, \Gamma) := \{ \zeta \in H_{\|}^{-1/2}(\Gamma) : \text{div}_\Gamma \zeta \in H^{-1/2}(\Gamma) \}, \]
\[ H_{\perp}^{-1/2}(\text{div}_\Gamma, 0, \Gamma) := \{ \zeta \in H_{\perp}^{-1/2}(\text{div}_\Gamma, \Gamma) : \text{div}_\Gamma \zeta = 0, \zeta \in H^{-1/2}(\Gamma) \}, \]
\[ H_{\perp}^{-1/2}(\text{curl}_\Gamma, \Gamma) := \{ \zeta \in H_{\perp}^{-1/2}(\Gamma) : \text{curl}_\Gamma \zeta \in H^{-1/2}(\Gamma) \}, \]
with the surface divergence operator \( \text{div}_\Gamma u := -\text{curl}_\Gamma (u \times n) \) and the surface curl operator \( \text{curl}_\Gamma u := \text{curl} u \cdot n \), see also [34, 35, 14]. We furthermore need the vectorial surface rotation for a scalar function \( \phi \) defined by \( \text{curl}_\Gamma \phi := \gamma^\times_t (\text{grad} \phi) \).

In the coupling formulation we will need integral operators to represent the exterior problem in (1.1)–(1.3). These operators are defined for \( x \in \Gamma \) as follows (for their properties see e.g. [14]):

\[ V(\lambda)(x) := \gamma_D V(\lambda)(x) = \gamma_D \int_{\Gamma} \Phi(x, y) \lambda(y) \, ds(y), \]
\[ K(\lambda)(x) := \gamma_D K(\lambda)(x) = \gamma_D \text{curl}_x \int_{\Gamma} \Phi(x, y) (n \times \lambda)(y) \, ds(y), \]
\[ \tilde{K}(\lambda)(x) := \gamma_N V(\lambda)(x) = (\gamma^\times_t) K(\lambda \times n)(x) = \gamma_N \int_{\Gamma} \Phi(x, y) \lambda(y) \, ds(y), \]
\[ W(\lambda)(x) := \gamma_N K(\lambda)(x) = (\gamma^\times_t) W(\lambda)(x) = \gamma_N \text{curl}_x \int_{\Gamma} \Phi(x, y) (n \times \lambda)(y) \, ds(y) \]
with Laplace kernel \( \Phi(x, y) = \frac{1}{4\pi|x-y|} \) and the limits \( \gamma_D \) and \( \gamma_N \) from \( \Omega_E \) onto \( \Gamma \) of the traces \( \gamma_D u := n \times (u \times n) =: u_{\Gamma} \) and \( \gamma_N u := \gamma^\times_t (\text{curl} u) \), where \( \gamma^\times_t u := u \times n \). Furthermore we need \( \gamma_n u := u \cdot n \).

After having collected the operators and spaces needed we formulate the coupled variational problem for the eddy current problem as ([14] and [18]):

**Find** \( u \in H(\text{curl}, \Omega), \lambda \in H_{\|}^{-1/2}(\text{div}_\Gamma, 0, \Gamma) \) such that

\[ (\mu^{-1} \text{curl} u, \text{curl} v)_{\Omega} + i\omega (\sigma u, v)_{\Omega} - (Wu_{\Gamma}, v_{\Gamma})_{\Gamma} + (\tilde{K} \lambda, v_{\Gamma})_{\Gamma} = -i\omega (J_0, v)_{\Omega}, \]
\[ (I - K) u_{\Gamma}, \zeta)_{\Gamma} + (W \lambda, \zeta)_{\Gamma} = 0 \]
for all \( v \in H(\text{curl}, \Omega), \zeta \in H_{\|}^{-1/2}(\text{div}_\Gamma, 0, \Gamma) \).

For brevity we write (2.1) as

\[ A(u, \lambda; v, \zeta) = L(v, \zeta). \]
The above formulation is obtained by using Green’s formula in $\Omega$ and a Stratton-Chu representation formula for $E$ in $\Omega_E$. The unknown $u$ corresponds to $E|_{\Omega}$, and the unknown $\lambda$ on the boundary corresponds to $\gamma_N E = -i\omega H_{\Omega_E} \times n$, which can indeed be seen to be surface divergence free. Due to the transmission conditions there holds $\lambda = \gamma_N u$. Note that the formulation (2.1) is block skew-symmetric. As observed by Hiptmair [14], the sesquilinear form $A$ is continuous and elliptic on $(H(\text{curl}, \Omega) \times H^{1/2}_|| (\text{div} 0, \Gamma))^2$. Thus, the variational formulation (2.1) admits a unique solution. Setting $E|_{\Omega} := u$, $E|_{\Omega_E} := \text{curl} V(n \times \gamma_D E) - V(\lambda)$ with the single layer potential $V$ with Laplace kernel and $H := \frac{1}{i\omega \mu} \text{curl} E$ gives a solution to the original problem (1.1)–(1.3) (in which the quantity $H$ is unique, as mentioned earlier).

3. THE GALERKIN METHOD

Let $T_h$ be a regular triangulation (with tetrahedral or hexahedral elements) of $\Omega$ and $K_h = \{ T \cap \Gamma : T \in T_h \}$ the induced triangulation on $\Gamma$. For the Galerkin method we use the finite element spaces suggested in [14], namely the well known $H(\text{curl}, \Omega)$-conforming finite element space $N^D(\Omega)$ of first kind Nédélec elements of first order [19] for discretization of the unknown $u \in H(\text{curl}, \Omega)$ and $R^T(K_h) := \{ \lambda_h \in R^T_1(K_h), \text{div}_T \lambda_h = 0 \}$ for the boundary unknown $\lambda \in H^{1/2}_|| (\text{div} 0, \Gamma)$, where $R^T_1(K_h)$ denotes the lowest order $H^{1/2}_|| (\text{div} \Gamma, \Gamma)$-conforming finite element space of Raviart-Thomas, which can be obtained as the image of $N^D(\Omega)$ under the mapping $\gamma^X$. Thus the Galerkin method reads:

Find $u_h \in N^D_1(T_h)$, $\lambda_h \in R^T_1(K_h)$ such that

$$(\mu^{-1} \text{curl} u_h, \text{curl} v_h)_\Omega + i\omega (\sigma u_h, v_h)_\Omega - \langle W \gamma_D u_h, \gamma_D v_h \rangle + \langle \tilde{\kappa} \lambda_h, \gamma_D v_h \rangle = -i\omega (J_0, v_h)_\Omega,$$

$$((I - \kappa) \gamma_D u_h, \zeta_h) + \langle \gamma \lambda_h, \zeta_h \rangle = 0$$

(3.1)

for all $v_h \in N^D_1(T_h)$, $\zeta_h \in R^T_1(K_h)$.

Now the conformity of the discrete spaces and the strong ellipticity of $A(\cdot, \cdot)$ imply that the Galerkin formulation (3.1) has a unique solution $(u_h, \lambda_h) \in N^D_1(T_h) \times R^T_1(K_h)$.

Next, we give an equivalent formulation of the above Galerkin method which is useful for the numerical implementation of the scheme. As $\Gamma$ is simply connected, we have $R^T_1(K_h) = \text{curl}_T S_1(K_h)$, where $S_1(K_h)$ denotes the finite element space of scalar, continuous and piecewise linear functions [36]. Instead of seeking $\lambda_h \in R^T_1(K_h)$, we now seek a function $\varphi_h \in S_1(K_h) := \{ \psi \in S_1(K_h), \int_\Gamma \psi ds(x) = 0 \}$ (and set $\lambda_h := \text{curl}_T \varphi_h$). We achieve this for a $\varphi_h \in S_1(K_h)$ by adding the equation $P(\varphi_h, \tau_h) := \int_\Gamma \varphi_h(x) ds(x) (\int_\Gamma \tau_h(x) ds(x)) = 0$ for all $\tau_h \in S_1(K_h)$ to our linear system. Note that the sesquilinear form $P(\varphi, \tau)$ is positive semidefinite ($P(\varphi, \varphi) = |\int_\Gamma \varphi(x) ds(x)|^2$), and that the corresponding matrix has rank 1. Thus the
alternative Galerkin method reads: \( \text{Find } \mathbf{u}_h \in \mathcal{N}D_1(T_h), \varphi_h \in S_1(K_h) \) such that

\[
\left(\mu^{-1} \text{curl } \mathbf{u}_h, \text{curl } \mathbf{v}_h\right)_\Omega + i\omega (\sigma \mathbf{u}_h, \mathbf{v}_h)_\Omega
- \langle \nabla \gamma_D \mathbf{u}_h, \gamma_D \mathbf{v}_h \rangle_\Gamma + \langle \hat{K} \text{curl}_\Gamma \varphi_h, \gamma_D \mathbf{v}_h \rangle_\Gamma
= -i\omega (\mathbf{J}_0, \mathbf{v}_h)_\Omega, \tag{3.2}
\]

for all \( \mathbf{v}_h \in \mathcal{N}D_1(T_h), \tau_h \in S_1(K_h) \).

Now, again, the conformity of the discrete spaces and the strong ellipticity of \( A(\cdot, \cdot) \) imply that the Galerkin formulation (3.2) has a unique solution \( (\mathbf{u}_h, \lambda_h) \in \mathcal{N}D_1(T_h) \times \text{curl}_\Gamma S_1(K_h) \).

\section{4. Finite element spaces}

We consider meshes with tetrahedral elements and with parallelepiped elements (which we will just call hexahedral elements) in \( \mathbb{R}^3 \), and we simply name them tetrahedral or hexahedral meshes. Analogously, we speak of triangular and quadrilateral meshes in \( \mathbb{R}^2 \), although quadrilateral elements are understood to be parallelograms.

In [19], Nédélec defines a family of conforming finite elements for \( \text{H(curl, } \Omega) \). For any element \( T \) of the regular tetrahedral mesh \( T_h \) define the local finite element space

\[
\mathcal{N}D_k(T) := (\mathbb{P}_{k-1}(T))^3 + \{ \mathbf{p} \in (\mathbb{P}_k(T))^3 : \mathbf{p}^T \cdot \mathbf{x} = 0 \} \subset (\mathbb{P}_k(T))^3,
\]

inducing the global finite element space

\[
\mathcal{N}D_k(T_h) := \left\{ \eta_h \in \text{H(curl, } \Omega) ; \eta_{h|T} \in \mathcal{N}D_k(T) \forall T \in T_h \right\}.
\]

\( \mathbb{P}_k(T) \) denotes the space of polynomials of order \( k \) (a monomial is of order \( k \) on a tetrahedron when the sum of the exponents equals \( k \)). The local degrees of freedom are given by

\[
\begin{align*}
(1) & \int_T \mathbf{u} \cdot \mathbf{t} \, q \, ds & \forall q \in \mathbb{P}_{k-1}, e \text{ edge of } T, \\
(2) & \int_T \mathbf{u} \times \mathbf{n} \cdot \mathbf{q} \, dS & \forall \mathbf{q} \in (\mathbb{P}_{k-2})^2, F \text{ face of } T, \\
(3) & \int_T \mathbf{u} \cdot \mathbf{q} \, dx & \forall \mathbf{q} \in (\mathbb{P}_{k-3})^3.
\end{align*}
\]

This choice of degrees of freedom ensures tangential continuity and thus \( \text{H(curl, } \Omega) \)-conformity [19]. If an element \( T \) is the image of another element \( \hat{T} \) under the affine transformation

\[
\mathbf{x} = \ell(\hat{\mathbf{x}}) := B\hat{\mathbf{x}} + \mathbf{d}, \quad B \in \mathcal{L}(\hat{T}, \mathbb{R}^3), \quad \mathbf{d} \in \mathbb{R}^3 \tag{4.1}
\]

and \( \{ \hat{\mathbf{b}}_j, j = 1, \ldots, n_k \} \) is a basis of \( \mathcal{N}D_k(\hat{T}) \), then a local basis on \( T \) is given by

\[
\mathbf{b}_j(\mathbf{x}) = (B^T)^{-1} \hat{\mathbf{b}}_j(\hat{\mathbf{x}}), \quad j = 1, \ldots, n_k. \tag{4.2}
\]

Global form functions are obtained by gluing together the local basis functions belonging to a common edge or face.
The construction of Nédélec finite elements on hexahedral meshes is very similar. Let $T_h$ be a regular hexahedral mesh on $\Omega$. Define the local space

$$\mathcal{N}D_k(T) = Q_{k-1,k,k}(T) \times Q_{k-1,k}(T) \times Q_{k,k,k-1}(T) \subset (Q_{k,k,k}(T))^3.$$ 

Here $Q_{k,l,m}(T)$ denotes the space of polynomials with maximum degree $k$ in $x$, $l$ in $y$ and $m$ in $z$. Then define the global finite element space $\mathcal{N}D_k(T_h)$ as before. The degrees of freedom are now given by

1. $\int_e u \cdot t \, q \, ds$ $\forall q \in Q_{k-1,e}$, $e$ edge of $T$,
2. $\int_F (u \times n) \cdot q \, dS$ $\forall q \in Q_{k-2,k,k-1} \times Q_{k-1,k-2}$, $F$ face of $T$,
3. $\int_T u \cdot q \, dx$ $\forall q \in Q_{k-1,k-2,k-2} \times Q_{k-2,k-1,k-2} \times Q_{k-2,k-2,k-1}$.

For the lowest order $p = 1$ we get the following basis functions associated to the edges of the reference element, see Figure 1.

$$b^{(e_0)} = \frac{1}{8}(1 - y)(1 - z)e_x, \quad b^{(e_1)} = \frac{1}{8}(1 + y)(1 - z)e_x, \quad b^{(e_2)} = \frac{1}{8}(1 - y)(1 + z)e_x,$$

$$b^{(e_3)} = \frac{1}{8}(1 + y)(1 + z)e_x, \quad b^{(e_4)} = \frac{1}{8}(1 - x)(1 - z)e_y, \quad b^{(e_5)} = \frac{1}{8}(1 + x)(1 - z)e_y,$$

$$b^{(e_6)} = \frac{1}{8}(1 - x)(1 + z)e_y, \quad b^{(e_7)} = \frac{1}{8}(1 + x)(1 + z)e_y, \quad b^{(e_8)} = \frac{1}{8}(1 - x)(1 - y)e_z,$$

$$b^{(e_9)} = \frac{1}{8}(1 + x)(1 - y)e_z, \quad b^{(e_{10})} = \frac{1}{8}(1 - x)(1 + y)e_z, \quad b^{(e_{11})} = \frac{1}{8}(1 + x)(1 + y)e_z.$$

We remark that the edge functions are constant on the edge which they are associated to.

Here are some examples of the 54 basis functions for the polynomial degree $p = 2$.

![Figure 1. Numbering of the edges and faces on the unit cube $[-1, 1]^3$.](image-url)
There are two edge functions associated to the edge $e_0$ ($y = -1, z = -1$).

\[
\begin{align*}
b_1^{(e_0)} &= \frac{1}{32} (3y + 1)(y - 1)(3z + 1)(z - 1) \mathbf{e}_x, \\
b_2^{(e_0)} &= \frac{3}{32} x(3y + 1)(y - 1)(3z + 1)(z - 1) \mathbf{e}_x.
\end{align*}
\]

The function $b_1^{(e_0)}$ is constant on the edge $e_0$ with the value $\frac{1}{2}$. Furthermore, the tangential component vanishes everywhere except on the two faces which are adjacent to the edge.

There are four face functions associated to the face $F_0$ ($z = -1$), two in each direction.

\[
\begin{align*}
b_1^{(F_0)} &= \frac{3}{32} (1 - y^2)(3z + 1)(z - 1) \mathbf{e}_x, \\
b_2^{(F_0)} &= \frac{9}{32} x (1 - y^2)(3z + 1)(z - 1) \mathbf{e}_x, \\
b_3^{(F_0)} &= -\frac{3}{32} (1 - x^2)(3z + 1)(z - 1) \mathbf{e}_y, \\
b_4^{(F_0)} &= -\frac{9}{32} y (1 - x^2)(3z + 1)(z - 1) \mathbf{e}_y.
\end{align*}
\]

The tangential component of the face function is only non-zero on its associated face.

There are six interior functions

\[
\begin{align*}
b_1^{(T)} &= \frac{9}{32} (1 - y^2)(1 - z^2) \mathbf{e}_x, \\
b_2^{(T)} &= \frac{27}{32} x (1 - y^2)(1 - z^2) \mathbf{e}_x, \\
b_3^{(T)} &= \frac{9}{32} (1 - x^2)(1 - z^2) \mathbf{e}_y, \\
b_4^{(T)} &= \frac{27}{32} y (1 - x^2)(1 - z^2) \mathbf{e}_y, \\
b_5^{(T)} &= \frac{9}{32} (1 - x^2)(1 - y^2) \mathbf{e}_z, \\
b_6^{(T)} &= \frac{27}{32} z (1 - x^2)(1 - y^2) \mathbf{e}_z.
\end{align*}
\]

The interior functions are zero on four faces and have a vanishing normal component on the other two faces.

$\mathcal{ND}_k$ is still invariant under the affine transformation (4.1) if we transform the basis functions using (4.2), and we obtain the global basis functions by gluing together the local basis functions as before.

For both mesh types, define $\Pi^{\mathcal{ND}_k(T)} u \in \mathcal{ND}_k(T)$ as the unique interpolate of $u \in (C^\infty(T))^3$ such that $\alpha(u - \Pi^{\mathcal{ND}_k(T)} u) = 0$ for all degrees of freedom $\alpha$. We then have the approximation property [19, Theorem 2]:

**Lemma 4.1.** For $u \in H^{k+1}(T) \subset H(\text{curl}, T)$ ($k \in \mathbb{N}_0$) and an element $T$ (tetrahedral or hexahedral) with diameter $h_T$ there holds

\[
\|u - \Pi^{\mathcal{ND}_k(T)} u\|_{H(\text{curl}, T)} \leq c h_T^k \|u\|_{H^{k+1}(T)}
\]

with a constant $c$ dependent only on $k$ and the regularity of the element $T$.

We now turn our attention to the discretization of the trace space $H^{1/2}(\text{div}_\Gamma, \Gamma)$. We assume $\Omega$ to be a polyhedron, so that $\Gamma$ is piecewise plane. According to [34, 35], we know that the space $H^{1/2}(\text{div}_\Gamma, \Gamma)$ is just the twisted tangential trace of $H(\text{curl}, \Omega)$. It is thus obvious to discretize $H^{1/2}(\text{div}_\Gamma, \Gamma)$ using the twisted tangential trace of the space of Nédélec
elements. It is well known [37] that this yields the two dimensional $H(\text{div}, \Omega)$-conforming space of Raviart-Thomas, i.e.

$$\gamma_t^\times : \mathcal{N}D_k(\mathcal{T}_h) \to \mathcal{RT}_k(\mathcal{K}_h). \quad (4.3)$$

Also, the degrees of freedom carry over [37], i.e. for an element $T \in \mathcal{T}_h$, a face $K$ of $T$ and $u \in (C^\infty(\mathcal{T}))^3$ we have the identity

$$\gamma_t^\times \Pi^{\mathcal{ND}_k(T)} u = \Pi^{\mathcal{RT}_k(K)} \gamma_t^\times u. \quad (4.4)$$

A definition of the Raviart-Thomas space $\mathcal{RT}_k$ can be found in [38, 19], but we will be content to define $\mathcal{RT}_k(\mathcal{K}_h)$ by the above characterization (4.3). The next lemma derives the quality of the approximation of $H^{-1/2}(\text{div}_\Gamma, \Gamma)$ by $\mathcal{RT}_k$ from the approximation property of $\mathcal{ND}_k$ given in Lemma 4.1.

**Lemma 4.2.** For $\lambda \in H^{k+1/2}(K) \subset H^{-1/2}(\text{div}_\Gamma, K)$ ($k \in \mathbb{N}_0$) and $K$ face of the element $T$ with diameter $h$ there holds

$$\|\lambda - \Pi^{\mathcal{RT}_k(K)} \lambda\|_{H^{-1/2}(\text{div}_\Gamma, K)} \leq c h^k \|\lambda\|_{H^{k+1/2}(K)}$$

with a constant $c$ dependent only on $k$ and the regularity of the element $T$.

**Proof.** The constant $c$ appearing in this proof is always to be regarded as a generic constant. Let $u \in H^{k+1}(T) \subset H(\text{curl}, T)$ with $(u \times n)_{|K} = \lambda$. Continuity of $\gamma_t^\times$, (4.4) and Lemma 4.1 yield

$$\|\lambda - \Pi^{\mathcal{RT}_k(K)} \lambda\|_{H^{-1/2}(\text{div}_\Gamma, K)} \leq c \|u - \Pi^{\mathcal{ND}_k(T)} u\|_{H(\text{curl}, T)} \leq c h^k \|u\|_{H^{k+1}(T)}. \quad (4.5)$$

Defining the $H^{k+1/2}$-norm on $K$ by

$$\|\phi\|_{H^{k+1/2}(K)} = \inf_{\Phi \in H^{k+1}(T), \Phi_{|K} = \phi} \|\Phi\|_{H^{k+1}(T)},$$

we obtain

$$\|n \times \lambda\|_{H^{k+1/2}(K)} = \inf_{u \in H^{k+1}(T)} \|u\|_{H^{k+1}(T)}.$$

Because of $\lambda \cdot n = 0$, every $u$ with $u_{|K} = \lambda \times n$ satisfies $(u \times n)_{|K} = \lambda$, so that (4.5) leads to

$$\|\lambda - \Pi^{\mathcal{RT}_k(K)} \lambda\|_{H^{-1/2}(\text{div}_\Gamma, K)} \leq c h^k \|n \times \lambda\|_{H^{k+1/2}(K)} \leq c h^k \|\lambda\|_{H^{k+1/2}(K)}.$$

□
5. T wo-leve l decompositions and p-hierarchica l error estimators

We seek stable two-level decompositions of the finite element spaces introduced in the last section for ultimately constructing hierarchical error estimators.

In [1], the authors consider a \( p \)-hierarchical two-level decomposition of \( \mathcal{N}^2 \) for tetrahedral grids and describe the construction of a hierarchical error estimator. Here, we will extend this result to hexahedral grids and then use the trace mapping (4.3) to derive \( H_{∥}^{-1/2}(\text{div}_Γ, Γ) \) -stable decompositions of \( \mathcal{R}^2(Γ_h) \), again producing hierarchical error estimators.

Here and in the rest of the paper, the symbol \( \leq \) signifies “\( \leq \) up to a multiplicative constant”. Such constants are always assumed to be independent of the mesh width \( h \) (if present in the context). The symbol \( \simeq \) means “\( \leq \) and \( \geq \)”.

5.1. Decomposition of \( \mathcal{N}^2 \). Let \( T_h \) be a regular grid on \( Ω \) with mesh width \( h \), and denote with \( M \) the number of edges, with \( N \) the number of faces and with \( L \) the number of elements. Further, let \( S_k \) denote the finite element space of scalar, continuous and piecewise polynomial functions of order \( k \), and let \( \tilde{S}_k := S_k \setminus S_{k-1} \) (the hierarchical surplus). The dimension of \( S_k(T) \) is \( \dim S_k(T) = \frac{1}{6}(k+1)(k+2)(k+3) \) for a tetrahedron \( T \) and \( \dim S_k(T) = (k+1)^3 \) for a hexahedron \( T \).

For tetrahedral grids, the decomposition given in [29, 1] reads

\[
\mathcal{N}^2(T_h) = \mathcal{N}^1(T_h) \oplus \text{grad} \tilde{S}_2(T_h) \oplus \mathcal{N}^2_{⊥}(T_h)
\]

(5.1)

where

\[
\mathcal{N}^2_{⊥}(T_h) := \{ u_h \in \mathcal{N}^2(T_h) : \int_e u_h \cdot t q ds = 0, \forall q \in \mathbb{P}_1, e \text{ edge of } T_h \},
\]

i. e. \( \mathcal{N}^2_{⊥}(T_h) \) is spanned by face functions only.

Counting degrees of freedom on element \( T \), one sees that (5.1) is a direct sum: the dimension of \( \mathcal{N}^1(T) \) equals the number of edges, i.e. six, and the dimension of \( \tilde{S}_2(T) \) is \( 10 - 4 = 6 \) (again equal to the number of edges). We write \( \text{grad} \tilde{S}_2(T_h) = \text{span}\{\text{grad} \phi(e_1), \ldots, \text{grad} \phi(e_M)\} \).

The space \( \mathcal{N}^2(T) \) has dimension 20, corresponding to two basis functions per edge and two per face of \( T \). The basis functions on the faces span the space \( \mathcal{N}^2_{⊥}(T) \), which thus has dimension eight. Accordingly, for a tetrahedral mesh we write

\[
\mathcal{N}^2_{⊥}(T_h) = \text{span}\{b_1(F_1), b_2(F_1), \ldots, b_1(F_N), b_2(F_N)\}
\]

for the space spanned by the face-orientated basis functions of \( \mathcal{N}^2(T_h) \). The decomposition (5.1) can then be written as:

\[
\mathcal{N}^2(T_h) = \mathcal{N}^1(T_h) \oplus \sum_{i=1}^M \text{span}\{\text{grad} \phi(e_i)\} \oplus \sum_{j=1}^N \text{span}\{b_1(F_j), b_2(F_j)\}.
\]

(5.2)

This construction cannot be extended offhand to the hexahedral case, for the decomposition defined in (5.1) is then no longer a direct sum. Counting degrees of freedom, we see that
\( \text{grad} \, \tilde{S}_2(T) \) and \( \mathcal{ND}_2^\perp(T) \) overlap: the dimension of \( \mathcal{ND}_1(T) \) equals the number of edges, i.e. 12, the dimension of \( \tilde{S}_2(T) \) is equal to \( 27 - 8 = 19 \) (corresponding to one function per edge, one per face and one inner function), and the dimension of \( \mathcal{ND}^\perp_2(T) \) is 30 (four functions per face and six inner functions). But the dimension of \( \mathcal{ND}_2(T) \) is 54, so that there must hold \( \dim(\text{grad} \, \tilde{S}_2(T) \cap \mathcal{ND}_2^\perp(T)) = 7 \). Hence, if we are to find a direct decomposition of \( \mathcal{ND}_2(T) \) for hexahedra, we must determine 7 functions to eliminate from \( \text{grad} \, \tilde{S}_2(T) \cap \mathcal{ND}_2^\perp(T) \).

Let us write

\[
\tilde{S}_2(T) = \text{span}\{\phi^{(e_0)}, \ldots, \phi^{(e_{11})}, \phi^{(F_0)}, \ldots, \phi^{(F_3)}, \phi^{(T)}\}
\]

with edge based functions \( \phi^{(e_i)} \), face based functions \( \phi^{(F_i)} \) and bubble function \( \phi^{(T)} \). Furthermore with face based functions \( b_i^{(F)} \) and suitable 'bubble' functions \( b_i^{(T)} \) we can write

\[
\mathcal{ND}_2^\perp(T) = \text{span}\{b_1^{(F_0)}, \ldots, b_4^{(F_0)}, \ldots, b_1^{(F_5)}, \ldots, b_4^{(F_5)}, b_1^{(T)}, \ldots, b_6^{(T)}\}.
\]

By explicitly computing the basis functions of \( \mathcal{ND}_2^\perp(T) \) for the reference element \( T = [-1, 1]^3 \) (according to the degrees of freedom given earlier), one ascertains that the face functions of \( \text{grad} \, \tilde{S}_2(T) \) can be described by functions of \( \mathcal{ND}_2^\perp(T) \), for example there holds

\[
\text{grad} \, \phi^{(F_3)} = \text{grad} (1 - x^2)(1 - y^2)(1 - z) = -\frac{32}{9} (b_2^{(F_0)} - b_4^{(F_0)} + b_2^{(T)} + b_4^{(T)} + b_5^{(T)}),
\]

and similarly for the other \( \text{grad} \, \phi^{(F_j)} \). There further holds

\[
\text{grad} \, \phi^{(T)} = \text{grad} (1 - x^2)(1 - y^2)(1 - z^2) = -\frac{64}{27} (b_2^{(T)} + b_4^{(T)} + b_6^{(T)}).
\]

With this information, there are now many ways to exchange the spaces \( \text{grad} \, \tilde{S}_2(T) \) and \( \mathcal{ND}_2^\perp(T) \) by reduced spaces \( \text{grad} \, \tilde{S}_2(T) \) and \( \mathcal{ND}_2^{-\perp}(T) \) to obtain a direct sum. We propose:

1. leave \( \tilde{S}_2(T) \) as it is,
2. substitute \( b_2^{(F_j)} + b_4^{(F_j)} \) for the face functions \( b_2^{(F_j)}, b_4^{(F_j)} (j = 0, \ldots, 5) \) and further substitute \( b_2^{(T)} - b_4^{(T)} \) and \( b_4^{(T)} - b_6^{(T)} \) for the interior functions \( b_2^{(T)}, b_4^{(T)}, b_6^{(T)} \); this changes \( \mathcal{ND}_2^\perp(T) \) to \( \mathcal{ND}_2^{-\perp}(T) \).

We then obtain the global space

\[
\mathcal{ND}_2^{-\perp}(T_h) := \text{span}\{b_1^{(F_j)}, b_3^{(F_j)}, b_2^{(F_j)} + b_4^{(F_j)}, b_1^{(T_k)}, b_3^{(T_k)}, b_5^{(T_k)}, b_2^{(T_k)} - b_4^{(T_k)}, b_4^{(T_k)} - b_6^{(T_k)},
\]

\[
j = 1, \ldots, N, k = 1, \ldots, L
\]

and the direct decomposition

\[
\mathcal{ND}_2(T_h) = \mathcal{ND}_1(T_h) \oplus \text{grad} \, \tilde{S}_2(T_h) \oplus \mathcal{ND}_2^{-\perp}(T_h)
\]
for hexahedral grids, which can be broken down to:

\[
\mathcal{N} \mathcal{D}_2(\mathcal{T}_h) = \mathcal{N} \mathcal{D}_1(\mathcal{T}_h) \oplus \sum_{i=1}^{M} \text{span}\{\text{grad } \phi^{(e_i)}\}
\]

\[
\oplus \sum_{j=1}^{N} \left( \text{span}\{\text{grad } \phi^{(F_j)}\} \oplus \text{span}\{b_1^{(F_j)}, b_3^{(F_j)}, b_2^{(F_j)} + b_4^{(F_j)}\} \right)
\]

\[
\oplus \sum_{k=1}^{L} \left( \text{span}\{\text{grad } \phi^{(T_k)}\} \oplus \text{span}\{b_1^{(T_k)}, b_3^{(T_k)}, b_5^{(T_k)}, b_2^{(T_k)} - b_4^{(T_k)}, b_4^{(T_k)} - b_6^{(T_k)}\} \right).
\]

(5.3)

In what follows the stability of the decompositions (5.2) and (5.3) is crucial for the derivation of hierarchical error indicators. To this aim, we define for tetrahedra the subspace projections

\[
P_1 : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \mathcal{N} \mathcal{D}_1(\mathcal{T}_h),
\]

\[
P^{(F)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{b_1^{(F)}, b_2^{(F)}\},
\]

\[
P^{(e)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{\text{grad } \phi^{(e)}\},
\]

and for hexahedra the projections

\[
\tilde{P}_1 : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \mathcal{N} \mathcal{D}_1(\mathcal{T}_h),
\]

\[
\tilde{P}^{(F)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{b_1^{(F)}, b_3^{(F)}, b_2^{(F)} + b_4^{(F)}\},
\]

\[
\tilde{P}^{(T)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{b_1^{(T_k)}, b_3^{(T_k)}, b_5^{(T_k)}, b_2^{(T_k)} - b_4^{(T_k)}, b_4^{(T_k)} - b_6^{(T_k)}\},
\]

\[
\tilde{R}^{(e)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{\text{grad } \phi^{(e)}\},
\]

\[
\tilde{R}^{(F)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{\text{grad } \phi^{(F)}\},
\]

\[
\tilde{R}^{(T)} : \mathcal{N} \mathcal{D}_2(\mathcal{T}_h) \to \text{span}\{\text{grad } \phi^{(T)}\},
\]

so that for \(u_2 \in \mathcal{N} \mathcal{D}_2(\mathcal{T}_h)\) the decompositions (5.2) and (5.3) can be written as

\[
u_2 = P_1 u_2 + \sum_{i=1}^{M} R^{(e_i)} u_2 + \sum_{j=1}^{N} P^{(F_j)} u_2
\]

(5.4)

and

\[
u_2 = \tilde{P}_1 u_2 + \sum_{i=1}^{M} \tilde{R}^{(e_i)} u_2 + \sum_{j=1}^{N} \left( \tilde{P}^{(F_j)} u_2 + \tilde{P}^{(F_j)} u_2 \right) + \sum_{k=1}^{L} \left( \tilde{R}^{(T_k)} u_2 + \tilde{R}^{(T_k)} u_2 \right).
\]

(5.5)

The following lemma is standard (see [39] for details). It is used in the proof of Lemma 5.2 which postulates the stability of the decompositions.

**Lemma 5.1.** Let \(T \in \mathcal{T}_h\) be an element with diameter \(h_T\) and \(\hat{T}\) be the reference element. Let \(q : \mathbb{R}^3 \to \mathbb{R}^3\), \(q : T \to \mathbb{R}^3\).
(1) Then with the $H(\text{curl})$-conforming transformation (cf. (4.1) and (4.2))

$$q(x) = (B^T)^{-1} \hat{q}(\ell^{-1}(x))$$  \hspace{1cm} (5.6)

there holds

$$\|q\|_{L^2(T)} \sim h^{1/2}\|\hat{q}\|_{L^2(\hat{T})},$$  \hspace{1cm} (5.7)

(2) Then with the $H(\text{div})$-conforming transformation

$$q(x) = \frac{1}{\det B} B q(\ell^{-1}(x))$$  \hspace{1cm} (5.8)

there holds

$$\|q\|_{L^2(T)} \sim h^{-1/2}\|\hat{q}\|_{L^2(\hat{T})},$$  \hspace{1cm} (5.9)

The next lemma states the stability result. For the sake of clarity, we will denote the $H(\text{curl}, \Omega)$-norm simply by $\| \|$.

**Lemma 5.2.** The decompositions (5.2) resp. (5.3) are stable with respect to the $H(\text{curl}, \Omega)$-norm, i.e. for all $u_2 \in N D_2(T_h)$ there holds

$$\|u_2\|^2 \simeq \|P_T u_2\|^2 + \sum_{i=1}^M \|R^{(e_i)} u_2\|^2 + \sum_{j=1}^N \|P^{(F_j)} u_2\|^2$$  \hspace{1cm} (5.10)

resp.

$$\|u_2\|^2 \simeq \|\bar{P}_T u_2\|^2 + \sum_{i=1}^M \|\bar{R}^{(e_i)} u_2\|^2 + \sum_{j=1}^N \left( \|\bar{R}^{(F_j)} u_2\|^2 + \|\bar{P}^{(F_j)} u_2\|^2 \right)$$  \hspace{1cm} (5.11)

$$\quad + \sum_{k=1}^L \left( \|\bar{R}^{(T_k)} u_2\|^2 + \|\bar{P}^{(T_k)} u_2\|^2 \right).$$

**Proof.** First let us consider the case of hexahedra, i.e. (5.11). First we observe that due to the uniqueness of the decomposition (5.3) the mapping $\| \|$ is a norm where $\| \|$ is defined by

$$\|u_2\|^2_{L^2(\Omega)} := \|P_T u_2\|^2_{L^2(\Omega)} + \sum_{i=1}^M \|R^{(e_i)} u_2\|^2_{L^2(\Omega)} + \sum_{j=1}^N \left( \|R^{(F_j)} u_2\|^2_{L^2(\Omega)} + \|P^{(F_j)} u_2\|^2_{L^2(\Omega)} \right)$$

$$\quad + \sum_{k=1}^L \left( \|R^{(T_k)} u_2\|^2_{L^2(\Omega)} + \|P^{(T_k)} u_2\|^2_{L^2(\Omega)} \right) =: \sum_{T \in T_h} \sum_{P_T} \|P_T u_2\|^2_{L^2(T)}$$

Since the $L^2$-Norm is local we conclude with (5.7) that there holds

$$\|u_2\|^2_{L^2(\Omega)} = \sum_{T \in T_h} \|u_2\|^2_{L^2(T)} = \sum_{T \in T_h} \sum_{P_T} \|P_T u_2\|^2_{L^2(T)} \sim \sum_{T \in T_h} h_T \sum_{P_T} \|P_T u_2\|^2_{L^2(T)}.$$  \hspace{1cm} (5.12)
where \( \tilde{\mathbf{v}}(\tilde{x}) = B^T\mathbf{v}(x) \) is the transformation of \( \mathbf{v} \) to the reference element \( \hat{T} \), and \( P_T \) denotes a projection operator that is related to the element \( T \). The constant in the equivalence relation depends only on the shape regularity of the mesh. Furthermore there holds for \( \hat{\mathbf{u}}_2 \in \mathcal{N}D_2(\hat{T}) \)

\[
\| \hat{\mathbf{u}}_2 \|_{L^2(\hat{T})} = \sum_{P_T} \| P_T \hat{\mathbf{u}}_2 \|_{L^2(\hat{T})} \sim \| \hat{\mathbf{u}}_2 \|_{L^2(\hat{T})},
\]

since all norms are equivalent on a finite dimensional space and the number of projection operators on an element is bounded. Here the constant in the equivalence relation depends only on the decomposition on \( \hat{T} \). With (5.12) and (5.7) we obtain

\[
\| \mathbf{u}_2 \|_{L^2(\Omega)}^2 \sim \sum_{T \in \mathcal{T}_h} h_T \| \hat{\mathbf{u}}_2 \|_{L^2(\hat{T})}^2 \sim \sum_{T \in \mathcal{T}_h} \| \mathbf{u}_2 \|_{L^2(T)}^2 = \| \mathbf{u}_2 \|_{L^2(\Omega)}^2.
\]

Now it is left to show that there holds

\[
\| \text{curl} \mathbf{u}_2 \|_{L^2(\Omega)} := \sum_{P} \| \text{curl} P \mathbf{u}_2 \|_{L^2(\Omega)}^2 \sim \| \text{curl} \mathbf{u}_2 \|_{L^2(\Omega)}^2.
\]

This follows with the same arguments as above, when we use relation (5.9) for the transformation to the reference element since \( \text{curl} \mathbf{u}_2 \in \mathcal{R}T_2(\mathcal{T}_h) \) for \( \mathbf{u}_2 \in \mathcal{N}D_2(\mathcal{T}_h) \). Note that also the following decomposition is unique:

\[
\text{curl} \mathbf{u}_2 = \text{curl} \hat{P}_1 \mathbf{u}_2 + \sum_{i=1}^M \text{curl} \hat{R}^{(e_i)} \mathbf{u}_2 + \sum_{j=1}^N (\text{curl} \hat{R}^{(F_j)} \mathbf{u}_2 + \text{curl} \hat{P}^{(F_j)} \mathbf{u}_2)
\]

\[
+ \sum_{k=1}^L (\text{curl} \hat{R}^{(T_k)} \mathbf{u}_2 + \text{curl} \hat{P}^{(T_k)} \mathbf{u}_2).
\]

This can be seen as follows.

Let \( \text{curl} \mathbf{u}_2 = 0 \). Then we have \( \text{curl} \hat{P}_1 \mathbf{u}_2 = \text{curl} \Pi^{\mathcal{N}D_1} \mathbf{u}_2 = \Pi^{\mathcal{R}T_1} \text{curl} \mathbf{u}_2 = 0 \), furthermore there holds \( \text{curl} \hat{R}^{(e_i)} \mathbf{u}_2 = \text{curl} \hat{R}^{(F_j)} \mathbf{u}_2 + \text{curl} \hat{P}^{(F_j)} \mathbf{u}_2 = 0 \) due to \( \text{curl} \text{grad} \equiv 0 \). The decomposition (5.5) yields \( \text{curl} \left( \sum_{j=1}^N \hat{P}^{(F_j)} \mathbf{u}_2 + \sum_{k=1}^L \hat{P}^{(T_k)} \mathbf{u}_2 \right) = 0 \). Therefore there exists \( \psi_2 \in \mathcal{S}_2(\mathcal{T}_h) \) with \( \sum_{j=1}^N \hat{P}^{(F_j)} \mathbf{u}_2 + \sum_{k=1}^L \hat{P}^{(T_k)} \mathbf{u}_2 = \text{grad} \psi_2 \). Now \( \text{grad} \psi_2 = 0 \) since the sum (5.3) is direct, hence \( \hat{P}^{(F_j)} \mathbf{u}_2 = 0 \) for all \( j \) and \( \hat{P}^{(T_k)} \mathbf{u}_2 = 0 \) for all \( k \). Especially there holds \( \text{curl} \hat{P}^{(F_j)} \mathbf{u}_2 = 0 \) for all \( j \) and \( \text{curl} \hat{P}^{(T_k)} \mathbf{u}_2 = 0 \) for all \( k \). Thus \( \text{curl} \mathbf{u}_2 = 0 \) implies \( \text{curl} P \mathbf{u}_2 = 0 \) for all projections \( P \). Altogether there holds, independently of the meshsize \( h \),

\[
\| \mathbf{u}_2 \|_{L^2(\Omega)} \sim \| \mathbf{u}_2 \|_{L^2(\Omega)}, \quad \| \text{curl} \mathbf{u}_2 \|_{L^2(\Omega)} \sim \| \text{curl} \mathbf{u}_2 \|_{L^2(\Omega)}.
\]

This gives the assertion of the lemma in case of a hexahedral mesh for the \( \mathbf{H} (\text{curl}, \Omega) \)-norm and the equivalent energy norm.

Similar arguments apply to the case of tetrahedral grids (see [1, Lemma 3 and Lemma 4]).
Now let
\[
a(u, v) := (\alpha \text{curl} u, \text{curl} v)_\Omega + (\beta u, v)_\Omega
\]
(\alpha, \beta \in \mathbb{C} \setminus \{0\}, \frac{\alpha}{\beta} \notin \mathbb{R}_{<0}) be such that \(a(\cdot, \cdot)\) is continuous on \((H(\text{curl}, \Omega))^2\) and satisfies there a Gårding inequality. Then define
\[
\|v\|_E := \|v\|_{E(\Omega)} := |a(v, v)|^{1/2},
\]
the energy norm induced by \(a\), equivalent to the \(H(\text{curl}, \Omega)\)-norm. Thus the stability proven above holds for the energy norm as well. We now arrive at the construction of a hierarchical error estimator for the Galerkin method to the variational problem

Find \(u \in H(\text{curl}, \Omega)\) such that
\[
a(u, v) = f(v)
\]
for all \(v \in H(\text{curl}, \Omega)\) for a given right hand side \(f \in H(\text{curl}, \Omega)'\).

Denote with \(u_h\) and \(u_2\) the solutions of the Galerkin formulations in \(\mathcal{N}D_1(\mathcal{T}_h)\) resp. in \(\mathcal{N}D_2(\mathcal{T}_h)\). A crucial requirement now needed is the saturation assumption: There exists a sequence \((\delta_h)_h\) with \(\delta_h \leq \delta < 1\) such that
\[
\|u - u_2\|_E \leq \delta_h \|u - u_h\|_E.
\]
(5.15)
One infers that the error \(\|u - u_h\|\) is equivalent to \(\|u_2 - u_h\|\):

**Lemma 5.3.** If the saturation assumption (5.15) holds, one has
\[
\|e_2\|_E \leq \|u - u_h\|_E \leq \frac{1}{1 - \delta} \|e_2\|_E
\]
with the error term \(e_2 := u_2 - u_h\).

Proof. See [1, Lemma 1] or [21, Equation (4.13)].

One thus seeks an estimate of \(\|e_2\|_E\), preferably of local type. Note that \(e_2 = u_2 - u_h\) satisfies the defect equation
\[
a(e_2, \eta) = r(\eta) := f(\eta) - a(u_h, \eta) \quad \forall \eta \in \mathcal{N}D_2(\mathcal{T}_h).
\]
(5.16)
Now let \(\tilde{a}\) be the decoupled sesquilinear form defined on \(\mathcal{N}D_2(\mathcal{T}_h) \times \mathcal{N}D_2(\mathcal{T}_h)\) through the sesquilinear form \(a\) and the decompositions (5.4) resp. (5.5), i.e.
\[
\tilde{a}(u_2, v_2) := a(P_1 u_2, P_1 v_2) + \sum_{i=1}^{M} a(R^{(e_i)} u_2, R^{(e_i)} v_2) + \sum_{j=1}^{N} a(P^{(F_j)} u_2, P^{(F_j)} v_2)
\]
(5.17)
for tetrahedra and
\[
\tilde{a}(u_2, v_2) := a(\tilde{P}_1 u_2, \tilde{P}_1 v_2) + \sum_{i=1}^{M} a(\tilde{R}^{(e_i)} u_2, \tilde{R}^{(e_i)} v_2)
\]
\[
+ \sum_{j=1}^{N} \left( a(\tilde{R}^{(F_j)} u_2, \tilde{R}^{(F_j)} v_2) + a(\tilde{P}^{(F_j)} u_2, \tilde{P}^{(F_j)} v_2) \right)
\]
\[
+ \sum_{k=1}^{L} \left( a(\tilde{R}^{(T_k)} u_2, \tilde{R}^{(T_k)} v_2) + a(\tilde{P}^{(T_k)} u_2, \tilde{P}^{(T_k)} v_2) \right)
\]  \hspace{1cm} (5.18)

for hexahedra. Lemma 5.2 states that \(\tilde{a}\) is equivalent to \(a\), i.e. there holds \(\tilde{a}(u_2, u_2) \simeq a(u_2, u_2)\). Hence, \(\tilde{a}\) is continuous on \((H(\text{curl}, \Omega))^2\) and satisfies there a Gårdings inequality. Now define the error term \(\tilde{e}_2 \in \mathcal{N}D_2(\mathcal{T}_h)\) by
\[
\tilde{a}(\tilde{e}_2, \eta) = a(e_2, \eta) = r(\eta) := f(\eta) - a(u_h, \eta) \hspace{1cm} \forall \eta \in \mathcal{N}D_2(\mathcal{T}_h).
\]  \hspace{1cm} (5.19)

We expect \(\tilde{e}_2\) to be a good approximation of \(e_2 \in \mathcal{N}D_2(\mathcal{T}_h)\), and indeed there holds

**Lemma 5.4.** For \(\tilde{e}_2\) defined by (5.19) there holds
\[
\|\tilde{e}_2\| \simeq \|e_2\|.
\]

Proof. See [1, Lemma 2]. \(\square\)

Now let \(P\) be a projection operator from (5.4) resp. (5.5) and let \(V_P \subset \mathcal{N}D_2(\mathcal{T}_h)\) be the corresponding subspace. Because of the decoupling character of the sesquilinear form \(\tilde{a}\), the defect equation (5.19) can be solved locally, yielding the defect equations
\[
a(P\tilde{e}_2, \eta) = r(\eta) := f(\eta) - a(u_h, \eta) \hspace{1cm} \forall \eta \in V_P \subset \mathcal{N}D_2(\mathcal{T}_h).
\]  \hspace{1cm} (5.20)

In particular, for \(V_P = \mathcal{N}D_1(\mathcal{T}_h)\) there holds the localized equation
\[
a(P_1 \tilde{e}_2, \eta) = f(\eta) - a(u_h, \eta) = 0 \hspace{1cm} \forall \eta \in \mathcal{N}D_1(\mathcal{T}_h).
\]

Thus \(P_1 \tilde{e}_2 = 0\), so that \(\tilde{e}_2\) actually lies in the hierarchical surplus \(\mathcal{N}D_2(\mathcal{T}_h) := (Id - \Pi_{\mathcal{N}D_1}) \mathcal{N}D_2(\mathcal{T}_h)\). Regarding the other subspaces for the tetrahedral case, (5.20) yields the one-dimensional problems

*For* \(i = 1, \ldots, M\) *find* \(\psi^{(e_i)} \in \text{span}\{\phi^{(e_i)}\}\) *such that*
\[
(\beta \text{grad} \psi^{(e_i)}, \text{grad} \phi^{(e_i)})_{\Omega} = r(\text{grad} \phi^{(e_i)})
\]  \hspace{1cm} (5.21)

and the two-dimensional problems

*For* \(j = 1, \ldots, N\) *find* \(\Psi^{(F_j)} \in \text{span}\{b_1^{(F_j)}, b_2^{(F_j)}\}\) *such that*
\[
a(\Psi^{(F_j)}, b_2^\perp) = r(b_2^\perp) \hspace{1cm} \forall b_2^\perp \in \text{span}\{b_1^{(F_j)}, b_2^{(F_j)}\}.
\]  \hspace{1cm} (5.22)
There holds $\nabla \psi(e) = R(e) \tilde{e}_2$ and $\Psi(F) = P(F) \tilde{e}_2$. We now define

\[ \Theta(e) := \| \nabla \psi(e) \| _e, \]
\[ \Theta(F) := \| \Psi(F) \| _e \]

and obtain by virtue of the stability result (5.10):

**Proposition 5.5** (Theorem 1 of [1]). *If the saturation assumption (5.15) is satisfied, then on a tetrahedral grid there holds*

\[ \eta \leq \| u - u_h \| _e \leq \frac{1}{1 - \delta} \eta \]

*with the local a posteriori error estimator*

\[ \eta^2 := \sum_{i=1}^{M} \left( \Theta(e_i) \right)^2 + \sum_{j=1}^{N} \left( \Theta(F_j) \right)^2. \]

**Proof.** The assertion follows from Lemmas 5.3 and 5.4. It is $\eta = \| \tilde{e}_2 \| _e$. \qed

Now let the edges of a tetrahedral element $T$ be numbered by the indices $i = 0, \ldots, 5$, and let the sides be numbered by $j = 0, \ldots, 3$. Then the local contribution on $T$ is given by

\[ \eta_T^2 = \sum_{i=0}^{5} \frac{1}{k_i} \left( \Theta(e_i) \right)^2 + \frac{1}{2} \sum_{j=0}^{3} \left( \Theta(F_j) \right)^2, \]

where $k_i$ denotes the number of elements sharing the edge with index $i$. An adaptive mesh refining algorithm would now consist of computing the local error estimators $\eta_T$ for every element $T$ of $\mathcal{T}_h$. The element is refined if this value exceeds a certain limit (usually depending on the mean value or the maximum of the $\eta_T$’s, according to the chosen strategy). Of course additional refinement must be performed to maintain mesh regularity.

One last simplification arises from the fact that the defect problem for $\nabla \phi(e_i)$ is one-dimensional: Simple computations yield

\[ \Theta(e) = \frac{|f(\nabla \phi(e)) - a(u_h, \nabla \phi(e))|}{\| \nabla \phi(e) \| _e}. \]

For the two-dimensional problems we can write

\[ \Theta(F) = \| \Psi(F) \| _e = \| \kappa_1 b_1(F) + \kappa_2 b_2(F) \| _e, \]

where $(\kappa_1, \kappa_2)^T$ are solutions to the LSE

\[
\begin{pmatrix}
  a(b_1(F), b_1(F)) & a(b_2(F), b_1(F)) \\
  a(b_1(F), b_2(F)) & a(b_2(F), b_2(F))
\end{pmatrix}
\begin{pmatrix}
  \kappa_1 \\
  \kappa_2
\end{pmatrix}
= \begin{pmatrix}
  f(b_1(F)) - a(u_h, b_1(F)) \\
  f(b_2(F)) - a(u_h, b_2(F))
\end{pmatrix}.
\]

For hexahedra this procedure yields the following theorem; the various $\Theta$’s are defined thereafter.
Proposition 5.6. If the saturation assumption (5.15) is satisfied, then on a hexahedral grid there holds

\[ \eta \leq \| u - u_h \| e \leq \frac{1}{1 - \delta} \eta \]

with the local a posteriori error estimator

\[ \eta^2 := \sum_{i=1}^{M} (\Theta^{(e_i)})^2 + \sum_{j=1}^{N} \left( (\Theta_1^{(F_j)})^2 + (\Theta_2^{(F_j)})^2 \right) + \sum_{k=1}^{L} \left( (\Theta_1^{(T_k)})^2 + (\Theta_2^{(T_k)})^2 \right). \]

The local contributions on an element \( T \) are

\[ \eta_T^2 := \frac{1}{4} \sum_{i=0}^{11} (\Theta^{(e_i)})^2 + \frac{1}{2} \sum_{j=0}^{5} \left( (\Theta_1^{(F_j)})^2 + (\Theta_2^{(F_j)})^2 \right) + (\Theta_1^{(T)})^2 + (\Theta_2^{(T)})^2. \]

We have

\[ \Theta^{(e)} := \frac{|f(\text{grad} \phi^{(e)}) - a(u_h, \text{grad} \phi^{(e)})|}{\| \text{grad} \phi^{(e)} \| e}, \]

\[ \Theta_1^{(F)} := \frac{|f(\text{grad} \phi^{(F)}) - a(u_h, \text{grad} \phi^{(F)})|}{\| \text{grad} \phi^{(F)} \| e}, \]

\[ \Theta_1^{(T)} := \frac{|f(\text{grad} \phi^{(T)}) - a(u_h, \text{grad} \phi^{(T)})|}{\| \text{grad} \phi^{(T)} \| e}, \]

\[ \Theta_2^{(F)} := \| \kappa_1 b_1^{(F)} + \kappa_2 b_2^{(F)} + \kappa_3 b_3^{(F)} \| e, \]

where \( b_2^{(F)} := b_2^{(F)} + b_4^{(F)} \) and \((\kappa_1, \kappa_2, \kappa_3)^T\) is the solution of the algebraic system

\[
\begin{pmatrix}
    a(b_1^{(F)}, b_1^{(F)}) & a(b_2^{(F)}, b_1^{(F)}) & a(b_3^{(F)}, b_1^{(F)}) \\
    a(b_1^{(F)}, b_2^{(F)}) & a(b_2^{(F)}, b_2^{(F)}) & a(b_3^{(F)}, b_2^{(F)}) \\
    a(b_1^{(F)}, b_3^{(F)}) & a(b_2^{(F)}, b_3^{(F)}) & a(b_3^{(F)}, b_3^{(F)})
\end{pmatrix}
\begin{pmatrix}
    \kappa_1 \\
    \kappa_2 \\
    \kappa_3
\end{pmatrix} =
\begin{pmatrix}
    f(b_1^{(F)}) - a(u_h, b_1^{(F)}) \\
    f(b_2^{(F)}) - a(u_h, b_2^{(F)}) \\
    f(b_3^{(F)}) - a(u_h, b_3^{(F)})
\end{pmatrix},
\]

and

\[ \Theta_2^{(T)} := \| \sum_{\ell=1}^{5} \kappa_\ell \tilde{b}_\ell^{(T)} \| e, \]

where \( \tilde{b}_1^{(T)} := b_1^{(T)}, \tilde{b}_2^{(T)} := b_2^{(T)} - b_4^{(T)}, \tilde{b}_3^{(T)} := b_3^{(T)}, \tilde{b}_4^{(T)} := b_4^{(T)} - b_6^{(T)}, \tilde{b}_5^{(T)} := b_5^{(T)} \) and \((\kappa_1, \ldots, \kappa_5)^T\) is the solution of the algebraic system

\[
\begin{pmatrix}
    a(b_k^{(T)}, \tilde{b}_\ell^{(T)})
\end{pmatrix}_{k,\ell=1,\ldots,5}(\kappa_\ell)_{\ell=1,\ldots,5} = (f(\tilde{b}_k^{(T)}) - a(u_h, \tilde{b}_k^{(T)}))_{k=1,\ldots,5}.
5.2. Decomposition of $\mathcal{RT}_2(K_h)$. We now turn our attention to the trace space $H^{-1/2}(\text{div} \Gamma, \Gamma)$. We aim to find a $H^{-1/2}(\text{div} \Gamma, \Gamma)$-stable decomposition of $\mathcal{RT}_2(K_h)$ using the results of the last section. Let $m$ denote the number of edges and $n$ the number of elements in $K_h$, the triangular or quadrilateral trace mesh of $\mathcal{T}_h$. We apply the trace mapping (4.3) to decomposition (5.1) for tetrahedra and obtain the decomposition

$$\mathcal{RT}_2(K_h) = \mathcal{RT}_1(K_h) \oplus \text{curl}_\Gamma \tilde{S}_2(K_h) \oplus \mathcal{RT}_2\perp(K_h)$$

(5.23)

for triangles, where

$$\mathcal{RT}_2\perp(K_h) := \{ \lambda_h \in \mathcal{RT}_2(K_h) : \int_E \lambda_h \cdot n \, q \, ds = 0, \forall q \in \mathbb{P}_1, \text{ e side of } K_h \}$$

and $\tilde{S}_k(K_h) := S_k(K_h) \setminus S_{k-1}(K_h)$. Here $S_k(K_h)$ is the space of piecewise polynomials in two dimensions of degree $k$.

For $K \in K_h$ there holds $|S_k(K)| = \frac{1}{2}(k + 1)(k + 2)$, and the dimension of $\mathcal{RT}_2(K)$ is $|\mathcal{RT}_2(K)| = 8$, corresponding to two basis functions per side and two inner functions. If $K \in K_h$ is the face of the element $T \in \mathcal{T}_h$, then its three sides are three edges of $T$, so that the three basis functions spanning $\mathcal{RT}_1(K)$ are the images of the the three basis functions of $N'D_1(T)$ corresponding to those edges under the mapping $\gamma^\perp$. The three basis functions of $\tilde{S}_2(K)$ are the images of the three basis functions of $S_2(T)$ corresponding to those edges and the two basis functions spanning $\mathcal{RT}_2\perp(K)$ are the images of the two basis functions of $\tilde{N'D}_1(T)$ corresponding to the face $K$. Counting the basis functions yields that (5.23) is a direct sum. We write $\tilde{S}_2(K_h) = \text{span}\{\varphi^{(e)}_1, \ldots, \varphi^{(e)}_m\}$ and $\mathcal{RT}_2\perp(K_h) = \text{span}\{\lambda_1^{(K)}, \lambda_2^{(K)}, \ldots, \lambda_1^{(K)}, \lambda_2^{(K)}\}$. Localization as before yields

$$\mathcal{RT}_2(K_h) = \mathcal{RT}_1(K_h) \oplus \sum_{i=1}^{m} \text{span}\{\text{curl}_\Gamma \varphi^{(e)}_i\} \oplus \sum_{j=1}^{n} \text{span}\{\lambda_1^{(K)}, \lambda_2^{(K)}\}.\quad (5.24)$$

For the trace mesh of a hexahedral grid we obtain the decomposition

$$\mathcal{RT}_2(K_h) = \mathcal{RT}_1(K_h) \oplus \text{curl}_\Gamma \tilde{S}_2(K_h) \oplus \mathcal{RT}_{2\perp}(K_h),$$

(5.25)

with $\text{curl}_\Gamma \tilde{S}_2(K) = \{\text{curl}_\Gamma \varphi^{(e)}_0, \ldots, \text{curl}_\Gamma \varphi^{(e)}_3, \text{curl}_\Gamma \varphi^{(K)}\}$ (with a suitable bubble function $\varphi^{(K)}$) and $\mathcal{RT}_{2\perp}(K) = \{\lambda_1^{(K)}, \lambda_3^{(K)}, \lambda_2^{(K)} + \lambda_4^{(K)}\}$ where $\lambda_i^{(K)}$ ($i = 1, \ldots, 4$) are the images of the basis functions $b_i^{(F)}$ in $\tilde{N'D}_2(T)$ corresponding to the face $K$. Again, (5.25) constitutes a direct sum (there holds dim $\mathcal{RT}_2(K) = 12$ for quadrilateral elements), and its
localization reads
\[ \mathcal{RT}_2(K_h) = \mathcal{RT}_1(K_h) \oplus \sum_{i=1}^{m} \text{span}\{\text{curl}_\Gamma \varphi^{(e_i)}\} \]
\[ \oplus \sum_{j=1}^{n} \left( \text{span}\{\text{curl}_\Gamma \varphi^{(K_j)}\} \oplus \text{span}\{\lambda_1^{(K_j)}, \lambda_2^{(K_j)}, \lambda_3^{(K_j)} + \lambda_4^{(K_j)}\} \right). \] (5.26)

The task at issue is to show the stability of (5.24) resp. (5.26). To this aim, define for the tetrahedral case the projection operators
\[ p_1 : \mathcal{RT}_2(K_h) \rightarrow \mathcal{RT}_1(K_h), \]
\[ p^{(K)} : \mathcal{RT}_2(K_h) \rightarrow \text{span}\{\lambda_1^{(K)}, \lambda_2^{(K)}\}, \]
\[ r^{(e)} : \mathcal{RT}_2(K_h) \rightarrow \text{span}\{\text{curl}_\Gamma \varphi^{(e)}\} \]
and for quadrilaterals the projections
\[ \tilde{p}_1 : \mathcal{RT}_2(K_h) \rightarrow \mathcal{RT}_1(K_h), \]
\[ \tilde{p}^{(K)} : \mathcal{RT}_2(K_h) \rightarrow \text{span}\{\lambda_1^{(K)}, \lambda_2^{(K)}, \lambda_3^{(K)} + \lambda_4^{(K)}\}, \]
\[ \tilde{r}^{(e)} : \mathcal{RT}_2(K_h) \rightarrow \text{span}\{\text{curl}_\Gamma \varphi^{(e)}\}, \]
\[ \tilde{r}^{(K)} : \mathcal{RT}_2(K_h) \rightarrow \text{span}\{\text{curl}_\Gamma \varphi^{(K)}\}, \]
so that the decompositions (5.24) resp. (5.26) can then be written as
\[ \lambda_2 = p_1 \lambda_2 + \sum_{i=1}^{m} r^{(e_i)} \lambda_2 + \sum_{j=1}^{n} p^{(K_j)} \lambda_2 \] (5.27)
resp.
\[ \lambda_2 = \tilde{p}_1 \lambda_2 + \sum_{i=1}^{m} \tilde{r}^{(e_i)} \lambda_2 + \sum_{j=1}^{n} \left( \tilde{r}^{(K_j)} \lambda_2 + \tilde{p}^{(K_j)} \lambda_2 \right). \] (5.28)

Now the stability of these \( \mathcal{RT}_2 \)-decompositions can be proven via the stability of the \( \mathcal{ND}_2 \)-decompositions, as we will show in the following lemma. For the sake of clarity, we will denote the \( \mathbf{H}^{-1/2}(\text{div}_\Gamma, \Gamma) \)-norm simply by \( \| \cdot \| \) in the statement of the lemma.

**Lemma 5.7.** Under the assumption that there exists a continuous extension from \( \mathcal{RT}_p(K_h) \) to \( \mathcal{ND}_p(T_h) \) which also preserves the basis functions, the decompositions (5.24) resp. (5.26) are stable with respect to the \( \mathbf{H}^{-1/2}(\text{div}_\Gamma, \Gamma) \)-norm, i.e. for all \( \lambda_2 \in \mathcal{RT}_2(K_h) \) there holds
\[ \| \lambda_2 \|^2 \simeq \| p_1 \lambda_2 \|^2 + \sum_{i=1}^{m} \| p^{(e_i)} \lambda_2 \|^2 + \sum_{j=1}^{n} \| p^{(K_j)} \lambda_2 \|^2 \] (5.29)
\[ \| \mathbf{\lambda}_2 \|^2 \simeq \| \tilde{p} \mathbf{\lambda}_2 \|^2 + \sum_{i=1}^{m} \| \tilde{p}^{(e_i)} \mathbf{\lambda}_2 \|^2 + \sum_{j=1}^{n} \left( \| \tilde{p}^{(K_j)} \mathbf{\lambda}_2 \|^2 + \| \tilde{q}^{(K_j)} \mathbf{\lambda}_2 \|^2 \right). \] (5.30)

**Remark 1.** There exists a continuous extension operator from \( \mathcal{RT}_p(\mathcal{K}_h) \) to \( \mathcal{ND}_p(\mathcal{T}_h) \), see Alonso & Valli [40]. But this is only valid for the whole spaces \( \mathcal{ND}_p(\mathcal{T}_h) \) and \( \mathcal{RT}_p(\mathcal{K}_h) \). Although, we know that for every basis functions \( \phi \in \mathcal{RT}_p(\mathcal{K}_h) \) there exists a basis function \( \mathbf{b} \in \mathcal{ND}_p(\mathcal{T}_h) \) with \( \gamma^X_t(\mathbf{b}) = \phi \) it is not clear if the estimate

\[ \| \mathbf{b} \|_{\mathbb{H}(\text{curl}, \Omega)} \leq C \| \phi \|_{\mathbb{H}^{-1/2}(\text{div}_T, \Gamma)} \]

is independent of the mesh size \( h \).

**Proof.** (of Lemma 5.7) We take an arbitrary \( \mathbf{\lambda}_2 \in \mathcal{RT}_2(\mathcal{K}_h) \). We decompose \( \mathbf{\lambda}_2 \) according to (5.27) resp. (5.28) into \( \mathbf{\lambda}_2 = \sum_{r=0}^r \mathbf{\lambda}_{2,r} \) (where \( r = m + n \) for a triangular mesh and \( r = m + 2n \) for a quadrilateral mesh). From [40] we know that there exists a \( \mathbf{u}_2 \in \mathcal{ND}_2(\mathcal{K}_h) \) with \( \gamma^X_t \mathbf{u}_2 = \mathbf{\lambda}_2 \) and \( \| \mathbf{u}_2 \|_{\mathbb{H}(\text{curl}, \Omega)} \simeq \| \mathbf{\lambda}_2 \|_{\mathbb{H}^{-1/2}(\text{div}_T, \Gamma)} \). Thus, \( \mathbf{u}_2 \) owns a stable decomposition according to Lemma 5.2 \( \mathbf{u}_2 = \sum_{j=0}^J \mathbf{u}_{2,j} \) with \( K = M + 2N + 2L \). We now assume that for every \( \mathbf{\lambda}_{2,i} \) there exists a \( \mathbf{u}_{2,j} \) of the decomposition with \( \gamma^X_t \mathbf{u}_{2,j} = \mathbf{\lambda}_{2,i} \) and \( \| \mathbf{u}_{2,j} \|_{\mathbb{H}(\text{curl}, \Omega)} \simeq \| \mathbf{\lambda}_{2,i} \|_{\mathbb{H}^{-1/2}(\text{div}_T, \Gamma)} \). Using the continuity of \( \gamma^X_t \) we then obtain the equivalences

\[ \| \mathbf{u}_2 \|_{\mathbb{H}(\text{curl}, \Omega)} \simeq \| \mathbf{\lambda}_2 \|_{\mathbb{H}^{-1/2}(\text{div}_T, \Gamma)}, \]

\[ \| \mathbf{u}_{2,i} \|_{\mathbb{H}(\text{curl}, \Omega)} \simeq \| \mathbf{\lambda}_{2,i} \|_{\mathbb{H}^{-1/2}(\text{div}_T, \Gamma)}, \quad i = 1, \ldots, r. \]

This, together with the \( \mathcal{ND}_2 \)-stability \( \sum_{i=0}^r \| \mathbf{u}_{2,i} \|_{\mathbb{H}(\text{curl}, \Omega)} \simeq \| \mathbf{u}_2 \|_{\mathbb{H}(\text{curl}, \Omega)} \) in Lemma 5.2 proves the statement of the lemma.

Now let \( \mathcal{V}(V) \) denote the vectorial (scalar) single layer potential operator for the Laplace equation defined for vector (scalar) functions \( \mathbf{\lambda}(\lambda) \) by

\[
\mathcal{V}(\lambda)(x) := \gamma_t \int_{\Gamma} \Phi(x, y) \mathbf{\lambda}(x) \, dS(y), \quad x \in \Gamma, \]

\[
V(\lambda)(x) := \int_{\Gamma} \Phi(x, y) \lambda(x) \, dS(y), \quad x \in \Gamma, \]

with the Laplace-kernel \( \Phi(x, y) := \frac{1}{4\pi |x - y|} \), then we can define on \( \mathbb{H}^{-1/2}(\text{div}_T, \Gamma) \) a continuous sesquilinear form \( b \) which satisfies a Gårding inequality by

\[
b(\lambda, w) = \alpha(V \text{ div}_T \lambda, \text{ div}_T w)_\Gamma + \beta(\mathcal{V} \lambda, w)_\Gamma \] (5.31)
with \( \alpha, \beta \in \mathbb{C} \setminus \{0\}, \frac{\alpha}{\beta} \notin \mathbb{R}_{<0} \). We will consider
\[
\|\lambda\|_e := |b(\lambda, \lambda)|^{1/2},
\]
the energy norm induced by \( b \), equivalent to the \( H^{-1/2}(\text{div}_\Gamma, \Gamma) \)-norm. We now search a \( p \)-
hierarchical error estimator for the Galerkin method using Raviart-Thomas elements for the
problem
\[
\text{Find } \lambda \in H^{-1/2}(\text{div}_\Gamma, \Gamma) \text{ such that } b(\lambda, w) = g(w)
\]
for all \( w \in H^{-1/2}(\text{div}_\Gamma, \Gamma) \).

with a right-hand side \( g \in H^{-1/2}(\text{div}_\Gamma, \Gamma)' \). Having proven the stability estimates (5.29) and (5.30), the principal work has already been done. We now must simply proceed analogously to the construction of the error estimator for Nédélec elements.

Let \( \lambda_h \) and \( \lambda_2 \) denote the solutions to the Galerkin formulations in \( RT_1(K_h) \) and \( RT_2(K_h) \). We again require the saturation assumption
\[
\|\lambda - \lambda_2\|_e \leq \delta_h \|\lambda - \lambda_h\|_e
\]
(5.33) to hold with \( \delta_h \leq \delta < 1 \). Exactly as before in Lemma 5.3 we then have:

**Lemma 5.8.** If the saturation assumption (5.33) holds, one has
\[
\|\varepsilon_2\|_e \leq \|\lambda - \lambda_h\|_e \leq \frac{1}{1 - \delta}\|\varepsilon_2\|_e
\]
with the error term \( \varepsilon_2 := \lambda_2 - \lambda_h \).

We now define a decoupled sesquilinear form \( \tilde{b} \) on \( RT_2(K_h) \times RT_2(K_h) \) according to the decompositions (5.27) and (5.28) via
\[
\tilde{b}(\lambda_2, w_1) := b(p_1 \lambda_2, p_1 w_1) + \sum_{i=1}^m b(r^{(e_i)} \lambda_2, r^{(e_i)} w_1) + \sum_{j=1}^n b(p^{(K_j)} \lambda_2, p^{(K_j)} w_1)
\]
(5.34) for triangles and
\[
\tilde{b}(\lambda_2, w_1) := b(\tilde{p}_1 \lambda_2, \tilde{p}_1 w_1) + \sum_{i=1}^m b(\tilde{r}^{(e_i)} \lambda_2, \tilde{r}^{(e_i)} w_1)
\]
\[
+ \sum_{j=1}^n \left( b(\tilde{r}^{(K_j)} \lambda_2, \tilde{r}^{(K_j)} w_1) + b(\tilde{p}^{(K_j)} \lambda_2, \tilde{p}^{(K_j)} w_1) \right)
\]
(5.35) for quadrilaterals. Thanks to Lemma 5.7, \( \tilde{b} \) is equivalent to \( b \) and thus continuous on \( H^{-1/2}(\text{div}_\Gamma, \Gamma) \) and satisfies a Gårdings inequality. Define the error term \( \tilde{\varepsilon}_2 \in RT_2(K_h) \) by
\[
\tilde{b}(\tilde{\varepsilon}_2, \zeta) = b(\varepsilon_2, \zeta) = g(\eta) - b(\lambda_h, \zeta) \quad \forall \zeta \in RT_2(K_h).
\]
(5.36)
Just as before in Lemma 5.4 there now holds

Lemma 5.9. For $\tilde{\epsilon}_2$ defined by (5.36) there holds

$$\|\tilde{\epsilon}_2\|_e \simeq \|\epsilon_2\|_e.$$  

Setting $\eta = \|\epsilon\|_e$, the last two lemmas immediately give the following two theorems:

Proposition 5.10. If the saturation assumption (5.33) is satisfied, then on a triangular grid there holds

$$\eta \lesssim \|\lambda - \lambda_h\|_e \lesssim \frac{1}{1 - \delta} \eta$$

with the local a posteriori estimator

$$\eta^2 := \sum_{i=1}^{m} \left( \vartheta(e_i) \right)^2 + \sum_{j=1}^{n} \left( \vartheta(K_j) \right)^2.$$

The local contribution on a triangle $K$ (with sides corresponding to the indices $i = 0, 1, 2$) is

$$\eta^2_K := \frac{1}{2} \sum_{i=0}^{2} \left( \vartheta(e_i) \right)^2 + \left( \vartheta(K) \right)^2$$

with

$$\vartheta(e) := \frac{|g(\text{curl}_\Gamma \varphi(e)) - b(\lambda_h, \text{curl}_\Gamma \varphi(e))|}{\|\text{curl}_\Gamma \varphi(e)\|_e},$$

$$\vartheta(K) := \|\kappa_1 \lambda_1^{(K)} + \kappa_2 \lambda_2^{(K)}\|_e,$$

where $(\kappa_1, \kappa_2)^T$ is the solution of the LSE

$$\begin{pmatrix} b(\lambda_1^{(K)}, \lambda_1^{(K)}) & b(\lambda_2^{(K)}, \lambda_1^{(K)}) \\ b(\lambda_1^{(K)}, \lambda_2^{(K)}) & b(\lambda_2^{(K)}, \lambda_2^{(K)}) \end{pmatrix} \begin{pmatrix} \kappa_1 \\ \kappa_2 \end{pmatrix} = \begin{pmatrix} g(\lambda_1^{(K)}) - b(\lambda_h, \lambda_1^{(K)}) \\ g(\lambda_2^{(K)}) - b(\lambda_h, \lambda_2^{(K)}) \end{pmatrix}.$$

The analogous statement for the quadrilateral case reads:

Proposition 5.11. If the saturation assumption (5.33) is satisfied, then on a quadrilateral grid there holds

$$\eta \lesssim \|\lambda - \lambda_h\|_e \lesssim \frac{1}{1 - \delta} \eta$$

with the local a posteriori error estimator

$$\eta^2 := \sum_{i=1}^{m} \left( \vartheta(e_i) \right)^2 + \sum_{j=1}^{n} \left( \vartheta_1(K_j) \right)^2 + \left( \vartheta_2(K_j) \right)^2.$$

Here, the local contribution on an element $K$ (whose sides correspond to the indices $i = 1, 2, 3, 4$) is

$$\eta^2_K := \frac{1}{2} \sum_{i=1}^{4} \left( \vartheta(e_i) \right)^2 + \left( \vartheta_1(K) \right)^2 + \left( \vartheta_2(K) \right)^2,$$
with

\[ \vartheta^{(e)} := \frac{|g(\text{curl}_\Gamma \varphi(e)) - b(\lambda, \text{curl}_\Gamma \varphi(e))|}{\|\text{curl}_\Gamma \varphi(e)\|_\epsilon}, \]

\[ \vartheta^{(K)}_1 := \frac{|g(\text{curl}_\Gamma \varphi(K)) - b(\lambda_h, \text{curl}_\Gamma \varphi(K))|}{\|\text{curl}_\Gamma \varphi(K)\|_\epsilon}, \]

\[ \vartheta^{(K)}_2 := \|\kappa_1 \lambda^{(K)}_1 + \kappa_2 \tilde{\lambda}^{(K)}_2 + \kappa_3 \lambda^{(K)}_3\|_\epsilon, \]

where \( \tilde{\lambda}^{(K)}_2 := \lambda^{(K)}_2 + \lambda^{(K)}_4 \) and \((\kappa_1, \kappa_2, \kappa_3)^T\) is the solution of the algebraic system

\[
\begin{pmatrix}
 b(\lambda^{(K)}_1, \lambda^{(K)}_1) & b(\lambda^{(K)}_2, \lambda^{(K)}_1) & b(\lambda^{(K)}_3, \lambda^{(K)}_1) \\
 b(\lambda^{(K)}_1, \lambda^{(K)}_2) & b(\lambda^{(K)}_2, \lambda^{(K)}_2) & b(\lambda^{(K)}_3, \lambda^{(K)}_2) \\
 b(\lambda^{(K)}_1, \lambda^{(K)}_1) & b(\lambda^{(K)}_2, \lambda^{(K)}_1) & b(\lambda^{(K)}_3, \lambda^{(K)}_1)
\end{pmatrix}
\begin{pmatrix}
 \kappa_1 \\
 \kappa_2 \\
 \kappa_3
\end{pmatrix}
= \begin{pmatrix}
 g(\lambda^{(K)}_1) - b(\lambda_h, \lambda^{(K)}_1) \\
 g(\lambda^{(K)}_2) - b(\lambda_h, \lambda^{(K)}_2) \\
 g(\lambda^{(K)}_3) - b(\lambda_h, \lambda^{(K)}_3)
\end{pmatrix}.
\]

(5.37)

6. APPLICATION TO THE COUPLING FORMULATION

We will now apply the theory of the last section to the symmetric eddy current formulation coupling finite elements in a bounded domain and boundary elements on the boundary for the homogeneous exterior domain as described in Section 3. To derive a \( p \)-hierarchical error estimator for the Galerkin method (3.1), let \( X := H(\text{curl}, \Omega) \times H^{-1/2}_\|= (\text{div}_\Gamma , 0, \Gamma) \) denote the continuous space of the variational formulation, \( X_h := N D_1 (T_h) \times \text{curl}_\Gamma \tilde{S}_1 (K_h) \) the finite element space of the Galerkin formulation and \( X_2 := N D_2 (T_h) \times \text{curl}_\Gamma \tilde{S}_2 (K_h) \) the higher order finite element space, and let

\[
\mathcal{A}(u, \lambda; v, \zeta) := (\mu^{-1} \text{curl}_\Gamma u, \text{curl}_\Gamma v)_\Omega + \text{i} \omega (\sigma u, v)_\Omega - (\mathcal{W} u, v)_\Gamma
\]

\[
+ (\tilde{K} \lambda, v)_\Gamma + (\{1 - \tilde{K}\)u, \zeta)_\Gamma + (\mathcal{V} \lambda, \zeta)_\Gamma,
\]

(6.1)

be the sesquilinear form on \( X \times X \) from (2.1) and \( \mathcal{L}(v, \zeta) \in X' \) the right hand side given by (2.1); there holds \( \mathcal{L}(0, \zeta) = 0 \). Theorem 7.1 in [14] implies that the energy norm induced by \( \mathcal{A} \) is equivalent to the natural norm \( \|\cdot\|_X \) on \( X' \). Let us define on \( X \times X \) the sesquilinear form

\[
\mathcal{Q}(u, \lambda; v, \zeta) := a(u, v) + b(\lambda, \zeta)
\]

with \( a(u, v) := (\mu^{-1} \text{curl}_\Gamma u, \text{curl}_\Gamma v)_\Omega + \text{i} \omega (\sigma u, v)_\Omega \) and \( b(\lambda, \zeta) := (\mathcal{V} \lambda, \zeta)_\Gamma \) and the energy norms

\[
\|v\|^2 := |a(v, v)|, \quad \|\zeta\|^2 := |b(\zeta, \zeta)|
\]

on \( H(\text{curl}, \Omega) \) resp. \( H^{-1/2}_\|= (\text{div}_\Gamma , 0, \Gamma) \). Note that the sesquilinear form \( a \) corresponds to the \( a \) from (5.13) with \( \alpha = \mu^{-1} \) and \( \beta = \text{i} \omega \sigma \) and the sesquilinear form \( b \) corresponds to \( b \) from
(5.31) with \( \beta = 1 \). The \( \alpha \langle \nabla \operatorname{div} \lambda, \operatorname{div} \zeta \rangle \Gamma \)-part from (5.31) does not appear here, as we are dealing with divergence-free functions on \( \Gamma \). We further define the “decoupled” sesquilinear forms

\[
\tilde{Q}(u, \lambda; v, \zeta) = \tilde{a}(u, v) + \tilde{b}(\lambda, \zeta)
\]

with \( \tilde{a} \) from (5.17) resp. (5.18) and \( \tilde{b} \) from (5.34) resp. (5.35).

Now let \( (u, \lambda, \zeta) \in X \) be the solution of (2.1), \( (u_h, \lambda_h) \in X_h \) the Galerkin solution of (3.1) and \( (u_2, \lambda_2) \in \tilde{X}_2 \) the Galerkin solution on the higher order finite element space. As before, \( M \) denotes the number of edges in \( T_h \), \( m < M \) the number of edges in \( K_h \) (those on \( \Gamma \)), \( N \) the number of faces in \( T_h \), \( n < N \) the number of faces in \( K_h \) (those on \( \Gamma \)) and \( L \) the number of elements in \( T_h \). We proceed as before:

Define the error terms \( \epsilon_2, \epsilon_2 \in \tilde{X}_2 \) by

\[
Q(\epsilon_2, \epsilon_2; v, \zeta) = L(v, \zeta) - A(u_h, \lambda_h; v, \zeta) \quad \forall (v, \zeta) \in \tilde{X}_2
\]

and \( (\tilde{e}_2, \tilde{e}_2) \in \tilde{X}_2 \) by

\[
\tilde{Q}(\tilde{e}_2, \tilde{e}_2; v, \zeta) = Q(\epsilon_2, \epsilon_2; v, \zeta) \quad \forall (v, \zeta) \in \tilde{X}_2.
\]

Using the notation of Section 5, define for tetrahedral grids the quantities

\[
\Theta^{(e)} := \| P^{(e)} \tilde{e} \|_e, \quad i = 1, \ldots, M,
\]

\[
\Theta^{(F)} := \| R^{(F)} \tilde{e} \|_e \quad j = 1, \ldots, N,
\]

\[
g^{(e)} := \| r^{(e)} \tilde{e} \|_e \quad i = 1, \ldots, m.
\]

There then holds (again using notation from Section 5)

\[
\Theta^{(e)} = \frac{| L(\nabla \phi^{(e)}, 0) - A(u_h, \lambda_h; \nabla \phi^{(e)}, 0) |}{\| \nabla \phi^{(e)} \|_e},
\]

\[
\Theta^{(F)} = \| \kappa_1 b_1^{(F)} + \kappa_2 b_3^{(F)} \|_e,
\]

where \( (\kappa_1, \kappa_2)^T \) is the solution of the LSE

\[
\begin{pmatrix}
    a(b_1^{(F)}, b_1^{(F)}) & a(b_2^{(F)}, b_1^{(F)}) \\
    a(b_1^{(F)}, b_2^{(F)}) & a(b_2^{(F)}, b_2^{(F)})
\end{pmatrix}
\begin{pmatrix}
    \kappa_1 \\
    \kappa_2
\end{pmatrix}
= \begin{pmatrix}
    L(b_1^{(F)}, 0) - A(u_h, \lambda_h; b_1^{(F)}, 0) \\
    L(b_2^{(F)}, 0) - A(u_h, \lambda_h; b_2^{(F)}, 0)
\end{pmatrix},
\]

and further

\[
g^{(e)} = \frac{| A(u_h, \lambda_h; 0, \nabla \phi^{(e)}) |}{\| \nabla \phi^{(e)} \|_e}.
\]

The quantities \( g^{(F)} := \| p^{(F)} \tilde{e} \|_e \) do not appear here, as \( \tilde{e}_2 \in \nabla \Gamma \tilde{S}_2 \) (i.e. \( p^{(F)} \tilde{e}_2 = 0 \)).

As usual, we now require the saturation assumption

\[
\|(u - u_2, \lambda - \lambda_2)\|_X \leq \delta_h \|(u - u_h, \lambda - \lambda_h)\|_X
\]

with \( \delta_h \leq \delta < 1 \). There holds
Theorem 6.1. If the saturation assumption (6.2) is satisfied, then on a tetrahedral grid there holds

\[ \eta \leq \| (u - u_h, \lambda - \lambda_h) \|_{\mathcal{X}} \leq \frac{1}{1 - \delta} \eta \]

with the local a posteriori estimator

\[ \eta^2 := \sum_{i=1}^{M} \left( \Theta(e_i) \right)^2 + \sum_{j=1}^{N} \left( \Theta(F_j) \right)^2 + \sum_{i=1}^{m} \left( \vartheta(e_i) \right)^2. \]

Proof. From the continuity and coercitivity of \( \mathcal{A} \) we have

\[ \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}}^2 \leq \mathcal{A}(u_2 - u_h, \lambda_2 - \lambda_h; u_2 - u_h, \lambda_2 - \lambda_h) \]
\[ = \mathcal{L}(u_2 - u_h, \lambda_2 - \lambda_h) - \mathcal{A}(u_2 - u_h, \lambda_2 - \lambda_h) \]
\[ = \mathcal{Q}(e_2, \varepsilon_2; u_2 - u_h, \lambda_2 - \lambda_h) \]
\[ \leq \|(e_2, \varepsilon_2)\|_{\mathcal{X}} \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}}. \]

Hence there holds \( \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}} \leq \|(e_2, \varepsilon_2)\|_{\mathcal{X}}. \) We obtain the reverse inequality in the following way:

\[ \|(e_2, \varepsilon_2)\|_{\mathcal{X}}^2 = \mathcal{Q}(e_2, \varepsilon_2; e_2, \varepsilon_2) \]
\[ = \mathcal{L}(e_2, \varepsilon_2) - \mathcal{A}(u_h, \lambda_h; e_2, \varepsilon_2) \]
\[ = \mathcal{A}(u_2 - u_h, \lambda_2 - \lambda_h; e_2, \varepsilon_2) \]
\[ \leq \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}} \|(e_2, \varepsilon_2)\|_{\mathcal{X}}, \]

i.e. \( \|(e_2, \varepsilon_2)\|_{\mathcal{X}} \leq \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}}, \) and so we have proven the equivalence

\[ \|(e_2, \varepsilon_2)\|_{\mathcal{X}} \simeq \|(u_2 - u_h, \lambda_2 - \lambda_h)\|_{\mathcal{X}}. \]

Now Lemmas 5.4 and 5.9 yield \( \eta := \|(e_2, \varepsilon_2)\|_{\mathcal{X}} \simeq \|(e_2, \varepsilon_2)\|_{\mathcal{X}}, \) and we have thus proven the statement of the theorem. \( \square \)

The same procedure for hexahedral grids yields the quantities

\[ \Theta^{(e)} := \frac{|\mathcal{L}(\text{grad } \phi^{(e)}, 0) - \mathcal{A}(u_h, \lambda_h; \text{grad } \phi^{(e)}, 0)|}{\|\text{grad } \phi^{(e)}\|_\varepsilon}, \]
\[ \Theta^{(F)}_1 := \frac{|\mathcal{L}(\text{grad } \phi^{(F)}, 0) - \mathcal{A}(u_h, \lambda_h; \text{grad } \phi^{(F)}, 0)|}{\|\text{grad } \phi^{(F)}\|_\varepsilon}, \]
\[ \Theta^{(T)}_1 := \frac{|\mathcal{L}(\text{grad } \phi^{(T)}, 0) - \mathcal{A}(u_h, \lambda_h; \text{grad } \phi^{(T)}, 0)|}{\|\text{grad } \phi^{(T)}\|_\varepsilon}, \]
\[ \Theta^{(F)}_2 := \|\kappa_1 b_1^{(F)} + \kappa_2 \tilde{b}_2^{(F)} + \kappa_3 b_3^{(F)}\|_\varepsilon, \]
where $\tilde{b}_2^{(F)} := b_2^{(F)} + b_4^{(F)}$ and $(\kappa_1, \kappa_2, \kappa_3)^T$ is the solution of the LSE

$$
\begin{pmatrix}
  a(b_1^{(F)}, b_1^{(F)}) & a(\tilde{b}_2^{(F)}, b_1^{(F)}) & a(b_3^{(F)}, b_1^{(F)}) \\
  a(b_2^{(F)}, b_2^{(F)}) & a(\tilde{b}_2^{(F)}, b_2^{(F)}) & a(b_3^{(F)}, b_2^{(F)}) \\
  a(b_1^{(F)}, b_3^{(F)}) & a(\tilde{b}_2^{(F)}, b_3^{(F)}) & a(b_3^{(F)}, b_3^{(F)})
\end{pmatrix}
\begin{pmatrix}
  \kappa_1 \\
  \kappa_2 \\
  \kappa_3
\end{pmatrix}
= \begin{pmatrix}
  \mathcal{L}(b_1^{(F)}, 0) - \mathcal{A}(u_h, \lambda_h; b_2^{(F)}, 0) \\
  \mathcal{L}(b_2^{(F)}, 0) - \mathcal{A}(u_h, \lambda_h; b_2^{(F)}, 0) \\
  \mathcal{L}(b_3^{(F)}, 0) - \mathcal{A}(u_h, \lambda_h; b_3^{(F)}, 0)
\end{pmatrix},
$$

and further

$$\Theta_2^{(T)} := \sum_{\ell=1}^5 \kappa_\ell \delta e^{(T)},$$

where $\tilde{b}_1^{(T)} := b_1^{(T)}$, $\tilde{b}_2^{(T)} := b_2^{(T)} - b_4^{(T)}$, $\tilde{b}_3^{(T)} := b_3^{(T)}$, $\tilde{b}_4^{(T)} := b_4^{(T)} - b_6^{(T)}$, $\tilde{b}_5^{(T)} := b_5^{(T)}$, and $(\kappa_1, \ldots, \kappa_5)^T$ is the solution of the algebraic system

$$\left( a(b_k^{(T)}, \tilde{b}_\ell^{(T)}) \right)_{k,\ell=1,\ldots,5} (\kappa_\ell)_{\ell=1,\ldots,5} = \left( \mathcal{L}(\tilde{b}_k^{(T)}, 0) - \mathcal{A}(u_h, \lambda_h; \tilde{b}_k^{(T)}, 0) \right)_{k=1,\ldots,5},$$

and further

$$\vartheta^{(e)} := \frac{|\mathcal{A}(u_h, \lambda_h; 0, \text{curl}\varphi^{(e)})|}{|\text{curl}\varphi^{(e)}|_e},$$

$$\vartheta^{(F)} := \frac{|\mathcal{A}(u_h, \lambda_h; 0, \text{curl}\varphi^{(F)})|}{|\text{curl}\varphi^{(F)}|_e}.$$
7. NUMERICAL EXPERIMENTS

We perform some numerical tests on hexahedral meshes to see if the error estimator gives a reliable and efficient estimate of the Galerkin error and to test its usefulness for an adaptive refinement scheme.

Let \( X := H(\text{curl}, \Omega) \times H^{-1/2}(\text{div} \Gamma_0, \Gamma) \) and \( X_h := ND_1(T_h) \times \text{curl}_\Gamma S_1(K_h) \) the finite element space as described above. Furthermore, we denote by \( X_2 := ND_2(T_h) \times \text{curl}_\Gamma S_1(K_h) \) the higher order finite element space. Here, we just consider a mesh of hexahedrons.

We define the energy norms on \( H(\text{curl}, \Omega) \) and \( H^{-1/2}(\text{div} \Gamma_0, \Gamma) \) by

\[
\|v\|_E^2 := \left\langle (\mu^{-1} \text{curl} u, \text{curl} u)^\Omega + i\omega(\sigma u, u)^\Omega, \right\rangle \\
\|\lambda\|_e^2 := |\langle V\lambda, \lambda \rangle^\Gamma|.
\]

In the first experiment we compute the solution to the Galerkin system as given in (3.1) with \( \Omega = (-1, 1)^3, \Gamma = \partial \Omega \), on a series of uniform hexahedral meshes, obtained by dividing each edge of \( \Omega \) into \( n \) equal parts. On grid \( n \) we thus have a meshwidth of \( h = \frac{2}{n} \). We then compare the energy norm \( \sqrt{\|u - u_h\|_E^2 + \|\lambda - \lambda_h\|_e^2} \) of the Galerkin error with the value of the error estimator. In the second example we will use the error estimator to perform adaptive mesh refinements. In the tables, \( n \) will denote the mesh number (as defined above) and \( N_u \) and \( N_\lambda \) the number of degrees of freedom for the fem resp. the bem variable. The choice of \( \Omega \) in both examples is only for simplicity; note that our above analysis is not restricted to convex domains \( \Omega \). All computations were performed using the program package maiprogs [41] (for further details see [20]).

Example 1. We choose the exact solution

\[
u(x) = \text{curl}(G\rho)(x) := \text{curl}\int_{\Omega} \frac{1}{\|x - y\|} \rho(y) \, dy
\]

with the density function \( \rho(x) = \rho(x) \left( \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right) \), where

\[
\rho(x) = \left( (1 - x_1^2)(1 - x_2^2)(1 - x_3^2) \right)^2 x_1 x_2 x_3.
\]

We compute the Galerkin method for \( n = 1, \ldots, 13 \) with hexahedral elements. In Figure 2 one sees that the error indicator \( \eta \) behaves nearly the same as the error in energy norm, the effectivity indices \( q = \frac{\eta}{e} \), calculated in Table 1, are nearly constant.

Example 2. We now use the error estimator to construct an adaptive mesh. We use hexahedral elements without hanging nodes (with the drawback that the resulting mesh is no longer form-regular). Our geometry remains the cube \( \Omega = (-1, 1)^3 \). We set \( \mu = 1 \) in \( \Omega \) and choose a discontinuous \( \sigma \), namely

\[
\sigma = \begin{cases} 
0.1, & \frac{1}{3} < x_1, x_2, x_3 < 1 \\
1, & \text{else}
\end{cases}
\]

For our right hand side in (2.1) we choose the function \( J_0 = (1, 1, 1) \) in \( \Omega \) and \( J_0 = 0 \) in \( \Omega_E \). Note that also in this case (2.1) holds. We start by computing the Galerkin solution for the
FIGURE 2. Energy norm $e$ of the Galerkin error and the 2-level hierarchical error estimator $\eta$ of Example 1.

<table>
<thead>
<tr>
<th>n</th>
<th>$h$</th>
<th>DOF</th>
<th>$e$</th>
<th>$\eta$</th>
<th>$q = \frac{\eta}{e}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>80</td>
<td>0.30987</td>
<td>0.15081</td>
<td>0.4867</td>
</tr>
<tr>
<td>3</td>
<td>0.667</td>
<td>200</td>
<td>0.30369</td>
<td>0.06440</td>
<td>0.2121</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>398</td>
<td>0.23548</td>
<td>0.05420</td>
<td>0.2302</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>692</td>
<td>0.18994</td>
<td>0.03879</td>
<td>0.2042</td>
</tr>
<tr>
<td>6</td>
<td>0.333</td>
<td>1100</td>
<td>0.15938</td>
<td>0.02969</td>
<td>0.1863</td>
</tr>
<tr>
<td>7</td>
<td>0.143</td>
<td>1640</td>
<td>0.13748</td>
<td>0.02410</td>
<td>0.1753</td>
</tr>
<tr>
<td>8</td>
<td>0.25</td>
<td>2330</td>
<td>0.12095</td>
<td>0.02037</td>
<td>0.1684</td>
</tr>
<tr>
<td>9</td>
<td>0.222</td>
<td>3188</td>
<td>0.10800</td>
<td>0.01770</td>
<td>0.1639</td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>4232</td>
<td>0.09755</td>
<td>0.01568</td>
<td>0.1607</td>
</tr>
<tr>
<td>11</td>
<td>0.091</td>
<td>5480</td>
<td>0.08894</td>
<td>0.01409</td>
<td>0.1584</td>
</tr>
<tr>
<td>12</td>
<td>0.083</td>
<td>6950</td>
<td>0.08172</td>
<td>0.01281</td>
<td>0.1568</td>
</tr>
<tr>
<td>13</td>
<td>0.077</td>
<td>8660</td>
<td>0.07558</td>
<td>0.01175</td>
<td>0.1555</td>
</tr>
</tbody>
</table>

TABLE 1. Energy norm $e$ of the Galerkin error, the 2-level hierarchical error estimator $\eta$ and the effectivity indices $q = \frac{\eta}{e}$ of Example 1.

uniform mesh with $n = 3$. The refinement algorithm then proceeds by first refining the 10% of
the elements on which the local contributions of the hierarchical error estimator are the largest and by then further refining in order to eliminate hanging nodes, since our algorithm yet cannot handle hanging nodes for 2nd order finite elements. We expect the algorithm to refine the mesh near the $\sigma$-discontinuity interface between $\Omega^{(1)} = \left(\frac{1}{3}, 1\right)^3$ and $\Omega^{(0)} = \Omega \setminus \Omega^{(1)}$, and especially close to the vertex $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Figure 3 shows adaptively generated meshes. Figure 4 shows that the adaptive refinement gives an improvement compared to uniform meshes. We expect even faster convergence when hanging nodes are allowed which avoid unnecessary refinement.

REFERENCES


FIGURE 4. Energy norm $e$ of the Galerkin error and the error estimator $\eta$ for Example 2. 0 indicates uniform refinement, + indicates adaptive refinement.


CIRCLE APPROXIMATION BY QUARTIC $G^2$ SPLINE USING ALTERNATION OF ERROR FUNCTION

SOO WON KIM$^1$ AND YOUNG JOON AHN$^{1\dagger}$

$^1$DEPARTMENT OF MATHEMATICS EDUCATION, CHOSUN UNIVERSITY, SOUTH KOREA

ABSTRACT. In this paper we present a method of circular arc approximation by quartic Bézier curve. Our quartic approximation method has a smaller error than previous quartic approximation methods due to the alternation of the error function of our quartic approximation. Our method yields a closed form of error so that subdivision algorithm is available, and curvature-continuous quartic spline under the subdivision of circular arc with equal-length until error is less than tolerance. We illustrate our method by some numerical examples.

1. INTRODUCTION

Circular arc and conic section have been widely used in CAD/CAM or Computer Vision. But these curves cannot be expressed by polynomial curve. Thus circle approximation and conic approximation by spline curve are important tasks in CAGD(Computer Aided Geometric Design) or Geometric Modeling. In recent twenty years a lot of methods of circle approximation and conic approximation by Bézier curve or spline have been presented.

The methods of Circle and Conic approximation by quadratic Bézier curve are simple and easy to calculate error. Mørken[13] presented the best approximation method of the circular arc by quadratic Bézier curve. Lee et al.[11] introduced some approximation methods of the circular arc by quadratic Bézier curves to obtain the offset approximation of planar spline curve using convolution of the quadratic approximant and the planar spline curve. Floater[6] showed that the quadratic approximant of conic section is curvature continuous under the subdivision of shoulder point of conic, and presented the upper bound of the Hausdorff distance between the conic section and the quadratic approximant.

The methods of circle and conic approximation by spline of odd degree are as follows. Dokken et al.[3] and Goldapp[8] proposed the best $G^k$ cubic approximations of circular arc for $k = 0, 1, 2$. Floater[7] presented a great approximation method of conic section by spline of odd degree $n$ having approximation order $2n$, which is the optimal approximation order. Fang[4, 5] gave $G^k$ quintic approximation of circular arc and conic section for $k \geq 2$.

Received by the editors April 3 2013; Accepted June 4 2013.

2000 Mathematics Subject Classification. 41A05, 41A15, 65D05, 65D07, 65D17, 68U07.

Key words and phrases. circular arc; quartic Bézier curve; spline; Hausdorff distance; approximation order; geometric contact.

$\dagger$ Corresponding author. ahn@chosun.ac.kr.

This study was supported by research funds from Chosun University, 2013.

In §2, previous methods for circle approximation by quartic spline are introduced. In §3, our quartic approximation of the circular arc is presented and the closed form of the Hausdorff distance between the circular arc and the quartic approximation curve is obtained. We have the subdivision algorithm and some numerical examples in §4, and summary our results in §5.

2. PRELIMINARIES FOR QUARTIC APPROXIMATION OF THE CIRCULAR ARC

In this section we propose the quartic Bézier approximation of the circular arc with angle \(0 < \alpha \leq 2\pi\) and radius 1. The unit circular arc \(c : [0, \alpha] \rightarrow \mathbb{R}^2\) can be parametrized by

\[
c(\theta) := (\cos \theta, \sin \theta), \quad 0 \leq \theta \leq \alpha,
\]
as shown in Figure 1. Let \(B^4_t(i)\) be the quartic Bernstein polynomial

\[
B^4_t(i) = \frac{4!}{i!(4-i)!} t^i (1-t)^{4-i}.
\]
The quartic Bézier approximation curve \( b : [0, 1] \rightarrow \mathbb{R}^2 \) of the circular arc \( c(\theta) \) is given by

\[
b(t) := (x(t), y(t)) := \sum_{i=0}^{4} b_i B_i^4(t)
\] (2.1)

with its control points \( b_i := (x_i, y_i), 0 \leq i \leq 4 \)

\[
b_0 = (1, 0), \quad b_1 = (1, u), \quad b_2 = v(\cos \frac{\alpha}{2}, \sin \frac{\alpha}{2})
\]

\[
b_3 = (\cos \alpha, \sin \alpha) + u(\sin \alpha, -\cos \alpha), \quad b_4 = (\cos \alpha, \sin \alpha)
\] (2.2)

so that the quartic approximation \( b(t) \) is a \( G^1 \) endpoint interpolation of the circular arc \( c(\theta) \) for \( u > 0 \).

For the circular arc \( c \), the Hausdorff distance \( d_H(b, c) \) between two curves \( c \) and \( b \) is

\[
d_H(b, c) = \max_{t \in [0, 1]} |\sqrt{x^2(t) + y^2(t)} - 1|
\]

Ahn and Kim[2] used the error function \( \psi(t) \) by

\[
\psi(t) := x^2(t) + y^2(t) - 1.
\] (2.3)

to calculate \( d_H(b, c) = \sqrt{\| \psi(t) \|_\infty + 1} - 1 \) for nonnegative \( \psi \), where the uniform norm of \( \psi \) on \([0, 1]\) is denoted by

\[
\| \psi(t) \|_\infty := \max_{t \in [0, 1]} |\psi(t)|.
\]

If \( \psi(t) \geq -1 \) for all \( t \in [0, 1] \), then

\[
d_H(b, c) = \max\{\sqrt{\max_{t \in [0, 1]} \psi(t) + 1} - 1, 1 - \sqrt{1 - \min_{t \in [0, 1]} \psi(t)}\}.
\]

Equations (2.1)-(2.3) yields \( \psi(t) = 4t^2(t - 1)^2\zeta(t) \) where

\[
\zeta(t) = \left(-9v^2 - (24u \sin \frac{\alpha}{2} + 18 \cos \frac{\alpha}{2})v - 16u^2 \sin^2 \frac{\alpha}{2} - 12u \sin \alpha - 9 \cos^2 \frac{\alpha}{2}\right)(t - \frac{1}{2})^4
\]

\[
+ \left(\frac{9}{2}v^2 - 9 \cos \frac{\alpha}{2}v - 8u^2 \sin^2 \frac{\alpha}{2} - 8u \sin \alpha + 4 - \frac{17}{2} \cos^2 \frac{\alpha}{2}\right)(t - \frac{1}{2})^2
\]

\[
- \frac{9}{16} v^2 - \frac{3}{2} u \sin \frac{\alpha}{2} + \frac{15}{8} \cos \frac{\alpha}{2}v - u^2 \sin^2 \frac{\alpha}{2} - \frac{5}{4} u \sin \alpha + 4 - \frac{25}{16} \cos^2 \frac{\alpha}{2} \right).
\] (2.4)

As shown in Table 1, Ahn and Kim[2] proposed the approximations \( b_{u3} \) whose error function \( \psi \) has quadruple-zero at both end points, \( t = 0, 1 \), and \( b_{\mu2} \) triple-zero at both end points and double-zero at midpoint, \( t = \frac{1}{2} \). Our approximation curve \( b(t) \) has contact with the circular arc \( c \) at the midpoint. Solving \( \zeta(1/2) = 0 \), we have two solutions

\[
v_i = -\frac{5}{3} \cos(\alpha/2) - \frac{4}{3} u \sin(\alpha/2) + (-1)^i \frac{8}{3}
\]
After choosing $v$ and $b$ than tolerance. (Refer to \[10, 7\].) arc $c = 1$

| Table 1. The circular arc approximation by quartic Bézier approximations $b_{u_3}$, $b_{μ_2}$, $b$, and $\tilde{b}$ are proposed by Ahn and Kim[2], Kim and Ahn[10], and Liu[12]. They have different zeros of error function $ψ(t)$. At the last line, $b_{u_2,2}$ is presented by our method.

If $v > v_i$, then $ζ(t) = 4(t - 1/2)^2η_i(t)$ where

$η_1(t) = (16(cos^{α/2}) + 1)(2sin^{α/2}u^2 + 2sin^{α/2}u + 1 + cos^{α/2})t(t - 1) + 4u^2 - 2u sin α - (cos^{α/2} + 1)(5sin^{α/2} + 3)$

$η_2(t) = 2^6 sin^{α/4}u cos^{α/4} - sin^{α/4}2^2t(t - 1) + 4u^2 - 2u sin α - 4sin^{α/4}2^2(1 - 5sin^{α/4})$.

After choosing $v = v_i$ the quartic approximation $b(t)$ is depend only on one parameter $u$. Kim and Ahn[10] proposed the $G^2$ quartic approximation $b$ with $η_2(t)$ having double-zero at the midpoint, and Liu et al.[12] $\tilde{b}$ with $η_2(t)$ having zero at $t = 1/3, α/4$.

If $u > 0$, then the quartic Bézier curve $b_u(t)$ is an $G^1$ endpoint interpolation of the circular arc $c(θ)$. Furthermore, by symmetry, $b_u(t)$ have the same curvature at both end points $b_u(0)$ and $b_u(1)$. Hence if $u > 0$, the method of quartic approximation $b_u(t)$ of circular arc yields $G^2$ quartic spline under the subdivision of equal-length of circular arc until the error is less than tolerance. (Refer to [10, 7].)

3. $G^2$ QUARTIC SPLINE APPROXIMATION OF CIRCULAR WITH ALTERNATION OF ERROR FUNCTION

Lemma 3.1. The eighth-degree monic polynomial $f(t) = t^2(t - 1)^2(t - 1/2)^2(t - a)(t - (1 - a))$, $t ∈ [0, 1]$ satisfy

$‖f‖_∞ = \max_{t ∈ [0, 1]} f(t) = - \min_{t ∈ [0, 1]} f(t)$

if $a = 1/2 ± 1/6\sqrt{6 - 4\sqrt{3} + 2\sqrt{6}\sqrt{\sqrt{3} - 1}}$.
Figure 2. The eighth-degree monic polynomial $f(t) = t^2(t - 1)^2(t - 1/2)^2(t - a)(t - (1 - a))$, for $a = \frac{1}{2} \pm \frac{1}{6} \sqrt{6 - 4\sqrt{3} + 2\sqrt{6\sqrt{3}} - 1}$. $f(-b) = \max f(x) = -\min f(x) = \|f\|_\infty$.  

Proof. Solving the equation $f'(t) = 0$, the solutions are $t = (4 \pm \sqrt{12a^2 - 20a + 9} \pm \sqrt{12a^2 - 4a + 1})/8$, 0, 1, 1/2. Put $b_i = (4 - \sqrt{12a^2 - 20a + 9} + (-1)^i \sqrt{12a^2 - 4a + 1})/8$, $i = 1, 2$. The polynomial $f(t)$ has the maximum $f(b_1) = f(1 - b_1)$ and the minimum $f(b_2) = f(1 - b_2)$ on $[0, 1]$, if $0 < a < \frac{1}{2}$ or $\frac{1}{2} < a < 1$, as shown in Figure 2. Solving the equation $f(b_1) = -f(b_2)$ with respect to $a$, we obtain two real solution $\frac{1}{2} \pm \frac{1}{6} \sqrt{6 - 4\sqrt{3} + 2\sqrt{6\sqrt{3}} - 1}$. Any of them satisfies Equation (3.1). □ 

Numerically, we have $a \approx 0.199$ and $\|f\|_\infty = f(b_1) \approx 9.01 \times 10^{-5}$, as shown in Figure 2. 

Solving equation $\eta_1(a) = 0$ with respect to $u$, there are four solutions $u_{i,j}$, $i, j = 1, 2$. Since for $u = u_{1,j}$, $j = 1, 2$, the quadratic polynomial $\eta_1(t)$ is 

$$\eta_1(t) = (16(cos \frac{\alpha}{2} + 1))(2 \sin^2 \frac{\alpha}{4} u^2 + 2 \sin \frac{\alpha}{2} u + 1 + \cos \frac{\alpha}{2})(t - a)(t - (1 - a))$$ 

and $\|\psi\|_\infty = 2^{10} f(b_1) + O(a^2)$, the approximation $b_{u_{1,j}}$, $j = 1, 2$ has approximation order zero, and so $b_{u_{1,j}}$ could not be a good approximation. Now, we consider two approximations $b_{u_{2,j}}$, $j = 1$ and 2 and compare their errors. The two parameters are $u_{2,j} = (\xi_1 + (-1)^j \sqrt{\xi_2})/2\xi$ where 

$$\xi = 4a(a - 1) \sin^2 \frac{\alpha}{2} + 1$$ 

$$\xi_1 = \sin \frac{\alpha}{2} (\cos \frac{\alpha}{2} + 16a(a - 1) \sin^2 \frac{\alpha}{4})$$ 

$$\xi_2 = 16 \sin^6 \frac{\alpha}{4} (2 - (2a - 1)^2 \sin^2 \frac{\alpha}{4}).$$
PROPOSITION 3.2. \( b_{u_{2,2}} \) is better approximation than \( b_{u_{2,1}} \).

**Proof.** If \( u = u_{2,j} \), then \( \eta_2(t) = \eta_{2,j}(t) \) is a quadratic polynomial with zero at \( t = a, 1 - a \). Thus

\[
\eta_{2,j}(t) = 2^6 \sin^2 \frac{\alpha}{4} (u_{2,j} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 (t - a)(t - (1 - a))
\]

and we have

\[
(u_{2,1} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 - (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 = \frac{\sin \frac{\alpha}{2} \sin^2 \frac{\alpha}{4} (2 \cos^2 \frac{\alpha}{4} + 1) \sqrt{\xi_2}}{\xi^2} \geq 0.
\]

Hence \( b_{u_{2,2}} \) is better approximation than \( b_{u_{2,1}} \). \( \square \)

PROPOSITION 3.3. For \( u = u_{2,2} \), the error function is given by

\[
\psi(t) = 2^{10} \sin^2 \frac{\alpha}{4} (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 t^2 (t - 1)^2 (t - \frac{1}{2})^2 (t - a)(t - (1 - a)) \quad (3.3)
\]

and

\[
\| \psi \|_{\infty} = 2^{10} \sin^2 \frac{\alpha}{4} (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 f(b_1) = \frac{(2\sqrt{3} - 3)^2}{2^6} f(b_1) \alpha^8 + \mathcal{O}(\alpha^{10}). \quad (3.4)
\]

**Proof.** Equation (3.3) follows from the equations \( \psi(t) = 4t^2(t - 1)^2 \zeta(t), \zeta(t) = 4(t - 1/2)^2 \eta_2(t), \) and \( \eta_2(t) = 2^6 \sin^2 \frac{\alpha}{4} (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 (t - a)(t - (1 - a)) \). Since \( u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4} = \frac{2\sqrt{3} - 3}{2^6} \alpha^3 + \mathcal{O}(\alpha^5) \), we obtain Equation (3.4). \( \square \)

We present a closed form of \( d_H(q, b) \) and show that the approximation order of \( b_{u_{2,2}} \) is eight, as follows.

PROPOSITION 3.4. The Hausdorff distance between the quartic approximation \( b \) and the circular arc \( c \) is given by

\[
d_H(b, c) = 1 - \sqrt{1 - 2^{10} \sin^2 \frac{\alpha}{4} (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 f(b_1)} \quad (3.5)
\]

and its asymptotic behavior is

\[
d_H(b, c) = \frac{(2\sqrt{3} - 3)^2}{2^7} f(b_1) \alpha^8 + \mathcal{O}(\alpha^{10}). \quad (3.6)
\]

**Proof.** Since the range of \( \psi(t) \) is \([-\|\psi\|_{\infty}, \|\psi\|_{\infty}] \), that of \( \sqrt{\psi(t) + 1 - 1} \) is \([\sqrt{1 - \|\psi\|_{\infty}} - 1, \sqrt{1 + \|\psi\|_{\infty}} - 1] \). Thus the Hausdorff distance \( d_H(b, c) \) between the circular arc \( c \) and the approximation \( b \) is

\[
d_H(b, c) = \max\{|\sqrt{1 - \|\psi\|_{\infty}} - 1|, |\sqrt{1 + \|\psi\|_{\infty}} - 1|\}. \quad (3.7)
\]

Since \(|\sqrt{1 - s} - 1| \geq |\sqrt{1 + s} - 1| \) for any real number \( 0 < s < 1 \), we have

\[
d_H(b, c) = 1 - \sqrt{1 - 2^{10} \sin^2 \frac{\alpha}{4} (u_{2,2} \cos \frac{\alpha}{4} - \sin \frac{\alpha}{4})^2 f(b_1)}. \quad (3.8)
\]

Equation (3.6) follows from Equations (3.4)-(3.5). \( \square \)
CIRCLE APPROXIMATION BY QUARTIC $G^2$ SPLINE

Figure 3. (a) The unit circle(red) and our quartic Bézier approximation(blue). The dash-lines(blue) are control polygon. The Hausdorff distance is $4.72 \times 10^{-2}$. (b) The circle(red) of radius 10 and our quartic $G^2$ spline approximation(blue) using four segment of quartic Bézier curve. At each junction points the spline curve is curvature-continuous. The Hausdorff distance is $7.60 \times 10^{-6}$.

The Hausdorff distance $d_H(b, c)$ between the unit circular arc $c$ of angle $\alpha$ and our quartic approximation $b_{u2,2}$ in Equation (3.5) is now depend only on $\alpha$. Thus we denote the Hausdorff distance $d_H(b, c)$ by $\varepsilon(\alpha)$.

4. ALGORITHM AND NUMERICAL EXAMPLES

Using our circle approximation method by quartic Bézier curve, we present a subdivision algorithm for the quartic $G^2$ spline approximation of the circular arc within the given tolerance as follows.

<table>
<thead>
<tr>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Input: the radius $r$ and angle $\phi$ of circular arc, and tolerance $TOL$.</td>
</tr>
<tr>
<td>2. Find the smallest positive integer $k$ satisfying $r\varepsilon(\frac{\phi}{k}) &lt; TOL$.</td>
</tr>
<tr>
<td>3. Calculate the control points $b_i$ of the quartic approximation $b(t)$ for unit circle using Equations (2.1)-(3.2) for $\alpha = \frac{\phi}{k}$.</td>
</tr>
<tr>
<td>4. Output: the control points $T^jrb_i$, $i = 0, \cdots, 4$, and $j = 0, \cdots, k - 1$ of quartic spline approximation, where $T$ is the rotation transformation of angle $\alpha$.</td>
</tr>
</tbody>
</table>
Figure 3(a) shows the unit full circle (red color) and its quartic Bézier approximation (blue color) using our method. The quartic Bézier curve have the same curvature at the meet point of start-point and end-point. The control polygon is plotted by dash-lines. The Hausdorff distance between the circle and the quartic Bézier curve is $4.72 \times 10^{-2}$.

If the radius of circle is given by $r = 10$ and the error tolerance $TOL$ is $10^{-5}$, the algorithm yields $k = 4$ and the quartic $G^2$ spline as shown in Figure 3(b). The Hausdorff distance between the unit circle and quartic spline is $7.60 \times 10^{-6} < TOL$. The control points $b_i$ of quartic Bézier approximation of unit circle is obtained from Equations (2.1)-(3.2) and $\alpha = \frac{\pi}{2}$. All control points of the quartic spline approximation are also obtained from $T^3r b_i$, $j = 0, \cdots, 3$ using the rotation transformation $T$ of angle $\frac{\pi}{2}$.

5. COMMENTS AND FUTURE WORK

In this paper we presented a method of circle approximation by quartic spline curve. Our circle approximation by quartic $G^2$ spline has some merits. Our method has smaller error than other previous methods of circle approximation by quartic Bézier curve. Also our method yields a closed form of error and the curvature-continuous quartic spline as previous methods. As a future work, we will extend the circle approximation by quartic spline to the conic approximation and surface approximation such as ellipsoid and torus.

REFERENCES


ANALYTIC APPROACH FOR THE STUDY OF AIR AND/OR LIQUID FILLED GEOMEMBRANE TUBE SECTIONS ON A HORIZONTAL

YOON-RAK CHOI

SCHOOL OF NAVAL ARCHITECTURE AND OCEAN ENGINEERING, UNIVERSITY OF ULSAN, KOREA
E-mail address: yrchoi@ulsan.ac.kr

ABSTRACT. This study considers an air and liquid-filled geomembrane tube section resting on a horizontal foundation. All quantities are normalized to obtain geometrically similar solutions in the static equilibrium condition. Analytic solutions are expressed in closed form. The solution for the air or liquid-filled tube section is derived systematically as an extreme case of the air and liquid-filled tube section. The validity of these solutions is confirmed by comparing to previous study, and some results are shown for the characteristic parameters and shapes of air and/or liquid-filled cases. Using the result of present study, one can estimate the shape and characteristic parameters of a tube section without numerical integrations or iterations.

1. INTRODUCTION

Geomembrane structures are composed of thin flexible sheets, and can be inflated with air, liquid, grain, slurry, or sand. Geomembrane structures are widely utilized in various applications because they are inexpensive and easy to install and dismantle. These structures can be used to store and transport fluids. In the field of civil engineering, geomembrane structures are employed for purposes such as increasing the height of existing dams or spillways, temporary dykes for flood control, breakwaters in beaches, foundations of emergency bridges, and coastal erosion prevention.

Inflated geomembrane structures are generally tube-shaped, so structural analysis is commonly performed about the tube section. Analytic approaches have been focused on the analysis of the structures resting on the horizontal foundations. Demiray and Levinson [1] formulated the governing equations according to the equilibrium of force and the geometric boundary conditions, and obtained the analytic solution expressed in terms of elliptic integrals. Wang and Watson [2] studied the shapes of the structures in lightly or heavily pressurized conditions. Namias [3] found the approximate solutions, which covered a wider range of the low and high pressures than that of Wang and Watson [2]. Plaut and Suherman [4] reviewed the analytic and approximate solutions in terms of the pressure at the bottom of the tube and at the top. Meanwhile, Plaut and Cotton [5] numerically studied the statics...
and vibration of the tube section of air-filled structures using the shooting method. Antman and Schagerl [6] analyzed extensible elastic membranes containing incompressible liquids and compressible gases, and inextensible membrane as a limit case of extensible membrane. However, the authors neglected the weight of membranes. Recently, Ghavanloo and Daneshmand [7] carried out semi-analytic approaches for the air and/or liquid-filled tube sections (that is, the solutions were expressed in integral form). Choi [8] derived the analytic solution in closed form for an air-filled, heavy membrane tube section on an incline.

In the present study, an analytic approach is applied to an air and liquid-filled geomembrane tube resting on a horizontal foundation. Based on the semi-analytic approach by Ghavanloo and Daneshmand [7], the integral terms of their results are evaluated analytically. The problem is divided into two parts: liquid-filled and air-filled. Each part is solved analytically and then matched to obtain the complete solution. The analytic solution for the air or liquid-filled tube is derived systematically using the solution of the air and liquid-filled tube. A comparison with a semi-analytic approach is performed, and characteristic parameters and shapes are shown for the air and/or liquid-filled tube sections.

![Cross section of the air and liquid-filled geomembrane tube](image)

**FIGURE 1.** Cross section of the air and liquid-filled geomembrane tube.

### 2. MATHEMATICAL FORMULATION OF THE PROBLEM

Figure 1 shows the cross section of an air and liquid-filled geomembrane tube resting on a rigid horizontal foundation. The geomembrane material is assumed to be inextensible and perfectly flexible. Both the friction between the geomembrane tube and the liquid, and that between the geomembrane tube and the rigid foundation are neglected. The origin of the coordinate system O is located at the right separation point. X and Y are the
SOLUTIONS OF AIR AND/OR LIQUID-FILLED

horizontal and vertical axes, respectively, $\theta$ is the tangential angle with respect to the horizontal, and $S$ is the arc length from the origin. The contact length is $\Xi$; the mass per unit length is $\lambda$; the density of the liquid is $\rho$; the height of the liquid level is $H$; the maximum width is $W$; the transmural pressure difference in the air-filled part is $P_0$; the intersection point of the air, liquid, and tube section is $C$; and the top point of the tube is $Q$. The analysis can be performed in the right half section due to the geometrical symmetry.

From the free-body diagram of Figure 2, the static equilibrium gives

$$
\frac{dT}{dS} = \lambda g \sin \theta, \quad (2.1)
$$

$$
T \frac{d\theta}{dS} = P + \lambda g \cos \theta, \quad (2.2)
$$

where $P$ is the transmural pressure difference. $P$ in the air- and liquid-filled parts is expressed as

$$
P = \begin{cases} 
P_0 + \rho g (H - Y) & \text{if } 0 \leq S \leq S_C, \\
P_0 & \text{if } S_C \leq S \leq \frac{L - \Xi}{2},
\end{cases} \quad (2.3)
$$

where $L$ is the perimeter of the tube section.

To simplify the analysis, dimensional quantities are nondimensionalized using $L$ and $\rho$ as follows:
\[ x = \frac{X}{L}, \quad y = \frac{Y}{L}, \quad s = \frac{S}{L}, \quad \xi = \frac{\Xi}{L}, \quad h = \frac{H}{L}, \]

\[ w = \frac{W}{L}, \quad v = \frac{V}{L^2}, \quad t = \frac{T}{\rho g L}, \quad p = \frac{P_0}{\rho g L}, \quad \mu = \frac{\lambda}{\rho L}, \]

(2.4)

where \( V \) is the sectional area of the tube. In (2.4), \( \mu \) indicates the weight ratio of geomembrane structure to the liquid. Using these quantities and (2.3), (2.1) and (2.2) can be rewritten as follows:

\[ \frac{dt}{ds} = 0, \quad t \frac{d\theta}{ds} = p - h - y \quad \text{if} \quad 0 \leq s \leq s_C, \quad (2.5a,b) \]

\[ \frac{dt}{ds} = \mu \sin \theta, \quad t \frac{d\theta}{ds} = p + \mu \cos \theta \quad \text{if} \quad s_C \leq s \leq \frac{1 - \xi}{2}. \quad (2.6a,b) \]

In (2.5), the effect of the tube weight is neglected in the liquid-filled part because \( \mu \) is generally very small. Hence, it can be noted that the tension is constant in this part.

The geometric relations are

\[ \frac{dx}{ds} = \cos \theta, \quad \frac{dy}{ds} = \sin \theta. \quad (2.7a,b) \]

The same equations can be found in Ghavanloo and Daneshmand [7].

This problem is governed by the system of non-linear differential equations in (2.5)-(2.7). The boundary conditions of this problem at points \( O, C, \) and \( Q \) are as follows:

\[ x = y = \theta = 0 \quad \text{at} \quad s = 0, \quad (2.8) \]

\[ x = x_C, \quad y = h, \quad \theta = \theta_C \quad \text{at} \quad s = s_C, \quad (2.9) \]

\[ x = -\frac{\xi}{2}, \quad \theta = \pi \quad \text{at} \quad s = \frac{1 - \xi}{2}. \quad (2.10) \]

3. ANALYTIC SOLUTIONS

As can be seen in (2.5) and (2.6), this problem can be broken into two problems: the liquid-filled part problem and the air-filled part problem. Therefore, each of the two problems may be analyzed independently. After solving the two problems with unknowns \( x_C, \theta_C, \) and \( s_C \) at point \( C \), matching the two solutions at point \( C \) yields the complete solution.

3.1. Solution of the liquid-filled part. This problem is composed of the governing
equations (2.5) and (2.7) with the boundary conditions (2.8) and (2.9). From (2.5a), tension in this part is constant and denoted as $t_0$. Differentiation of (2.5b) with respect to $s$ yields an integrable equation as follows:

$$\frac{d^2 \theta}{ds^2} = -\frac{\sin \theta}{t_0}. \quad (3.1)$$

Integration of the above equation results in

$$\frac{d \theta}{ds} = \sqrt{\frac{2}{t_0}} \sqrt{a + \cos \theta}, \quad (3.2)$$

where $a$ is defined as follows:

$$a = \frac{1}{2} \frac{p^2}{t_0} - \cos \theta_c = \frac{1}{2} \left( \frac{p + h}{t_0} \right)^2 - 1. \quad (3.3)$$

The relation in (3.3) is provided by the values in (2.5b) at points $O$ and $C$. Using (3.2), analytic integration can be performed with respect to $\theta$ (Gradshteyn and Ryzhik [9]).

$$s = \begin{cases} \sqrt{t_0} \sqrt{\frac{2}{a+1}F\left(\frac{\theta}{2},\sqrt{\frac{2}{a+1}}\right)} & \text{if } a > 1, \\ \sqrt{t_0} F\left(\sin^{-1}\sqrt{\frac{1-\cos \theta}{a+1}},\sqrt{\frac{a+1}{2}}\right) & \text{if } -\cos \theta_c \leq a \leq 1. \end{cases} \quad (3.4)$$

$$x = \sqrt{t_0} \left[ \frac{2}{a+1} (a+1)E\left(\frac{\theta}{2},\sqrt{\frac{2}{a+1}}\right) - aF\left(\frac{\theta}{2},\sqrt{\frac{2}{a+1}}\right) \right] \text{ if } a > 1, \quad (3.5)$$

$$y = \sqrt{2t_0} \left[ \sqrt{a+1} - \sqrt{a + \cos \theta} \right]. \quad (3.6)$$

In the above equations, $F$ and $E$ are the incomplete elliptic integrals of the first and second kind, respectively, which are defined in Gradshteyn and Ryzhik [9], and can be evaluated without numerical integrations (Press et al. [10]).

The sectional area of this part ($v^i$) is derived by integrating (2.5b).

$$v^i = 2t_0 \sin \theta_c + h \xi - 2px_c. \quad (3.7)$$
3.2. **Solution of the air-filled part.** This problem is composed of the governing equations (2.6) and (2.7) with the boundary conditions (2.9) and (2.10). To inflate the tube, the air pressure should be \( p > \mu \). Elimination of \( s \) from (2.6a) and (2.6b) yields

\[
t = t_0 \frac{p + \mu \cos \theta_c}{p + \mu \cos \theta}.
\]  

(3.8)

Using this result and the boundary condition (2.9), equations (2.7a), (2.7b), and (2.6b) can be integrated and expressed as functions of \( \theta \) \((\text{Gradshteyn and Ryzhik [9]})\).

\[
x = x_c + t_0 \frac{p + \mu \cos \theta_c}{p^2 - \mu^2} \left[ \frac{p \sin \theta}{p + \mu \cos \theta} - \frac{p \sin \theta_c}{p + \mu \cos \theta_c} - \frac{2\mu}{\sqrt{p^2 - \mu^2}} \left\{ \tan^{-1} \left( \sqrt{\frac{p - \mu}{p + \mu}} \tan \frac{\theta}{2} \right) - \tan^{-1} \left( \sqrt{\frac{p - \mu}{p + \mu}} \frac{\theta_c}{2} \right) \right\} \right].
\]  

(3.9)

\[
y = h + t_0 \frac{\cos \theta_c - \cos \theta}{p + \mu \cos \theta}.
\]  

(3.10)

\[
s = s_c - t_0 \frac{p + \mu \cos \theta_c}{p^2 - \mu^2} \left[ \frac{\mu \sin \theta}{p + \mu \cos \theta} - \frac{\mu \sin \theta_c}{p + \mu \cos \theta_c} - \frac{2p}{\sqrt{p^2 - \mu^2}} \left\{ \tan^{-1} \left( \sqrt{\frac{p - \mu}{p + \mu}} \tan \frac{\theta}{2} \right) - \tan^{-1} \left( \sqrt{\frac{p - \mu}{p + \mu}} \frac{\theta_c}{2} \right) \right\} \right].
\]  

(3.11)

The sectional area of this part \((v_a)\) is calculated as follows:

\[
v_a = -2 \int_{x_c}^{x_c + \xi} y dx - 2h(x_c + \xi) = -2t_0^2(p + \mu \cos \theta_c) \int_{\theta_c}^{\theta} \left\{ \frac{(\cos \theta_c - \cos \theta) \cos \theta}{(p + \mu \cos \theta)} \right\} d\theta.
\]  

(3.12)

The last integral term in (3.12) can be evaluated analytically (after some manipulations of the integration by parts). The result is then given as follows:

\[
v_a = 2t_0^2(p + \mu \cos \theta_c) \frac{p^2 + 3p \mu \cos \theta_c + 2 \mu^2}{(p^2 - \mu^2)^2} \left[ \frac{\pi}{2} - \tan^{-1} \left( \frac{p - \mu}{p + \mu} \frac{\theta_c}{2} \right) \right] + t_0^2 \frac{3p \mu + (p^2 + 2 \mu^2) \cos \theta_c}{(p^2 - \mu^2)^2} \sin \theta_c.
\]  

(3.13)

3.3. **Matching of two solutions.** For the liquid-filled part, the application of boundary conditions (2.9) to (3.4) and (3.5) results in the following relation.
On the other hand, the application of boundary conditions (2.10) to (3.9) and (3.11) results in the following relation for the air-filled part.

\[
\frac{1}{2} = s_c - x_c + t_0 \frac{p + \mu \cos \theta_c}{p - \mu} \left[ \frac{\sin \theta_c}{p + \mu \cos \theta_c} + \frac{2}{\sqrt{p^2 - \mu^2}} \left( \frac{\pi}{2} - \tan^{-1} \left( \frac{p - \mu \tan \frac{\theta_c}{2}}{\sqrt{p^2 - \mu^2}} \right) \right) \right].
\]  
(3.15)

From (3.3), \( t_0 \) can be expressed as follows:

\[
t_0 = \frac{h(2p - h)}{2(1 - \cos \theta_C)}.
\]  
(3.16)

Because \( t_0 \) and \( a \) are expressed in terms of the parameters \( p \), \( h \), and \( \theta_C \) as can be seen in (3.3) and (3.16), elimination of \( s_c - x_c \) from (3.14) and (3.15) yields an equation that represents the relation among those parameters. For given values of \( p \) and \( h \), one can find \( \theta_C \) as the root of this combined equation using various numerical schemes, such as the bisection method. The monotonic property of the combined equation to \( \theta_C \) confirms the uniqueness of the solution. The upper bound of \( h \) for a given \( p \) corresponds to the \( h \) of the fully liquid-filled case, which is described in the next section. After determining \( \theta_C \), the shape and all characteristic parameters can be determined for the given \( p \) and \( h \).

\[
\begin{align*}
x_c &= x(\theta = \theta_C), & x_{\text{max}} &= x(\theta = \frac{\pi}{2}), & y_{\text{max}} &= y(\theta = \pi), \\
-\frac{\xi}{2} &= x(\theta = \pi), & w &= \xi + 2x_{\text{max}}, & t_{\text{max}} &= t(\theta = \pi).
\end{align*}
\]  
(3.17)

All solutions solved in this study are geometrically similar solutions, because all quantities are normalized by the perimeter of the tube section and the density of the contained liquid.

4. TWO EXTREME CASES: LIQUID-FILLED AND AIR-FILLED

Using the solution of the air and liquid-filled geomembrane tube, the solution for the
liquid or air-filled tube can be derived.

4.1. **Liquid-filled case.** \( \theta_c = \pi \), \( h = y_{\text{max}} \), \( x_c = -\xi/2 \), and \( s_c = (1-\xi)/2 \) in the solutions of the liquid-filled part, and the tension is constant \( (t_0) \). In this case, \( p \) is the top pressure in the tube section. Noting (3.3), \( a \) is defined as follows, and should be \( a > 1 \).

\[
a = \frac{1}{2} \frac{p^2}{t_0} + 1 - \frac{1}{2} \frac{(p+h)^2}{t_0} - 1. \tag{4.1}
\]

Equation (3.16) is then rewritten as

\[
t_0 = \frac{1}{4} h(2p+h). \tag{4.2}
\]

Substitution of (4.1) and (4.2) into (3.14) yields

\[
\frac{1}{2} = (p+h) \left[ F \left( \frac{\pi}{2}, \sqrt{1-\left(\frac{p}{p+h}\right)^2} \right) - E \left( \frac{\pi}{2}, \sqrt{1-\left(\frac{p}{p+h}\right)^2} \right) \right]. \tag{4.3}
\]

\( h \) is determined as the root of (4.3) for a given \( p \). Therefore, \( h \) is dependent on \( p \). This value provides the upper bound of \( h \) in the air and liquid-filled case for the given \( p \).

After determining \( h \), the shape and characteristic parameters are calculated by (3.4)-(3.6) and (3.17). Equation (3.15) is automatically satisfied. These results are identical to those of previous studies (Demiray and Levinson [1] and Namias [3]).

From (3.7), the sectional area is

\[
v = v_i = \xi(h + p). \tag{4.4}
\]

4.2. **Air-filled case.** This case corresponds to \( \theta_c = 0 \), \( h = 0 \), \( x_c = 0 \), and \( s_c = 0 \) in the solutions of the air-filled part. In this case, the liquid density \( \rho \) is meaningless. Thus, new nondimensional parameters for the air pressure and tension are introduced as follows:

\[
\tilde{p} = \frac{p}{\mu} = \frac{P_0}{\lambda g}, \quad \tilde{t} = \frac{t}{\mu} = \frac{T}{\lambda g L}, \tag{4.5}
\]

where \( \tilde{p} > 1 \). Using the above definition, (3.15) yields

\[
\frac{1}{2} = \tilde{t}_0 \frac{\pi (\tilde{p}+1)}{(\tilde{p}-1)\sqrt{\tilde{p}^2-1}}. \tag{4.6}
\]
Hence, $\tilde{\varrho}_0$ is determined for a given $\tilde{\varrho}$ as follows:

$$
\tilde{\varrho}_0 = \frac{\tilde{\varrho} - 1}{2\pi} \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}}.
$$

(4.7)

Equation (3.14) is automatically satisfied. Using (4.5) and (4.7), the solutions of the air-filled part (3.8)-(3.11) yield the following results.

$$
\tilde{\varrho} = \frac{\tilde{\varrho} - 1}{2\pi} \sqrt{\frac{\tilde{\varrho}^2 - 1}{\tilde{\varrho} + \cos \theta}}.
$$

(4.8)

$$
x = \frac{1}{2\pi} \left[ \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}} \tilde{\varrho} \sin \theta - \frac{2}{\tilde{\varrho} + 1} \tan^{-1}\left( \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}} \tan \frac{\theta}{2} \right) \right].
$$

(4.9)

$$
y = \frac{\tilde{\varrho} - 1}{2\pi} \sqrt{\frac{1 - \cos \theta}{\tilde{\varrho} + \cos \theta}}.
$$

(4.10)

$$
s = \frac{1}{2\pi} \left[ \frac{2\tilde{\varrho}}{\tilde{\varrho} + 1} \tan^{-1}\left( \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}} \tan \frac{\theta}{2} \right) - \frac{\tilde{\varrho} - 1}{\tilde{\varrho} + \cos \theta} \sin \theta \right].
$$

(4.11)

The characteristic parameters are also derived as follows:

$$
\tilde{t}_{\text{max}} = \sqrt{\frac{\tilde{\varrho}^2 - 1}{2\pi}},
$$

(4.12)

$$
\xi = \frac{1}{\tilde{\varrho} + 1},
$$

(4.13)

$$
x_{\text{max}} = \frac{1}{2\pi} \left[ \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}} - \frac{2}{\tilde{\varrho} + 1} \tan^{-1}\left( \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}} \right) \right],
$$

(4.14)

$$
y_{\text{max}} = \frac{1}{\pi} \sqrt{\frac{\tilde{\varrho} - 1}{\tilde{\varrho} + 1}}.
$$

(4.15)

These closed-form solutions correspond to those of Choi [8] in the case of a horizontal foundation.

From (3.13), the sectional area is
5. NUMERICAL RESULTS AND DISCUSSION

Figure 3 shows the dependence of $\theta_c$ on the air pressure $p$ for several liquid levels $h$. To compare the results of present study to those of Ghavanloo and Daneshmand [7], the parameter $\mu$ is assumed hereafter to be 0.0035. $\theta_c = 180^\circ$ indicates the liquid-filled case. $\theta_c$ approaches the value of the perfectly circular shape for the given $h$ as $p$ increases. There is a minimum pressure required to keep a certain liquid level, which corresponds to the liquid-filled case. This minimum pressure increases as $h$ increases. The computing burden to find $\theta_c$ using the bisection method was negligibly small.

![Figure 3. $\theta_c$ related to the nondimensional air pressure $p$.](image)

To ensure the validity of the analytic solutions of this study, some characteristic parameters are compared with those of the semi-analytic approach by Ghavanloo and Daneshmand [7]. As shown in Table 1, close agreement is achieved between the two results. The solution of the semi-analytic approach contains the integral terms which should be computed numerically, while that of present study does not contain any integral terms.
### Table 1. Some characteristic parameters of the air and liquid-filled tube sections.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$h$</th>
<th>$\zeta$</th>
<th>$\theta_C$</th>
<th>$t_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.10</td>
<td>0.0785</td>
<td>1.3252</td>
<td>0.0396</td>
</tr>
<tr>
<td>0.25</td>
<td>0.15</td>
<td>0.1156</td>
<td>1.7585</td>
<td>0.0411</td>
</tr>
<tr>
<td>0.25</td>
<td>0.175</td>
<td>0.1306</td>
<td>1.9810</td>
<td>0.0422</td>
</tr>
<tr>
<td>0.25</td>
<td>0.20</td>
<td>0.1422</td>
<td>2.2148</td>
<td>0.0437</td>
</tr>
<tr>
<td>0.05</td>
<td>0.10</td>
<td>0.2415</td>
<td>1.7821</td>
<td>0.0083</td>
</tr>
<tr>
<td>0.10</td>
<td>0.10</td>
<td>0.1585</td>
<td>1.5111</td>
<td>0.0160</td>
</tr>
<tr>
<td>0.15</td>
<td>0.10</td>
<td>0.1182</td>
<td>1.4101</td>
<td>0.0238</td>
</tr>
<tr>
<td>0.20</td>
<td>0.10</td>
<td>0.0943</td>
<td>1.3575</td>
<td>0.0317</td>
</tr>
</tbody>
</table>

Figures 4(a) and 4(b) show the characteristic parameters for $h = 0.05$ and $0.2$, respectively. As expected, all values approach those of the circle for the given $h$ as $p$ increases. The approaches to those limit values are slower for larger values of $h$ because gravity tends to flatten more for higher liquid level. The values for the minimum air pressure for the given $h$ (that is, the pressure for $h = y_{\max}$) correspond to those of the liquid-filled case. Figures 5(a) and 5(b) depict the shapes for $h = 0.05$ and $0.2$, respectively. The dashed lines in the figures represent liquid levels $h$. The shapes of the air-filled part look like circular arcs because the parameter $\mu = 0.0035$ is very small (that is, very light geomembrane structure).

![Figure 4](image-url)
Figures 6 and 7 plot the characteristic parameters and shapes for the liquid-filled case. All characteristic parameters approach for the parameters of the perfectly folded tube section as $p \to 0$. Asymptotic values for $p \to \infty$ are those of the circle with unit perimeter.
Figures 8 and 9 show the characteristic parameters and shapes for the air-filled case. The characteristic features are similar to those of the liquid-filled case. The shapes are a little wide and flabby at the bottom section compared to those of the liquid-filled case.
6. CONCLUSIONS

The cross section of an air and liquid-filled geomembrane tube resting on a rigid horizontal foundation is analyzed using an analytic approach. Problems are solved separately for the air-filled and liquid-filled parts. The solution of the liquid-filled part is expressed in terms of elliptic integrals, while the solution of the air-filled part is described by trigonometric functions. The matching of two solutions is carried out to obtain the complete solution. The solution for the air or liquid-filled tube section is derived systematically as an extreme case of the air and liquid-filled tube section.

Because all solutions in this study are described in the closed form and are nondimensionalized, it is possible to determine the shapes and characteristic parameters for any size of geomembrane tube section without numerical iterations or integrations required in a direct numerical or semi-analytic approach, respectively. In addition to this, the closed-form solution is more amenable to interpreting physics and parametric study of the problem than a direct numerical or semi-analytic solution.

REFERENCES

[6] S. A. Antman and M. Schagerl, Slumping instabilities of elastic membranes holding liquids and gases,


COMPARISON OF DIFFERENT NUMERICAL SCHEMES FOR THE
CAHN–HILLIARD EQUATION

SEUNGGYU LEE1, CHAEYOUNG LEE1, HYUN GEUN LEE2, AND JUNSEOK KIM1†
1DEPARTMENT OF MATHEMATICS, KOREA UNIVERSITY, SOUTH KOREA
2INSTITUTE OF MATHEMATICAL SCIENCES, EWHA W. UNIVERSITY, SOUTH KOREA
E-mail address: cfdkim@korea.ac.kr

ABSTRACT. The Cahn–Hilliard equation was proposed as a phenomenological model for describing the process of phase separation of a binary alloy. The equation has been applied to many physical applications such as a morphological instability caused by elastic non-equilibrium, image inpainting, two- and three-phase fluid flow, phase separation, flow visualization and the formation of the quantum dots. To solve the Cahn–Hillard equation, many numerical methods have been proposed such as the explicit Euler’s, the implicit Euler’s, the Crank–Nicolson, the semi-implicit Euler’s, the linearly stabilized splitting and the non-linearly stabilized splitting schemes. In this paper, we investigate each scheme in finite-difference schemes by comparing their performances, especially stability and efficiency. Except the explicit Euler’s method, we use the fast solver which is called a multigrid method. Our numerical investigation shows that the linearly stabilized stabilized splitting scheme is not unconditionally gradient stable in time unlike the known result. And the Crank–Nicolson scheme is accurate but unstable in time, whereas the non-linearly stabilized splitting scheme has advantage over other schemes on the time step restriction.

1. INTRODUCTION

The Cahn–Hilliard (CH) equation was originated from a model of the phase separation which is called the spinodal decomposition in a binary alloy [1, 2]. Since the spinodal decomposition is one of few phase transformation models in solids, the equation has been applied to various problems in theoretical and experimental materials science fields such as a morphological instability caused by elastic non-equilibrium [3], image inpainting [4], multiphase fluid flow [5, 6, 7, 8], phase separation [9], flow visualization [10] and the formation of the quantum dots [11]. The CH equation is as follows:

\[
\frac{\partial c(x, t)}{\partial t} = \nabla \cdot \left[ M(c(x, t)) \nabla \mu(c(x, t)) \right], \quad x \in \Omega, \quad 0 < t \leq T, \tag{1.1}
\]

\[
\mu(c(x, t)) = F'(c(x, t)) - \epsilon^2 \Delta c(x, t), \tag{1.2}
\]

Received by the editors June 8 2013; Accepted July 1 2013.
2000 Mathematics Subject Classification. 65M06.
Key words and phrases. Cahn–Hilliard equation, Comparison study, Finite-difference method.
† Corresponding author.
where \( \Omega \subset \mathbb{R}^d \) \((d = 1, 2, 3)\) is a domain. The quantity \( c(x, t) \) is defined to be the difference between the concentrations of the two mixture components (e.g., \((m_1 - m_2)/(m_1 + m_2)\) where \( m_1 \) and \( m_2 \) are the masses of components 1 and 2 in a representative volume \( V \)) and \( \mu(c(x, t)) \) is identified as a chemical potential. The coefficient \( M(c) > 0 \) is a diffusional mobility depending on \( c \). \( F(c) = 0.25(1 - c^2)^2 \) is the Helmholtz free energy which has a double well potential (see Fig. 1) and the small constant \( \epsilon \) is the gradient energy coefficient related to the interfacial energy. The CH equation satisfies the zero Neumann condition or called the no flux boundary condition which is given as below:

\[
\frac{\partial c}{\partial n} = \frac{\partial \mu}{\partial n} = 0 \quad \text{on} \quad \partial \Omega, \tag{1.3}
\]

where \( n \) is the unit normal vector to a boundary of the domain \( \partial \Omega \), so \( \frac{\partial}{\partial n} \) denotes the normal derivative on \( \partial \Omega \). The CH equation arises from the Ginzburg–Landau free energy functional

\[
\mathcal{E}(c) := \int_{\Omega} \left( F(c) + \frac{\epsilon^2}{2} |\nabla c|^2 \right) \, dx. \tag{1.4}
\]

If we differentiate the energy functional \( \mathcal{E} \) and the total mass \( \int_{\Omega} c \, dx \) of mixture components with respect to time \( t \) and apply the no flux boundary condition (1.3), then we get the total energy non-increasing and the total mass conserving properties, i.e.,

\[
\frac{d}{dt} \mathcal{E}(t) \leq 0 \quad \text{and} \quad \frac{d}{dt} \int_{\Omega} c \, dx = 0,
\]

respectively. And further detail derivation of above results is presented in [12].

**FIGURE 1.** A double well potential \( F(c) = 0.25(1 - c^2)^2 \).

There have been development in many numerical algorithms to solving the CH equation such as phase-field [8, 13, 14], immersed boundary [15, 16], volume-of-fluid [17, 18], front-tracking [19, 20], boundary integral [21, 22], immersed interface [23, 24] and level-set [25, 26] methods. As shown in Eqs. (1.1) and (1.2), the system of equation is the fourth-order differential equation in space and it implies that there are some difficulties in numerical analysis for the CH equation.
Many spatial stencils are needed and the time step restriction is stringent, for example, \( \Delta t \sim h^4 \) for using the explicit methods. Moreover, at the lower order spatial derivatives, the nonlinear terms does not guarantee numerical stability. The explicit scheme is simple but less efficient due to severe time step restriction, whereas the implicit scheme is efficient but needs large linear systems of equations to solve. Likewise, each numerical method has its own unique advantages and disadvantages for specific needs. Therefore, comparison of different schemes have been discussed to use adequate schemes for specific problems. In this paper, we focus on six widely used schemes in numerical analysis such as the the explicit Euler’s, the implicit Euler’s, the Crank–Nicolson, the semi-implicit Euler’s, the linearly stabilized splitting and the non-linearly stabilized splitting schemes.

This paper is organized as follows. In Section 2, we describe several numerical schemes and introduce some properties related to the non-linearly stabilized splitting schemes. Numerical results are described in Section 3. Our discussion is presented in Section 4.

2. NUMERICAL ANALYSIS

In this section, we present fully-discrete finite-difference methods for the CH equation with six different schemes. We shall discretize the CH equation in two-dimensional domain \( \Omega = (a, b) \times (c, d) \). One- or three-dimensional discretizations are defined analogously. Let \( N_x \) and \( N_y \) be positive even integers which means a number of space step sizes in \( x \)- and \( y \)-directions, respectively, \( h = (b - a)/N_x = (d - c)/N_y \) be the uniform mesh size to both directions and the discretized domain \( \Omega_h = \{(x_i, y_j) : x_i = (i - 0.5)h, y_j = (j - 0.5)h, 1 \leq i \leq N_x, 1 \leq j \leq N_y\} \) be the set of cell-centered points. Let \( c^n_{ij} \) and \( \mu^n_{ij} \) be approximations of \( c(x_i, y_j, n\Delta t) \) and \( \mu(x_i, y_j, n\Delta t) \), respectively, where \( \Delta t = T/N_t \) is the time step, \( T \) is the final time and \( N_t \) is the total number of time steps. A vector valued phase field is defined as

\[
\mathbf{c}^n = \begin{pmatrix}
c^n_{11} & c^n_{12} & \cdots & c^n_{1N_y} \\
\vdots & \vdots & \ddots & \vdots \\
c^n_{N_x1} & c^n_{N_x2} & \cdots & c^n_{N_xN_y}
\end{pmatrix}.
\]

We first implement the no flux boundary condition (1.3) by requiring that for each \( n \),

\[
c^n_{0j} = c^n_{1j}, \quad c^n_{N_x+1,j} = c^n_{N_x,j}, \quad c^n_{i0} = c^n_{i1}, \quad c^n_{i,N_y+1} = c^n_{iN_y}.
\]

We define the discrete energy functional by

\[
\mathcal{E}^h(\mathbf{c}^n) = h^2 \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F(c^n_{ij}) + \frac{\epsilon^2}{2} \left( \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (c^n_{i+1,j} - c^n_{ij})^2 + \sum_{i=1}^{N_x} \sum_{j=0}^{N_y} (c^n_{i,j+1} - c^n_{ij})^2 \right)
\]

and the discrete Laplacian by the standard five-point stencil

\[
\Delta_d c_{ij} = \frac{c_{i-1,j} + c_{i+1,j} - 4c_{ij} + c_{i,j-1} + c_{i,j+1}}{h^2}.
\]
We also define the discrete infinite norm as
\[ \|c\|_\infty = \max_{1 \leq i \leq N_x, 1 \leq j \leq N_y} |c_{ij}|. \]

For \( M(c) \equiv 1 \) which is independent of the quantity \( c \) for simplicity and \( F'(c) \) which is denoted by \( f(c) \), we consider several numerical schemes of Eqs. (1.1) and (1.2):

1. Explicit Euler’s (EE) scheme [27]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \Delta_d \mu_{ij}^n,
\]
\[
\mu_{ij}^n = f(c_{ij}^n) - \epsilon^2 \Delta_d c_{ij}^n.
\]

2. Implicit Euler’s (IE) scheme [28]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \Delta_d \mu_{ij}^{n+1},
\]
\[
\mu_{ij}^{n+1} = f(c_{ij}^{n+1}) - \epsilon^2 \Delta_d c_{ij}^{n+1}.
\]

3. Crank–Nicolson (CN) scheme [5, 12, 29]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \frac{1}{2} \Delta_d (\mu_{ij}^{n+1} + \mu_{ij}^n),
\]
\[
\mu_{ij}^{n+1} = f(c_{ij}^{n+1}) - \epsilon^2 \Delta_d c_{ij}^{n+1}.
\]

4. Semi-implicit Euler’s (SIE) scheme [30, 31, 32, 33]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \Delta_d \mu_{ij}^{n+1},
\]
\[
\mu_{ij}^{n+1} = f(c_{ij}^n) - \epsilon^2 \Delta_d c_{ij}^{n+1}.
\]

5. Linearly stabilized splitting (LSS) scheme [4, 34]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \Delta_d \mu_{ij}^{n+1},
\]
\[
\mu_{ij}^{n+1} = f(c_{ij}^n) - 2c_{ij}^n + 2c_{ij}^{n+1} - \epsilon^2 \Delta_d c_{ij}^{n+1}.
\]

6. Non-linearly stabilized splitting (NLSS) scheme [35, 36, 37]
\[
\frac{c_{ij}^{n+1} - c_{ij}^n}{\Delta t} = \Delta_d \mu_{ij}^{n+1},
\]
\[
\mu_{ij}^{n+1} = f(c_{ij}^{n+1}) - \epsilon^2 \Delta_d c_{ij}^{n+1} + c_{ij}^{n+1} - c_{ij}^n.
\]

Compared with other general schemes 1–4, LSS and NLSS are known as having larger time step sizes [35]. Note that the NLSS scheme approximates the following viscous Cahn–Hilliard equation with an implicit Euler’s scheme:
\[
c_t = \Delta (f(c) - \epsilon^2 \Delta c + \nu c_t)
\]
where \( \nu \) is a viscosity. To see this, let us rewrite Eq. (2.2) as
\[ \mu^{n+1}_{ij} = f(c^{n+1}_{ij}) - \epsilon^2 \Delta d c^{n+1}_{ij} + \Delta t \frac{c^{n+1}_{ij} - c^n_{ij}}{\Delta t}. \] (2.3)

Therefore, the NLSS scheme approximates the viscous CH equation when \( \nu \) equals \( \Delta t \). When using a large time step, we effectively take a large viscous parameter.

In [35], Eyre proved that if \( c^{n+1} \) is a numerical solution of Eqs. (2.1) and (2.2) with a given \( c^n \), then

\[ \mathcal{E}^h(c^{n+1}) \leq \mathcal{E}^h(c^n). \] (2.4)

Furthermore, in [37], the authors showed the discrete energy decreasing property by using eigenvalues of the Hessian matrix of the energy functional. Using this decreasing property of the discrete total energy functional, we can show the boundedness of the numerical solution of Eqs. (2.1) and (2.2) [38]. If \( c^n \) is a numerical solution satisfying Eq. (2.4), then there exists a constant \( K \), independent of \( n \), such that

\[ \|c^n\|_{\infty} \leq K. \] (2.5)

Suppose Eq. (2.5) is false then there is an element \( c^{nK}_{ij} \) such that \( |c^{nK}_{ij}| > K \), where \( K = \sqrt{1 + 2 \sqrt{\mathcal{E}^h(c^0)/h^2}}. \) Since the total energy is non-increasing, we have \( \mathcal{E}^h(c^0) = h^2 F(K) < h^2 F(|c^{nK}_{ij}|) \leq \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h^2 F(c^{nK}_{ij}) \leq \mathcal{E}^h(c^{nK}) \leq \mathcal{E}^h(c^0). \) This contradiction implies that Eq. (2.5) should be satisfied.

3. NUMERICAL RESULTS

In this section, we perform the following numerical experiments: finding a relation between the \( \epsilon \) value and the width of the transition layer, the non-increase of the total energy, stability tests and comparison of the efficiency of CN and NLSS schemes.

3.1. The relation between the \( \epsilon \) value and the width of the transition layer. In the first numerical experiment, we consider a relation between the \( \epsilon \) value and the width of the transition layer for the CH equation. From our choice of the total energy density Eq. (1.4) and an equilibrium profile \( \text{tanh}(x/\sqrt{2\epsilon}) \) on the infinite domain, the concentration field varies from \(-0.9\) to \(0.9\) over a distance of about \(2\sqrt{2}\tanh^{-1}(0.9)\). Therefore, if we want this value to be about \(m\) grid points, the \( \epsilon \) value need to be taken as follows [13]:

\[ \epsilon_m = \frac{hm}{2\sqrt{2}\tanh^{-1}(0.9)}. \]

To confirm this, we run a simulation with the initial condition \( c(x, y, 0) = 0.01 \times \text{rand}(x, y) \) on the domain \( \Omega = (0, 64) \times (0, 64) \) with \( h = 1 \), \( \Delta t = 0.1 \) and \( \epsilon_4 \). Here, \( \text{rand}(x, y) \in [-1, 1] \) is a randomly generated number. As can be observed from Fig. 2, the transition layer from \( c = -0.9 \) to \( c = 0.9 \) is almost in 4 grid points regardless of type of scheme.
3.2. Non-increase of the total energy. As mentioned in Section 2, NLSS inherits the energy non-increasing property. In order to numerically demonstrate the energy non-increasing property, we consider the temporal evolution of the discrete total energy. In the simulation, we choose $h = 1/128$, $\Delta t = 0.01$, and $\epsilon_4$. In Fig. 3, the temporal evolution of the non-dimensional discrete total energy $E_h(t)/E_h(0)$ (solid line) of the numerical solutions with the initial state $c(x, y, 0) = 0.01 \times \text{rand}(x, y)$. As shown in Fig. 3, the energy is non-increasing during whole time evolution. This numerical result agrees well with the total energy non-increasing property.

3.3. Stability tests. We investigate the stability of the different schemes mentioned in Section 2. We consider numerical solutions with random initial condition $c(x, 0) = \text{rand}(x)$ and $c(x, y, 0) = \text{rand}(x, y)$ on the unit domain for one- and two-dimensional spaces, respectively. Define $\Delta t_c$ be the largest time step, which satisfies the gradient stable, i.e., $E_h(c^{n+1}) \leq E_h(c^n)$. The numerical simulations are performed on the uniform grids, $h = 1/2^n$ for $n = 5, 6, 7$ and 8. In Tables 1 and 2, we list the values of $\Delta t_c$ with different schemes for one- and two-dimensional spaces, respectively. From the results, we observe that EE, IE, CN and SIE schemes are not gradient stable when we use the time step larger than $\Delta t_c$, whereas both LSS and NLSS are unconditionally gradient stable.

Next, we also consider other numerical solutions to investigate gradient stability between LSS and NLSS with the initial data $c(x, 0) = 10 \times \text{rand}(x)$. For simplicity, we perform the comparison of LSS and NLSS in one-dimensional domain instead of two-dimensional one with larger random initial condition than previous simulations. In Table 3, we can observe that there is the different time step constraint in terms of numerical stability for NLSS and LSS; NLSS is still unconditionally gradient stable, but LSS is conditionally stable as shown in the reference [39].
Figure 3. The time dependent non-dimensional discrete total energy $\mathcal{E}^h(t)/\mathcal{E}^h(0)$ (solid line) of the numerical solutions with the initial state $c(x, y, 0) = 0.01 \times \text{rand}(x, y)$.

Table 1. $\Delta t_c$ with different schemes for the initial condition $c(x, 0) = \text{rand}(x)$.

<table>
<thead>
<tr>
<th>Case</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h = 1/32$</td>
<td>$h = 1/64$</td>
<td>$h = 1/128$</td>
<td>$h = 1/256$</td>
</tr>
<tr>
<td>EE</td>
<td>$8.0 \times 10^{-5}$</td>
<td>$1.9 \times 10^{-5}$</td>
<td>$4.7 \times 10^{-6}$</td>
<td>$1.1 \times 10^{-6}$</td>
</tr>
<tr>
<td>IE</td>
<td>$4.9 \times 10^{-3}$</td>
<td>$1.1 \times 10^{-3}$</td>
<td>$2.6 \times 10^{-4}$</td>
<td>$6.1 \times 10^{-5}$</td>
</tr>
<tr>
<td>CN</td>
<td>$5.3 \times 10^{-3}$</td>
<td>$1.1 \times 10^{-3}$</td>
<td>$1.1 \times 10^{-4}$</td>
<td>$2.6 \times 10^{-5}$</td>
</tr>
<tr>
<td>SIE</td>
<td>$2.0 \times 10^{-2}$</td>
<td>$2.7 \times 10^{-3}$</td>
<td>$8.8 \times 10^{-4}$</td>
<td>$2.2 \times 10^{-4}$</td>
</tr>
<tr>
<td>LSS</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>NLSS</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

3.4. **Comparison of the efficiency of CN and NLSS schemes.** We compare the efficiency of CN and NLSS schemes. In the early stages of spinodal decomposition, a rapid separation of two phases occurs because the free energy $\mathcal{E}$ is high. Thus we take a small time step size in these stages. However, after the free energy has been declined sufficiently, the phases separate leisurely in the late stages. Hence we may take a large time step near the equilibrium state. In this section, we perform two tests: the first case is that the CN is used over whole simulations and the second case is that the CN scheme is used over the early stages and we alter into
the NLSS that can adopt a large time step. The reason why we use the CN scheme at the early stages is to accurately evolve the rapid phase separation. We decide to change the scheme when $||c^{n+1} - c^n||_\infty < 2\epsilon - 4$. In this test, $||c^{n+1} - c^n||_\infty$ is less than $2\epsilon - 4$ within 10000 iterations. Thus, in second case, we turn on the NLSS schemes after 10000th iterations. We continue the computation until $||c^{n+1} - c^n||_\infty$ become less than $1\epsilon - 4$. In both cases, a $256 \times 256$ mesh is used on the domain $\Omega = (0, 1) \times (0, 1)$, $\epsilon_4$ is taken and a time step for the CN scheme, $\Delta t = 0.00001$ and for the NLSS scheme, $\Delta t = 0.0001$ was employed. Figure 4 shows the temporal evolution of each case. In the first case, we reach the $||c^{n+1} - c^n||_\infty < 1\epsilon - 4$ with 66000 iterations. In contrast, only 18000 iterations are needed to reach the same state in the second case. With less iterations, mixed scheme with the CN and the NLSS schemes can obtain the results same as those of the CN scheme.

4. CONCLUSIONS

In this paper, we investigated the performance of different schemes in terms of stability and efficiency. To compare the stability of each scheme, we performed numerical experiments such as finding the maximal time step which allows stable numerical computation with a stable initial condition. We observed that EE, IE, CN and SIE schemes are not gradient stable when we use the time step larger than the maximal time step, whereas both LSS and NLSS are unconditionally gradient stable. And, we considered evolution of numerical solution with an unstable initial condition using LSS and NLSS schemes. As a result, NLSS was still unconditionally gradient stable, but LSS was conditionally stable as shown in [39]. Finally, we compared the efficiency of CN and NLSS schemes. Owing to the advantage of NLSS which allows to use a

<table>
<thead>
<tr>
<th>Case</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$h = 1/256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>$2.0 \times 10^{-5}$</td>
<td>$5.0 \times 10^{-6}$</td>
<td>$1.2 \times 10^{-6}$</td>
<td>$3.2 \times 10^{-7}$</td>
</tr>
<tr>
<td>IE</td>
<td>$3.2 \times 10^{-3}$</td>
<td>$8.0 \times 10^{-4}$</td>
<td>$2.0 \times 10^{-4}$</td>
<td>$5.0 \times 10^{-5}$</td>
</tr>
<tr>
<td>CN</td>
<td>$5.0 \times 10^{-3}$</td>
<td>$1.3 \times 10^{-3}$</td>
<td>$3.5 \times 10^{-4}$</td>
<td>$1.0 \times 10^{-4}$</td>
</tr>
<tr>
<td>SIE</td>
<td>$9.1 \times 10^{-3}$</td>
<td>$1.6 \times 10^{-3}$</td>
<td>$3.5 \times 10^{-4}$</td>
<td>$9.0 \times 10^{-5}$</td>
</tr>
<tr>
<td>LSS</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
<tr>
<td>NLSS</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

TABLE 3. $\Delta t_e$ with different schemes for the initial condition $c(x, 0) = 10\text{rand}(x)$. 

<table>
<thead>
<tr>
<th>Case</th>
<th>$h = 1/32$</th>
<th>$h = 1/64$</th>
<th>$h = 1/128$</th>
<th>$h = 1/256$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSS</td>
<td>$8.1 \times 10^{-6}$</td>
<td>$2.0 \times 10^{-6}$</td>
<td>$3.6 \times 10^{-7}$</td>
<td>$8.8 \times 10^{-8}$</td>
</tr>
<tr>
<td>NLSS</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>
FIGURE 4. The temporal evolution of morphologies during a spinodal phase separation of a binary system with different time step size (a) $\Delta t = 0.00001$ using the Crank-Nicolson scheme and (b) $\Delta t = 0.00001$ using the Crank-Nicolson scheme until 10000th iteration and $\Delta t = 0.0001$ using the Nonlinearly stabilized splitting scheme after 10000th iteration, respectively. Both cases run until $||c^{n+1} - c^n||_\infty < 1\text{E}^{-4}$. With less iterations, mixed scheme with the CN and the NLSS schemes can obtain the results same as those of the the CN scheme.

larger time step, mixed scheme with the CN and the NLSS schemes can obtain the results same as those of the the CN scheme with less iterations.

ACKNOWLEDGMENT

The author (H.G. Lee) was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2009-0093827).

REFERENCES


EXTENSION OF AUSMPW+ SCHEME FOR TWO-FLUID MODEL

JIN SEOK PARK\(^1\) AND CHONGAM KIM\(^{1,2\dagger}\)

\(^{1}\)DEPARTMENT OF AEROSPACE ENGINEERING, SEOUL NATIONAL UNIVERSITY, SOUTH KOREA
\(E-mail\ address: \) seok82@snu.ac.kr

\(^{2}\)INSTITUTE OF ADVANCED AEROSPACE TECHNOLOGY, DEPARTMENT OF AEROSPACE ENGINEERING, SEOUL NATIONAL UNIVERSITY, SOUTH KOREA
\(E-mail\ address: \) chongam@snu.ac.kr

ABSTRACT. The present paper deals with the extension of AUSMPW+ scheme into two-fluid model for multiphase flow. AUSMPW+ scheme is the improvement of a single-phase AUSM+ scheme by designing pressure-based weighting functions to prevent oscillations near a wall and shock instability after a strong shock. Recently, Kitamura and Liou assessed a family of AUSM-type schemes with two-fluid model governing equations [K. Kitamura and M.-S. Liou, Comparative study of AUSM-Family schemes in compressible multi-phase flow simulations, ICCFD7-3702 (2012)]. It was observed that the direct application of the single-phase AUSMPW+ did not provide satisfactory results for most of numerical test cases, which motivates the current study. It turns out that, by designing pressure-based weighting functions, which play a key role in controlling numerical diffusion for two-fluid model, problems reported in can be overcome. Various numerical experiments validate the proposed modification of AUSMPW+ scheme is accurate and robust to solve multiphase flow within the framework of two-fluid model.

1. INTRODUCTION

Recently, analyzing multi-phase flow becomes popular due to various applications, such as cavitating flow in hydraulic machine, high-speed underwater projectiles and liquid-gas interactions. There have been many numerical studies, such as interface tracking methods and interface capturing methods. In order to handle complex geometry and phase transition, two interface capturing methods are available: homogeneous equilibrium model (HEM) and two-fluid model. While HEM is more efficient and provides reasonable solution for many applications, two-fluid model is more favorable in describing inter-phasic flow physics.

There are, however, many numerical difficulties to compute two-fluid model. Most notably, while governing equation for each phase is hyperbolic, combining them into two-fluid model becomes non-hyperbolic. In addition, wave speed for each phase is radically different, yielding a very stiff mathematical from. Finally, the non-conservative term may give erroneous results.
when the conservative scheme, which is typically used in single phase fluid, is applied. Thus, proper numerical algorithm for two-fluid model is necessary.

Various approaches are developed to overcome the above-mentioned difficulties. Recently, Liou et al. proposed a new approach based on stratified flow model [1, 2] and AUSM+-up scheme [3]. By successfully extending single-phase AUSM-type flux scheme into two-fluid model, this method does not require the exact Riemann solver.

Many AUSM-type flux schemes can be easily adopted and recently their performance were compared. Although AUSMPW+ scheme is one of the most shock-stable and accurate flux in compressible flows [4, 5], direct application to two-fluid model fails to compute most of test cases [6]. This is mainly attributed that the pressure-based weighting functions, the key part of AUSMPW+ scheme, is improperly working. These functions, based on the compressible flow physics, are originally devised to eliminate the overshoots behind strong shock and/or oscillations near a wall, which is a typical symptom appearing in most AUSM-type schemes. It is also studied that the modifications of the pressure-based weighting functions functions are necessary to extend AUSMPW+ scheme to HEM-based multiphase flows [7] and magnetohydrodynamics (MHD) [8].

The present paper deals with the extension of AUSMPW+ scheme for two-fluid model. At first, two-fluid model and its numerical treatment are briefly introduced as a baseline approach. Examining the shock-stability mechanism of AUSMPW+ scheme, the proper extension for two-fluid model is proposed. Extensive numerical experiments are conducted and conclusions are given.

2. Baseline Approach: Two-Fluid Model

2.1. Governing Equations. In the two-fluid model, the governing equations for each phase are directly solved.

\[
\frac{\partial Q_k}{\partial t} + \nabla \cdot F_k = P^{\text{int}}_k + S_k, \quad k = g, l
\]  

\[
Q_k = \begin{pmatrix}
\alpha \rho \\
\alpha \rho V \\
\alpha \rho E
\end{pmatrix}_k, \quad F_k = \begin{pmatrix}
\alpha \rho V \\
\alpha \rho (V \cdot V + \alpha p) \\
\alpha \rho VH
\end{pmatrix}_k, \quad P^{\text{int}}_k = p^{\text{int}} \left( \begin{matrix}
0 \\
\nabla \alpha \\
\frac{\partial \alpha}{\partial t}
\end{matrix} \right)_k,
\]

where \( g \) and \( l \) indicate the gas and liquid phase, respectively. \( Q_k, F_k \) and \( S_k \) are the conservative vector, inviscid flux vector and volumetric source vector (due to body forces, heat addition, phase change, etc.), respectively. \( \alpha_k \) is the volume fraction with \( \alpha_l + \alpha_g = 1 \). To describe the interaction between gas and liquid phase, the inter-facial pressure vector \( P^{\text{int}}_k \) is introduced.

It is assumed that the bulk pressure and inter-facial pressure of each phase are the same.

\[
p_l = p_g = p, \quad (2.2)
\]

\[
p^{\text{int}}_l = p^{\text{int}}_g = p^{\text{int}}. \quad (2.3)
\]

In addition, the inter-facial pressure is expressed with the bulk pressure and finite jump term [9].

\[
p^{\text{int}} = p - \Delta p^*. \quad (2.4)
\]
For liquid and gas system, it can be usually formulated as follows.

$$\Delta p^* = \sigma \frac{\alpha_l \rho_l}{\alpha_g \rho_g} |V_l - V_g|^2.$$  \hspace{1cm} (2.5)

Assuming that each phase is dispersed and continuous, it can be simplified as follows.

$$\Delta p^* = C_p^{*} \alpha_l |V_l - V_g|^2.$$ \hspace{1cm} (2.6)

While each phase are governed by hyperbolic conservation laws, the two-fluid model (Eq. (2.1)) may lose hyperbolicity. In order to keep the system hyperbolic, coefficient of the interfacial pressure ($\sigma, C_p^{*}$) becomes large \cite{2}. In the present computations, the interfacial pressure term is modeled by Eq. (2.6) with $C^{*} = 2.0$.

To close the system, the equation of state (EOS) for each phase is required. The ideal gas EOS and stiffened EOS \cite{11} are adopted for gas and liquid phase, respectively.

$$p_k = \rho_k \frac{\gamma_k - 1}{\gamma_k} C_{pk} T_k - p_{k\infty}, \hspace{0.5cm} e_k = \frac{C_{pk}}{\gamma_k} T_k + \frac{p_{k\infty}}{\rho_k},$$  \hspace{1cm} (2.7)

where $\gamma_g = 1.4$, $C_{pg} = 1004.5 \text{J/(kgK)}$, $p_{g\infty} = 0 \text{Pa}$ and $\gamma_l = 2.8$, $C_{pl} = 4186 \text{J/(kgK)}$, $p_{l\infty} = 8.5 \times 10^8 \text{Pa}$.

2.2. Stratified Flow Model and Spatial Discretization. In order to discretize the governing equation (Eq. (2.1)), we adopt the stratified flow model \cite{10, 1}. This model divides each phase region according to characteristic property, such as volume fraction. This concept is suitable to the finite volume discretization, and we can construct two control volumes for liquid and gas within a cell. It is assumed that the volume fraction $\alpha_k$ is piecewise continuous within the cell and allowed a jump at a cell interface. For each phase, the conservation laws are applied as in the single phase fluids, and the interactions at the sub-cell interface between gas and liquid phases are additionally considered. The present computations ignore phase transition and viscous effects, and thus only inter-facial pressure affects the momentum and energy conservation as follows.

$$\frac{\Delta x}{\Delta t} \delta^t \left( \begin{array}{c} \alpha_k \rho_k \\ \alpha_k \rho_k u_k \\ \alpha_k \rho_k E_k \end{array} \right)_j + \delta_x \left( \begin{array}{c} \alpha_k \rho_k u_k \\ \alpha_k \rho_k u_k^2 + \alpha_k p \\ \alpha_k \rho_k u_k H_k \end{array} \right)_j - p^{\text{int}}_j \left( \begin{array}{c} 0 \\ \frac{\Delta x}{\Delta t} \delta^t (\alpha_k) \end{array} \right)_j = 0.$$  \hspace{1cm} (2.8)

where $\delta^t(\cdot) = (\cdot)^{n+1} - (\cdot)^n$ and $\delta^t_x(\cdot) = (\cdot)_{j+1/2} - (\cdot)_{j-1/2}$.

The inter-facial pressure affects the momentum conservation if the properties (including $\alpha_k$) are discretized by at least a second-order accurate method. Here, a second-order accurate MUSCL interpolation is used. At the cell interface, monotone and accurate numerical flux scheme is required, which is going to be discussed in the next section.

2.3. Time Marching and Solution Update. In order to update solution variables, third-order accurate TVD Runge-Kutta method is used. The time-derivative term appearing in the inter-facial pressure vector is included in time marching, and thus the energy equation is computed
as follows.

$$\hat{Q}_{3,k} = Q_{3,k} + \alpha_k p^{int},$$  \hspace{1cm} (2.9)$$

$$\frac{\partial \hat{Q}_{3,k}}{\partial t} + \sum_{\text{face}} (F_{3,k} \cdot n) l = 0,$$  \hspace{1cm} (2.10)$$

where $n$ and $l$ are the normal vector and length of the face, respectively. After time marching, the decoding process is required to update $p^{n+1}$ and $\alpha_k^{n+1}$.

$$\begin{align*}
(p + \hat{a}_k) \alpha_k - \hat{A}_k &= 0, \\
\hat{a}_k &= \gamma_k p_{k,\infty} + (\gamma_k - 1)p^{int}, \\
\hat{A}_k &= (\gamma_k - 1) \left( \hat{Q}_{3,k} - \frac{Q_{2,k}^2}{2Q_{1,k}} \right).
\end{align*}$$  \hspace{1cm} (2.11)-(2.13)$$

Equation (2.11) are then solved by the Newton iteration method [6].

3. AUSMPW+ Scheme for Two-Fluid Model

3.1. Baseline Scheme: AUSMPW+. The AUSMPW+ scheme for a single gas phase with $k = g$ can be formulated as follows.

$$\begin{align*}
\Phi_{k,1/2,L/R} &= \frac{\tilde{M}_{k,L}^+ c_{k,1/2} \Phi_{k,L}}{2} + \frac{\tilde{M}_{k,R}^- c_{k,1/2} \Phi_{k,R}}{2} + \alpha_{k,1/2,L/R} (P_{k,L}^+ P_L + P_{k,R}^+ P_R),
\end{align*}$$  \hspace{1cm} (3.1)$$

where $\Phi = (\alpha, \varphi, \alpha \varphi V, \alpha \varphi H)^T$ and $P = (0, p n, 0)^T$.

For $m_{1/2} = m_{1/2} = M_{k,L}^+ + M_{k,R}^- \geq 0$,

$$\begin{align*}
\tilde{M}_{k,L}^+ &= M_{k,L}^+ + M_{k,R}^- \cdot [(1 - w_k) \cdot (1 + f_{k,R}) - F_{k,L}], \\
\tilde{M}_{k,R}^- &= M_{k,R}^- \cdot w_k \cdot (1 + f_{k,R}).
\end{align*}$$  \hspace{1cm} (3.2)$$

For $m_{1/2} \leq 0$,

$$\begin{align*}
\tilde{M}_{k,L}^+ &= M_{k,L}^+ \cdot w_k \cdot (1 + f_{k,L}), \\
\tilde{M}_{k,R}^- &= M_{k,R}^- + M_{k,L}^+ \cdot [(1 - w_k) \cdot (1 + f_{k,L}) - F_{k,R}].
\end{align*}$$  \hspace{1cm} (3.3)$$

We use the AUSM-type interpolating functions for the Mach number($M$) and pressure($P$) [4], as follows.

$$\begin{align*}
M_k^\pm &= \left\{ \begin{array}{ll}
\pm \frac{1}{4}(M_k \pm 1)^2, & |M_k| \leq 1, \\
\frac{1}{2}(M_k \pm |M_k|), & |M_k| > 1.
\end{array} \right. \\
P_{k,L/R}^\pm &= \left\{ \begin{array}{ll}
\pm \frac{1}{4}(M_k \pm 1)^2(2 \mp M_k) \pm \alpha M_k(M_k^2 - 1)^2, & |M_k| \leq 1, \\
\frac{1}{2}(1 \pm \text{sign}(M_k)), & |M_k| > 1.
\end{array} \right.
\end{align*}$$  \hspace{1cm} (3.4)-(3.5)$$

The speed of sound at a cell-interface is given by

$$c_{k,1/2} = \frac{c_{k,s}^2}{\text{max}(|U_{k,L/R}|, c_{k,s})}, c_{k,s} = \sqrt{\frac{2(\gamma - 1)}{(\gamma + 1)} H_{k,n}},$$

$$H_{k,n} = 0.5 \times (H_{k,L} - 0.5V_L^2 + H_{k,R} - 0.5V_R^2),$$  \hspace{1cm} (3.6)-(3.7)$$
where $U$ is the contravariant velocity and $V$ is the tangential velocity.

In order to treat oscillations near a wall and across a strong shock, two pressure-based functions, $f$ and $w$, are introduced as follows.

$$f_{k,L/R} = \begin{cases} (p_{k,s}^{L,R} - 1) w_2, & p_{k,s} \neq 0 \\ 0, & \text{elsewhere} \end{cases}, \quad w_2 = \min \left( 1, \frac{\min(p_{1,L},p_{1,R},p_{2,L},p_{2,R})}{\min(p_L,p_R)} \right)^2,$$

$$w(p_L,p_R) = 1 - \min \left( \frac{p_L}{p_R}, \frac{p_R}{p_L} \right)^3,$$

where $p_{k,s} = P_{k,L}^+ + P_{k,R}^-$. 

### 3.2. Analysis and Modification of Pressure-based Weighting Functions.

Equations (3.8) and (3.9) basically control the amount of numerical diffusion caused by the pressure difference and difference of advection term. Kitamura and Liou extended a single gas phase AUSMPW+ into two-fluid model without any modification [6], and it was observed that the two functions are unstable for two-fluid model, leading to divergence of computations in most test cases. According to authors’ experience on the extension of AUSMPW+ for HEM multiphase flow and MHD, it is essential to modify these terms based on its diffusion mechanism and its design principle. We start it by examining each numerical diffusion term.

For $0 \leq m_{1/2} \leq 1$, the diffusion term of AUSMPW+ scheme can be written as follows.

$$D_{AUSMPW+} = U_{k,R} \Delta \Phi_k + \frac{U_{k,L} + U_{k,R}}{2c_{k,1/2}} \Phi_{k,L} \Delta U_k - \left( P_{k,L}^+ P_L + P_{k,R}^- P_R \right)$$

$$+ \frac{c_{k,1/2}}{2} (M_{k,R} - 1)^2 w(1 + f_{k,R}) \Delta \Phi_k + \frac{c_{k,1/2}}{2} (M_{k,R} - 1)^2 \Phi_{k,L} \Delta f_k.$$

Comparing Eq. (3.10) to the diffusion term of AUSM+up [3] scheme, we can observe following characteristics.

$$D_{AUSM+up} = U_{k,R} \Delta \Phi_k + \frac{U_{k,L} + U_{k,R}}{2c_{k,1/2}} \Phi_{k,L} \Delta U_k - M_{pk} c_{k,1/2} \Phi_{k,L} - (P_{k,L}^+ P_L + P_{k,R}^- P_R - p_{uk}).$$

$$M_{pk} = -\frac{K_p}{f_a} \max(1 - \bar{M}_k^2,0) \frac{\Delta p}{\bar{\rho}_k c_{k,1/2}^2}, p_{uk} = -K_u \bar{P}_{k,L} |P_{k,R}^-| \bar{\rho}_k f_a c_{k,1/2} \Delta U_k,$$

where $K_p$ and $K_u$ are coefficients. $f_a$ is a scaling function for all-speed extension but it is needless for two-fluid model (i.e. $f_a = 1$).

i) Both scheme (AUSMPW+ and AUSM+up) have the numerical diffusion due to pressure difference, and they are similar to each other as follows.

$$D_{AUSMPW+}^{(p)} = \frac{c_{k,1/2}}{2} (M_{k,R} - 1)^2 \Phi_{k,L} \Delta f_k = \frac{c_{k,1/2}}{2} (M_{k,R} - 1)^2 w_2 \Phi_{k,L} \Delta p_{pk,s},$$

$$D_{AUSM+up}^{(p)} = -M_{pk} c_{k,1/2} \Phi_{k,L} = \frac{K_p}{f_a} c_{k,1/2} (1 - \bar{M}_k^2) \Phi_{k,L} \Delta p_{\bar{\rho}_k c_{k,1/2}^2}.$$
While AUSMPW+ scheme has additional diffusion mechanism due to jump in advection variable \((\Delta \Phi_k)\), AUSM+-up scheme adjusts the pressure flux by the difference in interface normal velocity.

The role of \(f\) of AUSMPW+ is a kind of bridge between AUSM+ and AUSMD. While it consists of pressure ratio, it actually reflects the density difference via the definition of speed of sound (see Eqs (14-17) in Ref [4]).

\[
(pu)_{k,1/2,AUSMPW} = M_{k,L}^+ c_{k,1/2} \rho_{k,L} + M_{k,R}^- c_{k,1/2} \rho_{k,L} \left( \frac{\rho_{k,R}}{\rho_{k,L}} \right),
\]

\[
\approx M_{k,L}^+ c_{k,1/2} \rho_{k,L} + M_{k,R}^- c_{k,1/2} \rho_{k,L} \left( \frac{p_R}{p_L} \right).
\]

(3.15)

(3.16)

To be consistent with the stiffened EOS [10], the pressure ratio of \(f\) (Eq. (3.8)) has to be written in terms of \(\bar{p} = p + p_\infty\).

\[
f_{k,L/R} = \begin{cases} 
    (\frac{\bar{p}_{L/R}}{\bar{p}_s} - 1) w_2 & p_s \neq 0 \\
    0 & \text{else}
\end{cases}
\]

(3.17)

Furthermore, the term \(w_2\) in Eq. (3.8), which is designed to detect stagnation region of boundary layer or shock in multi-dimensional situation, also controls the magnitude of numerical diffusion by pressure jump (see Eq. (3.13)). The problem is that, while \(w_2\) successfully controls the diffusion by pressure jump for gas phase, it does not provide adequate diffusion for liquid phase or transition of phase. From this perspective, \(w_2\) is also computed using \(\bar{p}\).

\[
w_2 = \min \left( 1, \frac{\min(\bar{p}_1, L, \bar{p}_{1, R}, \bar{p}_{2, L}, \bar{p}_{2, R})}{\min(\bar{p}_L, \bar{p}_R)} \right)^2
\]

(3.18)

On the other hand, the role of \(w\) is to detect a large pressure jump (especially around shock) and introduces additional diffusion to remove spurious oscillation (see the fourth term in Eq. (3.10)). If this term is also replaced by \(\bar{p}\), \(w\) cannot detect pressure jump in liquid phase and cannot suppress oscillations by pressure jump. Thus, \(w\) should be kept as the original form Eq. (3.9), and thus it is the same in both gas and liquid phase.

While AUSMPW+ for HEM model additionally introduces a term based on mixture density in phase transition (Eq. (46) in Ref. [7]) to control the fourth term in Eq. (3.10), AUSMPW+ for two-fluid model does not require such device. AUSM+-up and other AUSM-type scheme (for example, SLAU [12]) control the pressure flux by velocity difference (Eq.(3.12)). Though not essential, \(p_{uk}\) could be a little bit helpful in suppressing a slight kink after shock in liquid phase. Thus, we add \(p_{uk}\) in the pressure flux of AUSMPW+.

In summary, by examining the behavior of numerical diffusion, AUSMPW+ for two-fluid model can be obtained by replacing \(p\) of Eq. (3.8) by \(\bar{p}\). The usage of \(p_{uk}\) of Eq. (3.12) is optional.
4. Numerical Experiments

We use a two-dimensional structured finite volume solver. To account for the inter-facial pressure, a second-order MUSCL interpolation with minmod limiter is applied to the primitive variables \( Q_p = (\alpha_g, u_g, v_g, v_l, P_g, T_l) \). With the developed AUSMPW+ scheme, we compute four test cases. Without the proposed modification, it is reported that AUSMPW+ fails to solve them [6]. All computed results are compared with AUSM+ and AUSM+-up(1,1). Hereafter, AUSM+-up indicates AUSM+-up(1,1).

4.1. Moving Phase Discontinuity. The first test case is to examine the capability to resolve phase discontinuity maintaining the pressure. Computational domain is \([0, 10] \times [0, 1]\) with \(\Delta x = \Delta y = 0.05\). Initially, air and water is separated at \(x = 5\).

\[
(p, \alpha_g, u_k, v, k, T_k)_{L} = (10^5\text{Pa}, 1 - \epsilon, 100\text{m/s}, 0\text{m/s}, 300\text{K}) \text{ for } x \leq 5\text{m},
\]

\[
(p, \alpha_g, u_k, v, k, T_k)_{R} = (10^5\text{Pa}, \epsilon, 100\text{m/s}, 0\text{m/s}, 300\text{K}) \text{ else},
\]

where \(k = l, g\) and \(\epsilon = 1.0 \times 10^{-7}\).

Figure 1 shows the distributions of pressure and void fraction along the centerline. All flux schemes can compute the moving phase transition successfully.

![Figure 1. Distribution of pressure and volume of fraction along the centerline at \(t = 0.03\text{s}\).](image)

4.2. Shock Tube Problem. In order to examine robustness across underwater shock waves, two kinds of shock tube problems are considered. We conduct the propagation of shock wave from air to water and vice versa. Computational domain is \([0, 10] \times [0, 1]\) and \(\Delta x = \Delta y = 0.02\). Two initial conditions are imposed as follows.

Air-to-Water Shock Tube

\[
(p, \alpha_g, u_k, v_k, T_k)_L = \begin{cases} (10^9 \text{pa}, 1 - \epsilon, 100 \text{m/s}, 0 \text{m/s}, 308.15 K) & \text{for } x \leq 5 \text{m}, \\ (10^5 \text{pa}, 1 - \epsilon, 100 \text{m/s}, 0 \text{m/s}, 308.15 K) & \text{else} \end{cases}, \tag{4.2}
\]

Water-to-Air Shock Tube

\[
(p, \alpha_g, u_k, v_k, T_k)_L = \begin{cases} (10^7 \text{pa}, \epsilon, 100 \text{m/s}, 0 \text{m/s}, 308.15 K) & \text{for } x \leq 5 \text{m}, \\ (5 \times 10^6 \text{pa}, 1 - \epsilon, 100 \text{m/s}, 0 \text{m/s}, 308.15 K) & \text{else} \end{cases}, \tag{4.3}
\]

where \(k = l, g\) and \(\epsilon = 1.0 \times 10^{-7}\).

Figures 2 and 3 show the pressure distribution of air-to-water shock tube and water-to-air shock tube, respectively. The dotted line represents the gas volume fraction. AUSM+ suffers spurious oscillations around shock region in water, especially for water-to-air shock tube. AUSMPW+ with the proposed modification and AUSM+-up(1,1) can successfully resolve these wave structures.

**Figure 2.** Distribution of pressure along the centerline at \(t = 0.002\text{s}\). (Right: Close-up view around shock region)

4.3. One-dimensional Cavitation Problem. This is a simplified cavitation problem, proposed by Saurel and Abgrall [13]. After two opposite rarefaction waves are propagating in tube, velocity reaches zero at the center of domain. Thus, liquid flow containing 1% gas creates a cavitation zone at the center. Computational domain is \([0, 10] \times [0, 1]\) with \(\Delta x = \Delta y = 0.05\). The initial condition is

\[
(p, \alpha_g, u_k, v_k, T_k)_L = \begin{cases} (10^5 \text{pa}, -100 \text{m/s}, 0 \text{m/s}, 300 K) & \text{for } x \leq 5 \text{m}, \\ (10^5 \text{pa}, \epsilon, 100 \text{m/s}, 0 \text{m/s}, 300 K) & \text{else} \end{cases}, \tag{4.4}
\]

where \(k = l, g\) and \(\epsilon = 1.0 \times 10^{-2}\).
Figure 3. Distribution of pressure along the centerline at \( t = 0.002 \text{s} \). (Right: Close-up view around shock region)

Figure 4 shows the pressure and void fraction of AUSM+-up(1,1) and AUSMPW+ scheme. Both solutions are almost identical.

Figure 4. Distribution of pressure and void fraction along the centerline at \( t = 10 \mu \text{s} \)

4.4. Shock/Water-Column Interaction. The last test is more challenging and realistic problem. When shock wave impinges on the water column, shock wave transmits into water and travels faster. Shock wave in water also reflects at the rear interface. Considering the flow symmetry, only the upper half region is simulated. Computational domain is \([-15mm, 20mm] \times \)
[0mm, 15mm], and discretized by $400 \times 200$ cells. The initial condition is provided as follows.

\[
(p, \alpha_g, u_k, v_k, T_k) = \begin{cases} 
(2.35438 \times 10^5 \text{pa}, \epsilon, 225.86 \text{m/s}, 0 \text{m/s}, 381.85 \text{K}) & \text{for } x \leq -4 \text{mm}, \\
(10^5 \text{pa}, 1 - \epsilon, 0 \text{m/s}, 0 \text{m/s}, 293.15 \text{K}) & \text{for } x \geq -4 \text{mm}, \\
\sqrt{(x^2 + y^2)} \leq 3.2 \text{mm}, \\
(10^5 \text{pa}, \epsilon, 0 \text{m/s}, 0 \text{m/s}, 293.15 \text{K}) & \text{else}, 
\end{cases}
\]

where $k = l, g$ and $\epsilon = 1 \times 10^{-5}$.

Figure 5 shows the pressure contours at $t = 10\mu s$. The solid line is the boundary of water column. AUSMPW+ successfully captures the transmission and diffraction of the impinging shock wave and the transmitted shock reflected from the rear water column. These results are quite similar to other AUSM-type results [2, 6].

5. CONCLUSION

In this paper, we successfully extend AUSMPW+ scheme to two-fluid model. By examining the behavior of numerical diffusion, AUSMPW+ scheme for two-fluid model can be obtained by simply changing the pressure-based weighting function $f$ only. All computed results for critical test cases confirms that AUSMPW+ provides robust and accurate solutions of compressible two-fluid multiphase flow with shock waves.
ACKNOWLEDGMENTS

Authors appreciate the financial supports provided by NSL (National Space Lab.) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (Grant 2012-0009099) and by the EDISON program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (Grant 2012-0006661).

REFERENCES

Instruction for Authors

Manuscript Preparation
Beginning January 1, 2008, Journal of the Korean Society for Industrial and Applied Mathematics (J.KSIAM) uses a new paper format. The templates for the new paper format (MS word template and LaTeX template) can be downloaded from the KSIAM website (URL: http://www.ksiam.org).
To keep the review time as short as possible, J.KSIAM requests all authors to submit their manuscripts online via the journal Online Submission System, except for very special circumstances that J.KSIAM recognizes.
All manuscripts submitted to J.KSIAM should follow the instructions for authors and use the template files. Manuscripts should be written in English and typed on A4-sized papers. The first page of the manuscript must include (1) the title which should be short, descriptive and informative, (2) the name(s) and address(es) of the author(s) along with e-mail address(es), (3) the abstract which should summarize the manuscript and be at least one complete sentence and at most 300 words. The manuscript should also include as a footnote the 2000 Mathematics Subject Classification by the American Mathematical Society, followed by a list of key words and phrases.

Final Manuscript Submission
To facilitate prompt turnarounds of author proofs and early placements of the accepted manuscripts in the future issues, J.KSIAM requests all authors of the accepted manuscripts to submit the electronic text/graphics file(s).
The final manuscript must be submitted in either the LaTeX format or the MS word format. We recommend that all authors use the template files which can be downloaded at http://www.ksiam.org/jksiam. In the case that the corresponding author does not submit all relevant electronic text/graphics, further producing process regarding his/her manuscript would be stopped. The final manuscripts for publication should be sent to an appropriate managing editor via the journal Online Submission System.
Please note that the submission of the final manuscript implies that the author(s) are agreeing to be bound by the KSIAM Provisions on Copyright. Author check list and Copyright transfer can be found during the submission process via homepage.

Offprints
J. KSIAM supplies a final version of published paper in PDF form to a corresponding author. Authors are allowed to reproduce their own papers from the PDF file. Reprints of published paper can be also supplied at the approximate cost of production on request. For details, please contact one of managing editors.

Peer Review
Suitability for publication in J.KSIAM is judged by the editorial policy of the Editorial Board.
A submitted paper is allocated to one of the Managing Editors, who has full directorship for the reviewing process of the submission. The managing editor selects one of Associate Editors, and the selected Associate Editor chooses two qualified referees to review the paper. The Associate Editor him/herself may act as a referee. The managing Editor handles all correspondence with the author(s). The anonymity of the reviewers is always preserved. The reviewers should examine the paper and return it with their reports to the Associate Editor within 4 weeks from the date of the initial review request. The reviewers recommend acceptance, rejection or revision through a report. Based on the reports, the Associate Editor makes a recommendation to the managing Editor whether the paper should be accepted, rejected or needs to be returned to the author(s) for revision.
Papers needing revision will be returned to the author(s), and the author(s) must submit a revised manuscript to the online system within one month from the date of the revision request; otherwise it will be assumed that the paper has been withdrawn. The revision is assigned for review to the managing Editor who handled its initial submission. The managing Editor sends the revised manuscript to the original Associate Editor to check if the manuscript is revised as suggested by the reviewers in the previous review. The Associate Editor makes a recommendation for the revision to the managing Editor within 3 weeks from the date of the Associate Editor assignment.
Final decision by the managing Editor is usually made within 3 months from the time of initial submission. The length of time from initial submission to final decision may vary, depending on the time spent for review and revision. A letter announcing a publication date is sent to author(s) after a manuscript has been accepted by the managing Editor.

**Research and Publication Ethics**
Research published in Journal of KSIAM must have followed institutional, national and international guidelines. For the policies on the research and publication ethics that are not stated in these instructions, the Guidelines on Good Publication (http://www.publicationethics.org.uk/guidelines) can be applied.

**Publication Charge**
The publication fee is US $200, and extra charge could be necessary for color printings. The publication cost is subjected to change according to the society’s financial situation

**Forms of Publication**
- Original papers: this form of publication represents original research articles on research findings.
- Review article: this form do not cover original research but rather accumulate the results of many different articles on a particular topic
- Erratum/Revision/Retraction: these kinds of editorial notice may be published.

**Format of References**
The Vancouver style system is the preferred reference system for stipulations not otherwise described below: References should be listed at the end of the paper and conforming to current Journal style. Corresponding bracketed numbers are used to cite references in the text [1]. For multiple citations, separate reference numbers with commas [2, 3], or use a dash to show a range [4-7]. The DOI (Digital Object Identifier) should be incorporated in every reference for which it is available (see the last reference example).

- **Journal**

- **Book**

- **Proceedings**
Copyright Transfer Agreement

Copyright in the unpublished and original article, including the abstract, entitled -

________________________________________ (Title of Article)

Submitted by the following author(s)

________________________________________ (Names of Authors)

is hereby assigned and transferred to Korean Society for Industrial and Applied Mathematics(KSIAM) for the full term thereof throughout the world, subject to the term of this Agreement and to acceptance of the Article for publication in Journal of Korean Society for Industrial and Applied Mathematics.

KSIAM shall have the right to publish the Article in any medium or form, or by any means, now known or later developed.

KSIAM shall have the right to register copyright to the Article in its name as claimant whether separately or as part of the journal issue or other medium in which the Article is included. The Author(s) reserve all proprietary right other than copyright, such as patent rights. The Author(s) represent and warrant:

(1) that the Article is original with them;
(2) that the Article does not infringe any copyright or other rights in any other work, or violate any other rights;
(3) that the Author(s) own the copyright in the Article or are authorized to transfer it;
(4) that all copies of the Article the Author(s) make or authorize will include a proper notice of copyright in KSIAM’s name;

If each Author’s signature does not appear below, the signing Author(s) represent that they sign this Agreement as authorized agents for and on behalf of all the Authors, and that this Agreement and authorization is made on behalf of all the Authors.

Authors ____________________________________________

Signature ___________________ Date ________________
Author’s checklist

☐ This manuscript has never been submitted to or published in other journals.

☐ All citation references are correct and comply with the submission rule.

☐ I will supply a manuscript in either TeX or MS-Word format as well as all artworks once the article is accepted.

☐ I filled out and signed the proper copyright transfer agreement.

☐ Research published in the Journal has followed institutional, national, and international guidelines of ethnics.

☐ I checked whether the manuscript contained all authors’ names, affiliation, full address, e-mail. And also whether it contained an abstract and keywords.

Paper title

Authors

Signature__________________ Date__________________
This journal was supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST).

Journal of the Korean Society for Industrial and Applied Mathematics

제 17권 3호

2013년 9월 25일 인쇄
2013년 9월 25일 발행

발행처 한국산업응용수학회
편집인 편집위원회
발행인 임철호

대전광역시 유성구 대학로 291
KAIST 한국산업응용수학회
TEL: 042-350-5708
FAX: 042-350-5710
A P-HIERARCHICAL ERROR ESTIMATOR FOR A FEM-BEM COUPLING OF AN EDDY CURRENT PROBLEM IN $\mathbb{R}^3$
FLORIAN LEYDECKER, MATTHIAS MAISCHAK, ERNST P. STEPHAN, AND MATTHIAS TELTSCHER ............................................................... 139

CIRCLE APPROXIMATION BY QUARTIC $G^2$ SPLINE USING ALTERNATION OF ERROR FUNCTION
SOO WON KIM AND YOUNG JOON AHN ......................................................... 171

ANalytic approach for the study of air and/or liquid filled geomembrane tube sections on a horizontal
Yoon-Rak Choi ........................................................................................................ 181

Comparison of different numerical schemes for the Cahn-Hilliard equation
Seunghyu Lee, Chaeyoung Lee, Hyun Geun Lee, and Junseok Kim ......................... 197

Extension of AUSMPW+ scheme for two-fluid model
Jin Seok Park and Chongam Kim ............................................................................ 209

ISSN 1226-9433(print)
ISSN 1229-0645(electronic)