



**HAL**  
open science

## Guaranteed error estimates for finite element discretizations of Helmholtz problems

Théophile Chaumont-Frelet, Alexandre Ern, Martin Vohralík

### ► To cite this version:

Théophile Chaumont-Frelet, Alexandre Ern, Martin Vohralík. Guaranteed error estimates for finite element discretizations of Helmholtz problems. Numerical Waves, Oct 2021, Nice, France. hal-03404014

**HAL Id: hal-03404014**

**<https://inria.hal.science/hal-03404014v1>**

Submitted on 26 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Guaranteed error estimates for finite element discretizations of Helmholtz problems

---

T. Chaumont-Frelet<sup>\*,†</sup>, A. Ern<sup>‡,§</sup>, M. Vohralík<sup>§,‡</sup>

October 8, 2021

\* Inria Sophia-Antipolis, † Laboratoire J.A. Dieudonné, ‡ CERMICS, § Inria Paris

# Motivations

---

# Motivations

---

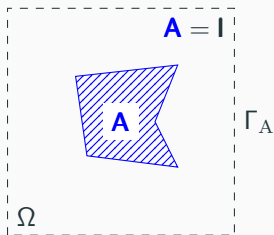
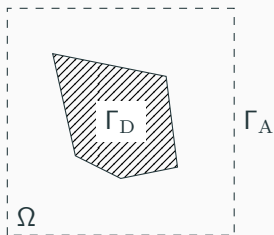
Model problem

# Model problem

Given  $f : \Omega \rightarrow \mathbb{C}$ , find  $u : \Omega \rightarrow \mathbb{C}$  such that

$$\begin{cases} -\omega^2 \mu u - \nabla \cdot (\mathbf{A} \nabla u) = \mu f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \mathbf{A} \nabla u \cdot \mathbf{n} - i\omega\gamma u = 0 & \text{on } \Gamma_A \end{cases}$$

where  $\mu$ ,  $\mathbf{A}$  and  $\gamma$  are given coefficients strictly positive coefficients.



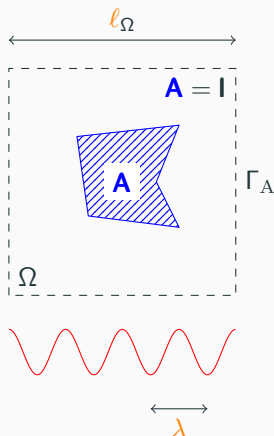
# What do I mean by high-frequency?

The physical meaning of  $\mu$  and  $\mathbf{A}$  depends on the application, but the wavespeed is always given by:

$$c := \sqrt{\frac{\sigma_{\min}(\mathbf{A})}{\mu}}$$

The (minimal) wavelength is then given by:

$$\lambda := \frac{2\pi}{\omega} c_{\min}.$$



The “important” quantity is  $N_\lambda := \ell_\Omega / \lambda$ . High-frequency means that

$$\frac{\omega \ell_\Omega}{c_{\min}} \simeq N_\lambda$$

is “large” (a few tens or hundreds).

# Variational formulation

Recall the Helmholtz problem in strong form

$$\begin{cases} -\omega^2 \mu u - \nabla \cdot (\mathbf{A} \nabla u) = \mu f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \mathbf{A} \nabla u \cdot \mathbf{n} - i\omega \gamma u = 0 & \text{on } \Gamma_A. \end{cases}$$

Assuming  $f \in L^2(\Omega)$ , we seek  $u \in H_D^1(\Omega)$  such that

$$b(u, v) = (\mu f, v) \quad \forall v \in H_D^1(\Omega),$$

with

$$b(u, v) := -\omega^2 (\mu u, v)_\Omega - i\omega (\gamma u, v)_{\Gamma_A} + (\mathbf{A} \nabla u, \nabla v)_\Omega.$$

# Finite element approximation

We consider a mesh  $\mathcal{T}_h$  of  $\Omega$  into tetrahedral element  $K$ .

The elements  $K \in \mathcal{T}_h$  are “small” ( $h_K := \text{diam } K \leq h$ ).

The coefficients  $\mu, \gamma, \mathbf{A}$  are constant inside each element/face.

We introduce a “finite element” discretization space

$$V_h := \{v_h \in H_D^1(\Omega) \mid v_h|_K \in \mathbb{P}_p(K) \forall K \in \mathcal{T}_h\}$$

with  $p \geq 1$ .



P.G. Ciarlet, 2002.

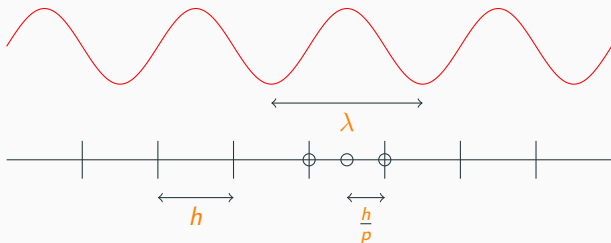


# What do I mean by refined mesh?

When I am saying that “the mesh is fine” , I mean that

$$N_{\text{dofs}/\lambda} = \lambda / \frac{h}{p} \simeq \left( \frac{\omega h}{C_{\min} p} \right)^{-1}$$

is large.



# Finite element approximation

Recall that  $u$  is the only element of  $H_D^1(\Omega)$  such that

$$b(u, v) = (\mu f, v) \quad \forall v \in H_D^1(\Omega).$$

Analogously, we seek a discrete a discrete solution  $u_h \in V_h$  such that

$$b(u_h, v_h) = (\mu f, v_h) \quad \forall v_h \in V_h. \quad (1)$$

Problem (1) corresponds to a finite dimensional linear system, that we can numerically solve.

In this talk, we are especially interested in measuring the error

$$e_h := u - u_h$$

# Motivations

---

**A priori error estimates**

# A priori error estimates

## A priori estimates

Assume that  $\omega h/c_{\min}^p \leq \mathcal{C}_1$ , then

$$\|\nabla e_h\|_{\mathbf{A},\Omega} \leq \mathcal{C}_2 \left( \frac{\omega h}{c_{\min}^p} \right)^p.$$



F. Ihlenburg and I. Babuška, *SIAM J. Numer. Anal.*, 1997.



J.M. Melenk and S.A. Sauter, *Math. Comp.*, 2010.



T. Chaumont-Frelet and S. Nicaise, *IMA J. Numer. Anal.*, 2019.

Some limitations:

- The above result requires important regularity assumptions.
- The error estimate is not always applicable.
- The constants  $\mathcal{C}_1$  and  $\mathcal{C}_2$  are *not* computable in general.

# A priori error estimates

A priori estimates provide qualitative upper bounds.

They are important as they show that the method converges.  
They also indicate how fast the convergence happens.

They are not suited to *quantitatively* estimate the error in practice.

# Motivations

---

**A posteriori error estimation**

# A posteriori estimates

## (ideal) A posteriori estimates

$$\|\nabla e_h\|_{\mathbf{A},\Omega} \leq \eta$$

Here  $\eta$  is a fully-computable real number called an “error estimator”. This quantity is computed as a post-processing of  $u_h$ , i.e.  $\eta = \eta(u_h)$ . There is *no generic constants*. We have a *guaranteed* error estimate.

# Outline

- 1 Low frequencies
- 2 High frequencies
- 3 Controlling the pre-factors
- 4 Numerical illustrations



## Low frequencies

---

## The low frequency case

We first consider the low frequency limit where  $\omega = 0$ .

The problem then reads: find  $u : \Omega \rightarrow \mathbb{C}$  such that

$$\begin{cases} -\nabla \cdot (\mathbf{A} \nabla u) &= \mu f & \text{in } \Omega, \\ u &= 0 & \text{on } \Gamma_D, \\ \mathbf{A} \nabla u \cdot \mathbf{n} &= 0 & \text{on } \Gamma_A. \end{cases}$$

For the sake of simplicity, we will assume that  $f = f_h \in \mathbb{P}_p(\mathcal{T}_h)$ .

# Variational formulation and discrete solution

Consider the sesquilinear form

$$a(\mathbf{u}, v) = (\mathbf{A}\nabla\mathbf{u}, \nabla v)_{\Omega}, \quad \mathbf{u}, v \in H_{\Gamma_D}^1(\Omega).$$

We can characterize  $\mathbf{u}$  as the unique element of  $H_{\Gamma_D}^1(\Omega)$  such that

$$a(\mathbf{u}, v) = (\mu f_h, v)_{\Omega}$$

for all  $v \in H_{\Gamma_D}^1(\Omega)$ .

$\mathbf{u}_h$  is the unique element of  $V_h$  satisfying

$$a(\mathbf{u}_h, v_h) = (\mu f_h, v_h)_{\Omega}$$

for all  $v_h \in V_h$ .

## Low frequencies

---

Key ideas behind a posteriori estimation

## Key ideas

Second-order problems typically arise from two physical laws, e.g., “Faraday’s law + Gauss’ law = Poisson problem”.

The continuous solution  $u$  is uniquely determined by the condition that

$$u \in H^1(\Omega), \quad u = 0 \text{ on } \Gamma_D \quad (2a)$$

and for the “flux”  $\sigma := -A\nabla u \in \mathbf{H}(\text{div}, \Omega)$

$$\sigma \cdot \mathbf{n} = 0 \text{ on } \Gamma_A, \quad \nabla \cdot \sigma = \mu f_h \text{ in } \Omega. \quad (2b)$$

The discrete solution  $u_h$  satisfies (2a) but not (2b) in general.

# Key ideas

Consider the minimization problem

$$\boldsymbol{\sigma} := \arg \min_{\substack{\boldsymbol{\tau} \in \mathbf{H}(\text{div}, \Omega) \\ \boldsymbol{\tau} \cdot \mathbf{n} = 0 \text{ on } \Gamma_A \\ \nabla \cdot \boldsymbol{\tau} = \mu f_h \text{ in } \Omega}} \|\mathbf{A}^{-1} \boldsymbol{\tau} + \nabla u_h\|_{\mathbf{A}, \Omega}.$$

If the above minimum is 0, then  $\boldsymbol{\sigma} = -\mathbf{A} \nabla u_h \in \mathbf{H}(\text{div}, \Omega)$  with

$$\boldsymbol{\sigma} \cdot \mathbf{n} = 0 \text{ on } \Gamma_A \quad \nabla \cdot \boldsymbol{\sigma} = \mu f_h \text{ in } \Omega$$

i.e.  $u_h$  satisfies (2b), which implies that  $u = u_h$ .

Otherwise, it measures how “non-conforming”  $u_h$  is.

## Low frequencies

---

### The Prager-Synge theorem

# Prager-Syngé inequality

Assume that  $\boldsymbol{\sigma} \in \mathbf{H}(\text{div}, \Omega)$  satisfies

$$\boldsymbol{\sigma} \cdot \mathbf{n} = 0 \text{ on } \Gamma_A, \quad \nabla \cdot \boldsymbol{\sigma} = \mu f_h \text{ in } \Omega.$$

$\boldsymbol{\sigma} = -\mathbf{A}\nabla u$  is one example.

Then

$$\begin{aligned} a(e_h, v) &= (\mu f_h, v)_\Omega - (\mathbf{A}\nabla u_h, \nabla v)_\Omega \\ &= (\nabla \cdot \boldsymbol{\sigma}, v)_\Omega - (\mathbf{A}\nabla u_h, \nabla v)_\Omega \\ &= -(\boldsymbol{\sigma} + \mathbf{A}\nabla u_h, \nabla v)_\Omega. \end{aligned}$$

## Prager-Syngé inequality

$$|a(e_h, v)| \leq \|\mathbf{A}^{-1}\boldsymbol{\sigma} + \nabla u_h\|_{\mathbf{A}, \Omega} \|\nabla v\|_{\mathbf{A}, \Omega}$$



# Upper bound via equilibrated flux

Recall that whenever  $\nabla \cdot \boldsymbol{\sigma} = \mu f_h$  in  $\Omega$  and  $\boldsymbol{\sigma} \cdot \mathbf{n} = 0$  on  $\Gamma_A$

$$|a(\mathbf{e}_h, v)| \leq \|\mathbf{A}^{-1} \boldsymbol{\sigma} + \nabla u_h\|_{\mathbf{A}, \Omega} \|\nabla v\|_{\mathbf{A}, \Omega} \quad \forall v \in H_{\Gamma_D}^1(\Omega).$$

Picking in particular  $v = \mathbf{e}_h$ , we have

$$\|\nabla \mathbf{e}_h\|_{\mathbf{A}, \Omega}^2 = |a(\mathbf{e}_h, \mathbf{e}_h)| \leq \|\mathbf{A}^{-1} \boldsymbol{\sigma} + \nabla u_h\|_{\mathbf{A}, \Omega} \|\nabla \mathbf{e}_h\|_{\mathbf{A}, \Omega}.$$

## Guaranteed upper bound

$$\|\nabla \mathbf{e}_h\|_{\mathbf{A}, \Omega} \leq \|\mathbf{A}^{-1} \boldsymbol{\sigma} + \nabla u_h\|_{\mathbf{A}, \Omega}$$



W. Prager and J.L. Synge, *Quart. Appl. Math.*, 1947.

# Low frequencies

---

## Practical flux construction

# Practical flux construction

Equilibrated fluxes satisfy  $\nabla \cdot \boldsymbol{\sigma} = \mu f_h$  in  $\Omega$  and  $\boldsymbol{\sigma} \cdot \mathbf{n} = 0$  on  $\Gamma_A$ .  
They readily provide guaranteed error bounds.

Here, because  $\mu f_h \in \mathbb{P}_p(\mathcal{T}_h)$  we can actually find discrete fluxes  $\boldsymbol{\sigma}_h$ .

The correct tool to do that is the Raviart–Thomas finite element space

$$\mathbf{W}_h := \left\{ \mathbf{w}_h \in \mathbf{H}_{\Gamma_A}(\text{div}, \Omega) \mid \mathbf{w}_h|_K \in [\mathbb{P}_p(K)]^3 + \mathbf{x}\mathbb{P}_{p-1}(K) \right\}.$$



P.A. Raviart and J.M. Thomas, 1977.

## How to select the flux?

We are going to select a discrete flux  $\sigma_h \in \mathbf{W}_h$ .

The “equilibration” constraint on the flux is

$$\nabla \cdot \sigma_h = \mu f_h \text{ in } \Omega.$$

The estimate the flux provides us is

$$\|\nabla e_h\|_{\mathbf{A},\Omega} \leq \|\mathbf{A}^{-1} \sigma_h + \nabla u_h\|_{\mathbf{A},\Omega}.$$

# Optimal discrete flux

Recall that  $\boldsymbol{\sigma}_h \in \mathbf{W}_h$  has to satisfy

$$\nabla \cdot \boldsymbol{\sigma}_h = \mu f_h \text{ in } \Omega$$

and gives us

$$\|\nabla e_h\|_{\mathbf{A},\Omega} \leq \|\mathbf{A}^{-1} \boldsymbol{\sigma}_h + \nabla u_h\|_{\mathbf{A},\Omega}.$$

The “optimal” choice is then

$$\boldsymbol{\sigma}_h := \arg \min_{\substack{\boldsymbol{\tau}_h \in \mathbf{W}_h \\ \nabla \cdot \boldsymbol{\tau}_h = \mu f_h \text{ in } \Omega}} \|\mathbf{A}^{-1} \boldsymbol{\tau}_h + \nabla u_h\|_{\mathbf{A},\Omega}.$$

# How to compute it and how expensive it is?

After introducing a Lagrange multiplier, there exists a unique pair  $(\boldsymbol{\sigma}_h, \mathbf{q}_h) \in \mathbf{W}_h \times \mathbb{P}_p(\mathcal{T}_h)$  such that

$$\begin{cases} (\mathbf{A}^{-1}\boldsymbol{\sigma}_h, \mathbf{w}_h)_\Omega + (\mathbf{q}_h, \nabla \cdot \mathbf{w}_h)_\Omega &= -(\nabla u_h, \mathbf{w}_h)_\Omega & \forall \mathbf{w}_h \in \mathbf{W}_h, \\ (\nabla \cdot \boldsymbol{\sigma}_h, r_h) &= (\mu f_h, r_h) & \forall r_h \in \mathbb{P}_p(\mathcal{T}_h). \end{cases}$$

This square linear system can be solved to compute the optimal flux.

Unfortunately, it is more expensive than solving the original problem, so we avoid that in practice by using a localization trick.



P. Destuynder and B. Métivet, *Math. Comp.*, 1999.

At the continuous level, the ideal flux is  $\boldsymbol{\sigma} := -\mathbf{A}\nabla u$ .

A characterization is

$$\boldsymbol{\sigma} = \arg \min_{\substack{\boldsymbol{\tau} \in \mathbf{H}_{\mathbf{r}_A}(\text{div}, \Omega) \\ \nabla \cdot \boldsymbol{\tau} = \mu f_h \text{ in } \Omega}} \|\mathbf{A}^{-1}\boldsymbol{\tau} + \nabla u\|_{\mathbf{A}, \Omega}.$$

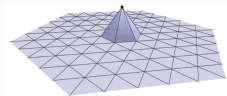
The “optimal” flux directly mimicks this definition at the discrete level

$$\boldsymbol{\sigma}_h := \arg \min_{\substack{\boldsymbol{\tau}_h \in \mathbf{W}_h \\ \nabla \cdot \boldsymbol{\tau}_h = \mu f_h \text{ in } \Omega}} \|\mathbf{A}^{-1}\boldsymbol{\tau}_h + \nabla u_h\|_{\mathbf{A}, \Omega}.$$

# Localization

Consider the set of “hat functions”  $\{\psi^a\}_{a \in \mathcal{V}_h}$  of the mesh. We then have

$$\sum_{a \in \mathcal{V}_h} \psi^a = 1.$$



The ideal flux  $\boldsymbol{\sigma} := -\mathbf{A}\nabla u$  can be decomposed as

$$\boldsymbol{\sigma} = \sum_{a \in \mathcal{V}_h} \boldsymbol{\sigma}^a, \quad \boldsymbol{\sigma}^a = -\psi^a \mathbf{A}\nabla u.$$

Easy computations show that

$$\boldsymbol{\sigma}^a \cdot \mathbf{n} = 0 \text{ on } \partial\omega^a \quad \nabla \cdot \boldsymbol{\sigma}^a = \psi^a \mu f_h - \mathbf{A}\nabla u \cdot \nabla \psi^a \text{ in } \omega^a$$



We have shown that

$$\sigma = \sum_{a \in \mathcal{V}_h} \sigma^a$$

and

$$\sigma^a = \arg \min_{\substack{\tau \in H_0(\text{div}, \omega^a) \\ \nabla \cdot \tau = \psi^a \mu f_h - \mathbf{A} \nabla u \cdot \nabla \psi^a \text{ in } \omega^a}} \|\mathbf{A}^{-1} \tau + \nabla u\|_{\mathbf{A}, \omega^a}.$$

We can mimick that on the discrete level!

# Localization

We thus set

$$\boldsymbol{\sigma}_h := \sum_{a \in \mathcal{V}_h} \boldsymbol{\sigma}_h^a$$

with

$$\boldsymbol{\sigma}_h^a := \arg \min_{\substack{\boldsymbol{\tau}_h \in \mathbf{H}_0(\operatorname{div}; \omega^a) \cap \mathbf{W}_h \\ \nabla \cdot \boldsymbol{\tau}_h = \psi^a \mu f_h - \mathbf{A} \nabla u_h \cdot \nabla \psi^a \text{ in } \omega^a}} \|\mathbf{A}^{-1} \boldsymbol{\tau}_h + \nabla u_h\|_{\mathbf{A}, \omega^a}.$$

The compatibility condition

$$(\psi^a \mu f_h - \mathbf{A} \nabla u_h \cdot \nabla \psi^a, 1)_{\omega^a} = (\mu f_h, \psi^a)_{\Omega} - (\mathbf{A} \nabla u_h, \nabla \psi^a)_{\Omega} = 0$$

holds true since  $u_h$  is the discrete solution and  $\psi^a \in V_h$ .

# Summary of the localization process

Step 1: solve a set of small, uncoupled linear systems

$$\sigma_h^a := \arg \min_{\substack{\tau_h \in \mathbf{H}_0(\operatorname{div}, \omega^a) \cap \mathbf{W}_h \\ \nabla \cdot \tau_h = \psi^a \mu_{f_h} - \mathbf{A} \nabla u_h \cdot \nabla \psi^a \text{ in } \omega^a}} \|\mathbf{A}^{-1} \tau_h + \nabla u_h\|_{\mathbf{A}, \omega^a}.$$

Step 2: assemble these local contributions

$$\sigma_h := \sum_{a \in \mathcal{V}_h} \sigma_h^a.$$

Step 3: compute the estimator

$$\eta := \|\mathbf{A}^{-1} \sigma_h + \nabla u_h\|_{\mathbf{A}, \Omega}.$$

Step 4: enjoy the guaranteed estimate

$$\|\nabla e_h\|_{\mathbf{A}, \Omega} \leq \eta.$$

**Low frequencies**



**Efficiency**

For time reasons, I will not give any proofs, but we can show that

$$\eta \leq C_{\text{eff}} \|\nabla e_h\|_{\mathbf{A}, \Omega},$$

where  $C_{\text{eff}}$  only depends on:

- the “flatness” of the tetrahedra in the mesh,
- the “contrasts” in the coefficients.

A nice aspect is that  $C_{\text{eff}}$  does not depend on  $p$ .



P. Braess, V. Pillwein and J. Schöberl, *CMAME*, 2009.



A. Ern and M. Vohralík, *Math. Comp.*, 2020.

# Low frequencies

---

## Takeaways

# Takeaways

An equilibrated flux is an object  $\boldsymbol{\sigma} \in \mathbf{H}(\text{div}, \Omega)$  such that

$$\boldsymbol{\sigma} \cdot \mathbf{n} = 0 \text{ on } \Gamma_A \quad \nabla \cdot \boldsymbol{\sigma} = \mu f_h \text{ in } \Omega.$$

There exist efficient algorithms to build a discrete flux  $\boldsymbol{\sigma}_h \in \mathbf{W}_h$ .

The *guaranteed* error estimate

$$\|\nabla e_h\|_{\mathbf{A}, \Omega} \leq \eta$$

holds true with

$$\eta := \|\mathbf{A}^{-1} \boldsymbol{\sigma}_h + \nabla u_h\|_{\mathbf{A}, \Omega}.$$

This bound cannot be too loose:

$$\eta \leq C_{\text{eff}} \|\nabla e_h\|_{\mathbf{A}, \Omega}.$$

## High frequencies

---



# The high frequency case

Back to our original problem

$$\begin{cases} -\omega^2 \mu u - \nabla \cdot (\mathbf{A} \nabla u) = \mu f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \mathbf{A} \nabla u \cdot \mathbf{n} - i\omega \gamma u = 0 & \text{on } \Gamma_A. \end{cases}$$

We introduce

$$b(\mathbf{u}, v) := -\omega^2 (\mu \mathbf{u}, v)_\Omega - i\omega (\gamma \mathbf{u}, v)_{\Gamma_A} + (\mathbf{A} \nabla \mathbf{u}, \nabla v)_\Omega.$$

# Energy norm and lack of coercivity

We will consider the “balanced” norm

$$\|v\|_{\omega, \Omega}^2 := \omega^2 \|v\|_{\mu, \Omega}^2 + \|\nabla v\|_{\mathbf{A}, \Omega}^2.$$

The sesquilinear form  $b$  is not coercive.

Instead we have the “Gårding” inequality

$$\operatorname{Re} b(v, v) = \|\nabla v\|_{\mathbf{A}, \Omega}^2 - \omega^2 \|v\|_{\mu, \Omega}^2 = \|v\|_{\omega, \Omega}^2 - 2\omega^2 \|v\|_{\mu, \Omega}^2.$$

# High frequencies

---

Flux equilibration

## Definition of an equilibrated flux

Letting  $\boldsymbol{\sigma} := -\mathbf{A}\nabla u$ , we have

$$\boldsymbol{\sigma} \cdot \mathbf{n} = -i\omega\gamma u \text{ on } \Gamma_A \quad \nabla \cdot \boldsymbol{\sigma} = \mu f_h + \omega^2 \mu u \text{ in } \Omega.$$

Hence, natural requirements for  $\boldsymbol{\sigma}_h$  are

$$\boldsymbol{\sigma}_h \cdot \mathbf{n} = -i\omega\gamma u_h \text{ on } \Gamma_A \quad \nabla \cdot \boldsymbol{\sigma}_h = \mu f_h + \omega^2 \mu u_h \text{ in } \Omega.$$

The “standard” reconstruction algorithms directly extend.

# Prager-Synge inequality

Let  $v \in H_{\Gamma_D}^1(\Omega)$ . We have

$$\begin{aligned} b(\mathbf{e}_h, v) &= (\mu \mathbf{f}_h, v)_\Omega - b(\mathbf{u}_h, v) \\ &= (\mu \mathbf{f}_h + \omega^2 \mu \mathbf{u}_h, v)_\Omega + i\omega(\gamma \mathbf{u}_h, v)_{\Gamma_A} - (\mathbf{A} \nabla \mathbf{u}_h, \nabla v)_\Omega \\ &= (\nabla \cdot \boldsymbol{\sigma}_h, v)_\Omega - (\boldsymbol{\sigma}_h \cdot \mathbf{n}, v)_{\Gamma_A} - (\mathbf{A} \nabla \mathbf{u}_h, \nabla v)_\Omega \\ &= -(\boldsymbol{\sigma}_h + \mathbf{A} \nabla \mathbf{u}_h, \nabla v)_\Omega. \end{aligned}$$

## Prager-Synge inequality

$$|b(\mathbf{e}_h, v)| \leq \eta \|\nabla v\|_{\mathbf{A}, \Omega} \quad \forall v \in H_{\Gamma_D}^1(\Omega)$$

So far, so good!

# What's the matter?

Here, we do not have

$$\|\nabla e_h\|_{\mathbf{A},\Omega}^2 \leq |b(e_h, e_h)|,$$

which is a major issue!

Instead, we only have the “Gårding” inequality

$$\operatorname{Re} b(e_h, e_h) \geq \|e_h\|_{\omega,\Omega}^2 - 2\omega^2 \|e_h\|_{\mu,\Omega}^2.$$

# High frequencies

---

A coarse error estimate

## Stability constant

For  $g \in L^2(\Omega)$ , let  $\mathcal{S}^*g$  denote the unique element of  $H_{\Gamma_D}^1(\Omega)$  such that

$$b(w, \mathcal{S}^*g) = 2\omega^2(\mu w, g)_\Omega \quad \forall w \in H_{\Gamma_D}^1(\Omega)$$

and let

$$\mathcal{C}_{\text{st}} := \frac{1}{\omega} \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu, \Omega} = 1}} \|\nabla(\mathcal{S}^*g)\|_{\mathbf{A}, \Omega}.$$

$\mathcal{C}_{\text{st}}$  is the best constant such that

$$\|\nabla(\mathcal{S}^*g)\|_{\mathbf{A}, \Omega} \leq \mathcal{C}_{\text{st}}\omega \|g\|_{\mu, \Omega} \quad \forall g \in L^2(\Omega).$$

It is closely related to resolvent estimates.



## Making up for the lack of coercivity

By definition, we have

$$b(w, \mathcal{S}^* e_h) = 2\omega^2(\mu w, e_h) \quad \forall w \in H_{\Gamma_D}^1(\Omega).$$

Hence, in particular,

$$b(e_h, \mathcal{S}^* e_h) = 2\omega^2 \|e_h\|_{\mu, \Omega}^2,$$

which is exactly the “bad” term the Gårding inequality:

$$\operatorname{Re} b(e_h, e_h) = \|e_h\|_{\omega, \Omega}^2 - 2\omega^2 \|e_h\|_{\mu, \Omega}^2.$$

# Making up for the lack of coercivity

Using Prager-Syngé inequality, we have

$$\|e_h\|_{\omega,\Omega}^2 = \operatorname{Re} b(e_h, e_h + \mathcal{I}^* e_h) \leq \eta \|\nabla(e_h + \mathcal{I}^* e_h)\|_{\mathbf{A},\Omega}.$$

It follows that

$$\begin{aligned} \|e_h\|_{\omega,\Omega}^2 &\leq \eta (\|\nabla e_h\|_{\mathbf{A},\Omega} + \|\nabla(\mathcal{I}^* e_h)\|_{\mathbf{A},\Omega}) \\ &\leq \eta (\|\nabla e_h\|_{\mathbf{A},\Omega} + \mathcal{C}_{\text{st}} \omega \|e_h\|_{\mu,\Omega}) \\ &\leq \eta \max(1, \mathcal{C}_{\text{st}}) \|e_h\|_{\omega,\Omega}, \end{aligned}$$

and

$$\|e_h\|_{\omega,\Omega} \leq \max(1, \mathcal{C}_{\text{st}}) \eta.$$

# Coarse error estimate

We obtained the following error estimate:

## Coarse error estimate

$$\|e_h\| \leq \max(1, \mathcal{C}_{\text{st}})\eta$$

$\mathcal{C}_{\text{st}}$  is the best constant such that:

## Stability constant

$$\|\nabla(\mathcal{I}^*g)\|_{\mathbf{A},\Omega} \leq \mathcal{C}_{\text{st}}\omega \|g\|_{\mu,\Omega} \quad \forall g \in L^2(\Omega).$$

# High frequencies

---

Efficiency

We can show that

$$\eta \leq C_{\text{eff}} \left( 1 + \max_{K \in \mathcal{T}_h} \frac{\omega h_K}{c_{\min K} \rho} \right) \|e_h\|_{\omega, \Omega},$$

where  $c_{\min K}$  is the wavespeed in the element  $K$ .

For any reasonable discretization, we have

$$\frac{\omega h_K}{c_{\min K} \rho} \leq 1,$$

so that in practice

$$\eta \leq C_{\text{eff}} \|e_h\|_{\omega, \Omega},$$

where  $C_{\text{eff}}$  only depends on the elements “flatness” and the contrasts.



T. Chaumont-Frelet, A. Ern and M. Vohralík, *Numer. Math.*, 2021.



W. Dörfler, S. Sauter, *Comput. Meth. Appl. Math.*, 2013.

# The problem with the coarse error estimate

Recall that

$$\eta \leq C_{\text{eff}} \|e_h\|_{\omega, \Omega} \quad \|e_h\|_{\omega, \Omega} \leq \max(1, \mathcal{C}_{\text{st}}) \eta.$$

We have  $C_{\text{eff}} \simeq 1$  and  $\mathcal{C}_{\text{st}} \gtrsim \omega \ell_{\Omega} / c_{\text{min}}$ , so that

$$\eta \lesssim \|e_h\|_{\omega, \Omega} \lesssim \frac{\omega \ell_{\Omega}}{c_{\text{min}}} \eta.$$

In practice, the coarse error estimate will largely overestimate the error in the high frequency regime.

**High frequencies**

---

**Sharp error estimate**

# The approximation factor

We introduce

$$\mathcal{C}_{\text{ap}} := \frac{1}{\omega} \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu, \Omega} = 1}} \min_{v_h \in V_h} \|\nabla(\mathcal{I}^* g - v_h)\|_{\mathbf{A}, \Omega}.$$

## Approximability

For all  $g \in L^2(\Omega)$ , there exists  $v_h^* \in V_h$  such that

$$\|\nabla(\mathcal{I}^* g - v_h^*)\|_{\mathbf{A}, \Omega} \leq \mathcal{C}_{\text{ap}} \omega \|g\|_{\mu, \Omega}.$$

This was called  $\eta$  in Markus' talk.



# It's better than the stability constant!

Recall that

$$\mathcal{C}_{\text{ap}} := \max_{\substack{\mathbf{g} \in L^2(\Omega) \\ \|\mathbf{g}\|_{\mu, \Omega} = 1}} \min_{\mathbf{v}_h \in V_h} \omega \|\mathcal{I} \mathbf{g} - \mathbf{v}_h\|_{\omega, \Omega}$$

since we can take  $\mathbf{v}_h = 0$ , we have

$$\mathcal{C}_{\text{ap}} \leq \max_{\substack{\mathbf{g} \in L^2(\Omega) \\ \|\mathbf{g}\|_{\mu, \Omega} = 1}} \omega \|\mathcal{I} \mathbf{g}\|_{\omega, \Omega} =: \mathcal{C}_{\text{st}}.$$

Besides, standard approximability results for FEM show that

$$\mathcal{C}_{\text{ap}} \leq \mathcal{C} \left( \frac{1}{p} \frac{h}{\ell_{\Omega}} \right)^s$$

for some  $s > 0$ , so that  $\mathcal{C}_{\text{ap}} \rightarrow 0$ .

# Using Galerkin orthogonality

Recall that

$$\|e_h\|_{\omega, \Omega}^2 = \operatorname{Re} b(e_h, e_h + \mathcal{I}^* e_h).$$

By Galerkin orthogonality, we have

$$\begin{aligned} \|e_h\|_{\omega, \Omega}^2 &= \operatorname{Re} b(e_h, e_h) + \operatorname{Re} b(e_h, \mathcal{I}^* e_h) \\ &= \operatorname{Re} b(e_h, e_h) + \operatorname{Re} b(e_h, \mathcal{I}^* e_h - v_h^*) \\ &\leq \eta (\|\nabla e_h\|_{\mathbf{A}, \Omega} + \|\nabla(\mathcal{I}^* e_h - v_h^*)\|_{\mathbf{A}, \Omega}) \\ &\leq \eta \max(1, \mathcal{C}_{\text{ap}}) \|e_h\|_{\omega, \Omega}. \end{aligned}$$

# Sharp error estimate

## Sharp error estimate

$$\|e_h\|_{\omega, \Omega} \leq \max(1, \mathcal{C}_{\text{ap}})\eta.$$

## Approximation factor

$$\mathcal{C}_{\text{ap}} := \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu, \Omega} = 1}} \min_{v_h \in V_h} \omega \| \mathcal{I}g - v_h \|_{\omega, \Omega} \rightarrow 0.$$



T. Chaumont-Frelet, A. Ern and M. Vohralík, *Numer. Math.*, 2021.



W. Dörfler, S. Sauter, *Comput. Meth. Appl. Math.*, 2013.

# High frequencies

---

## Takeaways

# Takeways

The “equilibration” technology is the same than for low frequencies.

## Coarse error estimate

$$\|e_h\|_{\omega, \Omega} \leq \max(1, \mathcal{C}_{st}) \eta \quad \mathcal{C}_{st} \gtrsim \omega \ell_{\Omega} / c_{\min}$$

## Sharp error estimate

$$\|e_h\|_{\omega, \Omega} \leq \max(1, \mathcal{C}_{ap}) \eta \quad \mathcal{C}_{ap} \rightarrow 0$$

## Efficiency

$$\eta \leq C_{\text{eff}} \left( 1 + \frac{\omega h}{c_{\min} p} \right) \|e_h\|_{\omega, \Omega}$$

## Controlling the pre-factors

---

# Controlling the pre-factors

---

The stability constant  $\mathcal{C}_{\text{st}}$

# The stability constant

The stability constant is defined by

$$\mathcal{C}_{\text{st}} := \max_{\substack{\mathbf{g} \in L^2(\Omega) \\ \|\mathbf{g}\|_{\mu, \Omega} = 1}} \|\nabla(\mathcal{I}^* \mathbf{g})\|_{\mathbf{A}, \Omega}.$$

It is only related to the PDE, and independent of the numerical scheme.



# Qualitative behaviour

It is known that we have “at least”:

$$\mathcal{C}_{\text{st}} \gtrsim \frac{\omega l \Omega}{c_{\text{min}}}.$$

For non-trapping settings (the “easier” scenario), we have

$$\mathcal{C}_{\text{st}} \lesssim \frac{\omega l \Omega}{c_{\text{min}}}.$$

If strong trapping happens, “extreme” behaviors can occur

$$\mathcal{C}_{\text{st}} \gtrsim \exp\left(\alpha \frac{\omega l \Omega}{c_{\text{min}}}\right)$$

for “some” frequencies. For “most frequency”

$$\mathcal{C}_{\text{st}} \gtrsim \left(\frac{\omega l \Omega}{c_{\text{min}}}\right)^\beta.$$



D. Lafontaine, E.A. Spence and J. Wunsch, *Comm. Pure Appl. Math.* 2021

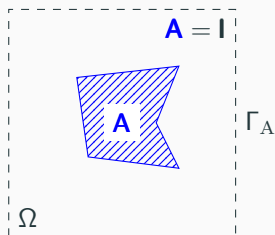
# Quantitative estimate for star-shaped non-trapping obstacles

$\Omega := (-\ell/2, \ell/2)^3$  is a cube centered at the origin.

$D \subset \Omega$  is star shaped with respect to the origin.

Assume that  $\gamma = 1$  and that

$\mu = 1$  and  $\mathbf{A} = \mathbf{I}$  in  $\Omega \setminus D$ .



Assume that  $\mu = \mu_D \geq 1$  and  $\mathbf{A} = \mathbf{A}_D \preceq \mathbf{I}$  in  $D$ .

This describes an obstacle made of a material with a “slow” wavespeed.

## Guaranteed upper bound

$$\mathcal{C}_{\text{st}} \leq 6 + \frac{3 + \sqrt{3}}{\sqrt{3}} \frac{\omega l_{\Omega}}{C_{\text{min}}}$$

The proof relies on a “Morawetz multiplier”:  
multiply the PDE by  $\mathbf{x} \cdot \nabla u$  and integrate by parts until it works!



C.S. Morawetz, *Comm. Pure Appl. Math.*, 1962.



J.M. Melenk, *PhD thesis*, 1995.



H. Barucq, T. Chaumont-Frelet and C. Gout, *Math. Comp.*, 2016.



T. Chaumont-Frelet, A. Ern, M. Vohralík, *Numer. Math.* 2021.

# Controlling the pre-factors

---

The approximation factor  $\mathcal{C}_{\text{ap}}$

# The approximation factor

The approximation factor is defined by

$$\mathcal{C}_{\text{ap}} := \max_{\substack{\mathbf{g} \in L^2(\Omega) \\ \|\mathbf{g}\|_{\mu, \Omega} = 1}} \min_{\mathbf{v}_h^* \in V_h} \|\nabla(\mathcal{I}^* \mathbf{g} - \mathbf{v}_h^*)\|_{\mathbf{A}, \Omega}.$$

It depends on both the PDE and the approximation space  $V_h$ .

Assuming that  $\mathbf{A} = \mathbf{I}$ ,  $\Omega$  is convex, and  $\mathcal{C}_{\text{st}}$  is known, we can control it.

## Idea one: explicit interpolation error



R. Arcangeli and J.L. Gout, RAIRO Numer. Anal., 1976.

If  $v \in H^2(\Omega)$ , let  $I_h^1 v \in V_h$  denotes its first-order Lagrange interpolant:

$$\|\nabla(v - I_h^1 v)\|_{\mathbf{A},\Omega} \leq \mathcal{C}_{\mathcal{T},i} h \|\nabla^2 v\|_{\Omega},$$

with a constant  $\mathcal{C}_{\mathcal{T},i}$  that is easily computable.

We then have

$$\begin{aligned} \mathcal{C}_{\text{ap}} &:= \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu,\Omega}=1}} \min_{v_h^* \in V_h} \|\nabla(\mathcal{S}^* g - v_h^*)\|_{\mathbf{A},\Omega} \\ &\leq \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu,\Omega}=1}} \|\nabla(\mathcal{S}^* g - I_h^1(\mathcal{S}^* g))\|_{\mathbf{A},\Omega} \\ &\leq \mathcal{C}_{\mathcal{T},i} h \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu,\Omega}=1}} \|\nabla^2(\mathcal{S}^* g)\|_{\Omega}. \end{aligned}$$

## Idea two: estimation of the Hessian norm



P. Grisvard, 1985.



T. Chaumont-Frelet, S. Nicaise and J. Tomezyk, *Comm. Pure Appl. Anal.*, 2020.

Because  $\Omega$  is convex and  $\gamma = 1$ , we have

$$\|\nabla^2(\mathcal{I}^*g)\|_{\Omega} \leq \|\Delta(\mathcal{I}^*g)\|_{\Omega}.$$

Then, we use the facts that

$$-\Delta(\mathcal{I}^*g) = 2\mu\omega^2g + \mu\omega^2\mathcal{I}^*g$$

and

$$\omega\|\mathcal{I}^*g\|_{\mu,\Omega} \leq 2\mathcal{C}_{\text{st}}\|g\|_{\mu,\Omega},$$

to show that

$$\|\nabla^2(\mathcal{I}^*g)\|_{\Omega} \leq 2\frac{\omega}{c_{\min}}(1 + \mathcal{C}_{\text{st}})\|g\|_{\mu,\Omega}.$$

# Explicit control of the approximation factor

Recall that

$$\mathcal{C}_{\text{ap}} \leq \mathcal{C}_{\mathcal{T},i} h \max_{\substack{g \in L^2(\Omega) \\ \|g\|_{\mu,\Omega}=1}} \|\nabla^2(\mathcal{I}^* g)\|_{\Omega}$$

and

$$\|\nabla^2(\mathcal{I}^* g)\|_{\Omega} \leq 2 \frac{\omega}{C_{\text{min}}} (1 + \mathcal{C}_{\text{st}}) \|g\|_{\mu,\Omega} \quad \forall g \in L^2(\Omega).$$

## Guaranteed bound

$$\mathcal{C}_{\text{ap}} \leq 2(1 + \mathcal{C}_{\mathcal{T},i}) \frac{\omega h}{C_{\text{min}}} \mathcal{C}_{\text{st}}$$



# Controlling the pre-factors

---

## Takeaways

# Takeaways

The estimator  $\eta$  needs to be “pre-factored” by  $\mathcal{C}_{\text{st}}$  or  $\mathcal{C}_{\text{ap}}$ .

The “qualitative” behaviors of both quantities are relatively well known.

The behaviour of  $\mathcal{C}_{\text{st}}$  is only dictated by the PDE.

Explicit bounds are available for non-trapping star-shaped obstacles.

The approximation factor  $\mathcal{C}_{\text{ap}}$  depends on the PDE and  $V_h$ .

When  $\mathbf{A} = \mathbf{I}$ ,  $\Omega$  is convex and  $\mathcal{C}_{\text{st}}$  is known, we can bound it nicely.

## Numerical illustrations

---

# Numerical illustrations

---

A validation experiment

# Propagation of a plane wave

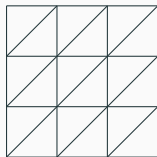
We consider the propagation of a plane wave in  $\Omega = (-1, 1)^2$

$$\begin{cases} -\omega^2 \mathbf{u} - \Delta \mathbf{u} = 0 & \text{in } \Omega, \\ \nabla \mathbf{u} \cdot \mathbf{n} - i\omega \mathbf{u} = \mathbf{g} & \text{on } \Gamma_A, \end{cases}$$

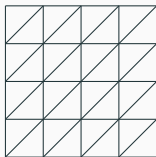
where

$$\mathbf{g} := \nabla \xi_\theta \cdot \mathbf{n} - i\omega \xi_\theta \quad \xi_\theta := e^{i\omega \mathbf{d} \cdot \mathbf{x}}$$

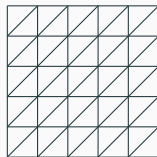
with  $\mathbf{d} := (\cos \theta, \sin \theta)$  and  $\theta = \pi/12$ . The solution is  $\mathbf{u} = \xi_\theta$ .



$$h = \sqrt{2} \times 2/3$$

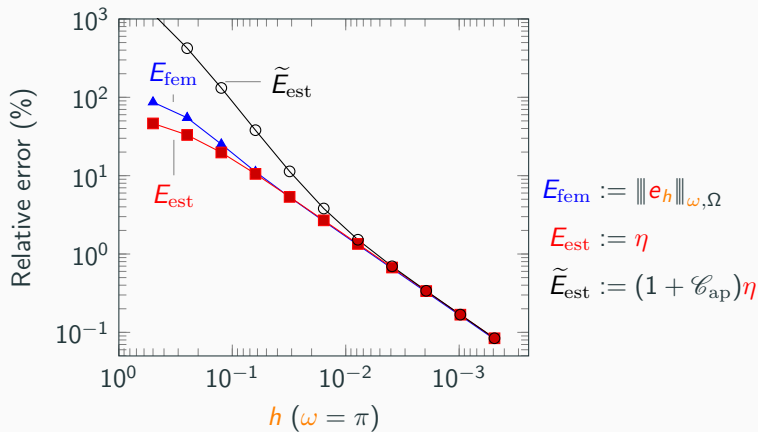


$$h = \sqrt{2} \times 1/2$$

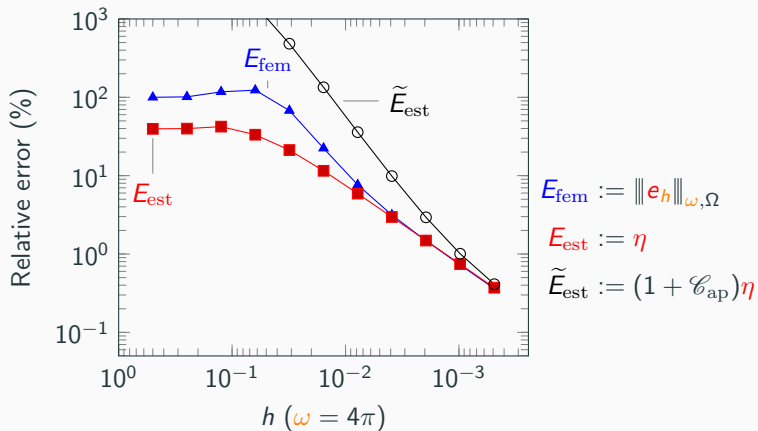


$$h = \sqrt{2} \times 2/5$$

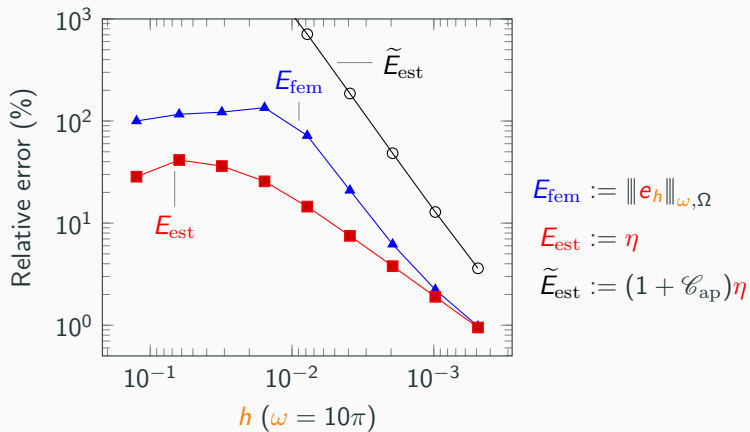
# Plane wave experiment $\rho = 1$ and $\omega = \pi$



# Plane wave experiment $\rho = 1$ and $\omega = 4\pi$

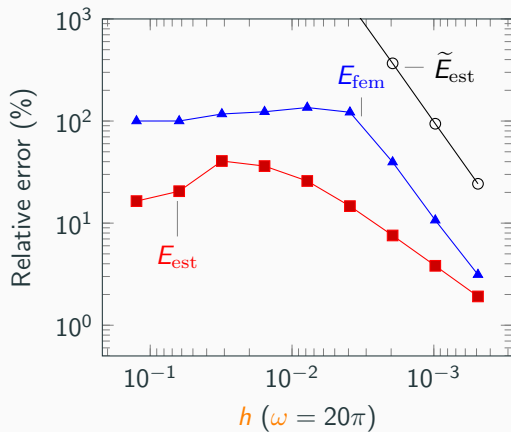


# Plane wave experiment $\rho = 1$ and $\omega = 10\pi$





# Plane wave experiment $p = 1$ and $\omega = 20\pi$

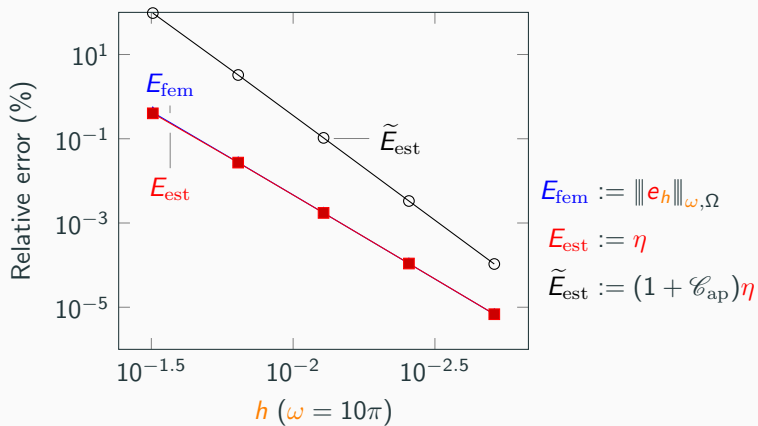


$$E_{\text{fem}} := \|e_h\|_{\omega, \Omega}$$

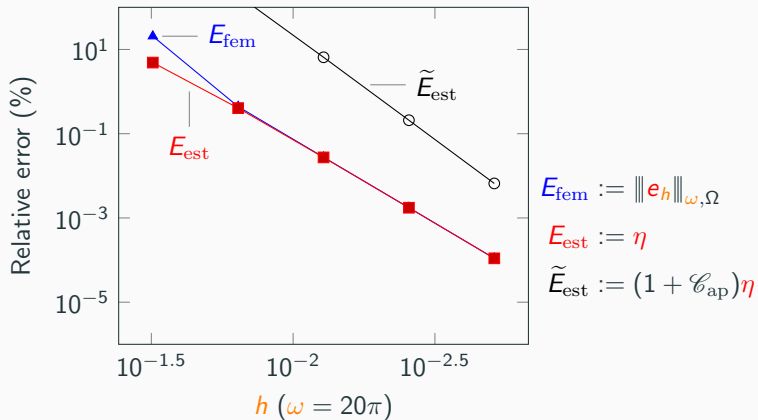
$$E_{\text{est}} := \eta$$

$$\tilde{E}_{\text{est}} := (1 + \mathcal{C}_{\text{ap}})\eta$$

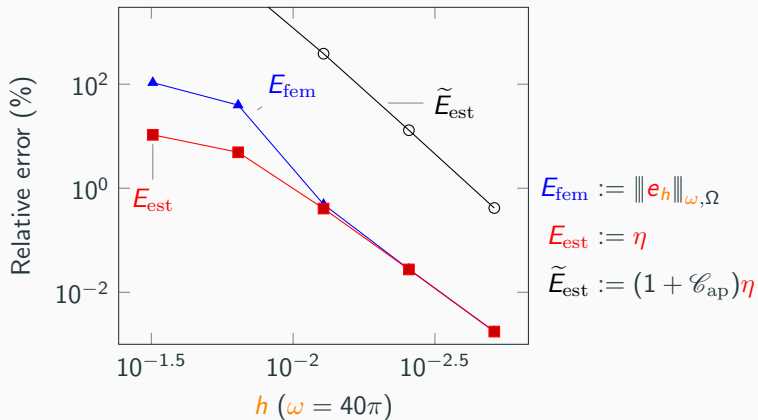
# Plane wave experiment $\rho = 4$ and $\omega = 10\pi$



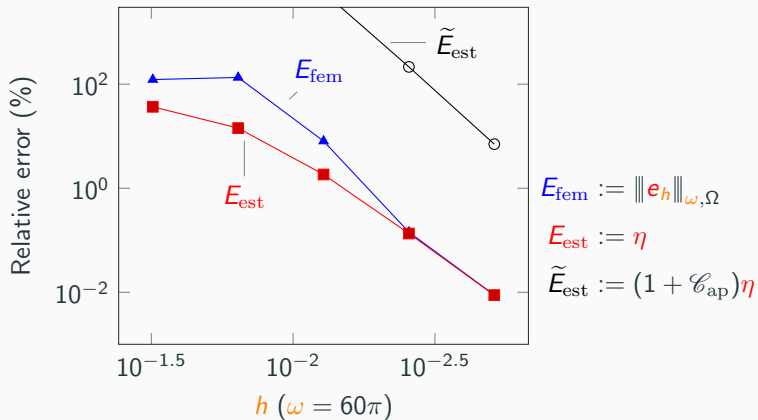
# Plane wave experiment $\rho = 4$ and $\omega = 20\pi$



# Plane wave experiment $p = 4$ and $\omega = 40\pi$



# Plane wave experiment $\rho = 4$ and $\omega = 60\pi$



# Numerical illustrations

---

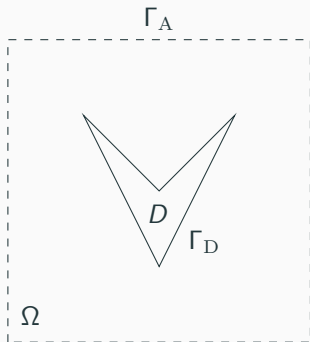
**A more realistic example**

# Scattering by an non-trapping obstacle

We now consider a scattering problem

$$\begin{cases} -\omega^2 \mathbf{u} - \Delta \mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = 0 & \text{on } \Gamma_D, \\ \nabla \mathbf{u} \cdot \mathbf{n} - i\omega \mathbf{u} = \mathbf{g} & \text{on } \Gamma_A, \end{cases}$$

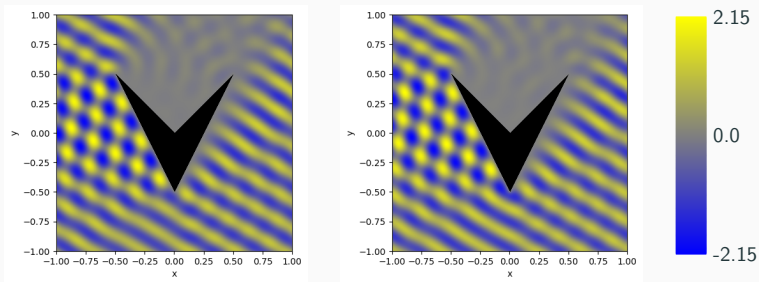
where again  $\mathbf{g} = \nabla \xi_\theta \cdot \mathbf{n} - i\omega \xi_\theta$ .



We fix the wavenumber  $\omega = 10\pi$  and employ  $\mathbb{P}_3$  elements.

We consider a sequence of meshes that are adaptively refined using  $\eta_K$ .

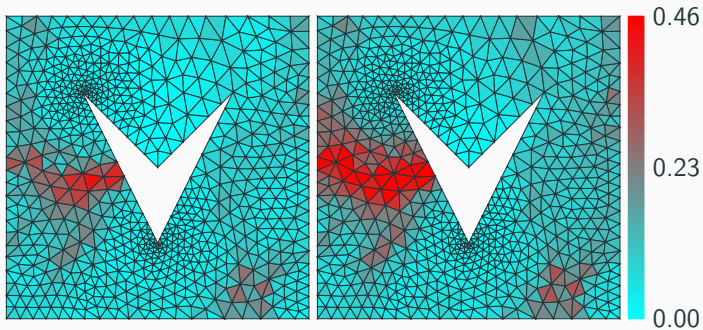
# Solution of the scattering problem



Real (left) and imaginary (right) parts of the solution

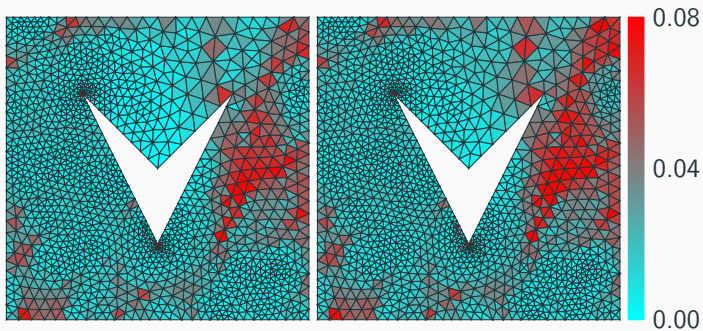


# Estimated error in mesh #1



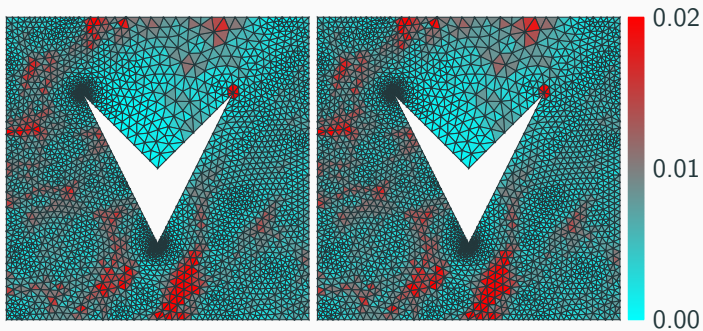
Estimator  $\eta_K$  (left) and elementwise error  $\|e_h\|_{\omega,K}$  (right)

## Estimated error in mesh #2



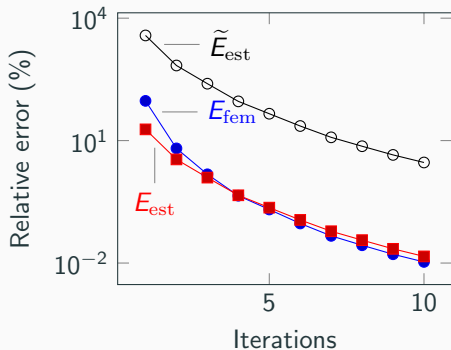
Estimator  $\eta_K$  (left) and elementwise error  $\|e_h\|_{\omega,K}$  (right)

## Estimated error in mesh #3



Estimator  $\eta_K$  (left) and elementwise error  $\|e_h\|_{\omega,K}$  (right)

# Behavior of the estimator through the adaptive procedure



$$E_{fem} := \|e_h\|_{\omega, \Omega}$$

$$E_{est} := \eta$$

$$\tilde{E}_{est} := (1 + \mathcal{C}_{st})\eta$$

Behaviors of the estimated and analytical errors in the adaptive procedure

## Concluding remarks

---

## Concluding remarks

---

Takeaways

# Takeaways

We construct an a posteriori error estimator  $\eta$  via flux equilibration.  
It directly provides guaranteed error estimates at low frequencies.

For high frequencies,  $\eta$  has to be pre-factored, either by  $\mathcal{C}_{\text{st}}$  and  $\mathcal{C}_{\text{ap}}$ .  
The estimates are asymptotically constant-free.

In specific situations, we can provide guaranteed bounds on  $\mathcal{C}_{\text{st}}$  and  $\mathcal{C}_{\text{ap}}$ .

There is still a long way toward fully reliable error estimation for high-frequency problems!



T. Chaumont-Frelet, A. Ern and M. Vohralík, *Numer. Math.*, 2021.

## **Concluding remarks**

---

**Extensions**



We can obtain guaranteed bounds on  $\mathcal{C}_{\text{st}}$  in weakly trapping geometry with “directional” Morawetz multiplier

$$(\mathbf{x}_d \mathbf{e}^d) \cdot \nabla u.$$



S.N. Chandler-Wilde et. al., *SIAM J. Math. Anal.*, 2020.



T. Chaumont-Frelet and E.A. Spence, *submitted*.



T. Chaumont-Frelet and Z. Kassali, *in progress*.

It is possible to obtain approximations for  $\mathcal{C}_{\text{ap}}$  in more general situations.

Everything I presented (painfully) extends to Maxwell's equations:



T. Chaumont-Frelet, A. Ern and M. Vohralík, *C.R. Math. Acad. Sci.*, 2020.



T. Chaumont-Frelet, A. Ern and M. Vohralík, *Math. Comp.*, 2021.



T. Chaumont-Frelet and M. Vohralík, *submitted*.



T. Chaumont-Frelet, *will be submitted soon!*

Thanks for your attention! :-)