Reduced order methods applied to aeroacoustic problems solved by integral equations

Fabien Casenave

UNIVERSITÉ PARIS-EST

THÈSE

présentée pour l'obtention du titre de

Docteur de l'Université Paris-Est

Spécialité : Mathématiques Appliquées

par **Fabien Casenave**

Sujet : Méthodes de réduction de modèles appliquées à des problèmes d'aéroacoustique résolus par équations intégrales

Soutenue le 05 12 2013
devant le jury composé de :

Rapporteurs :	Patrick Joly
	Anthony Patera
Examinateurs :	Martin Costabel
	Yvon Maday
	Anthony Nouy
	Isabelle Terrasse
Directeurs de thèse :	Alexandre Ern
	Tony Lelièvre

École Doctorale : Mathématiques et Sciences et Technologies de l'Information et de la Communication

le 11 décembre 2013

Méthodes de réduction de modèles appliquées à des problèmes d'aéroacoustique résolus par équations intégrales

Résumé :
Cette thèse s'articule autour de deux thématiques : les méthodes numériques pour la propagation d'ondes acoustiques sous écoulement et les méthodes de réduction de modèles. Dans la première thématique, nous développons une méthode de couplage d'éléments finis et d'éléments de frontière pour résoudre l'équation d'Helmholtz convectée, lorsque l'écoulement est uniforme à l'extérieur d'un domaine borné. En particulier, nous proposons une formulation bien posée à toutes les fréquences de la source. Dans la deuxième thématique, nous proposons une solution au problème classique d'accumulation d'arrondis machine qui survient en calculant l'estimateur d'erreur a posteriori dans la méthode des bases réduites. Par ailleurs, nous proposons une méthode non intrusive pour calculer une approximation sous forme séparée des systèmes linéaires résultant de l'approximation en dimension finie de problèmes aux limites dépendant d'un ou plusieurs paramètres.

Mots-clés : Équations intégrales, Aéroacoustique, Équation d'Helmholtz convectée, Couplage éléments finis et éléments de frontière, Formulation intégrale de type équation combinée, Méthode des bases réduites, Approximation certifiée, Estimateur d'erreur a posteriori, Méthode d'interpolation empirique, Approximation non intrusive

---

Reduced order methods applied to aeroacoustic problems solved by integral equations

Abstract :
This thesis has two topics : numerical methods for acoustic wave propagation in a flow and reduced order methods. In the first topic, we develop a coupled finite element and boundary element method to solve the convected Helmholtz equation, when the flow is uniform outside a bounded domain. In particular, we propose a formulation that is well-posed at all the frequencies of the source. In the second topic, we propose a solution to the classical problem of round-off error accumulation that occurs when computing the a posteriori error bound in the reduced basis method. Furthermore, we propose a nonintrusive method for the approximation, in a separated representation form, of linear systems resulting from the finite-dimensional approximation of boundary-value problems depending on one or several parameters.

Keywords : Integral equations, Aeroacoustics, Convected Helmholtz equation, Finite element and boundary element coupling, Combined fields integral equations, Reduced basis method, Certified approximation, A posteriori error bound, Empirical interpolation method, Nonintrusive approximation

Table des matières

1

Introduction générale

1.1 Contexte industriel

La réduction du bruit en aéronautique constitue un enjeu industriel considérable. En effet, face à l'augmentation substantielle du trafic aérien, les normes de certification acoustique sont fortement renforcées afin de protéger les riverains des nuisances sonores essentiellement occasionnées par les avions aux abords des aéroports. Cette volonté de réduire le bruit d'origine aéroportuaire s'accompagne d'un besoin en outil numérique permettant de prévoir la propagation dans un fluide du son produit par un avion avant même sa construction. Nous nous intéressons en particulier au bruit généré par les turboréacteurs. Or, aujourd'hui, peu d'outils numériques et de méthodes sont disponibles pour la simulation de la propagation du son dans un fluide. Il existe des méthodes issues de la mécanique des fluides basées sur les équations de Navier-Stokes. Cependant, ces méthodes sont lourdes en terme de puissance de calcul et sont peu adaptées (difficulté pour extraire la contribution de l'acoustique). Une autre possibilité consiste à utiliser les équations d'Euler linéarisées. Ces équations se ramènent, dans le cas d'un écoulement nul, à la classique équation d'Helmholtz, qui peut être résolue par équation intégrale. C'est cette deuxième approche que nous choisissons d'exploiter et d'enrichir afin de pouvoir prendre en compte la convection de l'écoulement dans lequel se propagent les ondes.

1.1.1 Bruit généré par un avion

Le bruit généré par un avion est créé soit par la turbulence de l'écoulement autour de certaines parties de l'avion, soit directement par certaines pièces mécaniques, voir figure 1.1.

La génération du bruit par un turboréacteur est liée à son fonctionnement, en particulier aux quatre phases qui ont lieu simultanément : l'admission, la compression, la combustion et la détente. Chacune de ces quatre phases est à l'origine d'une contribution dans la perturbation acoustique totale créée par le turboréacteur. Lors de l'admission au niveau de l'entrée d'air, le bruit de soufflante est généré. Viennent ensuite le bruit des compresseurs et le bruit généré par la combustion. La poussée contribue au bruit de jet. Enfin, la turbine qui assure la rotation des arbres du turboréacteur est à l'origine du bruit de turbine. Des mesures expérimentales montrent que, lors de l'atterrissage et du décollage, pour les turboréacteurs actuels à double flux, la perturbation prépondérante est le bruit de soufflante, voir figure 1.2.

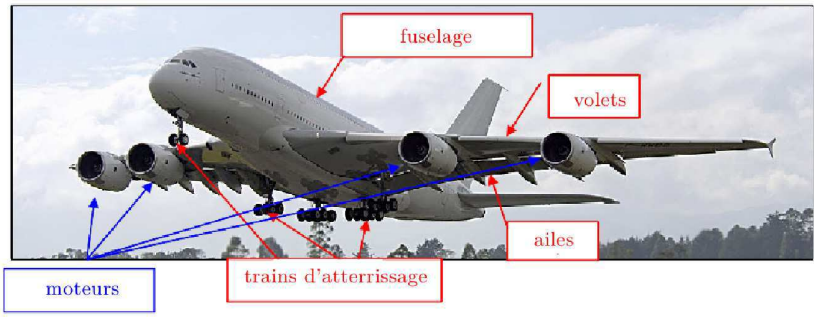

Fig. 1.1. Sources du bruit généré par un avion (image fournie par EADS)

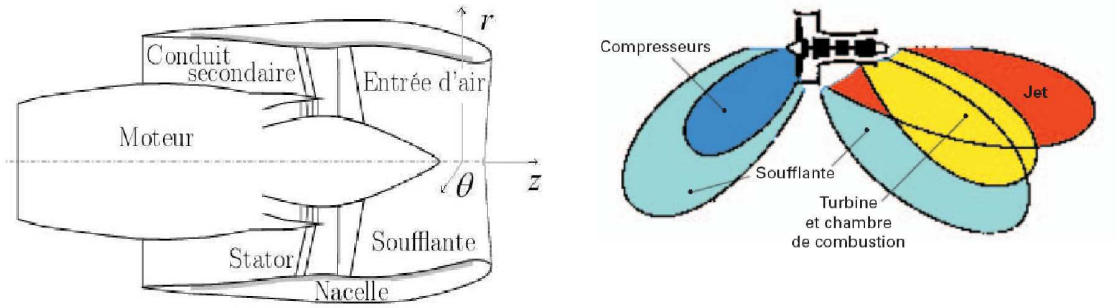

Fig. 1.2. Gauche : schéma d'une nacelle de turboréacteur (image issue de [40]), droite : principales composantes du bruit généré par un turboréacteur (image issue de [67]).

Lors de la phase d'admission, l'entrée d'air est assurée par la rotation des pales situées à l'avant du turboréacteur appelées rotor. Le spectre du bruit de soufflante se caractérise par la présence de raies harmoniques (voir figure 1.3) à des fréquences multiples de la fréquence de rotation des pales du rotor. Il s'agit du bruit tonal, appelé aussi bruit de raie.

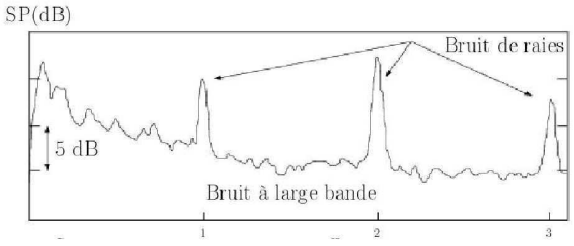

Fig. 1.3. Spectre du bruit de soufflante en régime subsonique (image issue de [40])

D'un point de vue acoustique, le moteur est modélisé par un cylindre semi-infini et la donnée de la source acoustique est décomposée sur une base de modes de propagation dans cette cavité.

1.1.2 Description de l'écoulement autour du turboréacteur

L'étude de la génération du bruit par un turboréacteur en présence d'un écoulement montre qu'il est possible de découper le domaine occupé par le fluide en zones correspondant chacune à un régime d'écoulement différent (écoulement uniforme, potentiel ou turbulent). Il en résulte que chaque zone est caractérisée par son propre type de problème acoustique, voir figure 1.4 et [40] pour une description détaillée.

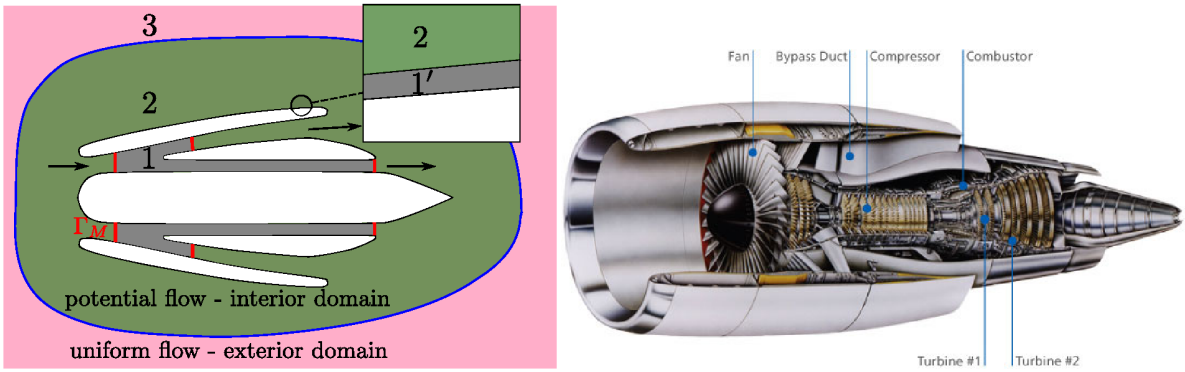

Fig. 1.4. Gauche : représentation schématique d'un turboréacteur, avec les trois zones de comportement identifiées (1,1' : écoulement complexe, 2 : écoulement potentiel, 3 : écoulement uniforme), droite : coupe réaliste d'un turboréacteur.

Les zones 1 et $1'$ correspondent à un écoulement complexe dans lequel il y a création de bruit dit aérodynamique. La zone 1 concerne l'intérieur du moteur et le jet à l'arrière du turboréacteur. La zone $1'$ désigne la couche limite. L'écoulement vérifie dans ces zones les équations de Navier-Stokes dans lesquelles la viscosité n'est pas négligeable. Il faut en outre une équation supplémentaire issue de la thermodynamique reliant la température et l'entropie. L'acoustique est non linéaire et il y a des interactions entre le fluide et les phénomènes acoustiques. Ensuite, dans la zone 2, la viscosité et les effets de conduction peuvent être négligés, et l'air peut être modélisé par un gaz parfait. Les équations de Navier-Stokes se simplifient alors pour aboutir aux équations d'Euler et l'équation de la thermodynamique évoquée plus haut traduit désormais une condition d'isentropie pour le fluide considéré comme parfait. Ces zones sont le siège de phénomènes acoustiques linéaires découplés de l'écoulement porteur. Nous considérons que l'acoustique est une perturbation linéaire faible devant l'écoulement porteur. Enfin, nous faisons, dans ces zones, l'hypothèse d'un écoulement irrotationnel et potentiel. De même, la perturbation acoustique est supposée potentielle. Pour finir, la zone 3 se situe assez loin du turboréacteur pour que l'écoulement puisse y être considéré comme uniforme. Pour la simulation d'ondes acoustiques autour du turboréacteur de la figure 1.4, nous supposons que le bruit généré dans la zone 1 est une donnée du problème connue au niveau des surfaces Γ_M indiquées sur la figure 1.4 à gauche. Ainsi, nous avons seulement besoin de résoudre la propagation des perturbations acoustiques dans les zones 2 et 3.

1.2 L'équation d'Helmholtz convectée

Pour cette introduction sur les phénomènes aéroacoustiques, nous nous référons à [91]. Dans cette section, nous expliquons comment est obtenue l'équation utilisée pour la simulation d'ondes acoustiques dans les zones 2 et 3 de la figure 1.4. Nous partons des équations de Navier–Stokes pour l'écoulement total :

$$\text{masse :} \qquad \frac{\partial}{\partial t}\rho + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}) = 0, \tag{1.1a}$$

$$\text{quantité de mouvement :} \qquad \frac{\partial}{\partial t}(\rho\mathbf{v}) + \boldsymbol{\nabla}\cdot(\rho\mathbf{v}\otimes\mathbf{v}) = -\boldsymbol{\nabla}p + \boldsymbol{\nabla}\cdot\boldsymbol{\tau}, \tag{1.1b}$$

$$\text{énergie :} \qquad \frac{\partial}{\partial t}(E) + \boldsymbol{\nabla}\cdot(E\mathbf{v}) = -\boldsymbol{\nabla}\cdot\mathbf{q} - \boldsymbol{\nabla}(p\mathbf{v}) + \boldsymbol{\nabla}(\boldsymbol{\tau}\cdot\mathbf{v}), \tag{1.1c}$$

où ρ est la densité, $\mathbf{v}$ est la vitesse, p est la pression, $\boldsymbol{\tau}$ est le tenseur des contraintes, E est l'énergie totale et q est le flux de chaleur dû à la conduction thermique. L'énergie totale se décompose de la façon suivante : $E = \rho e + \frac{1}{2}\rho v^2$, où e est la densité d'énergie interne et $\frac{1}{2}\rho v^2$ est la densité d'énergie cinétique. L'enthalpie est définie par $h = e + \frac{p}{\rho}$. En notant T le température, l'entropie s peut être introduite *via* la loi fondamentale de la thermodynamique pour les processus réversibles :

$$Tds = de + pd\left(\rho^{-1}\right) = dh - \rho^{-1}dp. \tag{1.2}$$

Introduisons la dérivée convective $\frac{D}{Dt} := \frac{\partial}{\partial t} + \mathbf{v}\cdot\boldsymbol{\nabla}$. Les équations de Navier–Stokes peuvent être réécrites sous la forme

$$\text{masse :} \qquad \frac{D}{Dt}\rho = -\rho\boldsymbol{\nabla}\cdot\mathbf{v}, \tag{1.3a}$$

$$\text{quantité de mouvement :} \qquad \rho\frac{D}{Dt}\mathbf{v} = -\boldsymbol{\nabla}p + \boldsymbol{\nabla}\cdot\boldsymbol{\tau}, \tag{1.3b}$$

$$\text{énergie :} \qquad \rho\frac{D}{Dt}e = -\boldsymbol{\nabla}\cdot\mathbf{q} - p\boldsymbol{\nabla}\cdot\mathbf{v} + \boldsymbol{\tau}:\boldsymbol{\nabla}\mathbf{v}, \tag{1.3c}$$

$$\text{enthalpie :} \qquad \rho\frac{D}{Dt}h = \frac{D}{Dt}p - \boldsymbol{\nabla}\cdot\mathbf{q} + \boldsymbol{\tau}:\boldsymbol{\nabla}\mathbf{v}, \tag{1.3d}$$

$$\text{entropie :} \qquad \rho T\frac{D}{Dt}s = -\boldsymbol{\nabla}\cdot\mathbf{q} + \boldsymbol{\tau}:\boldsymbol{\nabla}\mathbf{v}, \tag{1.3e}$$

où (1.1a) et (1.1b) ont été utilisées pour obtenir (1.3b), et les définitions de e, h et s ont été utilisées pour obtenir (1.3c)-(1.3d)-(1.3e). Les trois dernières équations ne sont pas indépendantes. En pratique, c'est celle sur l'entropie (1.3e) qui est utilisée pour des applications en acoustique.

Dans les conditions normales de température et de pression ($T = 273\ K$ et $p = 10^5\ Pa$), l'air vérifie la loi des gaz parfaits avec une grande précision. Cette approximation consiste à considérer que les molécules du gaz sont suffisamment éloignées les unes des autres pour négliger leurs interactions électrostatiques. Un gaz parfait vérifie la loi d'état

$$p = \rho RT, \tag{1.4}$$

où R est la constante spécifique du gaz parfait. Nous définissions C_P et C_V, respectivement, les capacités thermiques à volume et pression constants, par les relations suivantes :

$$de = C_V dT, \tag{1.5a}$$
$$dh = C_P dT. \tag{1.5b}$$

Les capacités thermiques vérifient la relation de Mayer des gaz parfaits : $C_P - C_V = R$. Un gaz parfait est dit de Laplace si les capacités thermiques sont constantes, alors que pour des gaz réels loin des conditions normales de température et de pression, elles peuvent dépendre significativement de la température. Nous considérons que l'air est un gaz parfait. Le coefficient de Laplace d'un gaz parfait est donné par $\gamma = \frac{C_P}{C_V}$. Pour l'air, $R \approx 286.73\ J.kg^{-1}.K^{-1}$ et $\gamma \approx 1.402$.

D'après (1.5), un gaz parfait vérifie

$$ds = C_V \frac{dp}{p} - C_P \frac{d\rho}{\rho}. \tag{1.6}$$

Les perturbations acoustiques dans l'air sont isentropiques. Elles se propagent à vitesse c, telle que

$$c^2 = \left(\frac{\partial p}{\partial \rho}\right)_s = \frac{\gamma p}{\rho} = \gamma R T. \tag{1.7}$$

Dans le cas des gaz parfaits de Laplace, nous pouvons intégrer les relations (1.5) et (1.6) pour obtenir

$$e = C_V T + e_{\text{init}}, \tag{1.8a}$$
$$h = C_p T + h_{\text{init}}, \tag{1.8b}$$
$$s = C_V \log p - C_P \log \rho + s_{\text{init}}, \tag{1.8c}$$

où e_{init}, h_{init} et s_{init} sont les constantes d'intégration, correspondant à un état de référence du gaz.

1.2.1 Approximation acoustique

Dans le domaine acoustique, les termes visqueux et turbulent ne vont jouer un rôle que dans les zones de création de bruit aérodynamique, tandis que les perturbations acoustiques sont trop rapides pour être affectées par les effets de conduction thermique. Pour le calcul de la propagation acoustique, nous négligeons donc les termes dus au tenseur des contraintes $\boldsymbol{\tau}$ et au flux de chaleur $\mathbf{q}$. Pour mieux comprendre cette approximation, nous adimensionnons les équations en introduisant les quantités adimensionnées telles que : $\mathbf{x} := L\tilde{\mathbf{x}}$, $\mathbf{v} := v_0\tilde{\mathbf{v}}$, $t := \frac{L}{v_0}\tilde{t}$, $\rho := \rho_0\tilde{\rho}$, $dp := \rho_0 v_0^2 d\tilde{p}$, $\boldsymbol{\tau} := \frac{\mu v_0}{L}\tilde{\boldsymbol{\tau}}$, $\mathbf{q} := \frac{\kappa \Delta T}{L}\tilde{\mathbf{q}}$, $T := T_0\tilde{T}$, $dT := \Delta T d\tilde{T}$ et $ds := \frac{C_P \Delta T}{T_0} d\tilde{s}$ avec les facteurs d'échelle L (longueur en m), v_0 (vitesse en $m.s^{-1}$), ρ_0 (densité en $kg.m^{-3}$), μ (viscosité dynamique en $Pa.s$), κ (conductivité thermique en $W.m^{-1}.K^{-1}$), T_0 (température en K), ΔT (variation de température en K), C_p (capacité thermique à pression constante en $J.kg^{-1}.K^{-1}$). Les équations de Navier–Stokes se réécrivent sous la forme

masse :
$$\frac{D}{D\tilde{t}}\tilde{\rho} = -\tilde{\rho}\tilde{\boldsymbol{\nabla}} \cdot \tilde{\mathbf{v}}, \tag{1.9a}$$

quantité de mouvement :
$$\tilde{\rho}\frac{D}{D\tilde{t}}\tilde{\mathbf{v}} = -\tilde{\boldsymbol{\nabla}}\tilde{p} + \frac{1}{\mathrm{Re}}\tilde{\boldsymbol{\nabla}} \cdot \tilde{\boldsymbol{\tau}}, \tag{1.9b}$$

entropie :
$$\tilde{\rho}\tilde{T}\frac{D}{D\tilde{t}}\tilde{s} = -\frac{1}{\mathrm{Pe}}\tilde{\boldsymbol{\nabla}} \cdot \tilde{\mathbf{q}} + \frac{\mathrm{Ec}}{\mathrm{Re}}\tilde{\boldsymbol{\tau}} : \tilde{\boldsymbol{\nabla}}\tilde{\mathbf{v}}, \tag{1.9c}$$

où $\mathrm{Re} = \frac{\rho_0 v_0 L}{\mu}$ est le nombre de Reynolds, $\mathrm{Pe} = \frac{\rho_0 C_P v_0 L}{\kappa}$ est le nombre de Péclet et $\mathrm{Ec} = \frac{v_0^2}{C_P \Delta T}$ est le nombre d'Eckert. Comme Pe = PrRe, où $\mathrm{Pr} = \frac{\mu C_p}{\kappa}$ est le nombre de Prandtl, qui est de l'ordre de 1 dans la plupart des gaz et fluides, Pe et Re sont du même ordre de grandeur. Ainsi, si Re est élevé et Ec n'est pas trop élevé, nous pouvons faire les approximations suivantes pour des simulations de propagation acoustique (nous revenons ici à des équations dimensionnées) :

$$\text{masse :} \qquad \frac{D}{Dt}\rho = -\rho \boldsymbol{\nabla} \cdot \mathbf{v}, \tag{1.10a}$$

$$\text{quantité de mouvement :} \qquad \rho \frac{D}{Dt}\mathbf{v} = -\boldsymbol{\nabla} p, \tag{1.10b}$$

$$\text{entropie :} \qquad \frac{D}{Dt} s = 0, \tag{1.10c}$$

où la dernière ligne signifie que l'entropie reste constante le long des lignes de courant. Nous faisons toujours l'hypothèse de gaz parfait de Laplace, ce qui conduit à compléter le système (1.10) par les relations (1.6) et (1.7). D'après (1.6), il vient $\frac{\rho}{p}\frac{dp}{dt} = \frac{C_P}{C_V}\frac{d\rho}{dt}$ et en utilisant (1.7), $\frac{d\rho}{dt} = c^2 \frac{dp}{dt}$. Si l'écoulement est homentropique (s est uniformément constante), il vient $\frac{\partial \rho}{\partial t} = c^2 \frac{\partial p}{\partial t}$, et il existe une constante K telle que

$$p = K \rho^{\gamma}. \tag{1.11}$$

Maintenant que nous avons simplifié les équations de Navier–Stokes dans le régime de propagation acoustique qui nous intéresse, nous allons séparer la description de l'écoulement moyen et celle de la perturbation acoustique.

1.2.2 Perturbation autour d'un écoulement moyen

Nous considérons que l'écoulement total est la somme d'un écoulement moyen stationnaire et d'une perturbation acoustique instationnaire, tels que

$$\mathbf{v} = \mathbf{v}_0 + \mathbf{v}', \ p = p_0 + p', \ \rho = \rho_0 + \rho', \ \text{et } s = s_0 + s', \tag{1.12}$$

où l'indice 0 désigne l'écoulement moyen et l'apostrophe désigne la perturbation acoustique. Nous linéarisons dans le régime des petites perturbations et obtenons pour l'écoulement moyen

$$\text{masse :} \qquad \boldsymbol{\nabla} \cdot (\rho_0 \mathbf{v}_0) = 0, \tag{1.13a}$$

$$\text{quantité de mouvement :} \qquad \rho_0 (\mathbf{v}_0 \cdot \boldsymbol{\nabla}) \mathbf{v}_0 = -\boldsymbol{\nabla} p_0, \tag{1.13b}$$

$$\text{entropie :} \qquad (\mathbf{v}_0 \cdot \boldsymbol{\nabla}) s_0 = 0, \tag{1.13c}$$

et les relations (1.6) et (1.7) deviennent

$$ds_0 = C_V \frac{dp_0}{p_0} - C_P \frac{d\rho_0}{\rho_0}, \tag{1.14}$$

et

$$c_0^2 = \frac{\gamma p_0}{\rho_0}. \tag{1.15}$$

Pour la perturbation acoustique, il vient au premier ordre

masse : $$\frac{\partial}{\partial t}\rho' + \boldsymbol{\nabla}\cdot\left(\mathbf{v}_0\rho' + \mathbf{v}'\rho_0\right) = 0, \quad (1.16a)$$

quantité de mouvement : $$\rho_0\left(\frac{\partial}{\partial t} + \mathbf{v}_0\cdot\boldsymbol{\nabla}\right)\mathbf{v}' + \rho_0\left(\mathbf{v}'\cdot\boldsymbol{\nabla}\right)\mathbf{v}_0 + \rho'\left(\mathbf{v}_0\cdot\boldsymbol{\nabla}\right)\mathbf{v}_0 = -\boldsymbol{\nabla}p', \quad (1.16b)$$

entropie : $$\left(\frac{\partial}{\partial t} + \mathbf{v}_0\cdot\boldsymbol{\nabla}\right)s' + \mathbf{v}'\cdot\boldsymbol{\nabla}s_0 = 0, \quad (1.16c)$$

et avec la convention $s_{\text{init}} = 0$,

$$s' = \frac{C_V}{p_0}p' - \frac{C_P}{\rho_0}\rho'. \quad (1.17)$$

1.2.3 Écoulement irrotationel et isentropique

Lorsque l'écoulement est irrotationel et que le domaine est simplement connexe, nous pouvons introduire un potentiel φ tel que $\mathbf{v} = \boldsymbol{\nabla}\varphi$. La conservation de la quantité de mouvement (1.10b) s'écrit alors

$$\rho\left(\frac{\partial}{\partial t}\left(\boldsymbol{\nabla}\varphi\right) + \left(\boldsymbol{\nabla}\varphi\right)\cdot\boldsymbol{\nabla}\left(\boldsymbol{\nabla}\varphi\right)\right) + \boldsymbol{\nabla}p = 0, \quad (1.18)$$

ou encore

$$\boldsymbol{\nabla}\left(\frac{\partial}{\partial t}\varphi + \frac{\|\boldsymbol{\nabla}\varphi\|^2}{2}\right) + \frac{1}{\rho}\boldsymbol{\nabla}p = 0. \quad (1.19)$$

Lorsque l'écoulement est isentropique partout (homentropique), nous pouvons utiliser (1.11) pour écrire

$$\frac{1}{\rho}\boldsymbol{\nabla}p = K\gamma\rho^{\gamma-2}\boldsymbol{\nabla}\rho = \boldsymbol{\nabla}\left(\frac{K\gamma}{\gamma-1}\rho^{\gamma-1}\right). \quad (1.20)$$

Nous pouvons alors intégrer (1.19) en espace :

$$\frac{\partial}{\partial t}\varphi + \frac{\|\boldsymbol{\nabla}\varphi\|^2}{2} + \frac{K\gamma}{\gamma-1}\rho^{\gamma-1} = f(t), \quad (1.21)$$

où $f(t)$ est une fonction du temps résultant de l'intégration en espace. Il est possible de prendre $f(t) = 0$, quitte à remplacer le potentiel φ par $\varphi - \int_0^t f(t)dt$, ce qui est possible dans la mesure où le potentiel n'est défini qu'à une fonction du temps près : $\boldsymbol{\nabla}\left(\varphi - \int_0^t f(t)dt\right) = \boldsymbol{\nabla}\varphi = \mathbf{v}$. Comme dans la Section 1.2.2, nous séparons la contribution stationnaire de la perturbation acoustique dans la définition du potentiel : $\varphi = \varphi_0 + \varphi'$ (il suffit de supposer que l'écoulement stationnaire moyen est potentiel pour que la perturbation acoustique soit également potentielle). Le troisième terme du membre de gauche de (1.21) vérifie au premier ordre

$$\frac{K\gamma}{\gamma-1}\rho^{\gamma-1} = \frac{K\gamma}{\gamma-1}\rho_0^{\gamma-1} + K\gamma\rho_0^{\gamma-2}\rho'. \quad (1.22)$$

En utilisant (1.11) et (1.15), il vient $K\gamma\rho_0^{\gamma-2}\rho' = \gamma\frac{p_0\rho'}{\rho_0^2} = c_0^2\frac{\rho'}{\rho_0}$. La linéarisation de (1.21) conduit donc à

$$\frac{\partial}{\partial t}\varphi' + \mathbf{v}_0\cdot\boldsymbol{\nabla}\varphi' + c_0^2\frac{\rho'}{\rho_0} = 0. \quad (1.23)$$

La formule (1.23) correspond à la relation de Bernoulli linéarisée pour les petites perturbations.

Nous supposons que les perturbations sont harmoniques. Cette hypothèse convient bien au profil de raies du bruit de soufflante du turboréacteur décrit dans la Section 1.1.1, où en pratique seules quelques pulsations décrivent la perturbation acoustique totale. Nous simplifions la description à une pulsation ω, telle que

$$\mathbf{v}' = \mathrm{Re}\left(\hat{\mathbf{v}}e^{-i\omega t}\right),\ \rho' = \mathrm{Re}\left(\hat{\rho}e^{-i\omega t}\right)\ \text{et}\ \varphi' = \mathrm{Re}\left(\hat{\varphi}e^{-i\omega t}\right). \tag{1.24}$$

Les relations (1.16a) et (1.23) deviennent alors

$$\text{masse :} \qquad -i\omega\hat{\rho} + \boldsymbol{\nabla}\cdot(\hat{\rho}\mathbf{v}_0 + \rho_0\boldsymbol{\nabla}\hat{\varphi}) = 0, \tag{1.25a}$$

$$\text{quantité de mouvement intégrée :} \qquad \frac{\rho_0}{c_0^2}\left(i\omega\hat{\varphi} - \mathbf{v}_0\cdot\boldsymbol{\nabla}\hat{\varphi}\right) = \hat{\rho}. \tag{1.25b}$$

L'équation sur le potentiel acoustique est alors obtenue en injectant l'expression de $\hat{\rho}$ fournie par (1.25b) dans l'équation de conservation de la masse (1.25a) :

$$-i\omega\frac{\rho_0}{c_0^2}\left(i\omega\hat{\varphi} - \mathbf{v}_0\cdot\boldsymbol{\nabla}\hat{\varphi}\right) + \boldsymbol{\nabla}\cdot\left[\frac{\rho_0}{c_0^2}\left(i\omega\hat{\varphi} - \mathbf{v}_0\cdot\boldsymbol{\nabla}\hat{\varphi}\right)\mathbf{v}_0 + \rho_0\boldsymbol{\nabla}\hat{\varphi}\right] = 0, \tag{1.26}$$

ou encore, en posant $k_0 := \frac{\omega}{c_0}$ et $\mathbf{M}_0 := \frac{\mathbf{v}_0}{c_0}$,

$$\rho_0\left(k_0^2\hat{\varphi} + ik_0\mathbf{M}_0\cdot\boldsymbol{\nabla}\hat{\varphi}\right) + \boldsymbol{\nabla}\cdot\left[\rho_0\left(\boldsymbol{\nabla}\hat{\varphi} - (\mathbf{M}_0\cdot\boldsymbol{\nabla}\hat{\varphi})\,\mathbf{M}_0 + ik_0\hat{\varphi}\mathbf{M}_0\right)\right] = 0. \tag{1.27}$$

L'équation (1.27) est connue sous le nom d'équation d'Helmholtz convectée (voir [9]). Un terme source non nul peut être ajouté dans le membre de droite du bilan de masse (1.25a) conduisant à (1.27) avec ce même membre de droite. Dans le cas où l'écoulement moyen stationnaire est nul, ρ_0 et k_0 sont uniformes et $\mathbf{M}_0 = \mathbf{0}$, conduisant à l'équation d'Helmholtz classique

$$\Delta\hat{\varphi} + k_0^2\hat{\varphi} = 0. \tag{1.28}$$

Dans la plupart des cas tests de cette thèse, nous considérons des objets plongés dans des écoulements potentiels et uniformes (correspondant aux zones 2 et 3 de la figure 1.4), et un terme source de type monopole acoustique. Un monopole acoustique localisé en $x_s \in \mathbb{R}^3$, d'amplitude A_s et de pulsation ω donne lieu à un terme source $g(x,t) := A_s\delta_{x_s}(x)\cos(\omega t)$, où $\delta_{x_s}(x)$ est la distribution de Dirac centrée en x_s. Pour simuler la propagation des perturbations acoustiques générées par une telle source dans un écoulement potentiel ou nul, il suffit d'ajouter $A_s\delta_{x_s}$ au membre de droite des équations (1.27) et (1.28). Dans cette thèse, nous supposerons toujours que l'écoulement dans lequel est plongé l'objet est nonuniforme au plus dans un domaine borné de l'espace. A l'extérieur de ce domaine, l'écoulement est uniforme et par un changement de variable et de fonction inconnue, il est possible (nous y reviendrons au chapitre 3) de transformer l'équation d'Helmholtz convectée (1.27) en l'équation d'Helmholtz classique (1.28). Ainsi, nous nous ramènerons toujours au cas où la fonction cherchée satisfait (1.28) à l'extérieur d'un domaine borné.

1.3 Méthode des équations intégrales pour l'équation d'Helmholtz classique

Dans cette section, nous présentons brièvement la méthode des équation intégrales. Une description plus détaillée et plus rigoureuse, notamment en ce qui concerne les espaces de définition, la régularité et les propriétés des opérateurs intégraux introduits, est donnée dans le chapitre 3. Nous considérons la géométrie représentée dans la figure 1.5, où Ω^- est un ouvert borné, $\Gamma = \partial\Omega^-$ et $\Omega^+ = \mathbb{R}^3 \backslash \overline{\Omega^-}$.

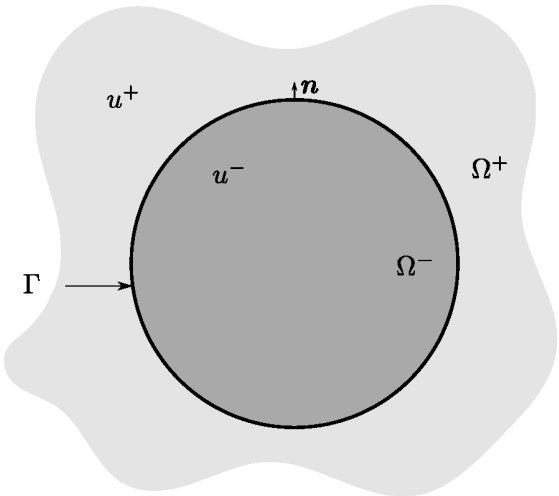

Fig. 1.5. Geometrie

Pour toute fonction u définie sur $\mathbb{R}^3$, nous notons respectivement $u|_{\Omega^+} := u^+$ et $u|_{\Omega^-} := u^-$ ses restrictions à Ω^+ et Ω^-. Nous notons respectivement γ_0^+ et γ_0^- les traces de Dirichlet intérieure et extérieure sur Γ. Si u est une fonction régulière définie sur $\Omega^+ \cup \Omega^-$, nous avons, pour $x \in \Gamma$, $\gamma_0^+ u(x) := u^+|_\Gamma(x)$ et $\gamma_0^- u(x) := u^-|_\Gamma(x)$. De même, les traces de Neumann intérieure et extérieure sur Γ sont notées respectivement γ_1^+ et γ_1^-, et pour $x \in \Gamma$, $\gamma_1^+ u(x) := (\boldsymbol{\nabla} u^+)|_\Gamma(x) \cdot \mathbf{n}(x)$ et $\gamma_1^- u(x) := (\boldsymbol{\nabla} u^-)|_\Gamma(x) \cdot \mathbf{n}(x)$, où $\mathbf{n}(x)$ est la normale à Γ en x dirigée vers l'extérieur (voir la figure 1.5). Le saut et la moyenne des traces de Dirichlet à travers Γ sont définis respectivement par $[\gamma_0 u]_\Gamma := \gamma_0^+ u - \gamma_0^- u$ et $\{\gamma_0 u\}_\Gamma := \frac{1}{2}\left(\gamma_0^+ u + \gamma_0^- u\right)$. De même, le saut et la moyenne des traces de Neumann à travers Γ sont définis respectivement par $[\gamma_1 u]_\Gamma := \gamma_1^+ u - \gamma_1^- u$ et $\{\gamma_1 u\}_\Gamma := \frac{1}{2}\left(\gamma_1^+ u + \gamma_1^- u\right)$.

Nous disons qu'une distribution u est une solution rayonnante par morceaux de l'équation d'Helmholtz lorsque

$$\begin{cases} \Delta u + k_\infty^2 u = 0, \quad \text{dans } \Omega^+ \cup \Omega^-, \\ \lim\limits_{r \to +\infty} r \left(\dfrac{\partial u}{\partial r} - i k_\infty u \right) = 0, \end{cases} \tag{1.29}$$

où le nombre d'onde est maintenant noté k et où la seconde ligne est la condition de radiation de Sommerfeld. Soit λ une fonction régulière définie sur Γ. Le potentiel de simple couche est défini par $\mathcal{S}(\lambda)(x) := \int_\Gamma E(y-x)\lambda(y) ds(y)$, $x \in \mathbb{R}^3 \backslash \Gamma$, où $E(x) := \frac{\exp(ik_\infty |x|)}{4\pi |x|}$ est la solution fondamentale de $\Delta u + k_\infty^2 u = 0$ dans $\mathbb{R}^3$ satisfaisant la condition de radiation de Sommerfeld.

De même, le potentiel de double couche est défini par $\mathcal{D}(\lambda)(\mathbf{x}) := \int_\Gamma \boldsymbol{\nabla}_\mathbf{y} E(\mathbf{y}-\mathbf{x})\lambda(\mathbf{y})ds(\mathbf{y})$, $\mathbf{x} \in \mathbb{R}^3 \backslash \Gamma$. D'après [80, Theorem 3.1.1], une solution rayonnante par morceaux de l'équation d'Helmholtz u est entièrement déterminée par le saut de ses traces de Dirichlet et de Neumann à travers Γ. Plus précisément,

$$u = -\mathcal{S}([\gamma_1 u]_\Gamma) + \mathcal{D}([\gamma_0 u]_\Gamma), \quad \text{dans } \Omega^+ \cup \Omega^-, \tag{1.30}$$

où (1.30) est connue sous le nom de formule de représentation. La trace de Dirichlet du premier terme du membre de droite est continue à travers Γ, et il en est de même de la trace de Neumann du deuxième terme.

Nous définissons les opérateurs intégraux de simple couche S, double couche D, le dual de l'opérateur de double couche $\tilde{D}$ et l'opérateur hypersingulier N par

$$\begin{aligned} S\lambda &:= \gamma_0 \left(\mathcal{S}\lambda\right), \\ D\lambda &:= \left\{\gamma_0 \left(\mathcal{D}\lambda\right)\right\}_\Gamma, \\ \tilde{D}\lambda &:= \left\{\gamma_1 \left(\mathcal{S}\lambda\right)\right\}_\Gamma, \\ N\lambda &:= -\gamma_1 \left(\mathcal{D}\lambda\right), \end{aligned} \tag{1.31}$$

où λ est une fonction définie sur Γ. D'après [80, Theorem 3.1.2], une solution rayonnante par morceaux de l'équation d'Helmholtz u vérifie

$$\begin{pmatrix} \frac{1}{2}I - D & S \\ N & \frac{1}{2}I + \tilde{D} \end{pmatrix} \begin{pmatrix} [\gamma_0 u]_\Gamma \\ [\gamma_1 u]_\Gamma \end{pmatrix} = - \begin{pmatrix} \gamma_0^- u^- \\ \gamma_1^- u^- \end{pmatrix}. \tag{1.32}$$

L'opérateur défini par bloc dans le membre de gauche de (1.32) est appelé projecteur de Calderón. Les deux relations dans (1.32) sont simplement des relations nécessaires vérifiées par une solution rayonnante par morceaux de l'équation d'Helmholtz u. Elles ne sont toutefois pas indépendantes (elles sont obtenues en pratique en prenant les traces de Neumann et Dirichlet de (1.30)). Dans un problème aux limites, un comportement peut être imposé à la surface Γ par le biais d'une condition aux limites qui, s'ajoutant à (1.32), peut conduire à un problème bien posé. Un exemple simple est celui où on suppose que $u^- \equiv 0$ et $\gamma_0^+ u = 0$. Alors, par unicité du problème extérieur de Dirichlet, $u^+ \equiv 0$ et $u \equiv 0$. En revanche, si l'on suppose seulement $\gamma_0^+ u = \gamma_0^- u = 0$, alors $u^+ \equiv 0$ est toujours vrai, mais u^- n'est pas déterminée de façon unique si k correspond à une fréquence propre du Laplacien de Dirichlet dans Ω^-. Deux autres exemples de condition aux limites sur Γ que nous étudierons en détail dans cette thèse sont les conditions de transmission et la condition de Robin. De façon générale, la méthode des équations intégrales consiste à utiliser les relations (1.32) et des conditions aux limites liées au problème physique pour obtenir un système d'équations bien posé d'inconnues $[\gamma_0 u]_\Gamma$ et $[\gamma_1 u]_\Gamma$, puis à utiliser la formule de représentation (1.30) pour retrouver la fonction inconnue u à partir du saut de ses traces sur Γ.

1.4 Méthodes de réduction de modèle

L'équation d'Helmholtz convectée (1.27) est l'équation à la base de notre modèle d'aéroacoustique. Une fois que nous disposons d'une méthode numérique fiable, nous souhaitons être capables de résoudre ce problème pour de nombreuses valeurs de certains paramètres. Ces

paramètres peuvent être la fréquence et la position de la source acoustique, ou encore le coefficient d'impédance des objets diffractants. Dans ce contexte, les méthodes de réduction de modèle prennent tout leur sens, car elles permettent d'obtenir très rapidement une approximation de la solution d'un problème. Ainsi, des études de propagation d'incertitude ou d'optimisation, requises en phase de conception d'un projet industriel et nécessitant d'évaluer la solution d'un problème pour de très nombreuses valeurs de certains paramètres, deviennent accessibles.

Dans cette section, nous présentons brièvement quelques méthodes de réduction de modèle.

1.4.1 Introduction : le fléau de la dimension

Le fléau de la dimension (*curse of dimensionality*) est une expression inventée par Bellman pour qualifier le comportement de l'augmentation de la complexité de description d'un espace avec l'ajout de dimensions supplémentaires, voir [11, 12]. Ce fléau est rencontré dans de nombreuses disciplines. L'exemple le plus simple est peut-être combinatoire : supposons que nous disposons de d pièces de monnaie que nous lançons simultanément. Le nombre de résultats ordonnés possible est 2^d : la taille du problème augmente de façon exponentielle avec la dimension. Un autre exemple est le probème d'échantillonage d'un domaine. Prenons l'hypercube $[0,1]^d$ que nous échantillonnons de façon uniforme dans chaque direction. Nous exigeons que la distance entre deux points soit inférieure à 0.1. En dimension 1, cela donne 11 points. En dimension quelconque, le nombre de points d'échantillonage vaut 11^d, et donc augmente de façon exponentielle avec la dimension.

Dans notre contexte d'aéroacoustique en aviation civile, nous disposons d'une méthode de simulation directe : à partir d'un modèle dépendant de certains paramètres, nous sommes capables de calculer le champ de pression acoustique diffracté. Maintenant, si nous sommes intéressés par l'exploration de l'ensemble des résultats engendrés par les solutions prises par un échantillonage de l'espace des paramètres, nous sommes victimes du fléau de la dimension : le problème aéroacoustique doit être résolu un nombre de fois qui augmente de façon exponentielle avec le nombre de paramètres. Une autre vision peut être de considérer les paramètres comme des variables du problème : notons $u(x, \mu_1, \mu_2, \cdots, \mu_d)$ le champ de pression acoustique, calculé pour les valeurs $\mu_1, \mu_2, \cdots, \mu_d$ des paramètres. La recherche de la solution u consiste alors à déterminer le champ de pression en tout point de l'espace et pour toutes les valeurs possibles des paramètres. Supposons pour simplifier que $x \in [0,1]$ et que $\mu_i \in [0,1]$ pour tout $1 \leq i \leq d$, et considérons u comme une fonction de $y = (x, \mu_1, \mu_2, \cdots, \mu_d) \in [0,1]^{d+1}$. Dans le cadre d'une méthode numérique, il faudra reconstruire une approximation $R(u)$ de u à partir d'un ensemble de N valeurs $\{u(y_i)\}_{1 \leq i \leq N}$, où $y_1, \cdots, y_N \in [0,1]^{d+1}$. Si $(y_i)_{1 \leq i \leq N}$ sont les nœuds d'une grille uniforme sur $[0,1]^{d+1}$ d'espacement $h > 0$ et si une reconstruction polynomiale est utilisée, il est connu que

$$\|u - R(u)\|_{L^\infty\left([0,1]^{d+1}\right)} \leq Ch^m, \tag{1.33}$$

où $C > 0$ est une constante indépendante de h et m l'ordre de la méthode. Comme le nombre de points d'échantillonnage N est en $O\left(h^{-(d+1)}\right)$, l'erreur d'approximation vérifie

$$\|u - R(u)\|_{L^\infty\left([0,1]^{d+1}\right)} \leq CN^{-\frac{m}{d+1}}, \tag{1.34}$$

ce qui signifie que l'erreur de reconstruction converge d'autant plus lentement avec le nombre de points d'échantillonage que la dimension est grande. Il a été prouvé qu'il est impossible de

construire des schémas de reconstruction tels que l'erreur converge plus rapidement [35]. Dans le cas d'une approximation de Galerkin où les fonctions de base sont obtenues par tensorisation de fonctions de bases univariées, et où on suppose que l'on a N fonctions de base pour la dimension spatiale et N_μ pour chaque dimension paramétrique, la taille du problème discrétisé est NN_μ^d et augmente donc de façon exponentielle avec le nombre de paramètres. Plus de détails peuvent être trouvés dans [41].

La dépendance exponentielle de la complexité du problème en fonction du nombre de paramètres rend impossible en pratique les approches brutales considérées ci-dessus. Dans la plupart des problèmes rencontrés en pratique, certains paramètres vont avoir une influence sur le résultat beaucoup plus faible que d'autres paramètres. Soit $\epsilon > 0$, notons $\mu = (\mu_1, \mu_2, \cdots, \mu_d) \in [0,1]^d$ et supposons qu'il existe une matrice rectangulaire A de taille $m \times d$, $m < d$, et une fonction $\hat{u}$ telles que pour tout $x \in [0,1]$ et tout $\mu \in [0,1]^d$,

$$|u(x,\mu) - \hat{u}(x, A\mu)| \leq \epsilon. \tag{1.35}$$

Au prix d'une approximation sur le résultat, il est possible de réduire la dimension du problème de d à m. Le plus petit m pour lequel le cas d'égalité de (1.35) est valable pour $\epsilon = 0$ est appelé dimension paramétrique intrinsèque du problème. Intuitivement, on veut fournir des efforts uniquement dans les directions qui influent le plus sur la qualité du résultat. Les méthodes de réduction de modèles ont été développées pour identifier les structures de dépendance principales du résultat en les paramètres et calculer rapidement des approximations de la solution.

1.4.2 Les représentations en produits tensoriels

Certaines méthodes de représentation en produits tensoriels cherchent à approcher une fonction u en grande dimension, par une somme de produits tensoriels de fonctions univariées :

$$u(x, \mu_1, \cdots, \mu_d) \approx \sum_{k=1}^{n} r_k(x) s_k^{(1)}(\mu_1) \cdots s_k^{(d)}(\mu_d). \tag{1.36}$$

Le but de ces méthodes est de trouver la meilleure approximation possible, à nombre de termes de la somme n fixé.

Les algorithmes gloutons

Des présentations détaillées peuvent être trouvées dans [7, 37, 101]. Considérons un espace de Hilbert H muni du produit scalaire $\langle \cdot, \cdot \rangle_H$ et de la norme associée $\| \cdot \|_H$. Un ensemble $\mathcal{D}$ de fonctions de H est appelé dictionnaire si pour tout élément g de $\mathcal{D}$, $\|g\|_H = 1$, et $\overline{\mathrm{Span}}(\mathcal{D}) = H$.

Le problème générique consiste à chercher la meilleure approximation d'un élément u de H comme une combinaison linéaire d'au plus n éléments $g_1, \cdots, g_n \in \mathcal{D}$:

$$(g_1, \cdots, g_n) \in \underset{(d_1, \cdots, d_n) \in \mathcal{D}}{\operatorname{argmin}} \|u - P_{d_1, \cdots, d_n} u\|_H, \tag{1.37}$$

où $P_{d_1, \cdots, d_n}$ est la projection orthogonale sur $\mathrm{Span}\,(d_1, \cdots, d_n)$ pour le produit scalaire $\langle \cdot, \cdot \rangle_H$. Trouver la meilleure approximation est un problème difficile et sujet au fléau de la dimension ;

nous cherchons une approximation suffisamment bonne. L'idée des algorithmes gloutons est de choisir de façon itérative les éléments g_i du dictionnaire. Nous supposons que pour tout $u \in H$, il existe $g \in \mathcal{D}$ tel que

$$g \in \underset{d\in\mathcal{D}}{\operatorname{argmax}} \langle u, d\rangle_H. \tag{1.38}$$

Il est facile de voir que si g est une solution de (1.38), alors

$$(g, \langle u, g\rangle_H) \in \underset{(d,\lambda)\in\mathcal{D}\times\mathbb{R}}{\operatorname{argmin}} \|u - \lambda d\|_H. \tag{1.39}$$

Nous pouvons distinguer deux principaux types d'algorithmes gloutons : le *Pure Greedy Algorithm* (PGA) et l'*Orthogonal Greedy Algorithm* (OGA), présentés dans la Table 1.1.

PGA

1. Fixer $r_0^{\mathrm{PGA}} := u$, $u_0^{\mathrm{PGA}} := 0$, $n = 0$ et choisir $\epsilon > 0$.
2. Tant que $\|r_n^{\mathrm{PGA}}\|_H > \epsilon \|u_n^{\mathrm{PGA}}\|_H$,
   trouver $g_{n+1}^{\mathrm{PGA}} \in \underset{g\in\mathcal{D}}{\operatorname{argmax}} \left\langle r_n^{\mathrm{PGA}}, g\right\rangle_H$,
   définir $u_{n+1}^{\mathrm{PGA}} := u_n^{\mathrm{PGA}} + \left\langle r_n^{\mathrm{PGA}}, g_{n+1}^{\mathrm{PGA}}\right\rangle_H g_{n+1}^{\mathrm{PGA}}$ et
   $r_{n+1}^{\mathrm{PGA}} := r_n^{\mathrm{PGA}} - \left\langle r_n^{\mathrm{PGA}}, g_{n+1}^{\mathrm{PGA}}\right\rangle_H g_{n+1}^{\mathrm{PGA}}$,
   remplacer $n \leftarrow n + 1$.

OGA

1. Fixer $r_0^{\mathrm{OGA}} := u$, $u_0^{\mathrm{OGA}} := 0$, $n = 0$ et choisir $\epsilon > 0$.
2. Tant que $\|r_n^{\mathrm{OGA}}\|_H > \epsilon \|u_n^{\mathrm{OGA}}\|_H$,
   trouver $g_{n+1}^{\mathrm{OGA}} \in \underset{g\in\mathcal{D}}{\operatorname{argmax}} \left\langle r_n^{\mathrm{OGA}}, g\right\rangle_H$,
   définir $H_{n+1}^{\mathrm{OGA}} := \operatorname{Span}\{g_i^{\mathrm{OGA}}, 1 \le i \le n+1\}$,
   $u_{n+1}^{\mathrm{OGA}} := P_{H_{n+1}^{\mathrm{OGA}}}(u)$ et $r_{n+1}^{\mathrm{OGA}} := u - P_{H_{n+1}^{\mathrm{OGA}}}(u)$,
   remplacer $n \leftarrow n + 1$.

Table 1.1. Présentation des algorithmes PGA et OGA.

Ces algorithmes convergent dans le sens suivant. Nous considérons les algorithmes PGA et OGA présentés dans la Table 1.1 avec $\epsilon = 0$. Pour tout dictionnaire $\mathcal{D}$ et tout $u \in H$,

$$\begin{aligned} \|u - u_n^{\mathrm{PGA}}\|_H &\underset{n\to\infty}{\to} 0, \\ \|u - u_n^{\mathrm{OGA}}\|_H &\underset{n\to\infty}{\to} 0. \end{aligned} \tag{1.40}$$

Avec des hypothèses techniques supplémentaires sur les fonctions de H, on peut montrer des taux de convergence polynomiaux pour PGA et OGA, voir [41] pour plus de détails.

Proper Generalized Decomposition

Dans le contexte des équations aux dérivées partielles en grande dimension, la Proper Generalized Decomposition (PGD) est une méthode de réduction de modèle qui utilise les algorithmes gloutons présentés ci-dessus. Pour plus de détails, nous renvoyons à [61, 4, 81].

Soit $a : \Omega \times \mathcal{P} \to \mathbb{R}$, une fonction mesurable telle qu'il existe $\alpha, \beta > 0$ tels que

$$\forall (x, \mu) \in \Omega \times \mathcal{P},\ \alpha \le a(x, \mu) \le \beta. \tag{1.41}$$

Nous considérons le problème de diffusion paramétrique suivant : Trouver $u \in H := L^2(\mathcal{P}, H_0^1(\Omega))$ tel que

$$-\boldsymbol{\nabla}_x \cdot (a \boldsymbol{\nabla}_x u) = f. \tag{1.42}$$

Soit H un espace de Hilbert de fonctions multivariées $u(x, \mu_1, \cdots, \mu_d)$ et H_x, H_1, ..., H_d des espaces de Hilbert de fonctions univariées dépendant respectivement de x, μ_1, ..., μ_d. La PGD repose sur le choix de dictionnaire suivant :

$$\mathcal{D} := \left\{ r \otimes s^{(1)} \otimes \cdots \otimes s^{(d)} \mid r \in H_x, s^{(1)} \in H_1, \cdots, s^{(d)} \in H_d, \|r \otimes s^{(1)} \otimes \cdots \otimes s^{(d)}\|_H = 1 \right\}, \tag{1.43}$$

où $r \otimes s^{(1)} \otimes \cdots \otimes s^{(d)}(x, \mu_1, \cdots, \mu_d) = r(x)s^{(1)}(\mu_1) \cdots s^{(d)}(\mu_d)$ et où on suppose que $\overline{\text{Span}}\,(\mathcal{D}) = H$. Le problème (1.42) peut être écrit comme le problème de minimisation

$$u = \underset{v \in H}{\text{argmin}}\ \mathcal{E}(v), \tag{1.44}$$

où $\mathcal{E}(v) = \|v - u\|_H^2$ et $\|v\|_H^2 = \int_{\Omega \times \mathcal{P}} a(x, \mu) \left|\boldsymbol{\nabla}_x v(x, \mu)\right|^2 dx d\mu$. La PDG pour l'approximation de la solution de (1.42) consiste à déterminer itérativement $\left(r_n, s_n^{(1)}, \cdots, s_n^{(d)}\right) \in H_x \times H_1 \times \cdots \times H_d$ tel que

$$\left(r_n, s_n^{(1)}, \cdots, s_n^{(d)}\right) \in \underset{\left(r, s^{(1)}, \cdots, s^{(d)}\right) \in H_x \times H_1 \times \cdots \times H_d}{\text{argmin}} \mathcal{E}\left(\sum_{k=1}^{n-1} r_k \otimes s_k^{(1)} \otimes \cdots \otimes s_k^{(d)} + r \otimes s^{(1)} \otimes \cdots \otimes s^{(d)}\right). \tag{1.45}$$

En utilisant (1.39), nous remarquons que cela revient exactement à réaliser un algorithme PGA pour l'approximation de la solution de (1.42) dans H pour le choix de dictionnaire (1.43). En pratique, le calcul de $\left(r_n, s_n^{(1)}, \cdots, s_n^{(d)}\right)$ à chaque itération se fait en résolvant les équations d'Euler associées au problème de minimisation (1.45) par une procédure de point fixe.

Le fléau de la dimension est contourné dans la mesure où le calcul d'une approximation de (1.42) à n termes nécessite de résoudre n problèmes de taille $N_x + N_1 + \cdots + N_d$, où N_x et N_i, $1 \leq i \leq d$, sont les dimensions des espaces vectoriels approchant respectivement H_x et H_i.

Une condition nécessaire pour que la PGD soit valable est que le problème soit symétrique, comme c'est le cas pour (1.42), pour que la forme faible soit équivalente à la condition d'Euler d'un problème de minimisation.

1.4.3 La méthode des bases réduites

La méthode des bases réduites est celle que nous avons choisie d'étudier dans cette thèse. Nous en rappelons ici le principe général et nous donnons quelques résultats théoriques connus à ce jour.

Introduction et problème réduit

Par la suite, nous notons $\mathcal{P}$ l'ensemble des valeurs considérées pour un paramètre μ, et $\mathcal{P}_{\text{trial}}$ un échantillonage de $\mathcal{P}$. Considérons la formulation variationnelle suivante : Trouver $u_\mu \in \mathcal{V}$ tel que

$$a_\mu(u_\mu, u^t) = b_\mu(u^t), \quad \forall u^t \in \mathcal{V}, \tag{1.46}$$

où a_μ est, dans le cas idéal, une forme bilinéaire, continue et coercive (uniformément en le paramètre $\mu \in \mathcal{P}$, c'est-à-dire avec des constantes de continuité M et de coercivité α_{coer} indépendantes de μ), b_μ est une forme linéaire et $\mathcal{V}$ un espace vectoriel de dimension N, approximation conforme d'un espace de Hilbert H. La méthode des bases réduites consiste dans un premier temps à précalculer des solutions u_{μ_i} de (1.46) pour des valeurs bien choisies $(\mu_i)_{1\leq i\leq \hat{N}}$ du paramètre au cours de la phase *offline*. Nous définissions l'espace vectoriel $\hat{\mathcal{V}}_{\hat{N}} := \text{Vect}\{u_{\mu_1}, \cdots, u_{\mu_{\hat{N}}}\}$. Puis, au cours de la phase *online*, pour des nouvelles valeurs de $\mu \in \mathcal{P}$, des approximations $\hat{u}_\mu^{\hat{N}} \in \hat{\mathcal{V}}_{\hat{N}}$ de la solution de (1.46) sont obtenues en résolvant

$$a_\mu(\hat{u}_\mu^{\hat{N}}, u_{\mu_j}) = b_\mu(u_{\mu_j}), \qquad \forall j \in \{1, ..., \hat{N}\}. \tag{1.47}$$

La décomposition de $\hat{u}_\mu^{\hat{N}}$ dans $\hat{\mathcal{V}}_{\hat{N}}$ est notée

$$\hat{u}_\mu^{\hat{N}} = \sum_{i=1}^{\hat{N}} \gamma_i(\mu) u_{\mu_i}. \tag{1.48}$$

Estimateur d'erreur a posteriori

Soit $r_\mu^{\hat{N}} := \|G_\mu \hat{u}_\mu^{\hat{N}}\|_H$, où G_μ est l'application linéaire de H dans H telle que $H \ni u \mapsto G_\mu u := J\left(a_\mu(u, \cdot) - b_\mu\right) \in H$, avec J l'isomorphisme de Riesz de H' dans H tel que pour tout $l \in H'$ et tout $v \in \mathcal{V}$,

$$\langle Jl, v\rangle_H = l(v). \tag{1.49}$$

La quantité $r_\mu^{\hat{N}}$ correspond à la norme duale du résidu associé à la formulation (1.46) calculé en la solution réduite $\hat{u}_\mu^{\hat{N}}$. Dans le contexte de (1.46), on peut prouver que

$$\frac{1}{M} r_\mu^{\hat{N}} \leq \|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H \leq \frac{1}{\alpha_{\text{coer}}} r_\mu^{\hat{N}}. \tag{1.50}$$

Pour ce faire, considérons

$$\begin{aligned} a_\mu(\hat{u}_\mu^{\hat{N}} - u_\mu, \hat{u}_\mu^{\hat{N}} - u_\mu) &= a_\mu(\hat{u}_\mu^{\hat{N}}, \hat{u}_\mu^{\hat{N}} - u_\mu) - b_\mu(\hat{u}_\mu^{\hat{N}} - u_\mu) \\ &= \left\langle G_\mu \hat{u}_\mu^{\hat{N}}, \hat{u}_\mu^{\hat{N}} - u_\mu \right\rangle_H, \end{aligned} \tag{1.51}$$

où nous avons utilisé (1.46) dans la première ligne, et (1.49), avec la définition de G_μ, dans la deuxième ligne. En utilisant la coercivité de a_μ et l'inégalité de Cauchy-Schwarz sur la deuxième ligne de (1.51), il vient $\alpha_{\text{coer}}\|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H^2 \leq \left\langle G_\mu \hat{u}_\mu^{\hat{N}}, \hat{u}_\mu^{\hat{N}} - u_\mu \right\rangle_H \leq r_\mu^{\hat{N}} \|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H$, d'où l'inégalité de droite de (1.50) est directement obtenue. La formulation variationnelle (1.46) peut être étendue sur H' pour obtenir $a_\mu(\hat{u}_\mu^{\hat{N}} - u_\mu, v) = \langle a_\mu(u, \cdot) - b_\mu, v\rangle_{H',H}$, où $\langle \cdot, \cdot \rangle_{H',H}$ est le produit de dualité entre H' et H. Utilisons maintenant la continuité de a_μ, il vient $\langle a_\mu(u, \cdot) - b_\mu, v\rangle_{H',H} \leq M\|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H \|v\|_H$. En particulier, $r_\mu^{\hat{N}} = \|a_\mu(u, \cdot) - b_\mu\|_{H'} = \sup_{0\neq v\in\mathcal{H}} \frac{\langle a_\mu(u,\cdot) - b_\mu, v\rangle_{H',H}}{\|v\|_H} \leq M\|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H$, l'inégalité de gauche de (1.50) est directement obtenue.

Supposons que $\{u_{\mu_1}, \cdots, u_{\mu_{\hat{N}}}\}$ ont été construites, on définit $\mu_{\hat{N}+1} \in \mathcal{P}_{\text{trial}}$ tel que

$$M\|\hat{u}^{\hat{N}}_{\mu_{\hat{N}+1}} - u_{\mu_{\hat{N}+1}}\|_H \geq r^{\hat{N}}_{\mu_{\hat{N}+1}} := \max_{\mu\in\mathcal{P}_{\text{trial}}} r^{\hat{N}}_{\mu} \geq \alpha_{\text{coer}} \max_{\mu\in\mathcal{P}_{\text{trial}}} \|\hat{u}^{\hat{N}}_{\mu} - u_{\mu}\|_H, \tag{1.52}$$

et ainsi

$$\|\hat{u}^{\hat{N}}_{\mu_{\hat{N}+1}} - u_{\mu_{\hat{N}+1}}\|_H \geq \gamma \max_{\mu\in\mathcal{P}_{\text{trial}}} \|\hat{u}^{\hat{N}}_{\mu} - u_{\mu}\|_H, \tag{1.53}$$

où $\gamma := \frac{\alpha_{\text{coer}}}{M}$. Ainsi, le paramètre qui maximise $r^{\hat{N}}_{\mu}$ sur $\mathcal{P}_{\text{trial}}$ est celui qui maximise l'erreur $\|\hat{u}^{\hat{N}}_{\mu} - u_{\mu}\|_H$ sur $\mathcal{P}_{\text{trial}}$ à la constante γ près.

Efficacité online

On dit que $a_\mu(u,v)$ et $b_\mu(v)$ dépendent de μ de façon affine s'il existe $\mu \mapsto \alpha_k(\mu)$ et $a_k(u,v)$, $1 \leq k \leq d^a$, et $\mu \mapsto \beta_k(\mu)$ et $b_k(v)$, $1 \leq k \leq d^b$, tels que $a_\mu(u,v) = \sum_{k=1}^{d^a} \alpha_k(\mu) a_k(u,v)$ et $b_\mu(v) = \sum_{k=1}^{d^b} \beta_k(\mu) b_k(v)$.

Sous l'hypothèse de dépendance affine, $a_\mu(u_{\mu_i}, u_{\mu_j}) = \sum_{k=1}^{d^a} \alpha_k(\mu) a_k(u_{\mu_i}, u_{\mu_j})$ et $b_\mu(u_{\mu_j}) = \sum_{k=1}^{d^b} \beta_k(\mu) b_k(u_{\mu_j})$. Ainsi le problème réduit (1.47) peut être construit en complexité indépendante de N si les quantités $a_k(u_{\mu_i}, u_{\mu_j})$, $1 \leq k \leq d^a$, $1 \leq i,j \leq \hat{N}$, et $b_k(u_{\mu_j})$, $1 \leq k \leq d^b$, $1 \leq i,j \leq \hat{N}$, sont calculées et stockées pendant la phase *offline*. Sous l'hypothèse de dépendance affine, on montre également que (on suppose ici pour simplifier que H est un espace vectoriel réel)

$$\begin{aligned}\left(r^{\hat{N}}_{\mu}\right)^2 = &\sum_{k=1}^{d^b}\sum_{p=1}^{d^b} \beta_k(\mu)\beta_p(\mu) \left\langle Jb_k, Jb_p\right\rangle_H - 2\sum_{k=1}^{d^b}\sum_{l=1}^{d^a}\sum_{i=1}^{\hat{N}} \beta_k(\mu)\alpha_l(\mu)\gamma_i(\mu) \left\langle Jb_k, Ja_l(u_{\mu_i},\cdot)\right\rangle_H \\ &+ \sum_{k=1}^{d^a}\sum_{p=1}^{d^a}\sum_{i=1}^{\hat{N}}\sum_{j=1}^{\hat{N}} \alpha_k(\mu)\gamma_i(\mu)\alpha_p(\mu)\gamma_j(\mu) \left\langle Ja_k(u_{\mu_i},\cdot), Ja_p(u_{\mu_j},\cdot)\right\rangle_H,\end{aligned} \tag{1.54}$$

où chaque évaluation de J est calculée en résolvant (1.49) dans la phase *offline*, et où les fonctions $\alpha_k(\mu)$, $1 \leq k \leq d^a$, et $\beta_k(\mu)$, $1 \leq k \leq d^b$, sont connues, et les coefficients $\gamma_i(\mu)$, $1 \leq i \leq \hat{N}$ sont calculés en complexité indépendante de N dans la phase *online* en résolvant (1.47)-(5.3).

Il est ainsi possible de calculer $\hat{u}^{\hat{N}}_{\mu}$ et une borne supérieure de l'erreur $\|\hat{u}^{\hat{N}}_{\mu} - u_{\mu}\|_H$ en complexité indépendante de N.

Algorithme de la phase *offline*

L'algorithme 1 détaille les étapes de la phase *offline* de la méthode des bases réduites. Le choix a priori de $\mathcal{P}_{\text{trial}}$ est un problème difficile. Cependant, nous disposons d'un estimateur d'erreur a posteriori calculable en complexité indépendante de N. Pour tout $\mu \in \mathcal{P}$, il est possible, à chaque évaluation de la solution réduite $\hat{u}^{\hat{N}}_{\mu}$, de quantifier l'erreur faite par l'approximation des base réduite. Dans la phase *online*, s'il arrive que, pour une certain $\mu^* \in \mathcal{P}$, cette erreur soit supérieure à la tolérance fixée dans l'algorithme 1, il est possible d'incrémenter la base réduite en calculant u^*_{μ} et en ajoutant μ^* à $\mathcal{P}_{\text{select}}$. Par ailleurs, l'efficacité du calcul de l'estimateur a posteriori permet d'exécuter la ligne 29 de l'algorithme 1 en complexité indépendante de N et de considérer des $\mathcal{P}_{\text{trial}}$ de grande taille.

Algorithm 1 Phase *offline* de la méthode des bases réduites

for all k in $\{1, ..., d^b\}$ **do**
  Calculer Jb_k
end for
for all k in $\{1, ..., d^b\}$ **do**
  for all l in $\{1, ..., d^b\}$ **do**
    Calculer et sauvegarder $(Jb_k, Jb_l)_H$ [Terme de $r_\mu^{\hat{N}}$ indépendant de $\hat{N}$]
  end for
end for
Initialiser $\hat{N} = 1$
Choisir $\mu_1 \in \mathcal{P}_{\text{trial}}$ aléatoirement et initialiser $\mathcal{P}_{\text{select}} = \{\mu_1\}$
Calculer u_{μ_1}
Initialiser $\hat{\mathcal{V}}_1 = \text{Span}\{u_{\mu_1}\}$
for all k in $\{1, ..., d^b\}$ **do**
  Calculer et sauvegarder $b_k(u_{\mu_1})$ [Premier coefficient du membre de droite de (1.47)]
end for
for all k in $\{1, ..., d^a\}$ **do**
  Calculer et sauvegarder $a_k(u_{\mu_1}, u_{\mu_1})$ [Premier coefficient du membre de gauche de (1.47)]
  Calculer $Ja_k(u_{\mu_1}, \cdot)$
  for all l in $\{1, ..., d^b\}$ **do**
    Calculer et sauvegarder $(Jb_l, Ja_k(u_{\mu_1}, \cdot))_H$ [Terme de $r_\mu^{\hat{N}}$ linéaire en $\hat{N}$]
  end for
end for
for all k in $\{1, ..., d^a\}$ **do**
  for all p in $\{1, ..., d^a\}$ **do**
    Calculer et sauvegarder $(Ja_k(u_{\mu_1}, \cdot), Ja_p(u_{\mu_1}, \cdot))_H$ [Terme de $r_\mu^{\hat{N}}$ quadratique en $\hat{N}$]
  end for
end for
while $\max_{\mu \in \mathcal{P}_{\text{trial}}} \frac{r_\mu^{\hat{N}}}{\alpha_{\text{coer}}} \geq$ tolérance **do**
  Trouver $\mu_{\hat{N}+1} \in \underset{\mu \in \mathcal{P}_{\text{trial}}}{\text{argmax}} \left(r_\mu^{\hat{N}} \right)$ [Selection du paramètre qui maximise l'estimateur d'erreur]
  Incrémenter $\mathcal{P}_{\text{select}} = \mathcal{P}_{\text{select}} \cup \{\mu_{\hat{N}+1}\}$
  Calculer $u_{\mu_{\hat{N}+1}}$
  Incrémenter $\hat{\mathcal{V}}_{\hat{N}+1} = \text{Span}\{\hat{\mathcal{V}}_{\hat{N}}, u_{\mu_{\hat{N}+1}}\}$
  for all k in $\{1, ..., d^b\}$ **do**
    Calculer et sauvegarder $b_k(u_{\mu_{\hat{N}+1}})$ [Incrémenter le membre de droite de (1.47)]
  end for
  for all k in $\{1, ..., d^a\}$ **do**
    Calculer $Ja_k(u_{\mu_{\hat{N}+1}}, \cdot)$
  end for
  for all k in $\{1, ..., d^a\}$ **do**
    for all i in $\{1, ..., \hat{N}+1\}$ **do**
      Calculer et sauvegarder $a_k(u_{\mu_i}, u_{\mu_{\hat{N}+1}})$ et $a_k(u_{\mu_{\hat{N}+1}}, u_{\mu_i})$
      [Incrémenter le membre de gauche de (1.47)]
      for all l in $\{1, ..., d^b\}$ **do**
        Calculer et sauvegarder $(Jb_l, Ja_k(u_{\mu_{\hat{N}+1}}, \cdot))_H$ [Terme de $r_\mu^{\hat{N}}$ linéaire en $\hat{N}$]
      end for
      for all p in $\{1, ..., d^a\}$ **do**
        Calculer et sauvegarder $(Ja_k(u_{\mu_{\hat{N}+1}}, \cdot), Ja_p(u_{\mu_i}, \cdot))_H$ [Terme de $r_\mu^{\hat{N}}$ quadratique en $\hat{N}$]
      end for
    end for
  end for
  $\hat{N} \leftarrow \hat{N} + 1$ [Incrémenter la taille de la base réduite]
end while

Résultats de convergence

Les performances de la méthode des bases réduites sont quantifiées par la décroissance de l'erreur $\|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H$ avec $\hat{N}$ uniformément en μ. D'après le lemme de Cea,

$$\|\hat{u}_\mu^{\hat{N}} - u_\mu\|_H \leq \gamma^{-1} \inf_{v_{\hat{N}} \in \hat{\mathcal{V}}_{\hat{N}}} \|u_\mu - v_{\hat{N}}\|_H. \tag{1.55}$$

Pour que l'erreur soit faible, il faut donc que $\hat{\mathcal{V}}_{\hat{N}}$ approche bien l'ensemble $F := \{u_\mu, \mu \in \mathcal{P}\}$. L'épaisseur de Kolmogorov permet de quantifier cette notion :

$$d_{\hat{N}}(F) := \inf_{Y_{\hat{N}} \subset H,\ \dim(Y_{\hat{N}}) = \hat{N}} \sup_{u \in F} \inf_{v_{\hat{N}} \in Y_{\hat{N}}} \|u - v_{\hat{N}}\|_H. \tag{1.56}$$

Si $d_{\hat{N}}(F)$ décroît rapidement avec $\hat{N}$, la méthode des bases réduites est susceptible de donner une bonne approximation de u_μ, pour tout $\mu \in \mathcal{P}$ et pour $\hat{N} \ll N$. Des vitesses de convergence pour la méthode des bases réduites n'ont pas encore été démontrées dans le cas général, mais quelques résultats théoriques ont été proposés récemment dans [24, 15, 36]. Considérons

$$\sigma_{\hat{N}}(F) := \sup_{u \in F} \inf_{v_{\hat{N}} \in \hat{\mathcal{V}}_{\hat{N}}} \|u - v_{\hat{N}}\|_H, \tag{1.57}$$

où $\sigma_{\hat{N}}(F)$ représente l'erreur d'approximation faite lorsque $\hat{\mathcal{V}}_{\hat{N}}$ est construit par l'algorithme glouton considéré ici ; $d_{\hat{N}}(F)$ désigne alors la plus petite de ces erreurs d'approximation parmi les sous-espaces de F de dimension $\hat{N}$. Le résultat le plus récent et le plus fort connu à ce jour est le suivant ([36, Corollary 3.3 (i)]) :

Supposons que F est un sous-espace compact de H, alors pour tout $\hat{N} \geq 1$,

$$\sigma_{\hat{N}}(F) \leq \sqrt{2}\gamma^{-1} \min_{1 \leq m \leq \hat{N}} \{d_m(F)\}^{\frac{\hat{N}-m}{\hat{N}}}. \tag{1.58}$$

En particulier, $\sigma_{2\hat{N}}(F) \leq \sqrt{2}\gamma^{-1}\sqrt{d_{\hat{N}}(F)}$. Ce résultat contient et affine des résultats énoncés précédemment dans [24, 15], en explicitant certaines constantes ([36, Corollary 3.3 (ii)-(iii)]) :

Supposons que F est un sous-espace compact de H. S'il existe des constantes $C_0, \alpha > 0$ telles que

$$d_{\hat{N}}(F) \leq C_0 \hat{N}^{-\alpha}, \tag{1.59}$$

alors

$$\sigma_{\hat{N}}(F) \leq 2^{5\alpha+1}\gamma^{-2} C_0 \hat{N}^{-\alpha}. \tag{1.60}$$

S'il existe des constantes $C_0, c_0, \alpha > 0$ telles que

$$d_{\hat{N}}(F) \leq C_0 e^{-c_0 \hat{N}^\alpha}, \tag{1.61}$$

alors

$$\sigma_{\hat{N}}(F) \leq \sqrt{2C_0}\gamma^{-1} e^{-2^{-1-2\alpha} c_0 \hat{N}^\alpha}. \tag{1.62}$$

L'estimation (1.58) permet d'obtenir une vitesse de convergence dans le cas où les solutions de (1.46) sont bornées uniformément en μ dans un espace un peu plus régulier que H :

Supposons que $H = H^1(\Omega)$ ou $H = H^1_0(\Omega)$, qu'il existe une constante C indépendante de μ telle que pour tout $\mu \in \mathcal{P}$, $\|u_\mu\|_{H^2(\Omega)} \leq C$ et que F est un sous-espace compact de H. Alors il existe une constante $C' > 0$ telle que

$$\sup_{\mu\in\mathcal{P}} \|\hat{u}_\mu^{2\hat{N}} - u_\mu\|_{H^1(\Omega)} \leq C' \hat{N}^{-\frac{1}{2d_x}}, \tag{1.63}$$

où d_x est la dimension de l'espace d'arrivée des fonctions de $\mathcal{V}$. Nous prouvons maintenant ce résultat. En utilisant le lemme de Céa (1.55), il vient

$$\sup_{\mu\in\mathcal{P}} \|\hat{u}_\mu^{2\hat{N}} - u_\mu\|_{H^1(\Omega)} \leq \gamma^{-1} \sup_{\mu\in\mathcal{P}} \inf_{v_{2\hat{N}}\in\mathcal{V}^{2\hat{N}}} \|v_{2\hat{N}} - u_\mu\|_{H^1(\Omega)} = \gamma^{-1}\sigma_{2\hat{N}}(F). \tag{1.64}$$

D'après le théorème (1.58),

$$\sup_{\mu\in\mathcal{P}} \|\hat{u}_\mu^{2\hat{N}} - u_\mu\|_{H^1(\Omega)} \leq \sqrt{2}\gamma^{-2}\sqrt{d_{\hat{N}}(F)}. \tag{1.65}$$

Considérons une base infinie $e_1, e_2, \cdots$ de $H^1(\Omega)$ et désignons par $E_{\hat{N}}$ l'espace vectoriel engendré par $e_1, \cdots, e_{\hat{N}}$. En utilisant des résultats d'approximation classiques, il est possible de choisir cette base telle qu'il existe une constante C'' indépendante de $\hat{N}$ telle que pour tout $f \in H^2(\Omega)$, il existe $g \in E_{\hat{N}}$ tel que

$$\|f - g\|_{H^1(\Omega)} \leq C'' \hat{N}^{-\frac{1}{d_x}} \|f\|_{H^2(\Omega)}. \tag{1.66}$$

On peut par exemple prendre pour $E_{\hat{N}}$ l'espace des polynômes continus affines par morceaux sur un maillage régulier de taille de maille $h \sim \hat{N}^{-\frac{1}{d}}$. En particulier, pour tout $f \in H^2(\Omega)$,

$$\inf_{g\in E_{\hat{N}}} \|f - g\|_{H^1(\Omega)} \leq C'' \hat{N}^{-\frac{1}{d_x}} \|f\|_{H^2(\Omega)}. \tag{1.67}$$

Comme $F \subset H^2(\Omega)$ par hypothèse,

$$\sup_{f\in F} \inf_{g\in E_{\hat{N}}} \|f - g\|_{H^1(\Omega)} \leq C'' C \hat{N}^{-\frac{1}{d_x}}, \tag{1.68}$$

si bien que

$$d_{\hat{N}}(F) \leq C'' C \hat{N}^{-\frac{1}{d_x}}. \tag{1.69}$$

En utilisant cette borne dans (1.65), la relation (1.63) est obtenue avec $C' = \sqrt{2CC''}\gamma^{-2}$.

Nous remarquons que la vitesse de convergence est indépendante de la dimension du paramètre μ. Considérons les classes de problèmes elliptiques suivantes :

$$\begin{cases} L_\mu u = f_\mu, \text{ dans } \Omega, \\ \quad u = 0, \quad \text{sur } \partial\Omega, \end{cases} \tag{1.70}$$

dans $H^1_0(\Omega)$ et

$$\begin{cases} \quad L_\mu u = f_\mu, \text{ dans } \Omega, \\ \boldsymbol{\nabla} u \cdot \mathbf{n} = 0, \quad \text{sur } \partial\Omega, \end{cases} \tag{1.71}$$

dans $H^1(\Omega)$, où $\mathbf{n}$ est la normale sortante sur $\partial\Omega$ et

$$L_\mu u = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(a_{\mu_{ij}} \frac{\partial}{\partial x_i} u\right), \tag{1.72}$$

avec Ω ouvert borné de classe C^2. On suppose de plus que $a_{\mu_{ij}} \in C^1(\bar{\Omega})$ uniformément en μ et qu'il existe une constante $c > 0$ telle que pour tout $\mu \in \mathcal{P}$, $\|f_\mu\|_{L^2(\Omega)} \leq c$ (ou simplement que f_μ est indépendant de μ). D'après [22, Théorèmes IX.25 et IX.26], il existe une constante $c' > 0$ qui ne dépend que de Ω telle que $\|u_\mu\|_{H^2(\Omega)} \leq c'\|f_\mu\|_{L^2(\Omega)}$. Par injection compacte de $H^2(\Omega)$ dans $H^1(\Omega)$, F est compact dans $H^1(\Omega)$, et l'estimation (1.63) peut être appliquée aux problèmes (1.70) et (1.71) avec $C = cc'$.

1.5 Contenu de la thèse

Cette thèse contient 2 parties et une annexe. Certains chapitres de ce manuscrit sont issus d'articles soumis ou publiés. Ils correspondent donc à l'origine à des travaux indépendants les uns des autres, écrits en langue anglaise. L'auteur s'excuse des quelques répétitions et changements de notation qui en résultent.

1.5.1 Production scientifique

Articles parus ou à paraître dans des revues à comité de lecture

[Ar1] F.C., Accurate a posteriori error evaluation in the reduced basis method, Comptes Rendus Mathematique, 350(9-10):539 - 542, 2012.

[Ar2] F.C., M. Ghattassi et R. Joubaud, A multiscale problem in thermal science, ESAIM: PROCEEDINGS, décembre 2012, Vol. 38, p. 202-219.

[Ar3] F.C., A. Ern et T. Lelièvre, Accurate and online-efficient evaluation of the a posteriori error bound in the reduced basis method, accepté pour publication dans Mathematical Modelling and Numerical Analysis, 2013.

[Ar4] F.C., A. Ern et G. Sylvand, Coupled BEM-FEM for the convected Helmholtz equation with non-uniform flow in a bounded domain, Journal of Computational Physics 257-A (2014), 627-644.

Prépublications

[Pr1] F.C., A. Ern, T. Lelièvre et G. Sylvand, A nonintrusive method to approximate linear systems with nonlinear parameter dependence.

[Pr2] N. Balin, F.C., F. Dubois, E. Duceau, S. Duprey, I. Terrasse, Boundary element and finite element coupling for aeroacoustic simulations.

[Pr3] F.C., A. Ern et T. Lelièvre, A nonintrusive Reduced Basis Method applied to aeroacoustic simulations

1.5.2 Plan de la thèse

La partie I regroupe les travaux effectués en acoustique.

Le chapitre 2 présente la formulation intégrale utilisée par EADS-IW pour résoudre les problèmes de diffraction d'ondes acoustiques dans l'air au repos par un objet dit impédant (les ondes sont partiellement absorbées à la surface de l'objet), voir la figure 1.6. L'équation étudiée

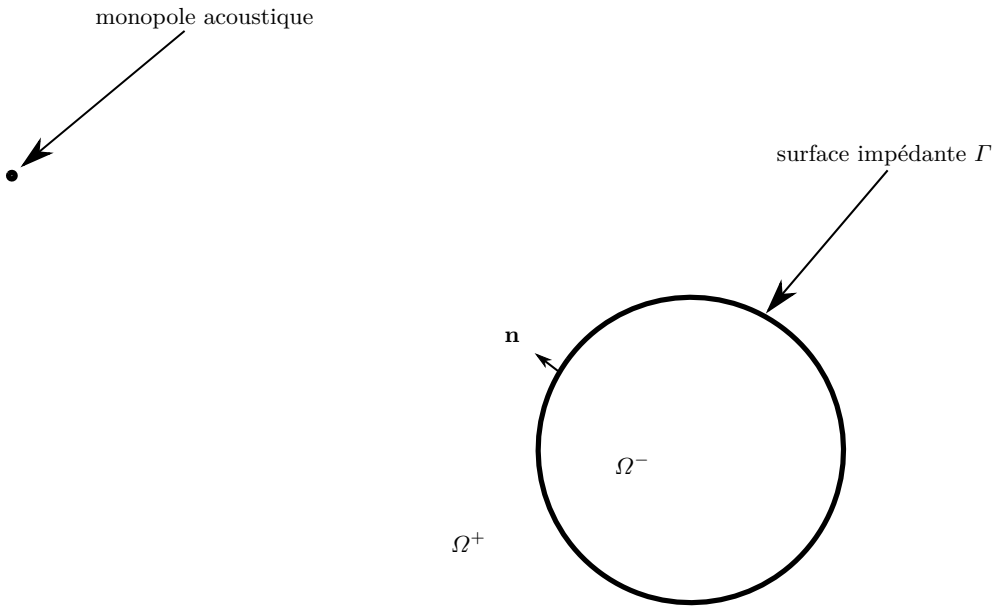

Fig. 1.6. Géométrie du cas test pour le problème impédant

est l'équation d'Helmholtz classique (1.28) et la méthodologie présentée dans la section 1.3 est utilisée. En particulier, la condition de Robin modélisant le caractère impédant de l'objet complète les équations (1.32) pour donner un système d'équations bien posé sur les potentiels acoustiques. La preuve d'existence et unicité du problème obtenu est présentée et la preuve de la stabilité inf-sup de l'approximation en dimension finie par éléments de frontière est rappelée en suivant les travaux de [55]. La propriété de stabilité inf-sup est essentielle pour pouvoir utiliser la méthode des bases réduites : elle généralise l'hypothèse de coercivité faite dans la présentation des bases réduites dans la section 1.4.3. Enfin, les problèmes liés à l'existence de fréquences de résonance dans l'utilisation de certaines formulations intégrales sont rappelés et une solution à ce problème, inspirée des travaux de [23], est présentée. Le contenu de ce chapitre n'est pas nouveau, mais sert à poser les bases des raisonnements présentés dans le chapitre suivant.

Le chapitre 3 reprend l'article [Ar4] et traite le problème de diffraction d'onde acoustique par un objet solide dans un écoulement potentiel dans un domaine borné Ω^- de l'espace et uniforme à l'extérieur de ce domaine, voir la figure 1.7. L'équation étudiée est l'équation d'Helmholtz convectée (1.27). La transformation de Prandtl–Glauert consiste en un changement de variable et un changement de fonction inconnue. Lorsque l'équation d'Helmholtz est convectée par un écoulement uniforme, il est connu qu'une transformation de Prandtl–Glauert particulière

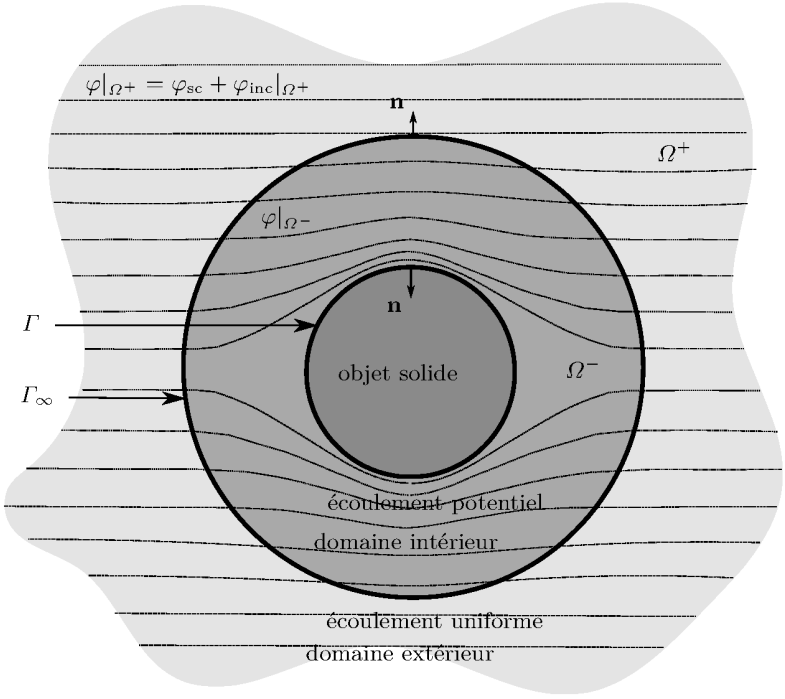

Fig. 1.7. Géométrie du cas test pour le problème aéroacoustique dans un écoulement potentiel dans un domaine borné Ω^- et uniforme à l'extérieur de ce domaine

permet de se ramener à l'équation d'Helmholtz classique. Nous disons alors que cette transformation de Prandtl–Glauert est adaptée à l'écoulement uniforme. Notre contribution consiste à appliquer la transformation de Prandtl–Glauert adaptée à l'écoulement uniforme du domaine extérieur à l'équation d'Helmholtz convectée par un écoulement potentiel dans le domaine Ω^-. Nous disposons donc de cette équation à coefficients dépendant de la variable spatiale dans le domaine Ω^-, où une méthode de résolution locale doit être utilisée (nous avons choisi la méthode des éléments finis tridimensionnels), et de l'équation d'Helmholtz classique dans le domaine Ω^+ où la méthode des équations intégrales pour l'équation d'Helmholtz classique présentée dans la section 1.3 peut être utilisée. Ces deux équations sont posées dans le même système de coordonnées et pour les mêmes fonctions inconnues, permettant un couplage direct. L'avantage de retrouver l'équation d'Helmholtz classique dans le domaine Ω^+ est essentiel, car il permet l'utilisation de codes existants et nous libère de la tâche de développer un code d'éléments de frontière résolvant Helmholtz convecté par un écoulement uniforme. Deux formulations couplées éléments de frontière / éléments finis sont proposées. La première est directement inspirée des premières méthodes de couplage proposées dans la littérature mais est instable à certaines fréquences dites de résonance. La deuxième formulation est stabilisée pour être bien posée à toutes les fréquences, en utilisant la méthode évoquée dans le chapitre 2. Les preuves d'existence et unicité des deux formulations sont données. En particulier, la première formulation est bien posée hors des fréquences de résonance et admet une infinité de solutions aux fréquences de résonance, tandis que la deuxième formulation est bien posée à toutes les fréquences.

Enfin, le chapitre 4 présente plusieurs expériences numériques à vocation de validation des formulations obtenues dans le chapitre 3 et d'illustration sur des cas tests industriels. En particulier, une application industrielle de la prépublication [Pr2] est présentée.

La partie II regroupe les travaux effectués sur la méthode des bases réduites.

Le chapitre 5 reprend la publication [Ar3], qui propose une amélioration de la méthode introduite dans la publication [Ar1] (cette dernière n'est pas reproduite dans le présent manuscrit). Comme expliqué dans la section 1.4.3, le succès de la méthode des bases réduites repose sur la formule (1.54) pour l'estimateur a posteriori, car les termes de cette formule sont soit précalculables dans la phase offline de l'algorithme, soit calculables en complexité indépendante de la taille du problème. Cependant, comme illustré dans [Ar1] et [Ar3], cette formule est très sensible aux erreurs d'arrondis machine : au fur et à mesure que la méthode des bases réduites converge, les valeurs prises par l'estimateur d'erreur évalué selon la formule (1.54) sont typiquement de plusieurs ordres de grandeur plus grandes que l'erreur réellement commise. La formule (1.54) devient alors inopérante pour calculer l'estimateur d'erreur de manière précise. Le problème rencontré est similaire à celui qui intervient lorsque l'on évalue un polynôme près de ses racines : le résultat renvoyé par l'ordinateur peut être très différent du résultat exact. Dans [Ar1], nous avons proposé une nouvelle formule pour l'estimateur d'erreur a posteriori, sous la forme d'une combinaison linéaire d'évaluations de l'estimateur à des valeurs données du paramètre, qui sont calculées avec la formule $\|G_\mu \hat{u}_\mu\|_{\mathcal{V}}$ pendant la phase offline. Cette dernière formule ne peut pas être utilisée dans la phase online, mais ne rencontre pas les problèmes de précision machine évoqués ci-dessus. Les coefficients de la combinaison linéaire dépendent de la valeur du paramètre en lequel nous souhaitons évaluer l'estimateur d'erreur et sont calculés en résolvant un système linéaire. La difficulté qui subsiste est que ce système linéaire peut être très mal conditionné dans certains cas. Dans [Ar3], nous résolvons ce problème en introduisant une nouvelle formule sur le même modèle que celle de [Ar1], mais dont le système linéaire à résoudre pour déterminer les coefficients de la combinaison linéaire est toujours bien conditionné. La stratégie utilisée est basée sur la méthode d'interpolation empirique. Un exemple numérique académique est proposé et la méthode est également testée sur le problème acoustique décrit dans le chapitre 2, avec comme paramètre le coefficient d'impédance des objets, et où la stabilité inf-sup a été prouvée. Dans ce problème, la constante inf-sup n'est pas indépendante du paramètre et une borne inférieure de cette constante n'est pas connue a priori, contrairement à ce qui a été supposé dans la section 1.4.3 où la constante de coercivité a été supposée indépendante du paramètre. Une possibilité pour calculer une borne inférieure de la constante inf-sup consiste à utiliser la méthode des contraintes successives [56], qui est rappelée à la fin de ce chapitre.

Le chapitre 6 reprend la prépublication [Pr1]. Dans la section 1.4.3, nous avons insisté sur le fait que la dépendance de l'opérateur et du second membre en les paramètres doit être affine pour pouvoir utiliser la méthode des bases réduites. Lorsque ce n'est pas le cas, une possibilité consiste à utiliser la méthode d'interpolation empirique pour rétablir l'hypothèse de dépendance affine de façon approchée. Cependant, cette méthode est intrusive, car elle nécessite de modifier les routines d'assemblage élémentaire du code considéré. Dans ce chapitre, nous proposons une méthode qui permet de rétablir l'hypothèse de dépendance affine de manière non intrusive, dans le sens où les routines d'assemblage élémentaires du code ne sont pas modifiées. Notre solution est également basée sur la méthode d'interpolation empirique et ne repose que sur l'hypothèse très raisonnable que l'utilisateur puisse récupérer les matrices complètes assemblées en certaines

valeurs du paramètre qu'il aura sélectionnées de manière judicieuse. Un exemple numérique académique est proposé,et la méthode est testée sur un problème intégral résolu sur un maillage d'avion complet, ainsi que sur la deuxième formulation du problème aéroacoustique décrit dans le chapitre 3.

Le chapitre 7 est basé sur la prépublication [Pr3]. Ce chapitre reprend l'idée de nonintrusivité introduite dans le chapitre 6. En particulier, de nombreuses variantes possibles pour obtenir une procédure non intrusive sont présentées, et les choix du chapitre 6 sont motivés. Enfin, la procédure est appliquée à la réduction de modèle par bases réduites aux problèmes aéroacoustiques présentés dans la partie I, et non simplement aux matrices du problème comme c'était le cas dans le chapitre 6. Les applications numériques de ce chapitre illustrent les contributions apportées par les deux parties de cette thèse.

Le chapitre 8 reprend la publication [Ar2]. Il décrit un travail réalisé à l'école d'été CEMRACS 2011 à Luminy. Il est motivé par la problématique industrielle de simulation rapide du champ de température dans une cabine d'avion, en présence de sources thermiques situées dans la soute et produites par des composants électroniques. Dans un premier temps, nous écrivons un solveur 2D en FreeFem++ pour résoudre les équations de Navier–Stokes incompressibles puis l'approximation de Boussinesq. Ensuite, nous considérons l'équation de la chaleur sous convection constante et appliquons la méthode des bases réduites en prenant comme paramètres la conductivité thermique de différents éléments des composants électroniques.

Enfin, le chapitre B en annexe élargit la perspective industrielle et présente l'étude d'un modèle d'incertitudes non paramétriques, dans un contexte de vibration de plaques, pour résoudre un problème d'optimisation sous contraintes en probabilité. Dans le contexte d'aéroacoustique en aviation civile, une fois que nous disposons d'une méthode pour simuler la propagation du bruit généré par le turboréacteur, nous pouvons étudier la puissance acoustique transmise aux passagers à travers le fuselage. Nous déterminons une condition nécessaire à imposer au modèle des plaques du fuselage pour garantir que la probabilité que la puissance acoustique transmise ne dépasse pas un certain seuil soit inférieure à un seuil de tolérance donné.

Part I

Two aeroacoustic problems solved by integral equations

2

Acoustic scattering by an impedant object

In this chapter, we consider the acoustic scattering in the air at rest by an object whose surface is impedant, meaning that any incident wave will be partially absorbed and partially scattered. In what follows, such an object is called impedant. The goal here is to give the mathematical justification we need in Part II to apply the Reduced Basis Method for this problem. We point out why the variational formulation cannot be thought in the natural Sobolev spaces. We show that the Fredholm alternative framework can be recovered in a different functional setting, leading to a well-posed variational formulation at all frequencies when the surface of the object is Lipschitz. We show that the classical discrete approximation is inf-sup stable. Finally, we discuss invertibility issues caused by the presence of resonant frequencies of some integral formulation and present a remedy taken from the literature.

2.1 Physical setting

We consider a bounded object $\Omega^- \subset \mathbb{R}^3$ with boundary Γ and $\Omega^+ := \mathbb{R}^3 \backslash \overline{\Omega^-}$, where $\overline{\Omega^-}$ denotes the closure of Ω^-. Any object such that Γ is Lipschitz can be considered, but we take a ball for simplicity of the presentation, see Figure 2.1. We consider a monopole source located in Ω^+. The surface of the ball is impedant: the proportion of absorbed and scattered parts is quantified by the impedance coefficient μ, which is used in a Robin boundary condition at Γ. We suppose that $\mu > 0$. We are interested in the computation of the scattered field p_{sc} in Ω^+. We denote p_{inc} the known pressure field of wave number k created by the source in the absence of the sphere; the total acoustic field in Ω^+ is the sum of p_{inc} and p_{sc}. The complex conjugation is denoted by $\bar{\cdot}$.

2.2 Weak formulation

2.2.1 Preliminaries

We recall that a distribution u ($u^{+,-} := u|_{\Omega^{+,-}}$) is a piecewise radiating Helmholtz solution, if

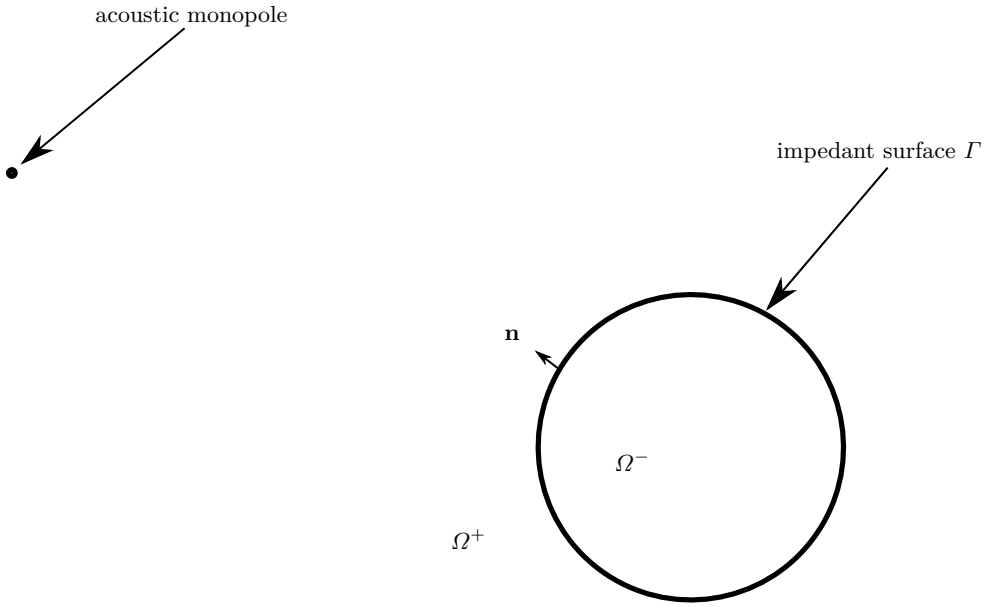

Fig. 2.1. Geometry for the three-dimensional acoustic scattering problem.

$$\begin{cases} \Delta u^{+,-} + k^2 u^{+,-} = 0 \quad \text{in } \Omega^+ \cup \Omega^-, \\ \lim\limits_{r\to+\infty} r\left(\dfrac{\partial u^+}{\partial r} - iku^+\right) = 0, \end{cases} \tag{2.1}$$

where the second equation is the Sommerfeld radiation condition. Let

$$M := \begin{pmatrix} N & \frac{1}{2}I + \tilde{D} \\ \frac{1}{2}I - D & S \end{pmatrix}, \tag{2.2}$$

where N, D, $\tilde{D}$ and S have been defined in (1.31). A piecewise radiating Helmholtz solution u verifies

$$M\begin{pmatrix} [\gamma_0 u] \\ [\gamma_1 u] \end{pmatrix} = -\begin{pmatrix} \gamma_1^- u^- \\ \gamma_0^- u^- \end{pmatrix}. \tag{2.3}$$

The operator M is not injective: define w in such a manner that $w|_{\Omega^-} = 0$ and $w|_{\Omega^+}$ is any non-zero field satisfying the Helmholtz equation and the Sommerfeld radiation condition at infinity. For instance, consider the problem

$$\begin{cases} \Delta v + k^2 v = 0 & \text{in } \Omega^+, \\ \gamma_0^+ v = g & \text{on } \Gamma, \\ \lim\limits_{r\to+\infty} r\left(\dfrac{\partial v}{\partial r} - ikv\right) = 0, \end{cases}$$

where $g \in H^{\frac{1}{2}}(\Gamma)$ is nonzero. This problem has a unique solution (see [76, Theorem 9.11 p.288]), which can be chosen for $w|_{\Omega^+}$. Since $w|_{\Omega^+}$ is non-zero, $[\gamma_0 w]$ and $[\gamma_1 w]$ cannot be both zero owing to the representation formula (1.30). However,

$$M\begin{pmatrix}[\gamma_0 w]\\ [\gamma_1 w]\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}. \tag{2.4}$$

Therefore, we have constructed a non-zero element of $\mathrm{Ker}(M)$. We can understand that M is not injective by observing that both relations in the system (2.3) being obtained from the same relation (1.30), they are not independent. In other words, the system (2.3) is not sufficient to determine the unknown potentials $[\gamma_0 u]$ and $[\gamma_1 u]$. It is merely a necessary relation satisfied by any radiating Helmholtz solution. Actually, no boundary condition has been enforced yet, and the system (2.3) is still valid for Dirichlet, Neumann or Robin boundary conditions on Γ. We will see that injecting a Robin boundary condition enables to recover the injectivity property.

2.2.2 The Robin boundary condition

The unknown field p_{sc} solves the following boundary value problem:

$$\begin{cases} \Delta p_{\mathrm{sc}} + k^2 p_{\mathrm{sc}} = 0 & \text{in } \Omega^+,\\ \gamma_1^+ p_{\mathrm{sc}} + i\dfrac{k}{\mu}\gamma_0^+ p_{\mathrm{sc}} = -\gamma_1 p_{\mathrm{inc}} - i\dfrac{k}{\mu}\gamma_0 p_{\mathrm{inc}} & \text{on } \Gamma,\\ \lim\limits_{r\to+\infty} r\left(\dfrac{\partial p_{\mathrm{sc}}}{\partial r} - ikp_{\mathrm{sc}}\right) = 0, \end{cases} \tag{2.5}$$

where p_{inc} is $\mathcal{C}^1$ in a neighborhood of Γ (leading to $\gamma_0 p_{\mathrm{inc}} := \gamma_0^+ p_{\mathrm{inc}} = \gamma_0^- p_{\mathrm{inc}}$ and the same for $\gamma_1 p_{\mathrm{inc}}$).

We define the distribution $v : \Omega^+ \cup \Omega^- \longrightarrow \mathbb{C}$ such that $v_{|\Omega^-} = -p_{\mathrm{inc}}$, $v_{|\Omega^+} = p_{\mathrm{sc}}$. We introduce the jumps $\chi := [\gamma_0 v]$ and $\lambda := [\gamma_1 v]$. We have $\chi = \gamma_0^+ p_{\mathrm{sc}} + \gamma_0 p_{\mathrm{inc}}$ and $\lambda = \gamma_1^+ p_{\mathrm{sc}} + \gamma_1 p_{\mathrm{inc}}$, so that, using the Robin boundary condition on Γ in the system (2.5), there holds

$$\lambda + i\frac{k}{\mu}\chi = 0. \tag{2.6}$$

Many choices for v are possible (for instance, $v_{|\Omega^-} = 0$). Our choice was motivated by the fact that the jumps χ and λ correspond to the exterior trace of the total acoustic potential. Other choices may have lead to jumps that have no physical meaning. Since v is a piecewise radiating Helmholtz solution, there holds

$$M\begin{pmatrix}\chi\\ \lambda\end{pmatrix} = -\begin{pmatrix}\gamma_1 p_{\mathrm{inc}}\\ \gamma_0 p_{\mathrm{inc}}\end{pmatrix}. \tag{2.7}$$

Injecting the Robin boundary condition, we obtain

$$\begin{pmatrix} N - \frac{ik}{2\mu}I & \tilde{D}\\ D & -S - \frac{i\mu}{2k}I\end{pmatrix}\begin{pmatrix}\chi\\ \lambda\end{pmatrix} = \begin{pmatrix}\gamma_1 p_{\mathrm{inc}}\\ -\gamma_0 p_{\mathrm{inc}}\end{pmatrix}. \tag{2.8}$$

The natural functional spaces of the involved integral operators are $N : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$, $S : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$, $D : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ and $\tilde{D} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ (more details are given in Section 3.3.2). The variational form is as follows: find $(\chi, \lambda) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ such that $\forall \left(\hat{\chi}, \hat{\lambda}\right) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$,

$$\begin{cases} \left\langle N\chi - \dfrac{ik}{2\mu}\chi, \hat{\chi} \right\rangle_{-\frac{1}{2},\frac{1}{2}} + \left\langle \tilde{D}\lambda, \hat{\chi} \right\rangle_{-\frac{1}{2},\frac{1}{2}} = \langle \gamma_1 p_{\text{inc}}, \hat{\chi} \rangle_{-\frac{1}{2},\frac{1}{2}}, \\ \left(D\chi, \hat{\lambda}\right) - \left(S\lambda + \dfrac{i\mu}{2k}\lambda, \hat{\lambda}\right) = -\left(\gamma_0 p_{\text{inc}}, \hat{\lambda}\right), \end{cases} \tag{2.9}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Gamma)$ inner product: $(\lambda,\chi) := \int_\Gamma \bar{\lambda}\chi$, and $\langle\cdot,\cdot\rangle_{-\frac{1}{2},\frac{1}{2}}$ denotes its extension to a duality pairing on $H^{-\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\Gamma)$.

Remark 2.1 *In the first equation of* (2.9), $\frac{ik}{2\mu}\chi$ *makes sense as an element of* $H^{\frac{1}{2}}(\Gamma) \subset H^{-\frac{1}{2}}(\Gamma)$. *In the second equation, we have taken* $\lambda, \hat{\lambda} \in L^2(\Gamma)$. *We cannot take* $\lambda, \hat{\lambda} \in H^{-\frac{1}{2}}(\Gamma)$ *since the product* $\left(\hat{\lambda}, \lambda\right)$ *would not make sense. Furthermore, taking* $\lambda, \hat{\lambda} \in H^{\frac{1}{2}}(\Gamma)$ *would make the equivalence property between the variational formulation and the boundary value problem fail, since for instance,* $\forall \hat{\lambda} \in H^{\frac{1}{2}}(\Gamma)$, $\left(D\chi, \hat{\lambda}\right) = 0$ *does not imply* $D\chi = 0$. *It would be the case if* $\left(\hat{\lambda}, D\chi\right) = 0$ *holds for all* $\hat{\lambda} \in H^{-\frac{1}{2}}(\Gamma)$.

2.3 Existence and uniqueness

Consider now $\mathcal{H} := H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ and $\mathcal{H}' := H^{-\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$. The norm $\|\cdot\|_{\mathcal{H}}$ is defined as $\|(\chi,\lambda)\|_{\mathcal{H}} := \left\{ \|\chi\|^2_{H^{\frac{1}{2}}(\Gamma)} + \|\lambda\|^2_{L^2(\Gamma)} \right\}^{\frac{1}{2}}$. We define a_0 and $\tilde{a}$ as the two sesquilinear forms such that

$$\begin{aligned} a_0\left((\chi,\lambda), \left(\hat{\chi},\hat{\lambda}\right)\right) &= \langle N_0\chi, \hat{\chi}\rangle_{-\frac{1}{2},\frac{1}{2}} - i\left(S_0\lambda, \hat{\lambda}\right) - i\tfrac{k}{2\mu}\left(\chi,\hat{\chi}\right) + \tfrac{\mu}{2k}\left(\lambda,\hat{\lambda}\right), \\ \tilde{a}\left((\chi,\lambda), \left(\hat{\chi},\hat{\lambda}\right)\right) &= \langle (N-N_0)\chi, \hat{\chi}\rangle_{-\frac{1}{2},\frac{1}{2}} - i\left((S-S_0)\lambda, \hat{\lambda}\right) + \left\langle \tilde{D}\lambda, \hat{\chi}\right\rangle_{-\frac{1}{2},\frac{1}{2}} + i\left(D\chi, \hat{\lambda}\right), \end{aligned} \tag{2.10}$$

and b as the linear form such that

$$b\left(\hat{\chi}, \hat{\lambda}\right) = \langle \gamma_1 p_{\text{inc}}, \hat{\chi}\rangle_{-\frac{1}{2},\frac{1}{2}} - i\left(\gamma_0 p_{\text{inc}}, \hat{\lambda}\right). \tag{2.11}$$

We have multiplied the second equation of (2.9) by i to obtain a required coercivity result, as we will see in what follows. The operators N_0, S_0, D_0 and $\tilde{D}_0$ refer respectively to N, S, D and $\tilde{D}$ taken at zero frequency. Let A_0, $\mathcal{H} \to \mathcal{H}'$, and $\tilde{A}$, $\mathcal{H} \to \mathcal{H}'$, be the bounded operators defined respectively from a_0 and $\tilde{a}$:

$$A_0 := \begin{pmatrix} N_0 - i\frac{k}{2\mu}I & 0 \\ 0 & \frac{\mu}{2k} - iS_0 \end{pmatrix}, \ \tilde{A} := \begin{pmatrix} N - N_0 & \tilde{D} \\ iD & -i(S-S_0) \end{pmatrix} \tag{2.12}$$

We denote $A := A_0 + \tilde{A}$.

The weak form (2.9) can also be written as follows: find $(\chi,\lambda) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ such that $\forall \left(\hat{\chi},\hat{\lambda}\right) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$,

$$a\left((\chi,\lambda), \left(\hat{\chi},\hat{\lambda}\right)\right) = b\left(\hat{\chi},\hat{\lambda}\right), \tag{2.13}$$

where $a = a_0 + \tilde{a}$.

Theorem 2.2 *Problem* (2.13) *has a unique solution.*

Proof. We apply the Fredholm alternative.

i. Continuity:

We denote the continuity constants using C with a subscript. It is well-known (see [76, Theorem 6.11]) that $\|N\chi\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C_N\|\chi\|_{H^{\frac{1}{2}}(\Gamma)}$, $\|D\chi\|_{H^{\frac{1}{2}}(\Gamma)} \leq C_D\|\chi\|_{H^{\frac{1}{2}}(\Gamma)}$, $\|S\lambda\|_{H^{\frac{1}{2}}(\Gamma)} \leq C_S\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma)}$, and $\|\tilde{D}\lambda\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C_{\tilde{D}}\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma)}$. These relations hold as well for at the zero-frequency particular case. Moreover, $\|D\chi\|_{L^2} \leq \|D\chi\|_{H^{\frac{1}{2}}(\Gamma)} \leq C_D\|\chi\|_{H^{\frac{1}{2}}(\Gamma)}$, $\|\tilde{D}\lambda\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C_{\tilde{D}}\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C_{\tilde{D}}\|\lambda\|_{L^2(\Gamma)}$, and $\|S\lambda\|_{L^2(\Gamma)} \leq \|S\lambda\|_{H^{\frac{1}{2}}(\Gamma)} \leq C_S\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C_S\|\lambda\|_{L^2(\Gamma)}$. Therefore,

$$\begin{aligned}
&|a_0\left((\chi,\lambda),\left(\hat{\chi},\hat{\lambda}\right)\right)| \leq \left(C_{N_0} + C_{S_0} + \tfrac{k}{2\mu} + \tfrac{\mu}{2k}\right)\|(\chi,\lambda)\|_{\mathcal{H}}\|(\hat{\chi},\hat{\lambda})\|_{\mathcal{H}},\\
&|\tilde{a}\left((\chi,\lambda),\left(\hat{\chi},\hat{\lambda}\right)\right)| \leq \left(C_{N_0} + C_N + C_{S_0} + C_S + C_D + C_{\tilde{D}}\right)\|(\chi,\lambda)\|_{\mathcal{H}}\|(\hat{\chi},\hat{\lambda})\|_{\mathcal{H}},\\
&|b\left(\hat{\chi},\hat{\lambda}\right)| \leq \left(\|\gamma_1 p_{\text{inc}}\|_{H^{-\frac{1}{2}}(\Gamma)} + \|\gamma_0 p_{\text{inc}}\|_{L^2(\Gamma)}\right)\|(\hat{\chi},\hat{\lambda})\|_{\mathcal{H}}.
\end{aligned} \tag{2.14}$$

ii. Coercivity up to a compact perturbation:

We denote the coercivity constants using K with a subscript. From [76, Theorem 8.12 p.261], $\langle S_0\lambda,\lambda\rangle_{\frac{1}{2},-\frac{1}{2}} \in \mathbb{R}$, $\forall\lambda \in H^{-\frac{1}{2}}(\Gamma)$. Therefore, $(S_0\lambda,\lambda) \in \mathbb{R}$, $\forall\lambda \in L^2(\Gamma)$. From [76, Theorem 8.21 p.267], $\langle N_0\chi,\chi\rangle_{-\frac{1}{2},\frac{1}{2}} \in \mathbb{R}$, $\forall\chi \in H^{\frac{1}{2}}(\Gamma)$ and N_0 is coercive on $H^{\frac{1}{2}}(\Gamma)$. Then,

$$\begin{aligned}
\operatorname{Re}\left(a_0\left((\lambda,\chi),(\lambda,\chi)\right)\right) &= \langle N_0\chi,\chi\rangle_{-\frac{1}{2},\frac{1}{2}} + \frac{\mu}{2k}(\lambda,\lambda)\\
&\geq \min\left(K_{N_0},\frac{\mu}{2k}\right)\|(\chi,\lambda)\|^2_{\mathcal{H}}.
\end{aligned} \tag{2.15}$$

Hence, a_0 is $\mathcal{H}$-coercive. Let us now examine the compactness properties of $\tilde{A}$. First, from [93, Lemma 3.9.8], $N - N_0$ is compact from $H^{\frac{1}{2}}(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$. Then, we saw that S and S_0 are continuous from $L^2(\Gamma)$ into $H^{\frac{1}{2}}(\Gamma)$, and the injection of $H^{\frac{1}{2}}(\Gamma)$ into $L^2(\Gamma)$ is compact (we denote it as $H^{\frac{1}{2}}(\Gamma) \hookrightarrow L^2(\Gamma)$), therefore $S - S_0$ is compact from $L^2(\Gamma)$ into $L^2(\Gamma)$. The operator D is continuous from $H^{\frac{1}{2}}(\Gamma)$ into $H^{\frac{1}{2}}(\Gamma)$, and again $H^{\frac{1}{2}}(\Gamma) \hookrightarrow L^2(\Gamma)$ implies that D is compact from $H^{\frac{1}{2}}(\Gamma)$ into $L^2(\Gamma)$. Finally, we prove that $\tilde{D}$ is compact from $L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$. From [54, Lemma 3.3], the operator $T := \tilde{D} - D^*$, where D^* is the adjoint of D, is compact from $H^{-\frac{1}{2}}(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$. In particular, T is compact from $L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$. We recall that since Γ is a closed surface, $H^{-\frac{1}{2}}(\Gamma)$ is the dual of $H^{\frac{1}{2}}(\Gamma)$. Then, by the Schauder theorem, D^* is compact from $L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$ as the adjoint of a compact operator. Therefore, $\tilde{D}$ is compact from $L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma)$ as the sum of two compact operators. Hence, $\tilde{A}$ is compact.

iii. Uniqueness:

The uniqueness property is shown in two steps. Suppose that (2.9) is verified for a zero right-hand side.

Step 1: Let us prove that $\lambda = \frac{ik}{\mu}\chi$.

Let $w(x) = -\mathcal{S}\lambda(x) + \mathcal{D}\chi(x),\ \forall x \in \mathbb{R}^3\backslash\Gamma$. The distribution w is a piecewise radiating Helmholtz solution. Using Proposition 3.26, there holds

$$\begin{cases} \gamma_1^+ w = -N\chi + \left(\frac{I}{2} - \tilde{D}\right)\lambda, \\ \gamma_0^+ w = \left(\frac{I}{2} + D\right)\chi - S\lambda. \end{cases} \tag{2.16}$$

Using (2.9) with zero right-hand side, we obtain

$$\begin{cases} \gamma_1^+ w = \frac{1}{2}\left(\lambda - \frac{ik}{\mu}\chi\right), \\ \gamma_0^+ w = \frac{i\mu}{2k}\left(\lambda - \frac{ik}{\mu}\chi\right). \end{cases} \tag{2.17}$$

Let B be a ball containing Ω^-. Using Green's formula,

$$\int_{B\backslash\Omega^-} \left(|\nabla w|^2 - k^2|w|^2\right) + \left(\gamma_1^+ w, \gamma_0^+ w\right) = \int_{\partial B} w\frac{\partial \bar{w}}{\partial n}. \tag{2.18}$$

The normals on Γ and ∂B point "outward" (to infinity), see Figure 2.2. Using (2.17), $(\gamma_1 w^+, \gamma_0 w^+) =$

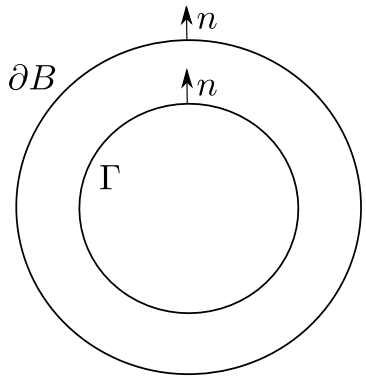

Fig. 2.2. Definition of the surface ∂B.

$\frac{i\mu}{4k}\left|\lambda - \frac{ik}{\mu}\chi\right|^2_{L^2(\Gamma)}$. Therefore, since $\mu > 0$, $\mathrm{Im}\left(\int_{\partial B} w^+ \frac{\partial \overline{w}^+}{\partial n}\right) = \frac{\mu}{4k}\left|\lambda - \frac{ik}{\mu}\chi\right|^2_{L^2(\Gamma)} \geq 0$. From a corollary of the Rellich's Lemma (see[76, Lemma 9.9]), $w|_{\mathbb{R}^3\backslash\bar{B}} \equiv 0$. From the Cauchy-Kowalevski theorem (unique analytic continuation), there holds $\gamma_0^+ w = 0$. Finally, from (2.17), $\lambda = \frac{ik}{\mu}\chi$.

Step 2: We now prove that $\lambda = -\frac{ik}{\mu}\chi$.

Using (2.3),

$$\begin{cases} \gamma_1^- w = \gamma_1^+ w - \lambda = -\frac{1}{2}\left(\lambda + \frac{ik}{\mu}\chi\right), \\ \gamma_0^- w = \gamma_0^+ w - \chi = \frac{i\mu}{2k}\left(\lambda + \frac{ik}{\mu}\chi\right). \end{cases} \tag{2.19}$$

Applying the Green's formula on w in Ω^- yields

$$\mathbb{R} \ni \int_{\Omega^-} \left(|\nabla w|^2 - k^2 |w|^2 \right) = \left(\gamma_1^- w, \gamma_0^- w \right). \tag{2.20}$$

From (2.19),

$$\left(\gamma_1^- w, \gamma_0^- w \right) = -\frac{i\mu}{4k} \left\| \lambda + \frac{ik}{\mu} \chi \right\|_{L^2(\Gamma)} \in i\mathbb{R}. \tag{2.21}$$

Using (2.20) and (2.21), we infer $\lambda = -\frac{ik}{\mu}\chi$. Combining this result with that of Step 1, we obtain $\lambda = \chi = 0$.

This proves uniqueness and concludes the whole proof. ♢

2.4 Inf-sup stability of the discrete formulation

Let V_h^1 and V_h^0 be respectively the Lagrange finite element spaces $\mathbb{P}_1$ and $\mathbb{P}_0$ on Γ built using a shape regular mesh of Γ. The product space $\mathcal{H}_h := V_h^1 \times V_h^0$ is a conforming approximation of $\mathcal{H} = H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$. The numerical resolution is carried out with a Galerkin procedure on $V_h^1 \times V_h^0$ using the boundary element method (BEM): Find $(\chi, \lambda) \in \mathcal{H}_h := V_h^1 \times V_h^0$ such that $\forall \left(\hat{\chi}, \hat{\lambda} \right) \in \mathcal{H}_h$,

$$\begin{cases} \left\langle N\chi - \dfrac{ik}{2\mu}\chi, \hat{\chi} \right\rangle_{-\frac{1}{2},\frac{1}{2}} + \left\langle \tilde{D}\lambda, \hat{\chi} \right\rangle_{-\frac{1}{2},\frac{1}{2}} = \langle \gamma_1 p_{\text{inc}}, \hat{\chi} \rangle_{-\frac{1}{2},\frac{1}{2}}, \\ \left(D\chi, \hat{\lambda} \right) - \left(S\lambda + \dfrac{i\mu}{2k}\lambda, \hat{\lambda} \right) = -\left(\gamma_0 p_{\text{inc}}, \hat{\lambda} \right), \end{cases} \tag{2.22}$$

where, for simplicity, we keep the same notation for the discretized unknowns.

Proposition 2.3 *The following approximation properties of V_h^1 and V_h^0 (see [93, 42]) hold: For all $(\chi, \lambda) \in H^1(\Gamma) \times H^{\frac{1}{2}}(\Gamma)$,*

$$\begin{aligned} \inf_{\chi_h \in V_h^1} \|\chi - \chi_h\|_{H^{\frac{1}{2}}(\Gamma)} &\leq C_1 h^{\frac{1}{2}} \|\chi\|_{H^1(\Gamma)}, \\ \inf_{\lambda_h \in V_h^0} \|\lambda - \lambda_h\|_{L^2(\Gamma)} &\leq C_2 h^{\frac{1}{2}} \|\lambda\|_{H^{\frac{1}{2}}(\Gamma)}, \end{aligned} \tag{2.23}$$

where C_1 and C_2 are constants independent of the mesh size h.

This leads to the following proposition: For all $(\chi, \lambda) \in H^1(\Gamma) \times H^{\frac{1}{2}}(\Gamma)$

$$\inf_{(\chi_h, \lambda_h) \in \mathcal{H}_h} \|(\chi, \lambda) - (\chi_h, \lambda_h)\|_{\mathcal{H}} \leq C h^{\frac{1}{2}} \left(\|\chi\|_{H^1(\Gamma)} + \|\lambda\|_{H^{\frac{1}{2}}(\Gamma)} \right), \tag{2.24}$$

where C is a constant independent of the mesh size h.

Proposition 2.4 *Since the sesquilinear form a defined by* (2.13) *satisfies a Garding inequality (because it is coercive up to an additive compact perturbation), is injective and $\mathcal{H}_h$ satisfies the*

approximation property (2.24)*, there exists* $\hat{h} > 0$ *and a constant* γ *such that for all mesh size* h *satisfying* $0 < h \leq \hat{h}$*, the following discrete inf-sup condition holds:*

$$\sup_{(0,0)\neq(\hat{\chi}_h,\hat{\lambda}_h)\in\mathcal{H}_h} \frac{\left|a\left((\chi_h,\lambda_h),(\hat{\chi}_h,\hat{\lambda}_h)\right)\right|}{\left\|(\hat{\chi}_h,\hat{\lambda}_h)\right\|_{\mathcal{H}}} \geq \gamma\|(\chi_h,\lambda_h)\|_{\mathcal{H}} \text{ for all } (\chi_h,\lambda_h)\in\mathcal{H}_h. \tag{2.25}$$

Proof. We follow the proof of [55, Theorem 14] and detail it for completeness of the presentation. We define $\tilde{\mathcal{H}} := H^1(\Gamma)\times H^{\frac{1}{2}}(\Gamma)$. We define G_{hA_0} as the Galerkin projection corresponding to the operator A_0 by $\mathcal{H}\ni x\mapsto G_{hA_0}x\in\mathcal{H}_h$, i.e. the solution of: Find $G_{hA_0}x\in\mathcal{H}_h$ such that, $\forall y_h\in\mathcal{H}_h$,

$$\langle A_0 G_{hA_0}x, y_h\rangle_{\mathcal{H}',\mathcal{H}} = \langle A_0 x, y_h\rangle_{\mathcal{H}',\mathcal{H}}\,. \tag{2.26}$$

We denote by C the continuity constant of A_0 and by α the coercivity constant of A_0. We observe that for all $y\in\mathcal{H}$, $\|G_{hA_0}y\|_{\mathcal{H}}\leq\frac{C}{\alpha}\|y\|_{\mathcal{H}}$ since $\alpha\|G_{hA_0}y\|_{\mathcal{H}}^2\leq|\langle A_0 G_{hA_0}y, G_{hA_0}y\rangle_{\mathcal{H}',\mathcal{H}}| = |\langle A_0 y, G_{hA_0}y\rangle_{\mathcal{H}',\mathcal{H}}| \leq C\|G_{hA_0}y\|_{\mathcal{H}}\|y\|_{\mathcal{H}}$. Additionally, $\|G_{hA_0}y-y\|_{\mathcal{H}}\leq(1+\frac{C}{\alpha})\inf_{z_h\in\mathcal{H}_h}\|y-z_h\|_{\mathcal{H}}$. Indeed, for all $z_h\in\mathcal{H}_h$, $\alpha\|G_{hA_0}y-z_h\|_{\mathcal{H}}^2\leq|\langle A_0(G_{hA_0}y-z_h), G_{hA_0}y-z_h\rangle_{\mathcal{H}',\mathcal{H}}| = |\langle A_0(y-z_h), G_{hA_0}y-z_h\rangle_{\mathcal{H}',\mathcal{H}}| \leq C\|y-z_h\|_{\mathcal{H}}\|G_{hA_0}y-z_h\|_{\mathcal{H}}$, so that $\|G_{hA_0}y-z_h\|_{\mathcal{H}}\leq\frac{C}{\alpha}\|y-z_h\|_{\mathcal{H}}$, and the assertion follows from the triangle inequality.

Step 1. Let us show that $\forall x\in\mathcal{H}$,

$$\|G_{hA_0}x-x\|_{\mathcal{H}}\underset{h\to0}{\longrightarrow}0. \tag{2.27}$$

Let $x\in\mathcal{H}$ and let $\epsilon>0$. Using the density of $\tilde{\mathcal{H}}$ in $\mathcal{H}$, there exists $y\in\tilde{\mathcal{H}}$ such that $\|x-y\|_{\mathcal{H}}\leq\frac{\epsilon}{2(1+\frac{C}{\alpha})}$. Using the approximation property, there exists $h_0>0$ such that $\forall h\leq h_0$, $\inf_{z_h\in\mathcal{H}_h}\|y-z_h\|_{\mathcal{H}}\leq\frac{\epsilon}{2(1+\frac{C}{\alpha})}$. Therefore, $\forall h\leq h_0$, $\|G_{hA_0}y-y\|_{\mathcal{H}}\leq\frac{\epsilon}{2}$. Furthermore, $\|G_{hA_0}(y-x)\|_{\mathcal{H}}\leq\frac{C}{\alpha}\|x-y\|_{\mathcal{H}}\leq\frac{C}{\alpha}\frac{\epsilon}{2(1+\frac{C}{\alpha})}$. To sum up, $\forall h\leq h_0$, there holds $\|G_{hA_0}x-x\|_{\mathcal{H}}\leq\|x-y\|_{\mathcal{H}}+\|y-G_{hA_0}y\|_{\mathcal{H}}+\|G_{hA_0}(y-x)\|_{\mathcal{H}}\leq\epsilon$.

Step 2. Let $L := I + A_0^{-1}\tilde{A}$ and $L_h := I + G_{hA_0}A_0^{-1}\tilde{A}$ mapping $\mathcal{H}$ to $\mathcal{H}$. We have $L-L_h = (I-G_{hA_0})A_0^{-1}\tilde{A}$. We now prove that, for h small enough, L_h^{-1} exists and there exists c_0 independent of h such that $|||L_h^{-1}|||\leq c_0$. We showed in Step 1 that $I-G_{hA_0}$ converges pointwise to 0 as $h\to0$. Using a corollary of the Banach-Steinhaus theorem, $I-G_{hA_0}$ converges to 0 in operator norm on compact sets. Moreover, since A_0^{-1} exists and is bounded, and $\tilde{A}$ is compact, then $A_0^{-1}\tilde{A}$ is compact. Therefore, we have

$$|||L-L_h|||\underset{h\to0}{\to}0. \tag{2.28}$$

Hence, since L^{-1} exists ($L^{-1}=A^{-1}A_0$ with A invertible), there exists $h_1>0$ such that $\forall0<h\leq h_1$, L_h^{-1} exists. From (2.28), there exists $0<h_2\leq h_1$ such that $\forall0<h\leq h_2$, $|||L-L_h|||\leq\frac{1}{2|||L^{-1}|||}$. We have $L_h=[I-(L-L_h)L^{-1}]L$. Then, using the Neumann series of $L-L_h$, $L_h^{-1}=L^{-1}\sum_{k=0}^{+\infty}[(L-L_h)L^{-1}]^k$. Hence, $\forall0<h\leq h_2$, $|||L_h^{-1}|||\leq\frac{|||L^{-1}|||}{1-|||L-L_h|||||L^{-1}|||}\leq2|||L^{-1}|||=:c_0$.

Step 3. We now prove the discrete inf-sup stability of a. Since $A=A_0L$, we obtain: $\forall x_h,\hat{x}_h\in\mathcal{H}_h$,

$$\mathrm{Re}(a\,(x_h,\hat{x}_h)) = \mathrm{Re}(\langle Ax_h,\hat{x}_h\rangle_{\mathcal{H}',\mathcal{H}}) = \mathrm{Re}(\langle A_0L_hx_h,\hat{x}_h\rangle_{\mathcal{H}',\mathcal{H}}) + \mathrm{Re}(\langle A_0(L-L_h)x_h,\hat{x}_h\rangle_{\mathcal{H}',\mathcal{H}}).$$

Hence,

$$|a\,(x_h,\hat{x}_h)\,| + |\,\langle A_0(L-L_h)x_h,\hat{x}_h\rangle_{\mathcal{H}',\mathcal{H}}\,| \geq \mathrm{Re}(\langle A_0L_hx_h,\hat{x}_h\rangle_{\mathcal{H}',\mathcal{H}}).$$

From the coercivity of A_0, taking $\hat{x}_h = L_hx_h$,

$$\mathrm{Re}(\langle A_0L_hx_h,L_hx_h\rangle_{\mathcal{H}',\mathcal{H}}) \geq \alpha\|L_hx_h\|^2_{\mathcal{H}} \geq \frac{\alpha}{c_0}\|x_h\|_{\mathcal{H}}\|L_hx_h\|_{\mathcal{H}},$$

where we used the uniform boundedness of L_h^{-1} in the last inequality. Moreover, still with $\hat{x}_h = L_hx_h$,

$$|\,\langle A_0(L-L_h)x_h,L_hx_h\rangle_{\mathcal{H}',\mathcal{H}}\,| \leq |||A_0|||\,|||L-L_h|||\,\|x_h\|_{\mathcal{H}}\|L_hx_h\|_{\mathcal{H}}.$$

Hence, for $L_hx_h \neq 0$,

$$\frac{|a\,(x_h,L_hx_h)\,|}{\|L_hx_h\|_{\mathcal{H}}} \geq \left(\frac{\alpha}{c_0} - |||A_0|||\,|||L-L_h|||\right)\|x_h\|_{\mathcal{H}}.$$

Since $|||L-L_h||| \underset{h\to 0}{\to} 0$, there exists $0 < h_3 \leq h_2$ such that, $\forall 0 < h \leq h_3$,

$$\left(\frac{\alpha}{c_0} - |||A_0|||\,|||L-L_h|||\right) \geq \frac{\alpha}{2c_0} = \frac{\alpha}{4|||A^{-1}A_0|||} =: \varsigma.$$

Therefore, $\forall 0 < h \leq h_3$,

$$\sup_{\|\hat{x}_h\|_{\mathcal{H}}\neq 0} \frac{|a\,(x_h,\hat{x}_h)\,|}{\|\hat{x}_h\|_{\mathcal{H}}} \geq \frac{|a\,(x_h,L_hx_h)\,|}{\|L_hx_h\|_{\mathcal{H}}} \geq \varsigma\|x_h\|_{\mathcal{H}}, \text{ for all } x_h \in \mathcal{H}_h.$$

♢

Remark 2.5 (Symmetry vs Hermitian symmetry) *Consider two normed spaces X and Y. The spaces $X^* := \mathcal{L}(X,\mathbb{C})$ and $Y^* := \mathcal{L}(Y,\mathbb{C})$ are respectively the dual spaces of X and Y. In what follows, $(\cdot,\cdot)_{X,X^*}$ denotes the duality pairing between X and X^*. For instance, if $u \in X$ and $g \in X^*$, $(u,g)_{X,X^*} = g(u)$. Consider an operator $A : X \to Y$. Its adjoint operator $A^* : Y^* \to X^*$ is defined by $(Au,v)_{Y,Y^*} = (u,A^*v)_{X,X^*}$, for all $u \in X$ and $v \in Y^*$. The boundary integral operators $D : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ and $\tilde{D} : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ are transpose (or dual) but not adjoint, that is, they satisfy $\tilde{D} = \overline{D^*}$. In a complex-valued functions context, the convenient notion of symmetry is the Hermitian symmetry, naturally inherited from the chosen inner product: $(u,v) = \int_\Gamma \bar{u}v = \int_\Gamma \overline{\overline{v}u} = \overline{(v,u)}$. However, when considering finite dimensional approximations by means of finite elements and boundary elements, the basis functions are chosen to be real-valued. Consider the matrix M_h obtained from (2.22), and the indices i,j such that $M_{hi,j}$ is in the lower-left extradiagonal block. Then, $M_{hi,j} = (D\psi_i,\varphi_j)$, where ψ_i is a basis function of V_h^1 and φ_j is a basis function of V_h^0. There holds $M_{hi,j} = (\psi_i,D^*\varphi_j) = (\psi_i,\overline{\tilde{D}}\varphi_j) = \overline{(\overline{\psi_i},\tilde{D}\overline{\varphi_j})} = (\tilde{D}\overline{\varphi_j},\overline{\psi_i})$. Then, since ψ_i and φ_j are real-valued, there holds $M_{hi,j} = M_{hj,i}$, which means that the two extradiagonal blocks of M_h are complex and symmetric (but not Hermitian symmetric). Actually, the whole matrix M_h is symmetric.*

2.5 Combined field integral equations (CFIE)

In this section, we no longer consider the problem (2.9), but Helmholtz exterior problems with Neumann and Dirichlet boundary condition. In that case, the unicity property fails for some values, and news strategies have to be considered.

To get a well posed systems of integral equations, we saw that a boundary condition, based on physical considerations, has to be added to Equation (1.32) involving the Calderón projector. The goal is to select the physical solution among the solutions of (1.32), by ensuring that the kernel of the considered system of boundary integral operators contains only the zero function. However, when dealing with the Helmholtz equation, this may not be sufficient, in the sense that after injecting the correct boundary condition, the kernel may still contain nonzero functions. This is due to the fact that there exist nontrivial solution to the homogeneous Laplace equation on bounded domains. Notice that for the scattering of an impedant surface considered in Section 2, the considered variational formulation (2.9) was unconditionally well posed. In what follows, we present the method used to recover an unconditionally invertible boundary integral operator known as combined field integral equations.

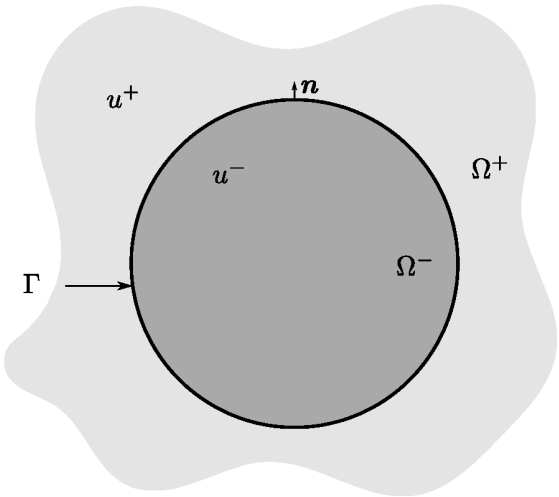

Fig. 2.3. Geometry

Consider the geometry illustrated in Figure 2.3. The boundary Γ is supposed Lipschitz. For any function u defined in $\mathbb{R}^3$, we denote its restrictions to Ω^- and Ω^+ by $u|_{\Omega^-} := u^-$ and $u|_{\Omega^+} := u^+$. We focus on Dirichlet boundary conditions; the extension to Neumann boundary conditions can be treated in the same way.

2.5.1 Eigenvalue problems in Ω^-

From [76, Theorems 4.12 and 8.2], there exist positive real numbers λ_n^D and nontrivial solutions ϕ_n^D of the following eigenvalue problem:

$$\begin{cases} -\Delta\phi_n^D = \lambda_n^D \phi_n^D & \text{in } \Omega^-, \\ \gamma_0^- \phi_n^D = 0 & \text{on } \Gamma, \end{cases} \tag{2.29}$$

where the exponent D stands for the Dirichlet boundary condition enforced on Γ and where γ_0^- stands for the interior Dirichlet trace on Γ. We denote $\{\lambda_n^D\}$ the set of eigenvalues and $\{\phi_n^D\}$ the set of eigenfunctions.

Likewise, there exist positive real numbers λ_n^N and nontrivial solutions ϕ_n^N of the following eigenvalue problem:

$$\begin{cases} -\Delta\phi_n^N = \lambda_n^N \phi_n^N & \text{in } \Omega^-, \\ \gamma_1^- \phi_n^N = 0 & \text{on } \Gamma, \end{cases} \tag{2.30}$$

where the exponent N stands for the Neumann boundary condition enforced on Γ and where γ_1^- stands for the interior Neumann trace on Γ. We denote $\{\lambda_n^N\}$ the set of eigenvalues and $\{\phi_n^N\}$ the set of eigenfunctions.

We observe that there exist geometries for which $\{\lambda_n^D\} \cap \{\lambda_n^N\} \neq \emptyset$. Indeed, fix a dimension $d \in \mathbb{N}^*$ and consider $\Omega^- = [0, L]^d$, with $L > 0$. Considering $\prod_{j=1}^d \sin(\frac{2\pi x_j}{L})$ as an eigenfunction of (2.29) and $\prod_{j=1}^d \cos(\frac{2\pi x_j}{L})$ as an eigenfunction of (2.30), we see that $\frac{4d\pi^2}{L^2} \in \{\lambda_n^D\} \cap \{\lambda_n^N\}$.

2.5.2 Kernel of boundary integral operators

Proposition 2.6 *For all $k^2 \in \{\lambda_n^D\}$, the interior Neumann traces of the eigenfunctions of the Dirichlet eigenvalue problem* (2.29) *are nonzero elements of* $\mathrm{Ker}(S)$.

Proof. Let ϕ_n^D be an eigenfunction of (2.29) and define the distribution $w_n \in \mathcal{D}'(\mathbb{R}^3)$ such that $w_n^- := \phi_n^D$ and $w_n^+ := 0$. Using the representation formula (1.30), $w_n = \mathcal{S}(\gamma_1^- \phi_n^D)$. Since $w_n = \mathcal{S}(\gamma_1^- \phi_n^D) \neq 0$, we infer $\gamma_1^- \phi_n^D \neq 0$. The interior Dirichlet trace of w_n is $\gamma_0^- \mathcal{S}(\gamma_1^- \phi_n^D) = S(\gamma_1^- \phi_n^D) = \gamma_0^- w_n^- = \gamma_0^- \phi_n^D = 0$. Therefore, $\lambda := \gamma_1^- \phi_n^D$ is a nontrivial solution of $S\lambda = 0$. ♢

Corollary 2.7 *For all $k^2 \in \{\lambda_n^D\}$,* $\mathrm{Ker}(S) \neq \{0\}$.

Proposition 2.8 *For all $k > 0$,* $\mathrm{Ker}(S) = \mathrm{Ker}(\tilde{D} - \frac{1}{2} I)$.

Proof. Let $\lambda \in H^{-\frac{1}{2}}(\Gamma)$ and define $w := \mathcal{S}\lambda$. From [80, Section 3.2.1], w is a radiating Helmholtz solution. The exterior Dirichlet and Neumann traces of w are $\gamma_0^+ w^+ = S\lambda$ and $\gamma_1^+ w^+ = (\tilde{D} - \frac{1}{2} I)\lambda$. Suppose $\lambda \in \mathrm{Ker}(S)$. Hence, $\gamma_0^+ w^+ = 0$, and by uniqueness of the solution to the exterior Helmholtz problem with Dirichlet boundary condition, $w^+ = 0$, and hence $(\tilde{D} - \frac{1}{2} I)\lambda = 0$. Therefore $\mathrm{Ker}(S) \subset \mathrm{Ker}(\tilde{D} - \frac{1}{2} I)$. Suppose now that $\lambda \in \mathrm{Ker}(\tilde{D} - \frac{1}{2} I)$. Hence, $\gamma_1^+ w^+ = 0$, and by uniqueness of the solution to the exterior Helmholtz problem with Neumann boundary condition, $w^+ = 0$, and hence $S\lambda = 0$. Therefore $\mathrm{Ker}(S) \supset \mathrm{Ker}(\tilde{D} - \frac{1}{2} I)$. ♢

Proposition 2.9 *For all $k^2 \in \{\lambda_n^D\}$, the nonzero elements of* $\mathrm{Ker}(S)$ *are interior Neumann traces of eigenfunctions of the problem* (2.29)*; for all $k^2 \notin \{\lambda_n^D\}$,* $\mathrm{Ker}(S) = \{0\}$.

Proof. Let $\lambda \in \mathrm{Ker}(S)$ and define $w := \mathcal{S}\lambda$. From [80, Section 3.2.1], w is a radiating Helmholtz solution. In particular, $\Delta w^- + k^2 w^- = 0$. The interior Dirichlet trace of w is $\gamma_0^- w^- = S\lambda = 0$. Therefore, w^- solves (2.29). Moreover, $\gamma_1^- w^- = (\tilde{D} + \frac{1}{2}I)\lambda = \lambda$ from Proposition 2.8. If $k_\infty \in \{\lambda_n^D\}$, from Corollary 2.7, we can suppose $\lambda \neq 0$. Hence, $w^- \neq 0$, and w^- is an eigenvalue of problem (2.29). If $k_\infty \notin \{\lambda_n^D\}$, problem (2.29) admits only zero as solution. Hence $w^- = 0$, and $\lambda = \gamma_1^- w^- = 0$. ♢

Corollary 2.10 *The nonzero elements of* $\mathrm{Ker}(S)$ *are exactly the interior Neumann traces of the eigenfunctions of the Dirichlet eigenvalue problem* (2.29).

Remark 2.11 *We can show in the same fashion that* $\mathrm{Ker}(N) = \mathrm{Ker}(D + \frac{1}{2}I) \neq \{0\}$, *and that the nonzero elements of* $\mathrm{Ker}(N)$ *are exactly the interior Dirichlet traces of the eigenfunctions of the Neumann eigenvalue problem* (2.30).

Remark 2.12 *For all* $k^2 \in \{\lambda_n^D\}$, $D - \frac{1}{2}I$ *is noninvertible, as the transpose (or dual) of the noninvertible operator* $\tilde{D} - \frac{1}{2}I$. *Indeed,* $(\mathrm{Im}(D - \frac{1}{2}I))^\perp = (\mathrm{Im}(\tilde{D} - \frac{1}{2}I)^T)^\perp = \mathrm{Ker}(\tilde{D} - \frac{1}{2}I) \neq \{0\}$, *hence* $D - \frac{1}{2}I$ *is not surjective. Likewise, for all* $k^2 \in \{\lambda_n^N\}$, $\tilde{D} + \frac{1}{2}I$ *is noninvertible, as the transpose of the noninvertible operator* $D + \frac{1}{2}I$.

2.5.3 CFIE for the exterior Helmholtz problem

Consider now the following exterior Helmholtz problem with a Dirichlet boundary condition:

$$\begin{cases} \Delta u + k^2 u = 0 & \text{in } \Omega^+ \\ \gamma_0^+ u = -\gamma_0^- f_{\mathrm{inc}} & \text{on } \Gamma \\ \lim\limits_{r \to +\infty} r \left(\dfrac{\partial u}{\partial r} - iku \right) = 0 \end{cases} \tag{2.31}$$

where we suppose that f_{inc} is such that $\gamma_0^- f_{\mathrm{inc}} \in H^{\frac{1}{2}}(\Gamma)$. This exterior Helmholtz problem admits a unique solution in $H(\Delta, \Omega^+)$ [93, Theorem 2.10.15], where $H(\Delta, \Omega^+) = \{v \in H^1(\Omega^+), \Delta v \in L^2\}$. We define the distribution v such that $v|_{\Omega^+} := u$ and $v|_{\Omega^-} := -f_{\mathrm{inc}}$. Using the representation formula (1.30) and the boundary integral operators, there holds

$$\begin{cases} S\lambda = -\gamma_0^- f_{\mathrm{inc}}, \\ \left(\dfrac{1}{2}I + \tilde{D} \right) \lambda = -\gamma_1^- f_{\mathrm{inc}}, \end{cases} \tag{2.32}$$

where $\lambda := [\gamma_1 v]_\Gamma$ is the surface unknown.

Remark 2.13 *Even if the exterior Helmholtz problem* (2.31) *has a unique solution at all frequencies, the integral equations* (2.32), *written as necessary relations verified by its solution, admit an infinity of solutions. This phenomenon is related to the eigenvalues and eigenfunctions of the complementary interior problem, and has no physical meaning. In* (2.32), *the first equation is not invertible for* $k^2 \in \{\lambda_n^D\}$, *whereas the second equation is not invertible for* $k^2 \in \{\lambda_n^N\}$. *Furthermore, there exist geometries and wave numbers for which both equations in* (2.32) *have infinitely many solutions. More precisely, consider a geometry for which*

$\{\lambda_n^D\} \cap \{\lambda_n^N\}$ is nonempty, let $k^2 \in \{\lambda_n^D\} \cap \{\lambda_n^N\}$ and consider the unique solution u of (2.31) at this wave number. Then, $\forall v^* \in \mathrm{Ker}(S)$, $S(\gamma_1^+ u + \gamma_1^- f_{\mathrm{inc}} + v^*) = -\gamma_0 f_{\mathrm{inc}}$ and $\forall w^* \in \mathrm{Ker}(\frac{1}{2}I + \tilde{D})$, $\left(\frac{1}{2}I + \tilde{D}\right)(\gamma_1^+ + \gamma_1^- f_{\mathrm{inc}} + w^*) = -\gamma_1 f_{\mathrm{inc}}$.

Consider a geometry for which $\{\lambda_n^D\} \cap \{\lambda_n^N\}$ is nonempty, and take $k^2 \in \{\lambda_n^D\} \cap \{\lambda_n^N\}$. Then, the normalized Dirichlet eigenfunction at k^2 is linealy independent of the normalized Neumann eigenfunction at k^2.

Proposition 2.14 *For all $k > 0$, there holds* $\mathrm{Ker}(S) \cap \mathrm{Ker}\left(\frac{1}{2}I + \tilde{D}\right) = \{0\}$.

Proof. Let $\lambda \in \mathrm{Ker}(S) \cap \mathrm{Ker}\left(\frac{1}{2}I + \tilde{D}\right) \setminus \{0\}$ and set $u = \mathcal{S}\lambda$. Then, u is a radiating Helmholtz solution, such that $\gamma_0^+ u^+ = S\lambda = 0$ and $\gamma_1^+ u^+ = \left(-\frac{1}{2}I + \tilde{D}\right)\lambda = -\lambda$. Therefore, u^+ solves an exterior Helmholtz equation with a homogeneous Dirichlet boundary condition at Γ. By uniqueness of the solution to this problem, $u^+ \equiv 0$, and hence $\lambda = -\gamma_1^+ u^+ = 0$. ♢

Remark 2.15 *After introducing the exterior Helmholtz problem with a Neumann boundary condition, we can show in the same fashion that for all $k > 0$,* $\mathrm{Ker(N)} \cap \mathrm{Ker(D} - \frac{1}{2}\mathrm{I}) = \{0\}$.

Only the physical solution $\gamma_1^+ u + \gamma_1^- f_{\mathrm{inc}}$ solves simultaneously both equations in (2.32). To avoid dealing with the overdetermined system (2.32), the idea of considering a linear combination of the equations in (2.32) was independently proposed in 1965 by Brakhage and Werner [20], Leis [65] and Panich [83]:

Proposition 2.16 *For all $k > 0$, for all $\eta \in \mathbb{C}$ such that* $\mathrm{Re}(\eta) \neq 0$, $\mathrm{Ker}\left(S + i\eta\left(\frac{1}{2}I + \tilde{D}\right)\right) = \{0\}$.

Proof. Let $\lambda \in \mathrm{Ker}\left(S + i\eta\left(\frac{1}{2}I + \tilde{D}\right)\right)$ and set $u = \mathcal{S}\lambda$. The interior traces of u are $\gamma_0^- u^- = S\lambda$ and $\gamma_1^- u^- = (\frac{1}{2}I + \tilde{D})\lambda$, hence

$$\gamma_0^- u^- + i\eta\gamma_1^- u^- = 0. \tag{2.33}$$

Since u is a radiating Helmholtz solution, there holds

$$\int_{\Omega^-} \left\{|\boldsymbol{\nabla} u|^2 - k^2|u|^2\right\} - \left(\gamma_1^- u^-, \gamma_0^- u^-\right)_\Gamma = 0. \tag{2.34}$$

Take the imaginary part of this relation. From $\mathrm{Re}(\eta) \neq 0$ and (2.33), there holds $\|\gamma_1^- u^-\|_{L^2(\Gamma)} = 0$, which leads to $\gamma_1^- u^- = 0$, and $\gamma_0^- u^- = 0$ using again (2.33). Then, $\lambda \in \mathrm{Ker}(S) \cap \mathrm{Ker}\left(\frac{1}{2}I + \tilde{D}\right) = \{0\}$. ♢

Likewise, there holds

Proposition 2.17 *For all $k > 0$, for all $\eta \in \mathbb{C}$ such that* $\mathrm{Re}(\eta) \neq 0$, $\mathrm{Ker}\left(\left(\frac{1}{2}I - D\right) + i\eta N\right) = \{0\}$.

Consider the equation $\left(S + i\eta\left(\frac{1}{2}I + \tilde{D}\right)\right)\lambda = g$. This equation is set in $H^{\frac{1}{2}}(\Gamma)$ and the unknown λ is sought in $H^{-\frac{1}{2}}(\Gamma)$. However, because of the term $\left(\frac{1}{2}I + \tilde{D}\right)$, the equation cannot be set in $H^{\frac{1}{2}}(\Gamma)$ anymore. This means that, to write a variational formulation, we have to deal either with a pairing in $H^{-\frac{1}{2}}(\Gamma)$, or with a Petrov-Galerkin numerical approximation. A third possibility, which follows here, is the method introduced in [23].

Consider the following Hermitian form:

$$\delta_\Gamma(p,q) := (\boldsymbol{\nabla}_\Gamma p, \boldsymbol{\nabla}_\Gamma q)_\Gamma + (p,q)_\Gamma\,, \tag{2.35}$$

for all $p, q \in H^1(\Gamma)$. We define the regularizing operator $M : H^{-1}(\Gamma) \mapsto H^1(\Gamma)$ through the following implicit relation:

$$\delta(M(p), q)_\Gamma = (p,q)_\Gamma, \tag{2.36}$$

for all $q \in H^1(\Gamma)$ (see [23, Equation (3.10)]), where $\boldsymbol{\nabla}_\Gamma$ denotes the surfacic gradient on Γ. From [54], $(\lambda, M\lambda)_\Gamma > 0$ for all $\lambda \in H^{-\frac{1}{2}}(\Gamma)\backslash\{0\}$ and M is compact from $H^{-\frac{1}{2}}(\Gamma)$ into $H^{\frac{1}{2}}(\Gamma)$. Applying the regularizing operator to the second equation in (2.32), and then writing the CFIE linear combination leads to the following operator:

$$H^{-\frac{1}{2}}(\Gamma) \ni \lambda \to \left(S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)\right)\lambda = -\gamma_0^- f_{\text{inc}} - i\eta M\gamma_1^- f_{\text{inc}} \in H^{\frac{1}{2}}(\Gamma). \tag{2.37}$$

Theorem 2.18 *For all $k > 0$, for all $\eta \in \mathbb{C}$ such that $\mathrm{Re}(\eta) \neq 0$, Equation (2.37) has a unique solution λ.*

Proof. Let $k > 0$ and $\eta \in \mathbb{C}$ such that $\mathrm{Re}(\eta) \neq 0$. First, we write $S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right) = S_0 + (S - S_0) + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)$, where S_0 is coercive, and $(S - S_0)$ (see [93, Lemma 3.9.8]) and $M\left(\frac{1}{2}I + \tilde{D}\right)$ are compact in the natural trace spaces. Second, we prove that $\mathrm{Ker}\left(S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)\right) = \{0\}$. The proof is close to that of Proposition 2.16. Let $\lambda \in \mathrm{Ker}\left(S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)\right)$ and set $u = \mathcal{S}\lambda$. Applying the interior trace operators to u yields $\gamma_0^- u^- = S\lambda$ and $\gamma_1^- u^- = \left(\tilde{D} + \frac{1}{2}I\right)\lambda$. Therefore,

$$\gamma_0^- u^- + i\eta M\gamma_1^- u^- = 0. \tag{2.38}$$

Since u is a radiating Helmholtz solution, there holds

$$\int_{\Omega^-}\left\{|\boldsymbol{\nabla} u|^2 - k^2|u|^2\right\} - \left(\gamma_1^- u^-, \gamma_0^- u^-\right)_\Gamma = 0. \tag{2.39}$$

Take the imaginary part of this relation. From $\mathrm{Re}(\eta) \neq 0$ and (2.38), there holds $\left(\gamma_1^- u^-, M\gamma_1^- u^-\right) = 0$, which leads to $\gamma_1^- u^- = 0$, and $\gamma_0^- u^- = 0$ using again (2.38). Then, $\lambda \in \mathrm{Ker}(S) \cap \mathrm{Ker}\left(\frac{1}{2}I + \tilde{D}\right) = \{0\}$. From the Fredholm alternative, Equation (2.37) has a unique solution λ. ♢

The operator M enables to recover the consistency in the functional spaces setting, with Γ Lipschitz, and well-posed Galerkin methods can then be derived.

Remark 2.19 *Other regularizing operators M are possible. It is possible to regularize exterior Helmholtz problems with Neumann boundary conditions in the same fashion (see [23, 54]).*

Remark 2.20 (Smooth boundary) *When Γ is smooth, which is never the case in practice when considering a mesh of Γ, D is compact from $L^2(\Gamma)$ into $L^2(\Gamma)$ (see [23]). The operator $\tilde{D}$ is compact from $L^2(\Gamma)$ into $L^2(\Gamma)$ as the transpose of a compact operator. Besides, since for all $\lambda \in H^{-\frac{1}{2}}(\Gamma)$, $\|S\lambda\|_{H^{\frac{1}{2}}(\Gamma)} \leq C\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma)} \leq C\|\lambda\|_{L^2(\Gamma)}$ and $H^{\frac{1}{2}}(\Gamma) \hookrightarrow L^2(\Gamma)$ is compact, the operator S is also compact from $L^2(\Gamma)$ into $L^2(\Gamma)$. From the Fredholm alternative, $S + i\eta\left(\frac{1}{2}I + \tilde{D}\right)$ is bijective from $L^2(\Gamma)$ into $L^2(\Gamma)$ without resorting to any regularizing operator, since $\frac{1}{i\eta}S + \tilde{D}$ is a compact perturbation of $\frac{1}{2}I$.*

Remark 2.21 (The particular case of the scattering by an impedant surface) *The CFIE consists in a complex linear combination of an equation posed in $H^{\frac{1}{2}}(\Gamma)$ and an equation posed in $H^{-\frac{1}{2}}(\Gamma)$. For the scattering by an impedant surface, such a combination is implicitely contained in the Robin boundary condition (2.6), which directly links χ, the surface unknown in $H^{\frac{1}{2}}(\Gamma)$, and λ, the unknown in $H^{-\frac{1}{2}}(\Gamma)$. In particular, no regularization is required, because the Robin boundary condition (2.6) induces some regularity on λ. This enabled us to consider a functional space for λ which is more regular that $H^{-\frac{1}{2}}(\Gamma)$, namely $L^2(\Gamma)$. Actually, the formulation (2.9) hides the principles of regularized CFIE.*

2.6 Numerical illustration

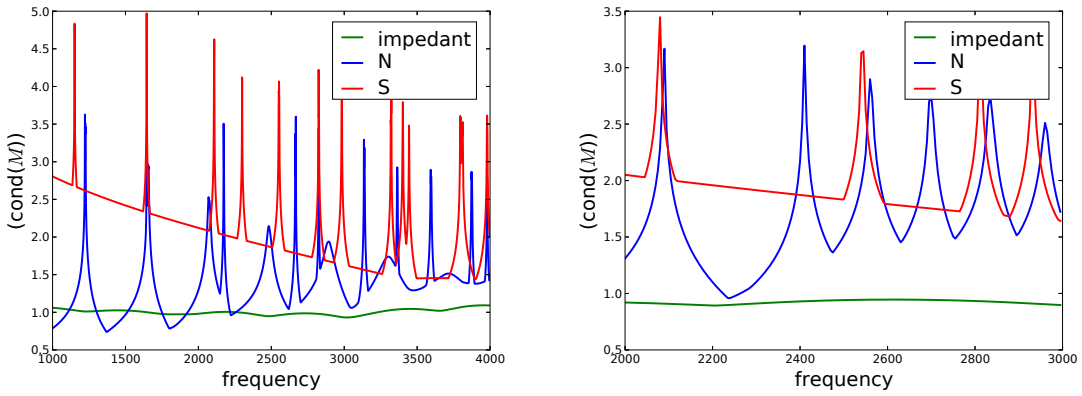

Fig. 2.4. Condition number (log scale) of the matrices of three boundary integral problems for a sphere (left) and a cube (right).

In Figure 2.4 we represent the condition numbers of matrices obtained after discretization of three integral equations, defined on very simple geometries. For the sphere, a mesh with 122 vertices has been used, and for the cube, a mesh with 221 vertices has been used. “Impedant” refers to the problem (2.9) of scattering by an impedant object. The two other problems deal

with the inversion of the boundary integral operators N and S, for which noninversibility at respectively the Neumann and Dirichlet eigenvalues of the Laplacian defined in the interior domain is seen in Section 2.5.2. The formulation (2.9) does not suffer from any noninversibility problem at any frequency.

3

A coupled FEM/BEM for the convected Helmholtz equation with non-uniform flow in a bounded domain

This chapter is an extended and detailed version of the article [Ar4].

Summary. We consider the convected Helmholtz equation modeling linear acoustic propagation at a fixed frequency in a subsonic flow around a scattering object. The flow is supposed to be uniform in the exterior domain far from the object, and potential in the interior domain close to the object. Our key idea is the reformulation of the original problem using the Prandtl–Glauert transformation on the whole flow domain, yielding (i) the classical Helmholtz equation in the exterior domain and (ii) an anisotropic diffusive PDE with skew-symmetric first-order perturbation in the interior domain such that its transmission condition at the coupling boundary naturally fits the Neumann condition from the classical Helmholtz equation. Then, efficient off-the-shelf tools can be used to perform the BEM-FEM coupling, leading to two novel variational formulations for the convected Helmholtz equation. The first formulation involves one surface unknown and can be affected by resonant frequencies, while the second formulation avoids resonant frequencies and involves two surface unknowns. Numerical simulations are presented to compare the two formulations.

3.1 Introduction

The scope of the present work is the computation of linear acoustic wave propagation at a fixed frequency in the presence of a flow. When the flow is at rest, the simplest model is the classical Helmholtz equation for the acoustic potential. This equation can be reduced to finding unknown functions defined on the surface of the scattering object and solving integral equations which can be effectively approximated by the Boundary Element Method (BEM) [93]. When the medium of propagation is non-uniform, a volumic resolution has to be considered using, e.g., a Finite Element Method (FEM). If such non-uniformities occur only in a given bounded domain, it is possible to benefit from the advantages of both a volumic resolution and an integral equation formulation. Coupling BEM and FEM at the boundary of the given bounded domain allows this. Coupled BEM-FEM can be traced back to McDonald and Wexler [75], Zienkiewicz, Kelly and Bettess [14], Johnson and Nédélec [58] and Jin and Liepa [57]. Over the last decade, such methods have been investigated, among others, for electromagnetic scattering [53, 66, 68], elasticity [28], and fluid-structure [38] or solid-solid interactions [74, 105]. Coupled BEM-FEM for the classical Helmholtz equation can present resonant frequencies, leading to infinitely many solutions. All these solutions deliver the same acoustic potential in the exterior domain, but the numerical procedure becomes ill-conditioned. This problem has been solved in [23, 54], where

a stabilization of the coupling, based on combined field integral equations (CFIE), has been proposed by introducing an additional unknown at the coupling surface.

When the medium of propagation is not at rest, the simplest governing equation is the convected Helmholtz equation resulting from the linearized harmonic Euler equations. Nonlinear interaction between acoustics and fluid mechanics is not considered herein; we refer to the early work of Lighthill for aerodynamically generated acoustic sources [69, 70], to [51] for a review on nonlinear acoustics, and to [102] for the coupling of Computational Aero Acoustic (CAA) and Computational Fluid Dynamics (CFD) solvers. Moreover, we assume that the flow is potential close to the scattering object and uniform far away from it. This geometric setup leads to a partition of the unbounded medium of propagation into two subdomains, the bounded interior domain near the scattering object where the flow is non-uniform and the unbounded exterior domain far away from the object where the flow is uniform. The main contribution of this work is the reformulation of the convected Helmholtz equation using the Prandtl–Glauert transformation on the whole flow domain, yielding (i) the classical Helmholtz equation in the exterior domain and (ii) an anisotropic diffusive PDE with skew-symmetric first-order perturbation in the interior domain such that its transmission condition at the coupling boundary naturally fits the Neumann condition from the classical Helmholtz equation. The Prandtl–Glauert transformation has been used in [39] for the uniformly convected Helmholtz equation. In the present case where the flow is non-uniform in the interior domain, this reformulation allows us to use efficient off-the-shelf tools to perform a BEM-FEM coupling. Namely, a FEM is utilized in the interior domain to discretize the anisotropic second-order PDE, a BEM is utilized for the classical Helmholtz equation in the exterior domain, and Dirichlet-to-Neumann maps are used for the coupling. We emphasize that the key advantage of using the Prandtl–Glauert transformation is that the BEM part of the resolution only involves integral operators corresponding to the classical Helmholtz equation. We consider two approaches for the coupling, leading, to the authors' knowledge, to two novel coupled BEM-FEM formulations for the convected Helmholtz equation. The first formulation involves one surface unknown and can be affected by resonant frequencies, while the second one uses the stabilized CFIE technique from [23, 54], avoids resonant frequencies, and involves two surface unknowns. Our numerical results show that the first formulation yields results polluted by spurious oscillations in the close vicinity of resonant frequencies, whereas the second formulation yields consistent solutions at all frequencies. This advantage of the second formulation is particularly relevant in practice at high frequencies, where the density of resonant frequencies is higher.

We briefly discuss alternative methods from the literature to solve the convected Helmholtz equation in unbounded domains. In some cases with simple geometries, the far-field solution is analytically known [89]. Boundary integral equations involving the Green kernel associated with the convected Helmholtz equation have been derived in [10]. Other numerical methods include infinite finite elements [13, 108] and the method of fundamental solutions [43]. An alternative approach to treat unbounded domains is to use Perfectly Matched Layers (PML), combined with a volumic resolution using, e.g., the FEM. Versions of PML for the convected Helmholtz equation are considered in [8, 78]. The use of PML allows one to consider unbounded domains of propagation, but the solution is only available within the domain of computation. This can be a drawback in the following situations: (i) when one is interested in the pressure field far away from the scattering object, or (ii) when scattering objects are located far away from each other so that the volumic resolution has to be carried out in a very large area. Instead, with

coupled BEM-FEM, the volumic resolution only takes place in the areas where the flow is non-uniform, and the pressure can be retrieved at any point of the exterior domain using known representation formulae, regardless of the distance of this point to the scattering objects. However, coupled BEM-FEM exhibit matrices with dense blocks for the unknowns on the boundary, and an additional treatment is sometimes needed to avoid resonant frequencies. These two points are addressed in this work.

The material is organized as follows: the problem of interest is presented in Section 3.2, and coupling procedures are detailed in Section 3.3, where the main mathematical results are stated. The finite-dimensional approximation of the coupled formulations is addressed in Section 3.4, along with a discussion on the structure of the linear systems and the algorithms to solve them effectively. Finally, numerical results are presented in Section 3.5, and some conclusions are drawn in Section 3.6. The proofs of the mathematical results stated in the previous sections are presented in Section 3.7.

3.2 Aeroacoustic problem

This section describes the problem of acoustic scattering by a solid object in a non-uniform convective flow, together with the underlying physical assumptions.

3.2.1 Notation and preliminaries

Figure 3.1 describes the geometric setup. The interior domain, corresponding to the area near the scattering object where the convective flow is non-uniform, is denoted by Ω^-. In the exterior domain, Ω^+, the convective flow is assumed to be uniform. The complete medium of propagation, denoted by $\Omega \subset \mathbb{R}^3$, is such that $\Omega := \Omega^+ \cup \Omega^- \cup \Gamma_\infty = \mathbb{R}^3 \backslash \overline{\{\text{solid object}\}}$, where $\Gamma_\infty := \partial\Omega^+ \cap \partial\Omega^-$ is the boundary between the interior and exterior domains. The surface Γ_∞ is assumed to be Lipschitz. Such an assumption is sufficiently large to include for instance polyhedric surfaces resulting from the use of a finite element mesh in Ω^-. The surface of the solid scattering object, $\partial\Omega^- \backslash \Gamma_\infty$, is denoted by Γ and is assumed to be Lipschitz.

The speed of sound when the medium of propagation is at rest is denoted by c, the wave number by k, the density by ρ, and the acoustic velocity and pressure, respectively, by $\boldsymbol{v}$ and p. The rescaled velocity is defined as $\boldsymbol{M} := c^{-1}\boldsymbol{v}$, where $M := |\boldsymbol{M}|$ is the Mach number. The subscript ∞ refers to uniform flow quantities related to Ω^+, whereas the subscript 0 refers to point-dependent flow quantities related to Ω^-, that is, $\rho_{|\Omega^-} = \rho_0(\boldsymbol{x})$, $\rho_{|\Omega^+} \equiv \rho_\infty$, $k_{|\Omega^-} = k_0(\boldsymbol{x})$, $k_{|\Omega^+} \equiv k_\infty$, $c_{|\Omega^-} = c_0(\boldsymbol{x})$, $c_{|\Omega^+} \equiv c_\infty$, $\boldsymbol{M}_{|\Omega^-} = \boldsymbol{M}_0(\boldsymbol{x})$, $\boldsymbol{M}_{|\Omega^+} \equiv \boldsymbol{M}_\infty$. The convective flow is continuous through Γ_∞ and tangential on Γ. Hence ρ, k and $\boldsymbol{M}$ are continuous through Γ_∞, and $\boldsymbol{M} \cdot \boldsymbol{n} = 0$ on Γ.

The source term g is time-harmonic with pulsation ω and is assumed to be located at Ω^+ for simplicity. This source term is an acoustic monopole located at $x_s \in \Omega^+$ of amplitude A_s, so that $g := A_s \delta_{x_s} \cos(\omega t)$, where δ denotes the Dirac mass distribution. The physical quantities are associated with complex quantities with the following convention on, for instance, the acoustic pressure: $p \leftrightarrow \mathrm{Re}\left(p \exp\left(-i\omega t\right)\right)$. In what follows, we always refer to the complex

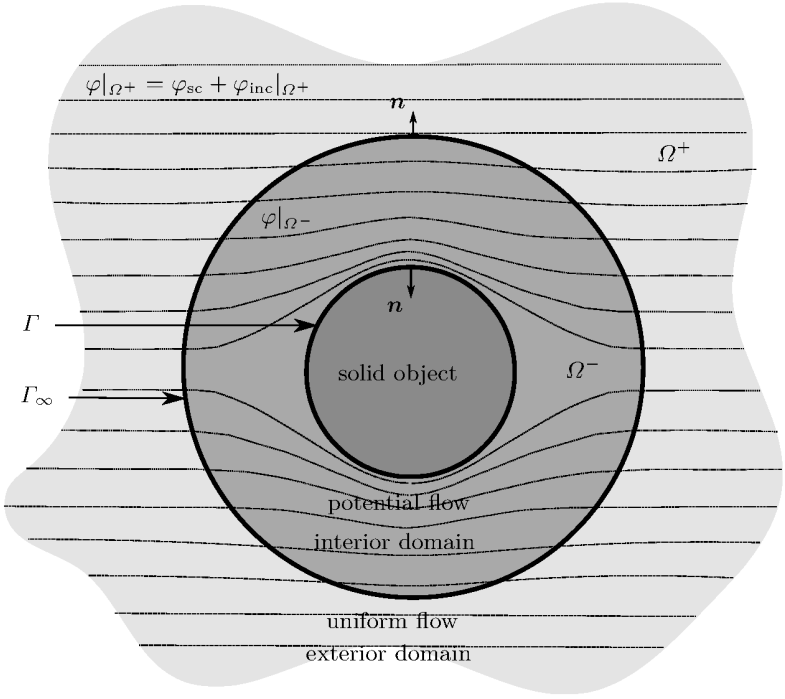

Fig. 3.1. Geometric setup for the coupled problem

quantity. Furthermore, the Hermitian product of two vectors $\boldsymbol{U}, \boldsymbol{W} \in \mathbb{C}^3$ is denoted by $\overline{\boldsymbol{U}} \cdot \boldsymbol{W} := \sum_{i=1}^{3} \overline{U_i} W_i$, where $\bar{\cdot}$ denotes the complex conjugate, and the associated Euclidian norm in $\mathbb{C}^3$ is denoted by $\|\cdot\|$.

3.2.2 The convected Helmholtz equation

In the interior domain Ω^-, the convective flow is supposed to be stationary, inviscid, isentropic, potential and subsonic. The acoustic effects are considered to be a first-order perturbation of this flow. With these assumptions, there exists an acoustic potential φ such that $\boldsymbol{v} = \boldsymbol{\nabla}\varphi$.

Following [91, Equation (F27)] and [49], and making use of the acoustic potential, the linearization of the Euler equations leads to

$$\rho\left(k^2\varphi + ik\boldsymbol{M}\cdot\boldsymbol{\nabla}\varphi\right) + \boldsymbol{\nabla}\cdot\left(\rho\left(\boldsymbol{\nabla}\varphi - (\boldsymbol{M}\cdot\boldsymbol{\nabla}\varphi)\,\boldsymbol{M} + ik\varphi\boldsymbol{M}\right)\right) = g \quad \text{in } \Omega, \tag{3.1}$$

where φ is the unknown acoustic potential, and ρ, k, $\boldsymbol{M}$, and $g = A_s\delta_{x_s}$ are known. Equation (3.1) is the convected Helmholtz equation. Under the assumption that the acoustic perturbations are perfectly reflected by the solid object, the acoustic potential verifies an homogeneous Neumann boundary condition on Γ:

$$\boldsymbol{\nabla}\varphi\cdot\boldsymbol{n} = 0 \quad \text{on } \Gamma. \tag{3.2}$$

Problem (3.1)-(3.2) is completed by a Sommerfeld-like boundary condition at infinity (see (3.18c) below) that selects outgoing waves to ensure uniqueness of the solution. In the exterior domain Ω^+ where the flow quantities are uniform, equation (3.1) simplifies into

$$\Delta\varphi + k_\infty^2\varphi + 2ik_\infty \boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\varphi - \boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\left(\boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\varphi\right) = g \quad \text{in } \Omega^+. \tag{3.3}$$

If there were no scattering object and if the convective flow were uniform in $\mathbb{R}^3$ (and thus equal to the flow at infinity), the source term g would create an acoustic potential denoted by φ_{inc} in $\mathbb{R}^3$. This potential, which solves (3.3) in $\mathbb{R}^3$, has an analytical expression, and φ_{inc} and $\boldsymbol{n} \cdot \boldsymbol{\nabla}\varphi_{\text{inc}}$ are continuous across Γ_∞. The acoustic potential scattered by the solid object is defined as $\varphi_{\text{sc}} := \varphi - \varphi_{\text{inc}}$ in Ω^+. Eliminating the known acoustic potential φ_{inc} created by the source yields

$$\Delta\varphi_{\text{sc}} + k_\infty^2\varphi_{\text{sc}} + 2ik_\infty \boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\varphi_{\text{sc}} - \boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\left(\boldsymbol{M}_\infty \cdot \boldsymbol{\nabla}\varphi_{\text{sc}}\right) = 0 \quad \text{in } \Omega^+. \tag{3.4}$$

3.2.3 The Prandtl–Glauert transformation

The Prandtl–Glauert transformation was introduced by Glauert in 1928 [47] to study the compressible effects of the air on the lift of an airfoil and was applied to subsonic aeroacoustic problems by Amiet and Sears in 1970 [3]. Herein, the Prandtl–Glauert transformation is applied in the complete medium of propagation and is based on the reduced velocity $\boldsymbol{M}_\infty$. This transformation consists in changing the space and time variables as

$$\begin{cases} \boldsymbol{x}' = \gamma_\infty \left(\hat{\boldsymbol{M}}_\infty \cdot \boldsymbol{x}\right) \hat{\boldsymbol{M}}_\infty + \left(\boldsymbol{x} - (\hat{\boldsymbol{M}}_\infty \cdot \boldsymbol{x})\hat{\boldsymbol{M}}_\infty\right) & \boldsymbol{x} \in \Omega, \\ t' = t - \dfrac{\gamma_\infty^2}{c_\infty} \boldsymbol{M}_\infty \cdot \boldsymbol{x} & t \in \mathbb{R}, \end{cases} \tag{3.5}$$

where $\gamma_\infty := \frac{1}{\sqrt{1-M_\infty^2}}$ and $\hat{\boldsymbol{M}}_\infty := M_\infty^{-1}\boldsymbol{M}_\infty$ with $M_\infty := |\boldsymbol{M}_\infty|$. The spatial transformation corresponds to a dilatation along $\hat{\boldsymbol{M}}_\infty$ of magnitude γ_∞, the component orthogonal to $\hat{\boldsymbol{M}}_\infty$ being unchanged. In what follows, we suppose that $M_\infty < 1$, so that the Prandtl–Glauert transformation is a $\mathcal{C}^\infty$-diffeomorphism from $\Omega \times \mathbb{R}$ to $\Omega' \times \mathbb{R}$, where Ω' denotes the transformed medium of propagation.

3.2.4 The transformed problem

Let f be such that $\varphi(\boldsymbol{x}) = f(\boldsymbol{x}') \exp\left(-ik_\infty\gamma_\infty\left(\boldsymbol{M}_\infty \cdot \boldsymbol{x}'\right)\right)$, $\boldsymbol{x}' \in \Omega'$; f_{inc} and f_{sc} are defined from φ_{inc} and φ_{sc} in the same fashion, so that f_{inc} is analytically known, and defined in $\mathbb{R}^3$. Let $\varsigma(\boldsymbol{x}') := \rho_\infty^{-1} \exp\left(ik_\infty\gamma_\infty\left(\boldsymbol{M}_\infty \cdot \boldsymbol{x}'\right)\right) g(\boldsymbol{x}')$, $\boldsymbol{x}' \in \Omega'$.

In what follows, the transformed geometry, unknowns and operators are considered unless specified otherwise. For brevity, primes are omitted.

To apply the the Prandtl–Glauert transformation to a PDE in the frequency domain, one has first to change the differential operators as

$$\boldsymbol{\nabla} u = \mathcal{N}\boldsymbol{\nabla}' u, \qquad \boldsymbol{\nabla} \cdot \boldsymbol{U} = \boldsymbol{\nabla}' \cdot \mathcal{N}\boldsymbol{U}, \tag{3.6}$$

for a scalar-valued function u and a vector-valued function $\boldsymbol{U}$. Here, $\mathcal{N} = I + C_\infty \boldsymbol{M}_\infty \boldsymbol{M}_\infty^T$ with $C_\infty = \frac{\gamma_\infty - 1}{M_\infty^2}$ and $\gamma_\infty = \frac{1}{\sqrt{1-M_\infty^2}}$, and $\boldsymbol{\nabla'}$ refers to derivatives with respect to the transformed variables $\boldsymbol{x}'$. Moreover, it is readily verified that

$$\mathcal{N}\boldsymbol{M} = \boldsymbol{M} + C_\infty P \boldsymbol{M}_\infty, \qquad \mathcal{N}\boldsymbol{M}_\infty = \gamma_\infty \boldsymbol{M}_\infty, \qquad \mathcal{N}\boldsymbol{M}\cdot\boldsymbol{M}_\infty = \mathcal{N}\boldsymbol{M}_\infty\cdot\boldsymbol{M} = \gamma_\infty P, \quad (3.7)$$

where $P = \boldsymbol{M}\cdot\boldsymbol{M}_\infty$. After changing the differential operators, one has to change the unknown function as $\varphi(\boldsymbol{x}) = \alpha(\boldsymbol{x}')f(\boldsymbol{x}')$, where $\alpha(\boldsymbol{x}') := \exp\left(-ik_\infty\gamma_\infty\left(\boldsymbol{M}_\infty\cdot\boldsymbol{x}'\right)\right)$.

We now show that, following the Prandtl–Glauert transformation, Equation (3.1) becomes

$$rk^2\beta f + irk\boldsymbol{V}\cdot\boldsymbol{\nabla}'f + \boldsymbol{\nabla}'\cdot\left(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla}'f\right) = \varsigma, \quad (3.8)$$

where $r = \frac{\rho}{\rho_\infty}$, $\beta = (1+qP)^2 - q^2M_\infty^2$, $\boldsymbol{V} = (1+qP)\mathcal{N}\boldsymbol{M} - q\gamma_\infty\boldsymbol{M}_\infty$, $q = \gamma_\infty^2\frac{k_\infty}{k}$, $\Xi = \mathcal{N}(I - \boldsymbol{M}\boldsymbol{M}^T)\mathcal{N}$, and $\varsigma = \rho_\infty^{-1}\alpha^{-1}g$. Dividing Equation (3.1) by ρ_∞ leads to $\alpha\varsigma = rk^2\varphi + irk\boldsymbol{M}\cdot\boldsymbol{\nabla}'\varphi + \boldsymbol{\nabla}'\cdot\left(r\left(\boldsymbol{\nabla}'\varphi - \left(\boldsymbol{M}\cdot\boldsymbol{\nabla}'\varphi\right)\boldsymbol{M} + ik\varphi\boldsymbol{M}\right)\right)$. Applying (3.6), it is inferred

$$\alpha\varsigma = rk^2\varphi + irk\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'\varphi + \boldsymbol{\nabla}'\cdot\left(r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'\varphi\right) - \boldsymbol{\nabla}'\cdot\left(r\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'\varphi\right)\mathcal{N}\boldsymbol{M}\right) + \boldsymbol{\nabla}'\cdot\left(irk\varphi\mathcal{N}\boldsymbol{M}\right).$$

Substituting φ for αf and expanding the derivatives with respect to α yields

$$\begin{aligned}\alpha\varsigma = {}& \alpha rk^2 f + \alpha irk\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f + \alpha rkk_\infty\gamma_\infty f(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{M}_\infty) + \boldsymbol{\nabla}'\cdot\left(\alpha r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\right)\\ & - \boldsymbol{\nabla}'\cdot\left(\alpha irk_\infty\gamma_\infty f\mathcal{N}\mathcal{N}\boldsymbol{M}_\infty\right) - \boldsymbol{\nabla}'\cdot\left(\alpha r\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f\right)\mathcal{N}\boldsymbol{M}\right)\\ & + \boldsymbol{\nabla}'\cdot\left(\alpha irk_\infty\gamma_\infty f\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{M}_\infty\right)\mathcal{N}\boldsymbol{M}\right) + \boldsymbol{\nabla}'\cdot\left(\alpha irkf\mathcal{N}\boldsymbol{M}\right),\end{aligned}$$

since $\boldsymbol{\nabla}'\alpha = -\alpha ik_\infty\gamma_\infty\boldsymbol{M}_\infty$. Using (3.7) and simplifying some terms leads to

$$\begin{aligned}\alpha\varsigma = {}& \alpha rk^2 f + \alpha irk\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f + \alpha rk^2qPf + \boldsymbol{\nabla}'\cdot\left(\alpha r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\right) - \boldsymbol{\nabla}'\cdot\left(\alpha irk\gamma_\infty qf\boldsymbol{M}_\infty\right)\\ & - \boldsymbol{\nabla}'\cdot\left(\alpha r\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f\right)\mathcal{N}\boldsymbol{M}\right) + \boldsymbol{\nabla}'\cdot\left(\alpha irk(1+qP)f\mathcal{N}\boldsymbol{M}\right).\end{aligned}$$

Expanding again the derivatives with respect to α yields

$$\begin{aligned}\varsigma = {}& rk^2 f + irk\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f + rk^2qPf + \boldsymbol{\nabla}'\cdot\left(r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\right) - irk_\infty\gamma_\infty\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\cdot\boldsymbol{M}_\infty\\ & - \boldsymbol{\nabla}'\cdot\left(irkq\gamma_\infty f\boldsymbol{M}_\infty\right) - rkk_\infty q\gamma_\infty^2 fM_\infty^2 - \boldsymbol{\nabla}'\cdot\left(r\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f\right)\mathcal{N}\boldsymbol{M}\right)\\ & + irk_\infty\gamma_\infty\left(\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{\nabla}'f\right)\mathcal{N}\boldsymbol{M}\cdot\boldsymbol{M}_\infty + \boldsymbol{\nabla}'\cdot\left(irk(1+qP)f\mathcal{N}\boldsymbol{M}\right) + rkk_\infty\gamma_\infty(1+qP)f\mathcal{N}\boldsymbol{M}\cdot\boldsymbol{M}_\infty.\end{aligned}$$

Using again (3.7) as well as the symmetry of $\mathcal{N}$, it is inferred

$$\begin{aligned}\varsigma = {}& rk^2 f + irk\mathcal{N}\boldsymbol{M}\cdot\boldsymbol{\nabla}'f + rk^2qPf + \boldsymbol{\nabla}'\cdot\left(r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\right) - irkq\gamma_\infty\boldsymbol{M}_\infty\cdot\boldsymbol{\nabla}'f\\ & - \boldsymbol{\nabla}'\cdot\left(irkq\gamma_\infty f\boldsymbol{M}_\infty\right) - rk^2q^2M_\infty^2 f - \boldsymbol{\nabla}'\cdot\left(r\left(\mathcal{N}\boldsymbol{M}\boldsymbol{M}^T\mathcal{N}\right)\nabla'f\right)\\ & + irkqP\mathcal{N}\boldsymbol{M}\cdot\boldsymbol{\nabla}'f + \boldsymbol{\nabla}'\cdot\left(irk(1+qP)f\mathcal{N}\boldsymbol{M}\right) + rk^2qP(1+qP)f.\end{aligned} \quad (3.9)$$

The terms are now reorganized with respect to the orders of derivation of f to obtain

$$\begin{aligned}\varsigma = {}& rk^2(1+qP-q^2M_\infty^2+qP(1+qP))f\\ & + irk\left((1+qP)\mathcal{N}\boldsymbol{M} - q\gamma_\infty\boldsymbol{M}_\infty\right)\cdot\boldsymbol{\nabla}'f\\ & + \boldsymbol{\nabla}'\cdot\left(irkf\left((1+qP)\mathcal{N}\boldsymbol{M} - q\gamma_\infty\boldsymbol{M}_\infty\right)\right)\\ & + \boldsymbol{\nabla}'\cdot\left(r\mathcal{N}\mathcal{N}\boldsymbol{\nabla}'f\right) - \boldsymbol{\nabla}'\cdot\left(r\left(\mathcal{N}\boldsymbol{M}\boldsymbol{M}^T\mathcal{N}\right)\boldsymbol{\nabla}'f\right),\end{aligned} \quad (3.10)$$

yielding (3.8).

We now show that, following the Prandtl–Glauert transformation, the boundary condition (3.2) becomes

$$(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla}'f)\cdot\boldsymbol{n}' = 0 \quad \text{on } \Gamma', \tag{3.11}$$

where Γ' denotes the transformed boundary Γ. The normals on the initial geometry are denoted by $\boldsymbol{n}$, and the normals on the transformed geometry by $\boldsymbol{n}'$. It is readily seen that

$$\boldsymbol{n} = K_\infty \mathcal{N}\boldsymbol{n}' \quad \text{on } \Gamma, \tag{3.12}$$

where K_∞ is a normalization factor that is not needed in what follows. Owing to (3.6) and (3.12), (3.2) becomes $\mathcal{N}\boldsymbol{\nabla}'\varphi\cdot\mathcal{N}\boldsymbol{n}' = 0$. Hence, $\mathcal{N}\boldsymbol{\nabla}'(\alpha f)\cdot\mathcal{N}\boldsymbol{n}' = 0$, leading to $(\mathcal{N}\boldsymbol{\nabla}'f - ik_\infty\gamma_\infty f\mathcal{N}\boldsymbol{M}_\infty)\cdot\mathcal{N}\boldsymbol{n}' = 0$. Since the flow is tangential on Γ, $\boldsymbol{M}\cdot\boldsymbol{n} = 0$ on Γ. Hence, $\boldsymbol{M}\cdot\mathcal{N}\boldsymbol{n}' = 0$ on Γ, so that

$$\left(\mathcal{N}\boldsymbol{\nabla}'f - (\mathcal{N}\boldsymbol{M}\cdot\boldsymbol{\nabla}'f)\,\boldsymbol{M} + ikf\left((1+qP)\boldsymbol{M} - \frac{k_\infty}{k}\gamma_\infty\mathcal{N}\boldsymbol{M}_\infty\right)\right)\cdot\mathcal{N}\boldsymbol{n}' = 0. \tag{3.13}$$

Using the symmetry of $\mathcal{N}$ and (3.7), (3.13) leads to (3.11).

An important observation is that in Ω^+, $\beta = \gamma_\infty^2$, $\boldsymbol{V} = \boldsymbol{0}$ and $\Xi = I$, so that (3.18a) becomes

$$\Delta f + \hat{k}_\infty^2 f = \varsigma \quad \text{in } \Omega^+, \tag{3.14}$$

where

$$\hat{k}_\infty := \gamma_\infty k_\infty. \tag{3.15}$$

Moreover, since $\mathrm{supp}(\varsigma)\subset\Omega^+$, f_{inc} satisfies

$$\Delta f_{\mathrm{inc}} + \hat{k}_\infty^2 f_{\mathrm{inc}} = \varsigma \quad \text{in } \Omega^+, \qquad \Delta f_{\mathrm{inc}} + \hat{k}_\infty^2 f_{\mathrm{inc}} = 0 \quad \text{in } \mathbb{R}^3\backslash\overline{\Omega^+}. \tag{3.16}$$

Eliminating f_{inc} in (3.14) yields,

$$\Delta f_{\mathrm{sc}} + \hat{k}_\infty^2 f_{\mathrm{sc}} = 0 \quad \text{in } \Omega^+. \tag{3.17}$$

This is the classical Helmholtz equation with modified wave number $\hat{k}_\infty$. Another important property is that the matrix Ξ is symmetric positive definite in Ω^- as well. To prove this, consider the matrices $\mathcal{N}$, $\mathcal{O}$ and Ξ, that are are all symmetric. If $M_0 < 1$, $\forall\boldsymbol{U}\in\mathbb{C}^3$, $\overline{\boldsymbol{U}}^T\Xi\boldsymbol{U} = \left(\overline{\mathcal{N}\boldsymbol{U}}\right)^T\mathcal{O}\left(\mathcal{N}\boldsymbol{U}\right) \geq (1-M_0^2)\,\|\mathcal{N}\boldsymbol{U}\|^2$. Now consider $\|\mathcal{N}\boldsymbol{U}\|^2 = \left\|\left(\hat{\boldsymbol{M}}_\infty^T\mathcal{N}\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty\right\|^2 + \left\|\mathcal{N}\boldsymbol{U} - \left(\hat{\boldsymbol{M}}_\infty^T\mathcal{N}\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty\right\|^2$, where $\hat{\boldsymbol{M}}_\infty := \frac{\boldsymbol{M}_\infty}{M_\infty}$. We have $\left(\hat{\boldsymbol{M}}_\infty^T\mathcal{N}\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty = \frac{1}{\sqrt{1-M_\infty^2}}\left(\hat{\boldsymbol{M}}_\infty^T\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty$ and $\mathcal{N}\boldsymbol{U} - \left(\hat{\boldsymbol{M}}_\infty^T\mathcal{N}\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty = \boldsymbol{U} - \left(\hat{\boldsymbol{M}}_\infty^T\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty$, therefore, for all $x\in\Omega^-$
$\overline{\boldsymbol{U}}^T\Xi(x)\boldsymbol{U} \geq (1-M_0^2(x))\left(\frac{1}{1-M_\infty^2}\left\|\left(\hat{\boldsymbol{M}}_\infty^T\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty\right\|^2 + \left\|\boldsymbol{U} - \left(\hat{\boldsymbol{M}}_\infty^T\boldsymbol{U}\right)\hat{\boldsymbol{M}}_\infty\right\|^2\right) \geq (1-M_0^2(x))\,\|\boldsymbol{U}\|^2$.
Moreover, for all $\boldsymbol{U},\boldsymbol{W}\in\mathbb{C}^3$, there holds $\overline{\boldsymbol{U}}\cdot\Xi(x)\boldsymbol{W} \leq \frac{1+M_0^2(x)}{1-M_\infty^2}\|\boldsymbol{U}\|\|\boldsymbol{W}\|$, $\forall x\in\Omega^-$.

In summary, the boundary value problem we consider is

$$rk^2\beta f + irk\boldsymbol{V}\cdot\boldsymbol{\nabla}f + \boldsymbol{\nabla}\cdot(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla}f) = \varsigma \quad \text{in } \Omega, \tag{3.18a}$$

$$(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla}f)\cdot\boldsymbol{n} = 0 \quad \text{on } \Gamma, \tag{3.18b}$$

$$\lim_{r\to+\infty} r\left(\frac{\partial(f-f_{\mathrm{inc}})}{\partial r} - i\hat{k}_\infty(f-f_{\mathrm{inc}})\right) = 0, \tag{3.18c}$$

where f is searched in $H^1_{\mathrm{loc}}(\Omega) := \{u \in H^1(K), \forall K \subset \Omega \text{ compact}\}$. Equation (3.18a) is the transformed convected Helmholtz equation, (3.18b) the transformed boundary condition, and the condition at infinity (3.18c) the classical Sommerfeld radiation condition that guarantees existence and uniqueness for Helmholtz exterior problems [76, Theorem 9.10]. In the general case, the Sommerfeld radiation condition is written for the scattered potential, since some incident acoustic potentials, e.g., plane waves, do not verify it.

3.3 Coupling procedure

Problem (3.18) is separated into an interior problem posed in Ω^- and an exterior problem posed in Ω^+ in view of using different numerical methods in each subdomain. Specifically, the problem in the interior domain is solved by means of finite elements, whereas the problem in the exterior domain is solved by means of boundary elements. The main purpose of this section is to derive two coupling procedures between the interior and exterior problems.

3.3.1 The transmission problem

The one-sided Dirichlet traces on Γ_∞ of a smooth function u in $\Omega^+ \cup \Omega^-$ are defined as $\gamma_0^\pm u^\pm := u^\pm|_{\Gamma_\infty}$, and the one-sided Neumann traces as $\gamma_1^\pm u^\pm := (\boldsymbol{\nabla} u^\pm)|_{\Gamma_\infty} \cdot \boldsymbol{n}$, where $u^\pm := u|_{\Omega^\pm}$ and where $\boldsymbol{n}$ is the unit normal vector to Γ_∞ conventionally pointing towards Ω^+ (see Figure 3.1). The one-sided normal traces on Γ_∞ of a smooth vector field $\boldsymbol{\sigma}$ in $\Omega^+ \cup \Omega^-$ are defined as $\gamma_n^\pm \boldsymbol{\sigma}^\pm := \boldsymbol{\sigma}^\pm|_{\Gamma_\infty} \cdot \boldsymbol{n}$, where $\boldsymbol{\sigma}^\pm := \boldsymbol{\sigma}|_{\Omega^\pm}$. These trace operators are extended to bounded linear operators $\gamma_0^\pm : H^1(\Omega^\pm) \to H^{\frac{1}{2}}(\Gamma_\infty)$, $\gamma_1^\pm : H(\Delta, \Omega^\pm) \to H^{-\frac{1}{2}}(\Gamma_\infty)$ and $\gamma_n^\pm : H(\mathrm{div}, \Omega^\pm) \to H^{-\frac{1}{2}}(\Gamma_\infty)$, where $H^1(\Omega^\pm)$, $H^{\frac{1}{2}}(\Gamma_\infty)$, and $H^{-\frac{1}{2}}(\Gamma_\infty)$ are the usual Sobolev spaces on $\Omega^\pm$ and Γ_∞, $H(\mathrm{div}, \Omega^\pm) = \{\boldsymbol{\sigma} \in \mathrm{L}^2(\Omega^\pm), \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} \in \mathrm{L}^2(\Omega^\pm)\}$, with $L^2(\Omega^\pm)$ the Lebesgue space of square integrable functions on $\Omega^\pm$, and $H(\Delta, \Omega^\pm) = \{u^\pm \in H^1(\Omega^\pm), \Delta u^\pm \in L^2(\Omega^\pm)\}$ (see [100, Lemma 20.2]). It is actually sufficient to consider functional spaces on compact subsets of Ω^+ to define exterior traces on Γ_∞. Let X denote the surface Γ or Γ_∞. The $L^2(X)$-inner product $\langle \cdot, \cdot \rangle_{L^2(X), L^2(X)} : L^2(X) \times L^2(X) \to \mathbb{C}$ is defined as

$$\langle \lambda, \mu \rangle_{L^2(X), L^2(X)} := \int_X \overline{\lambda}(\boldsymbol{y}) \mu(\boldsymbol{y}) ds(\boldsymbol{y}). \tag{3.19}$$

This inner product can be extended to a duality pairing on $H^{-\frac{1}{2}}(X) \times H^{\frac{1}{2}}(X)$ denoted by $\langle \cdot, \cdot \rangle_{H^{-\frac{1}{2}}(X), H^{\frac{1}{2}}(X)}$. Define now the product

$$(\lambda, \mu)_X := \begin{cases} \langle \lambda, \mu \rangle_{H^{-\frac{1}{2}}(X), H^{\frac{1}{2}}(X)} & \text{if } \lambda \in H^{-\frac{1}{2}}(X), \quad \mu \in H^{\frac{1}{2}}(X), \\ \overline{\langle \mu, \lambda \rangle}_{H^{-\frac{1}{2}}(X), H^{\frac{1}{2}}(X)} & \text{if } \lambda \in H^{\frac{1}{2}}(X) \ , \quad \mu \in H^{-\frac{1}{2}}(X). \end{cases} \tag{3.20}$$

We consider the following transmission problem where the one-sided normal trace $\gamma_{n,\Gamma}^-$ on Γ from Ω^- is used to formulate the boundary condition (3.18b):

$$rk^2\beta f^- + irk\boldsymbol{V}\cdot\boldsymbol{\nabla} f^- + \boldsymbol{\nabla}\cdot(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla} f)^- = 0 \quad \text{in } \Omega^-, \tag{3.21a}$$

$$\Delta f_{\rm sc} + \hat{k}_\infty^2 f_{\rm sc} = 0 \quad \text{in } \Omega^+, \tag{3.21b}$$

$$\gamma_{n,\Gamma}^-\,(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla} f)^- = 0 \quad \text{on } \Gamma, \tag{3.21c}$$

$$\gamma_0^+ f^+ - \gamma_0^- f^- = 0 \quad \text{on } \Gamma_\infty, \tag{3.21d}$$

$$\gamma_1^+ f^+ - \gamma_1^- f^- = 0 \quad \text{on } \Gamma_\infty, \tag{3.21e}$$

$$\lim_{r\to+\infty} r\left(\frac{\partial(f^+ - f_{\rm inc}^+)}{\partial r} - i\hat{k}_\infty(f^+ - f_{\rm inc}^+)\right) = 0. \tag{3.21f}$$

Proposition 3.1 *Problem* (3.18) *is equivalent to Problem* (3.21).

Proof. If f solves (3.18), it is clear that f solves (3.21). Conversely, let f be defined in $\Omega^+\cup\Omega^-$ such that f verifies (3.21). From [76, Lemma 4.19], since f solves (3.21a) and (3.21b), then f solves (3.18a) if and only if the jumps on Γ_∞ of its traces and of the normal component of $irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla} f$ vanish. The first condition is just (3.21d), while the second condition is $\gamma_1^+ f^+ - \gamma_n^-\,(irkf\boldsymbol{V} + r\Xi\boldsymbol{\nabla} f)^- = 0$, which, by continuity of the convective flow across Γ_∞, is just (3.21e). Finally, (3.18b) and (3.18c) are simply (3.21c) and (3.21f). ♢

Theorem 3.2 *Problem* (3.21) *is well-posed.*

Proof. See Section 3.7. ♢

3.3.2 Basic ingredients of the coupling procedure

The coupling procedure hinges on a weak formulation in the interior domain Ω^- and a Dirichlet-to-Neumann map (DtN) associated with the classical Helmholtz equation (3.21b) in the exterior domain Ω^+.

Derivation of Dirichlet-to-Neumann maps

For $u \in H^1(\Omega^+\cup\Omega^-)$, the jump and average of its Dirichlet traces across Γ_∞ are defined respectively as $[\gamma_0 u]_{\Gamma_\infty} := \gamma_0^+ u^+ - \gamma_0^- u^-$ and $\{\gamma_0 u\}_{\Gamma_\infty} := \frac{1}{2}\left(\gamma_0^+ u^+ + \gamma_0^- u^-\right)$. For $u \in H(\Delta, \Omega^+\cup\Omega^-)$, the jump and average of its Neumann traces across Γ_∞ are defined respectively as $[\gamma_1 u]_{\Gamma_\infty} := \gamma_1^+ u^+ - \gamma_1^- u^-$ and $\{\gamma_1 u\}_{\Gamma_\infty} := \frac{1}{2}\left(\gamma_1^+ u^+ + \gamma_1^- u^-\right)$. When a trace is single-valued at Γ_∞, we omit the superscripts $\pm$.

In what follows, Helmholtz equations, as well as corresponding boundary integral operators, are written for the transformed wave number $\hat{k}_\infty\gamma_\infty k_\infty$, cf. (3.15) . A function u defined over $\mathbb{R}^3$ is said to be a piecewise Helmholtz solution if $u|_{\Omega^+}$ and $u|_{\mathbb{R}^3\setminus\overline{\Omega^+}}$ solve the classical Helmholtz equation (3.21b) respectively in Ω^+ and $\mathbb{R}^3\setminus\overline{\Omega^+}$. A radiating piecewise Helmholtz solution is a piecewise Helmholtz solution that satisfies the Sommerfeld radiation condition (3.21f). For all $\lambda \in C^0(\Gamma_\infty)$, the single-layer potential is defined as $\mathcal{S}(\lambda)(\boldsymbol{x}) := \int_{\Gamma_\infty} E(\boldsymbol{y}-\boldsymbol{x})\lambda(\boldsymbol{y})ds(\boldsymbol{y})$,

$\boldsymbol{x} \in \mathbb{R}^3 \backslash \Gamma_\infty$, where $E(\boldsymbol{x}) := \frac{\exp(i\hat{k}_\infty |\boldsymbol{x}|)}{4\pi|\boldsymbol{x}|}$ is the fundamental solution of the classical Helmholtz equation (3.21b) with wave number $\hat{k}_\infty$ satisfying the Sommerfeld radiation condition (3.21f). For all $\mu \in C^0(\Gamma_\infty)$, the double-layer potential is defined as $\mathcal{D}(\mu)(\boldsymbol{x}) := \int_{\Gamma_\infty} \boldsymbol{\nabla}_{\boldsymbol{y}} E(\boldsymbol{y}-\boldsymbol{x})\mu(\boldsymbol{y}) ds(\boldsymbol{y})$, $\boldsymbol{x} \in \mathbb{R}^3 \backslash \Gamma_\infty$. From [93, Theorem 3.1.16], these operators can be extended to bounded linear operators $\mathcal{S} : H^{-\frac{1}{2}}(\Gamma_\infty) \to H^1_{\mathrm{loc}}(\mathbb{R}^3)$ and $\mathcal{D} : H^{\frac{1}{2}}(\Gamma_\infty) \to H^1_{\mathrm{loc}}(\mathbb{R}^3 \backslash \Gamma_\infty)$. Moreover, both map onto radiating piecewise Helmholtz solutions. Recalling [80, Theorem 3.1.1], a radiating piecewise Helmholtz solution u can be represented from its Dirichlet and Neumann jumps across Γ_∞ in the form

$$u = -\mathcal{S}([\gamma_1 u]_{\Gamma_\infty}) + \mathcal{D}([\gamma_0 u]_{\Gamma_\infty}) \quad \text{in } \Omega^+ \cup (\mathbb{R}^3 \backslash \overline{\Omega^+}). \tag{3.22}$$

The single-layer and double-layer potentials satisfy the following jump relations across Γ_∞ [80, Theorem 3.1.2]:

$$\begin{aligned} &[\gamma_0(\mathcal{S}\lambda)]_{\Gamma_\infty} = 0, && [\gamma_1(\mathcal{S}\lambda)]_{\Gamma_\infty} = -\lambda, && \forall \lambda \in H^{-\frac{1}{2}}(\Gamma_\infty), \\ &[\gamma_0(\mathcal{D}\mu)]_{\Gamma_\infty} = \mu, && [\gamma_1(\mathcal{D}\mu)]_{\Gamma_\infty} = 0, && \forall \mu \in H^{\frac{1}{2}}(\Gamma_\infty). \end{aligned} \tag{3.23}$$

The operators

$$\begin{aligned} &S : H^{-\frac{1}{2}}(\Gamma_\infty) \to H^{\frac{1}{2}}(\Gamma_\infty), && S\lambda := \gamma_0(\mathcal{S}\lambda), \\ &D : H^{\frac{1}{2}}(\Gamma_\infty) \to H^{\frac{1}{2}}(\Gamma_\infty), && D\mu := \{\gamma_0(\mathcal{D}\mu)\}_{\Gamma_\infty}, \\ &\tilde{D} : H^{-\frac{1}{2}}(\Gamma_\infty) \to H^{-\frac{1}{2}}(\Gamma_\infty), && \tilde{D}\lambda := \{\gamma_1(\mathcal{S}\lambda)\}_{\Gamma_\infty}, \\ &N : H^{\frac{1}{2}}(\Gamma_\infty) \to H^{-\frac{1}{2}}(\Gamma_\infty), && N\mu := -\gamma_1(\mathcal{D}\mu), \end{aligned} \tag{3.24}$$

are respectively the single-layer, double-layer, transpose (or dual) of the double-layer, and hypersingular boundary integral operators. The Dirichlet and Neumann traces are well-defined, and the mapping properties can be found in [76, Theorem 7.1]. The following trace identities are directly derived from (3.23):

$$\begin{aligned} &\gamma_0 \mathcal{S} = S, && \gamma_1^{\pm} \mathcal{S} = \tilde{D} \mp \frac{1}{2} I, \\ &\gamma_0^{\pm} \mathcal{D} = D \pm \frac{1}{2} I, && \gamma_1 \mathcal{D} = -N, \end{aligned} \tag{3.25}$$

Moreover, if u is a radiating piecewise Helmholtz solution, taking the interior traces of (3.22) and using (3.25) leads to

$$\begin{pmatrix} \frac{1}{2}I - D & S \\ N & \frac{1}{2}I + \tilde{D} \end{pmatrix} \begin{pmatrix} [\gamma_0 u]_{\Gamma_\infty} \\ [\gamma_1 u]_{\Gamma_\infty} \end{pmatrix} = - \begin{pmatrix} \gamma_0^- u^- \\ \gamma_1^- u^- \end{pmatrix}. \tag{3.26}$$

Let now f solve (3.21) and let v be the function defined by $v|_{\Omega^+} := f_{\mathrm{sc}}$ and $v|_{\mathbb{R}^3 \backslash \overline{\Omega^+}} := -f^-_{\mathrm{inc}}$. The function v is a radiating piecewise Helmholtz solution (on Ω^+ this follows from (3.21b) and (3.21f), and on $\mathbb{R}^3 \backslash \overline{\Omega^+}$ from (3.16)). Since f_{inc} is continuous across Γ_∞,

$$[\gamma_0 v]_{\Gamma_\infty} = \gamma_0^+ f_{\mathrm{sc}} + \gamma_0^- f^-_{\mathrm{inc}} = \gamma_0^+ f_{\mathrm{sc}} + \gamma_0^+ f^+_{\mathrm{inc}} = \gamma_0^+ f^+. \tag{3.27}$$

Likewise, $[\gamma_1 v]_{\Gamma_\infty} = \gamma_1^+ f^+$. Other choices can be made for v (in particular $v|_{\mathbb{R}^3 \backslash \overline{\Omega^+}} = 0$). Here, the jumps correspond to the traces of the total transformed acoustic potential, and a direct

coupling with the equation in Ω^- (having f as unknown), is then possible. In what follows, we drop $\pm$ superscripts for the Dirichlet and Neumann traces of f and f_{inc} since the traces are single-valued. Then, (3.26) applied to v yields

$$\begin{pmatrix} \frac{1}{2}I - D & S \\ N & \frac{1}{2}I + \tilde{D} \end{pmatrix} \begin{pmatrix} \gamma_0 f \\ \gamma_1 f \end{pmatrix} = \begin{pmatrix} \gamma_0 f_{\text{inc}} \\ \gamma_1 f_{\text{inc}} \end{pmatrix}. \tag{3.28}$$

Various identities relating $\gamma_0 f$ and $\gamma_1 f$ can be derived from (3.28), and these identities can be used to define DtN maps. For example, using the first line of (3.28) yields formally $\gamma_1 f = DtN_0(\gamma_0 f) := S^{-1}\left(\gamma_0 f_{\text{inc}} + \left(D - \frac{1}{2}I\right)\gamma_0 f\right)$, where the question of the inversibility of S has to be addressed. Two other examples are detailed in Sections 3.3.3 and 3.3.4 below.

Remark 3.3 *The block operator defined on the left-hand side of* (3.28) *is not injective.*

Weak formulation in the interior domain Ω^-

Let $\Phi := f|_{\Omega^-}$ where f solves (3.21). Multiplying (3.21a) by a test function $\Phi^t \in H^1(\Omega^-)$ and using a Green formula together with the boundary condition (3.21c) at Γ yields

$$\mathcal{V}(\Phi, \Phi^t) - \left(\gamma_1^- \Phi, \gamma_0^- \Phi^t\right)_{\Gamma_\infty} = 0, \tag{3.29}$$

with the sesquilinear form

$$\mathcal{V}(\Phi, \Phi^t) := \int_{\Omega^-} r\Xi \boldsymbol{\nabla}\overline{\Phi} \cdot \boldsymbol{\nabla}\Phi^t - \int_{\Omega^-} rk^2 \beta \overline{\Phi}\Phi^t + i\int_{\Omega^-} rk\boldsymbol{V} \cdot \left(\overline{\Phi}\boldsymbol{\nabla}\Phi^t - \Phi^t \boldsymbol{\nabla}\overline{\Phi}\right). \tag{3.30}$$

Using the transmission conditions (3.21d)-(3.21e), $\gamma_0^-\Phi = \gamma_0 f$ and $\gamma_1^-\Phi = \gamma_1 f$, so that the coupling with the exterior problem can be written as $\gamma_1^-\Phi = DtN(\gamma_0^-\Phi)$. This yields the following coupled formulation: Find $\Phi \in H^1(\Omega^-)$ such that $\forall \Phi^t \in H^1(\Omega^-)$,

$$\mathcal{V}(\Phi, \Phi^t) - \left(DtN(\gamma_0^-\Phi), \gamma_0^-\Phi^t\right)_{\Gamma_\infty} = 0. \tag{3.31}$$

3.3.3 Unstable coupled formulation

To carry out the coupling, a first classical DtN map is considered. Since this DtN map is not well-defined at some frequencies, the resulting coupled formulation is not well-posed at these frequencies, and is therefore called unstable. From (3.28), recalling $\gamma_0^-\Phi = \gamma_0 f$ and $\gamma_1^-\Phi = \gamma_1 f$, there holds

$$\begin{pmatrix} \frac{1}{2}I - D & S \\ N & \frac{1}{2}I + \tilde{D} \end{pmatrix} \begin{pmatrix} \gamma_0^-\Phi \\ \gamma_1^-\Phi \end{pmatrix} = \begin{pmatrix} \gamma_0 f_{\text{inc}} \\ \gamma_1 f_{\text{inc}} \end{pmatrix}. \tag{3.32}$$

Using the first line of (3.32), $\gamma_1^-\Phi = S^{-1}\left(\left(D - \frac{1}{2}I\right)\left(\gamma_0^-\Phi\right) + \gamma_0 f_{\text{inc}}\right)$. At this point, the inverse of S is written formally. Conditions of inversibility are discussed below. From the second line of (3.32), $\gamma_1^-\Phi = -N(\gamma_0^-\Phi) + \left(\frac{1}{2}I - \tilde{D}\right)\left(\gamma_1^-\Phi\right) + \gamma_1 f_{\text{inc}}$. Injecting into the right-hand side of this relation the expression of $\gamma_1^-\Phi$ derived above yields the DtN affine map: $DtN_{\text{unstab}} : H^{\frac{1}{2}}(\Gamma_\infty) \to H^{-\frac{1}{2}}(\Gamma_\infty)$ such that

$$\gamma_1^- \Phi = DtN_{\text{unstab}}(\gamma_0^- \Phi) := -N(\gamma_0^- \Phi) + \left(\frac{1}{2} I - \tilde{D}\right) S^{-1} \left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \Phi) + \gamma_0 f_{\text{inc}}\right) + \gamma_1 f_{\text{inc}}. \tag{3.33}$$

The operator inversion requires to introduce the auxiliary field $\lambda \in H^{-\frac{1}{2}}(\Gamma_\infty)$ such that

$$\left(D - \frac{1}{2} I\right)(\gamma_0^- \Phi) - S\lambda = -\gamma_0 f_{\text{inc}}, \tag{3.34}$$

yielding

$$DtN_{\text{unstab}}(\gamma_0^- \Phi) = -N(\gamma_0^- \Phi) + \left(\frac{1}{2} I - \tilde{D}\right)(\lambda) + \gamma_1 f_{\text{inc}}. \tag{3.35}$$

Injecting $DtN_{\text{unstab}}(\gamma_0^- \Phi)$ from (3.35) into the formulation (3.31) yields, using (3.34), the following coupled variational formulation: Find $(\Phi, \lambda) \in \mathcal{H}$ such that, $\forall\, (\Phi^t, \lambda^t) \in \mathcal{H}$,

$$\mathcal{V}(\Phi, \Phi^t) + \left(N(\gamma_0^- \Phi), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\left(\tilde{D} - \frac{1}{2} I\right)(\lambda), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} = \left(\gamma_1 f_{\text{inc}}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty}, \tag{3.36a}$$

$$\left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \Phi), \lambda^t\right)_{\Gamma_\infty} - \left(S(\lambda), \lambda^t\right)_{\Gamma_\infty} = -\left(\gamma_0 f_{\text{inc}}, \lambda^t\right)_{\Gamma_\infty}, \tag{3.36b}$$

with product space $\mathcal{H} := H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma_\infty)$ and inner product $\left((\Phi, \lambda), (\Phi^t, \lambda^t)\right)_{\mathcal{H}} := (\Phi, \Phi^t)_{H^1(\Omega^-)} + (\lambda, \lambda^t)_{H^{-\frac{1}{2}}(\Gamma_\infty)}$. The formulation (3.36) is called unstable since it admits infinitely many solutions at some frequencies of the source, leading to incorrect numerical results.

Remark 3.4 *The DtN_{unstab} affine map was proposed by Costabel to obtain a symmetric coupling in the case of self-adjoint operators [31]. The DtN_{unstab} map can be well-defined for certain operators: for instance, for transmission problems for the Laplace equation, this map leads to a well-defined symmetric system. In the system* (3.36)*, the only non-symmetric contribution results from the vector $\boldsymbol{V}$ in the sesquilinear form $\mathcal{V}$, cf.* (3.30)*. The system becomes symmetric when the flow is uniform everywhere. However, since D and $\tilde{D}$ are dual but not adjoint operators, there is no Hermitian symmetry.*

Proposition 3.5 *If f solves* (3.21)*, then $(f^-, \gamma_1 f)$ solves* (3.36)*. Conversely, if (Φ, λ) solves* (3.36)*, then $\mathcal{R}(\Phi, \lambda)$ solves* (3.21)*, where $\mathcal{R} : \mathcal{H} \to H^1_{\text{loc}}(\Omega \backslash \Gamma_\infty)$ is such that $\mathcal{R}(\Phi, \lambda)|_{\Omega^-} := \Phi$ and $\mathcal{R}(\Phi, \lambda)|_{\Omega^+} := (-\mathcal{S}(\lambda) + \mathcal{D}(\gamma_0^- \Phi) + f_{\text{inc}})|_{\Omega^+}$.*

Proof. See Section 3.7. ♢

The main difficulty with the unstable coupled formulation (3.36) stems from the fact that $\text{Ker}(S)$ depends on whether $-\hat{k}_\infty^2$ belongs to the set Λ of Dirichlet eigenvalues for the Laplacian on the bounded domain $\mathbb{R}^3 \backslash \overline{\Omega^+}$. Specifically, $\text{Ker}(S) = \{0\}$ if $-\hat{k}_\infty^2 \notin \Lambda$, while $\text{Ker}(S)$ contains nontrivial elements if $-\hat{k}_\infty^2 \in \Lambda$.

Proposition 3.6 *If f solves* (3.21)*, then for all $\lambda^* \in \text{Ker}(S)$, $(f^-, \gamma_1 f + \lambda^*)$ solves* (3.36)*.*

Proof. This is a direct consequence of $\text{Ker}(S) = \text{Ker}(\tilde{D} - \frac{1}{2} I)$. ♢

Theorem 3.7 *If $-\hat{k}_\infty^2 \notin \Lambda$, then problem* (3.36) *is well-posed. If $-\hat{k}_\infty^2 \in \Lambda$, then* (3.36) *admits infinitely many solutions of the form $(f^-, \gamma_1 f + \lambda^*)$, where f is the solution to* (3.21) *and λ^* is any element in* $\mathrm{Ker}(S)$.

Proof. See Section 3.7. ♢

Remark 3.8 *Let $-\hat{k}_\infty^2 \in \Lambda$. Owing to Proposition 3.5, for any couple (Φ, λ) solving* (3.36)*, $\mathcal{R}(\Phi, \lambda)$ solves* (3.21)*. However, even if our goal is to solve* (3.21)*, we will see in Section 3.5 that the numerical procedure to approximate* (3.36) *fails to the point that $\mathcal{R}(\Phi, \lambda)$ is dominated by numerical errors.*

The formulation (3.36) is written on a geometry and for unknown functions that has been transformed by the Prandtl–Glauert transformation. The physical acoustic potential in the canonical system of coordinates is obtained by applying the inverse Prandtl–Glauert transformation to $\mathcal{R}(\Phi, \lambda)$. The formulation (3.36) looks similar to the classical symmetrical coupled formulation proposed by Costabel [31], and recalled in the case of the Helmholtz transmission problem in [54, equation (15)]. In the formulation (3.36), the integral operators are written at the transformed wave number $\hat{k}_\infty$, which differ from the wavenumber k_∞ of the source. The local convection of the acoustic potential by the mean flow is taken into account through the Prandtl–Glauert transformation and the volumic term $\mathcal{V}$, which differ from classical coupled formulations written for the classical (nonconvected) Helmholtz equation.

3.3.4 Stable coupled formulation

The idea of considering a linear combination of S and $\frac{1}{2}I + \tilde{D}$ to derive well-posed boundary integral equations was independently proposed in 1965 by Brakhage and Werner [20], Leis [65] and Panich [83]. This is the so-called Combined Field Integral Equation (CFIE). However, S and D map $H^{-\frac{1}{2}}(\Gamma_\infty)$ into different spaces ($H^{\frac{1}{2}}(\Gamma_\infty)$ and $H^{-\frac{1}{2}}(\Gamma_\infty)$ respectively). This inconsistency in the functional setting can be solved by considering a regularizing compact operator from $H^{-\frac{1}{2}}(\Gamma_\infty)$ into $H^{\frac{1}{2}}(\Gamma_\infty)$, as observed by Buffa and Hiptmair [23]. We briefly recall the approach of [23] and apply it to the present setting. Let $\boldsymbol{\nabla}_{\Gamma_\infty}$ denote the surfacic gradient on Γ_∞. Consider the following Hermitian sesquilinear form: For all $p, q \in H^1(\Gamma_\infty)$,

$$\delta_{\Gamma_\infty}(p, q) := (\boldsymbol{\nabla}_{\Gamma_\infty} p, \boldsymbol{\nabla}_{\Gamma_\infty} q)_{\Gamma_\infty} + (p, q)_{\Gamma_\infty}\,, \tag{3.37}$$

and the regularizing operator $M : H^{-1}(\Gamma_\infty) \to H^1(\Gamma_\infty)$ is defined through the following implicit relation: For all $p \in H^1(\Gamma_\infty)$,

$$\delta_{\Gamma_\infty}(Mp, q) = (p, q)_{\Gamma_\infty}, \quad \forall q \in H^1(\Gamma_\infty). \tag{3.38}$$

It is readily seen that $M = (-\Delta_{\Gamma_\infty} + I)^{-1}$, where Δ_{Γ_∞} is the Laplace–Beltrami operator on Γ_∞.

Many choices of DtN maps based on CFIE strategies with the regularizing operator M lead to well-posed systems whatever the value of $\hat{k}_\infty$. The present choice hinges on the inversion of the operator $S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)$ mapping $H^{-\frac{1}{2}}(\Gamma_\infty)$ into $H^{\frac{1}{2}}(\Gamma_\infty)$ since, from [23, Lemma 4.1], this operator is bijective as long as the coupling parameter η is such that $\mathrm{Re}(\eta) \neq 0$. To

do so, the first line of (3.32) and the application of M to the second line of (3.32) are used to obtain

$$\begin{pmatrix} \left(\frac{1}{2}I - D\right) + i\eta MN & S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right) \\ N & \frac{1}{2}I + \tilde{D} \end{pmatrix} \begin{pmatrix} \gamma_0^- \Phi \\ \gamma_1^- \Phi \end{pmatrix} = \begin{pmatrix} \gamma_0 f_{\text{inc}} + i\eta M \gamma_1 f_{\text{inc}} \\ \gamma_1 f_{\text{inc}} \end{pmatrix}. \tag{3.39}$$

Then, using both equations in (3.39) in the same fashion as in Section 3.3.3 leads to $DtN_{\text{stab}} : H^{\frac{1}{2}}(\Gamma_\infty) \to H^{-\frac{1}{2}}(\Gamma_\infty)$ such that

$$\begin{aligned} \gamma_1^- \Phi = DtN_{\text{stab}}(\gamma_0^- \Phi) := -N(\gamma_0^- \Phi) + \left(\frac{1}{2}I - \tilde{D}\right) \left[S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)\right]^{-1} \\ \left(-\left[\left(\frac{1}{2}I - D\right) + i\eta MN\right](\gamma_0^- \Phi) + \gamma_0 f_{\text{inc}} + i\eta M \gamma_1 f_{\text{inc}}\right) + \gamma_1 f_{\text{inc}}. \end{aligned} \tag{3.40}$$

The operator inversion requires to introduce the auxiliary field $\lambda \in H^{-\frac{1}{2}}(\Gamma_\infty)$ such that

$$\left[S + i\eta M\left(\frac{1}{2}I + \tilde{D}\right)\right](\lambda) + \left[\left(\frac{1}{2}I - D\right) + i\eta MN\right](\gamma_0^- \Phi) = \gamma_0 f_{\text{inc}} + i\eta M(\gamma_1 f_{\text{inc}}), \tag{3.41}$$

so that

$$DtN_{\text{stab}}(\gamma_0^- \Phi) = -N(\gamma_0^- \Phi) + \left(\frac{1}{2}I - \tilde{D}\right)(\lambda) + \gamma_1 f_{\text{inc}}. \tag{3.42}$$

The evaluation of M involving an operator inversion as well, it requires to introduce another auxiliary field $p \in H^1(\Gamma_\infty)$ such that, for all $q \in H^1(\Gamma_\infty)$,

$$\delta_{\Gamma_\infty}(p, q) = \left(N(\gamma_0^- \Phi), q\right)_{\Gamma_\infty} + \left(\left(\frac{1}{2}I + \tilde{D}\right)(\lambda), q\right)_{\Gamma_\infty} - (\gamma_1 f_{\text{inc}}, q)_{\Gamma_\infty}, \tag{3.43}$$

so that equation (3.41) can be rewritten

$$S(\lambda) + \left(\frac{1}{2}I - D\right)(\gamma_0^- \Phi) + i\eta p = \gamma_0 f_{\text{inc}}. \tag{3.44}$$

Injecting $DtN_{\text{stab}}(\gamma_0^- \Phi)$ from (3.42) into the formulation (3.31) yields, using (3.43) and (3.44), the following stable coupled variational formulation: Find $(\Phi, \lambda, p) \in \mathbb{H}$ such that $\forall\, (\Phi^t, \lambda^t, p^t) \in \mathbb{H}$,

$$\mathcal{V}(\Phi, \Phi^t) + \left(N(\gamma_0^- \Phi), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\left(\tilde{D} - \frac{1}{2}I\right)(\lambda), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} = \left(\gamma_1 f_{\text{inc}}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty}, \tag{3.45a}$$

$$\left(\left(D - \frac{1}{2}I\right)(\gamma_0^- \Phi), \lambda^t\right)_{\Gamma_\infty} - \left(S(\lambda), \lambda^t\right)_{\Gamma_\infty} + i\overline{\eta}\left(p, \lambda^t\right)_{\Gamma_\infty} = -\left(\gamma_0 f_{\text{inc}}, \lambda^t\right)_{\Gamma_\infty}, \tag{3.45b}$$

$$\left(N(\gamma_0^- \Phi), p^t\right)_{\Gamma_\infty} + \left(\left(\tilde{D} + \frac{1}{2}I\right)(\lambda), p^t\right)_{\Gamma_\infty} - \delta_{\Gamma_\infty}(p, p^t) = \left(\gamma_1 f_{\text{inc}}, p^t\right)_{\Gamma_\infty}, \tag{3.45c}$$

with product space $\mathbb{H} := H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma_\infty) \times H^1(\Gamma_\infty)$ and inner product $\left((\Phi, \lambda, p), (\Phi^t, \lambda^t, p^t)\right)_{\mathbb{H}} := (\Phi, \Phi^t)_{H^1(\Omega^-)} + (\lambda, \lambda^t)_{H^{-\frac{1}{2}}(\Gamma_\infty)} + (p, p^t)_{H^1(\Gamma_\infty)}$.

Proposition 3.9 *If f solves (3.21), then $(f^-, \gamma_1 f, 0)$ solves (3.45). Conversely, if (Φ, λ, p) solves (3.45), then $\mathcal{R}(\Phi, \lambda)$ solves (3.21) and $p = 0$, where $\mathcal{R}$ is defined in Proposition 3.5.*

Proof. See Section 3.7. ♢

Theorem 3.10 *Problem* (3.45) *is well-posed at all frequencies.*

Proof. See Section 3.7. ♢

Remark 3.11 *Hiptmair and Meury [54] derived abstract trace transformation operators and generalized Calderón projectors for the Helmholtz transmission problem. By construction, integral operators written using these projectors enjoy the uniqueness property whatever the value of* $\hat{k}_\infty$*. The map* DtN_{stab} *corresponds to a particular choice of the trace transformation operator in the general setting of [54, Section 9]. Hence, Theorem 3.10 is an extension of [54, Theorem 7.3] to the case where a flow exists and is nonuniform in a bounded domain, for a particular choice of the trace transformation operator.*

The formulation (3.45) corresponds to the stabilization of the formulation (3.36), using a well-posed CFIE technique taken from the literature.

3.4 Finite-dimensional approximation

The coupled formulations (3.36) and (3.45) are approximated by finite element and boundary element methods. The underlying results are well-known from both theories and can be directly applied to the present setting.

3.4.1 Discrete finite element spaces

Let $\mathcal{M}$ be a shape-regular tetrahedral mesh of Ω^-. The mesh $\mathcal{F}_\infty$ of Γ_∞ is composed of the boundary faces of $\mathcal{M}$. Let $h_\mathcal{M} > 0$ denote the mesh size, $V^1_\mathcal{M}$ the space of continuous piecewise affine polynomials on $\mathcal{M}$, $S^0_\mathcal{M}$ the space of piecewise constant polynomials on $\mathcal{F}_\infty$, and $S^1_\mathcal{M}$ the space of continuous piecewise affine polynomials on $\mathcal{F}_\infty$. Let $\mathcal{H}_\mathcal{M} := V^1_\mathcal{M} \times S^0_\mathcal{M}$, and $\mathbb{H}_\mathcal{M} := V^1_\mathcal{M} \times S^0_\mathcal{M} \times S^1_\mathcal{M}$. The discretization of (3.36) reads: Find $(\Phi_\mathcal{M}, \lambda_\mathcal{M}) \in \mathcal{H}_\mathcal{M}$ such that, $\forall\, (\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}) \in \mathcal{H}_\mathcal{M}$,

$$a^{\mathrm{unstab}}\left((\Phi_\mathcal{M}, \lambda_\mathcal{M}), \left(\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}\right)\right) = b^{\mathrm{unstab}}\left(\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}\right), \tag{3.46}$$

with a^{unstab} and b^{unstab} readily deduced from (3.36), while the discretization of (3.45) reads: Find $(\Phi_\mathcal{M}, \lambda_\mathcal{M}, p_\mathcal{M}) \in \mathbb{H}_\mathcal{M}$ such that, $\forall\, (\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}, p^t_\mathcal{M}) \in \mathbb{H}_\mathcal{M}$,

$$a^{\mathrm{stab}}\left((\Phi_\mathcal{M}, \lambda_\mathcal{M}, p_\mathcal{M}), \left(\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}, p^t_\mathcal{M}\right)\right) = b^{\mathrm{stab}}\left(\Phi^t_\mathcal{M}, \lambda^t_\mathcal{M}, p^t_\mathcal{M}\right), \tag{3.47}$$

with a^{stab} and b^{stab} readily deduced from (3.45). Since $\mathcal{H}_\mathcal{M} \subset \mathcal{H}$ and $\mathbb{H}_\mathcal{M} \subset \mathbb{H}$, both approximations are conforming.

In what follows, $A \lesssim B$ denotes the inequality $A \leq cB$ with positive constant c independent of the mesh size and of the discrete and exact solutions. The following classical approximation properties are available (see [21, 42, 93]):

$$\begin{aligned} \inf_{\Phi_{\mathcal{M}} \in V^1_{\mathcal{M}}} \|\Phi - \Phi_{\mathcal{M}}\|_{H^1(\Omega^-)} &\lesssim h_{\mathcal{M}} \|\Phi\|_{H^2(\Omega^-)}, \\ \inf_{\lambda_{\mathcal{M}} \in S^0_{\mathcal{M}}} \|\lambda - \lambda_{\mathcal{M}}\|_{H^{-\frac{1}{2}}(\Gamma_\infty)} &\lesssim h_{\mathcal{M}} \|\lambda\|_{H^{\frac{1}{2}}(\Gamma_\infty)}, \\ \inf_{p_{\mathcal{M}} \in S^1_{\mathcal{M}}} \|p - p_{\mathcal{M}}\|_{H^1(\Gamma_\infty)} &\lesssim h_{\mathcal{M}} \|p\|_{H^2(\Gamma_\infty)}. \end{aligned} \tag{3.48}$$

Hence, the following approximation properties hold: $\forall (\Phi, \lambda) \in H^2(\Omega^-) \times H^{\frac{1}{2}}(\Gamma_\infty)$,

$$\inf_{(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}) \in \mathcal{H}_{\mathcal{M}}} \|(\Phi, \lambda) - (\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}})\|_{\mathcal{H}} \lesssim h_{\mathcal{M}} \left(\|\Phi\|_{H^2(\Omega^-)} + \|\lambda\|_{H^{\frac{1}{2}}(\Gamma_\infty)} \right), \tag{3.49}$$

and $\forall (\Phi, \lambda, p) \in H^2(\Omega^-) \times H^{\frac{1}{2}}(\Gamma_\infty) \times H^2(\Gamma_\infty)$,

$$\inf_{(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}, p_{\mathcal{M}}) \in \mathbb{H}_{\mathcal{M}}} \|(\Phi, \lambda, p) - (\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}, p_{\mathcal{M}})\|_{\mathbb{H}} \lesssim h_{\mathcal{M}} \left(\|\Phi\|_{H^2(\Omega^-)} + \|\lambda\|_{H^{\frac{1}{2}}(\Gamma_\infty)} + \|p\|_{H^2(\Gamma_\infty)} \right). \tag{3.50}$$

Remark 3.12 *Taking a polynomial approximation with one order less for $H^{-\frac{1}{2}}(\Gamma_\infty)$ than for $H^1(\Omega^-)$ and $H^1(\Gamma_\infty)$ enables all the approximations to be at the same order in $h_{\mathcal{M}}$.*

3.4.2 Discretization of the coupled formulations

Unstable formulation

Let $(\theta_i)_{1 \leq i \leq p}$ and $(\psi_i)_{1 \leq i \leq q}$ denote finite element bases for $V^1_{\mathcal{M}}$ and $S^0_{\mathcal{M}}$ respectively. These basis functions are real-valued. The decompositions of $\Phi_{\mathcal{M}} \in V^1_{\mathcal{M}}$ and $\lambda_{\mathcal{M}} \in S^0_{\mathcal{M}}$ on these bases are written in the form $\Phi_{\mathcal{M}} = \sum_{i=1}^p \Phi_{\mathcal{M}i} \theta_i$ and $\lambda_{\mathcal{M}} = \sum_{i=1}^q \lambda_{\mathcal{M}i} \psi_i$. Let

$$u^{\text{unstab}}_{\mathcal{M}} = \begin{pmatrix} (\Phi_{\mathcal{M}i})_{\ 1 \leq i \leq p} \\ (\lambda_{\mathcal{M}i})_{\ 1 \leq i \leq q} \end{pmatrix}, \qquad B^{\text{unstab}} = \begin{pmatrix} \left(\gamma_1 f_{\text{inc}}, \gamma_0^- \theta_i\right)_{\Gamma_\infty \ 1 \leq i \leq p} \\ -\left(\gamma_0 f_{\text{inc}}, \psi_i\right)_{\Gamma_\infty \ 1 \leq i \leq q} \end{pmatrix}, \tag{3.51}$$

$$A^{\text{unstab}} = \left(\begin{array}{c|c} \mathcal{V}(\theta_j, \theta_i) + \left(N(\gamma_0^- \theta_j), \gamma_0^- \theta_i\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \gamma_0^- \theta_i\right)_{\Gamma_\infty} \\ \hline \left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \theta_j), \psi_i\right)_{\Gamma_\infty} & -(S(\psi_j), \psi_i)_{\Gamma_\infty} \end{array} \right), \tag{3.52}$$

where in A^{unstab} the index i refers to the rows and the index j to the columns. The linear system resulting from (3.46) is

$$A^{\text{unstab}} u^{\text{unstab}}_{\mathcal{M}} = B^{\text{unstab}}. \tag{3.53}$$

To better understand the structure of the linear system (3.53), the basis functions $(\theta_i)_{1 \leq i \leq p}$ of $V^1_{\mathcal{M}}$ are separated into two sets: the basis function $(\theta_i^{\mathcal{F}_\infty})_{1 \leq i \leq p^{\mathcal{F}_\infty}}$ associated to the vertices of $\mathcal{F}_\infty$, and $(\theta_i^{\mathring{\mathcal{M}}})_{1 \leq i \leq p^{\mathring{\mathcal{M}}}}$, such that $p = p^{\mathcal{F}_\infty} + p^{\mathring{\mathcal{M}}}$. The matrix A^{unstab} is written

$$A^{\mathrm{unstab}} = \left(\begin{array}{c|c|c} \mathcal{V}(\theta_j^{\mathring{\mathcal{M}}}, \theta_i^{\mathring{\mathcal{M}}}) & \mathcal{V}(\theta_j^{\mathcal{F}\infty}, \theta_i^{\mathring{\mathcal{M}}}) & 0 \\ \hline \mathcal{V}(\theta_j^{\mathring{\mathcal{M}}}, \theta_i^{\mathcal{F}\infty}) & \mathcal{V}(\theta_j^{\mathcal{F}\infty}, \theta_i^{\mathcal{F}\infty}) + \left(N(\gamma_0^- \theta_j^{\mathcal{F}\infty}), \gamma_0^- \theta_i^{\mathcal{F}\infty}\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \gamma_0^- \theta_i^{\mathcal{F}\infty}\right)_{\Gamma_\infty} \\ \hline 0 & \left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \theta_j^{\mathcal{F}\infty}), \psi_i\right)_{\Gamma_\infty} & -(S(\psi_j), \psi_i)_{\Gamma_\infty} \end{array}\right). \tag{3.54}$$

The blocks of the matrix in (3.54) are denoted

$$A^{\mathrm{unstab}} = \left(\begin{array}{c|c|c} A_{1,1}^{\mathrm{unstab}} & A_{1,2}^{\mathrm{unstab}} & 0 \\ \hline A_{2,1}^{\mathrm{unstab}} & A_{2,2}^{\mathrm{unstab}} & A_{2,3}^{\mathrm{unstab}} \\ \hline 0 & A_{3,2}^{\mathrm{unstab}} & A_{3,3}^{\mathrm{unstab}} \end{array}\right). \tag{3.55}$$

All the blocks are complex-valued. The blocks $A_{1,1}^{\mathrm{unstab}}$, $A_{1,2}^{\mathrm{unstab}}$ and $A_{2,1}^{\mathrm{unstab}}$ are sparse. The block $A_{1,1}^{\mathrm{unstab}}$ is not symmetric, and the block $A_{1,2}^{\mathrm{unstab}}$ is neither the transpose nor the hermitian transpose of block $A_{2,1}^{\mathrm{unstab}}$. The block $A_{2,2}^{\mathrm{unstab}}$ has two contributions: one sparse and nonsymmetric and one dense and symmetric, therefore this block is dense and nonsymmetric. The blocks $A_{2,3}^{\mathrm{unstab}}$, $A_{3,2}^{\mathrm{unstab}}$ and $A_{3,3}^{\mathrm{unstab}}$ are dense. The block $A_{2,3}^{\mathrm{unstab}}$ is the transpose of the block $A_{3,2}^{\mathrm{unstab}}$, and the block $A_{3,3}^{\mathrm{unstab}}$ is symmetric.

Stable formulation

Let $(\xi_i)_{1\leq i\leq r}$ denote a finite element basis for $S^1_{\mathcal{M}}$. The decomposition of $p_{\mathcal{M}} \in S^1_{\mathcal{M}}$ on this basis is written in the form $p_{\mathcal{M}} = \sum_{i=1}^r p_{\mathcal{M}i}\xi_i$. Let

$$u_{\mathcal{M}}^{\mathrm{stab}} = \begin{pmatrix} (\Phi_{\mathcal{M}i})_{\ 1\leq i\leq p} \\ (\lambda_{\mathcal{M}i})_{\ 1\leq i\leq q} \\ (p_{\mathcal{M}i})_{\ 1\leq i\leq r} \end{pmatrix}, \qquad B^{\mathrm{stab}} = \begin{pmatrix} \left(\gamma_1 f_{\mathrm{inc}}, \gamma_0^- \theta_i\right)_{\Gamma_\infty\ 1\leq i\leq p} \\ -\left(\gamma_0 f_{\mathrm{inc}}, \psi_i\right)_{\Gamma_\infty\ 1\leq i\leq q} \\ \left(\gamma_1 f_{\mathrm{inc}}, \xi_i\right)_{\Gamma_\infty\ 1\leq i\leq r} \end{pmatrix}, \tag{3.56}$$

$$A^{\mathrm{stab}} = \left(\begin{array}{c|c|c} \mathcal{V}(\theta_j, \theta_i) + \left(N(\gamma_0^- \theta_j), \gamma_0^- \theta_i\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \gamma_0^- \theta_i\right)_{\Gamma_\infty} & 0 \\ \hline \left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \theta_j), \psi_i\right)_{\Gamma_\infty} & -(S(\psi_j), \psi_i)_{\Gamma_\infty} & i\overline{\eta}(\xi_j, \psi_i)_{\Gamma_\infty} \\ \hline \left(N(\gamma_0^- \theta_j), \xi_i\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \xi_i\right)_{\Gamma_\infty} & -\delta_{\Gamma_\infty}(\xi_j, \xi_i)_{\Gamma_\infty} \end{array}\right), \tag{3.57}$$

with the same convention on the indices i and j of A^{stab}. The linear system resulting from (3.47) is

$$A^{\mathrm{stab}} u_{\mathcal{M}}^{\mathrm{stab}} = B^{\mathrm{stab}}. \tag{3.58}$$

Like the previous section, the matrix of the linear system (3.58) is written

$$A^{\mathrm{stab}} = \left(\begin{array}{c|c|c|c} \mathcal{V}(\theta_j^{\mathring{\mathcal{M}}}, \theta_i^{\mathring{\mathcal{M}}}) & \mathcal{V}(\theta_j^{\mathcal{F}\infty}, \theta_i^{\mathring{\mathcal{M}}}) & 0 & 0 \\ \hline \mathcal{V}(\theta_j^{\mathring{\mathcal{M}}}, \theta_i^{\mathcal{F}\infty}) & \mathcal{V}(\theta_j^{\mathcal{F}\infty}, \theta_i^{\mathcal{F}\infty}) + \left(N(\gamma_0^- \theta_j^{\mathcal{F}\infty}), \gamma_0^- \theta_i^{\mathcal{F}\infty}\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \gamma_0^- \theta_i^{\mathcal{F}\infty}\right)_{\Gamma_\infty} & 0 \\ \hline 0 & \left(\left(D - \frac{1}{2} I\right)(\gamma_0^- \theta_j^{\mathcal{F}\infty}), \psi_i\right)_{\Gamma_\infty} & -(S(\psi_j), \psi_i)_{\Gamma_\infty} & i\overline{\eta}(\xi_j, \psi_i)_{\Gamma_\infty} \\ \hline 0 & \left(N(\gamma_0^- \theta_j^{\mathcal{F}\infty}), \xi_i\right)_{\Gamma_\infty} & \left(\left(\tilde{D} - \frac{1}{2} I\right)(\psi_j), \xi_i\right)_{\Gamma_\infty} & -\delta_{\Gamma_\infty}(\xi_j, \xi_i)_{\Gamma_\infty} \end{array}\right). \tag{3.59}$$

The blocks of the matrix in (3.59) are denoted

$$A^{\mathrm{stab}} = \left(\begin{array}{c|c|c|c} A^{\mathrm{stab}}_{1,1} & A^{\mathrm{stab}}_{1,2} & 0 & 0 \\ \hline A^{\mathrm{stab}}_{2,1} & A^{\mathrm{stab}}_{2,2} & A^{\mathrm{stab}}_{2,3} & 0 \\ \hline 0 & A^{\mathrm{stab}}_{3,2} & A^{\mathrm{stab}}_{3,3} & A^{\mathrm{stab}}_{3,4} \\ \hline 0 & A^{\mathrm{stab}}_{4,2} & A^{\mathrm{stab}}_{4,3} & A^{\mathrm{stab}}_{4,4} \end{array}\right). \tag{3.60}$$

All the blocks are complex-valued. The blocks $A^{\mathrm{stab}}_{1,1}$, $A^{\mathrm{stab}}_{1,2}$, $A^{\mathrm{stab}}_{1,3}$, $A^{\mathrm{stab}}_{2,1}$, $A^{\mathrm{stab}}_{2,2}$, $A^{\mathrm{stab}}_{2,3}$, $A^{\mathrm{stab}}_{3,2}$ and $A^{\mathrm{stab}}_{3,3}$ are the same as their corresponding counterparts in matrix (3.55). The blocks $A^{\mathrm{stab}}_{3,3}$ and $A^{\mathrm{stab}}_{4,4}$ are sparse, whereas the blocks $A^{\mathrm{stab}}_{4,2}$ and $A^{\mathrm{stab}}_{4,3}$ are dense. The block $A^{\mathrm{stab}}_{4,4}$ is symmetric.

3.4.3 Inf-sup stability of the discretized formulations

From the Fredholm setting and the approximation properties (3.49) and (3.50), the following discrete inf-sup conditions can be derived following [55, Theorem 14].

Proposition 3.13 *If $-\hat{k}_\infty^2 \notin \Lambda$ and $h_{\mathcal{M}}$ is small enough, there holds, for all $(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}) \in \mathcal{H}_{\mathcal{M}}$,*

$$\sup_{(0,0)\neq\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}}\right)\in\mathcal{H}_{\mathcal{M}}} \frac{\left|a^{\mathrm{unstab}}\left((\Phi_{\mathcal{M}},\lambda_{\mathcal{M}}),(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}})\right)\right|}{\|(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}})\|_{\mathcal{H}}} \gtrsim \|(\Phi_{\mathcal{M}},\lambda_{\mathcal{M}})\|_{\mathcal{H}}. \tag{3.61}$$

At all frequencies and if $h_{\mathcal{M}}$ is small enough, there holds, for all $(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}, p_{\mathcal{M}}) \in \mathbb{H}_{\mathcal{M}}$,

$$\sup_{(0,0,0)\neq\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}},p^t_{\mathcal{M}}\right)\in\mathbb{H}_{\mathcal{M}}} \frac{\left|a^{\mathrm{stab}}\left((\Phi_{\mathcal{M}},\lambda_{\mathcal{M}},p_{\mathcal{M}}),(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}},p^t_{\mathcal{M}})\right)\right|}{\|(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}},p^t_{\mathcal{M}})\|_{\mathbb{H}}} \gtrsim \|(\Phi_{\mathcal{M}},\lambda_{\mathcal{M}},p_{\mathcal{M}})\|_{\mathbb{H}}. \tag{3.62}$$

3.4.4 Convergence

From the inf-sup stability of the discrete problems, the following error estimates easily follow from [55, Theorem 13].

Proposition 3.14 *If $-\hat{k}_\infty^2 \notin \Lambda$ and $h_{\mathcal{M}}$ is small enough, the discrete problem* (3.46) *has a unique solution $(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}) \in \mathcal{H}_{\mathcal{M}}$, and the following optimal error estimate holds:*

$$\|(\Phi,\lambda)-(\Phi_{\mathcal{M}},\lambda_{\mathcal{M}})\|_{\mathcal{H}} \lesssim \inf_{\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}}\right)\in\mathcal{H}_{\mathcal{M}}} \|(\Phi,\lambda)-\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}}\right)\|_{\mathcal{H}}, \tag{3.63}$$

where (Φ, λ) is the unique solution of (3.36)*. At all frequencies and if $h_{\mathcal{M}}$ is small enough, the discrete problem* (3.47) *has a unique solution $(\Phi_{\mathcal{M}}, \lambda_{\mathcal{M}}, p_{\mathcal{M}}) \in \mathbb{H}_{\mathcal{M}}$, and the following optimal error estimate holds:*

$$\|(\Phi,\lambda,p)-(\Phi_{\mathcal{M}},\lambda_{\mathcal{M}},p_{\mathcal{M}})\|_{\mathbb{H}} \lesssim \inf_{\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}},p^t_{\mathcal{M}}\right)\in\mathbb{H}_{\mathcal{M}}} \|(\Phi,\lambda,p)-\left(\Phi^t_{\mathcal{M}},\lambda^t_{\mathcal{M}},p^t_{\mathcal{M}}\right)\|_{\mathbb{H}}, \tag{3.64}$$

where (Φ, λ, p) is the unique solution of (3.45)*.*

Remark 3.15 *The constant in* (3.63) *depends on $\hat{k}_\infty$, and its value explodes as $-\hat{k}_\infty^2$ tends to an element of Λ. The constant in* (3.64) *depends on $\hat{k}_\infty$ as well, but remains bounded on any bounded set of frequencies.*

3.4.5 Numerical resolution

Both the unstable and stable formulations have been implemented in the EADS in-house boundary element software called ACTIPOLE [33, 34]. This software can treat general three-dimensional geometries. The iterative solver is a block-GMRES [63, 82, 92] with no restart, suitable for non-symmetric linear systems. The "block" means that the solver treats all the right-hand sides simultaneously. The restart is an option that enables to save memory, but generally degrades the convergence. Since the solver stores on hard drive created during the iterations on hard drive, the memory usage is not an issue. In the iterative solver, the specificity of each block is taken into account. Matrix-vector products involving sparse blocks are optimized accordingly, and matrix-vector products involving boundary integral terms can be accelerated using a fast multipole method and out-of-core parallelization techniques. The preconditioner uses a combination of a sparse approximate inverse (SPAI) preconditioner [26, 27] and the sparse direct solver MUMPS [2]. More precisely, for each of the dense diagonal blocks $A_{2,2}^{\text{unstab}}$, $A_{3,3}^{\text{unstab}}$ and $A_{2,2}^{\text{stab}}$, $A_{3,3}^{\text{stab}}$, the SPAI preconditioner searches for an approximation of the inverse of these blocks. Consider any these blocks, denoted A in the following. A is made sparse by keeping, in each column, the interaction terms between the corresponding basis function and the ones in its vicinity (in the sense that the corresponding vertices or faces are nearby). The result of this operation is denoted by A^{sp}, and define $\mathcal{S}_{A^{\text{sp}}} = \{M \in \mathbb{C}^{n,n} \mid \forall\ 1 \leq i,j \leq n \text{ such that } A_{i,j}^{\text{sp}} = 0, M_{i,j} = 0\}$. The SPAI preconditioner of A is given by $P := \underset{\mathcal{S}_{A^{\text{sp}}}}{\text{argmin}} \|A^{\text{sp}}M - I\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm, and n the number of rows of A. Notice that for the blocks $A_{2,2}^{\text{unstab}}$ and $A_{2,2}^{\text{stab}}$, the SPAI preconditioner is computed ignoring the volumic contributions. For the sparse diagonal blocks, the preconditioner is taken as the inverse of each blocks. The inverse is not actually computed: since MUMPS provides a factorization of each of these blocks, each time a product preconditioner-vector is needed when constructing the Krylov vectors of the iterative method, two triangular systems are efficiently solved using this factorization. The preconditioner for the whole system is block diagonal, each bloc being a SPAI or MUMPS preconditioner.

3.5 Numerical results

The purpose of this section is the comparison between the unstable formulation (3.36) and the stable formulation (3.45) with the coupling parameter $\eta = 1$. Both formulations have been implemented in the EADS in-house boundary element software called ACTIPOLE [33, 34]. This software can treat general three-dimensional geometries. The iterative solver is a block generalized conjugate residual method [82, 63] based on a generalized minimal residual method [92], suitable for non-symmetric systems. A sparse approximate inverse preconditioner [26, 27] is used.

Consider an ellipsoid with major axis directed along the z-axis. This object is included inside a larger ball. The external border of the ball after discretization is the surface Γ_∞. A potential flow is computed around the ellipsoid and inside the ball, such that the flow is uniform outside the ball, of Mach number 0.3 and directed along the z-axis. An acoustic monopole source lies upstream of the ball, on the z-axis as well. Four different meshes are considered, see Table 3.1. For accuracy reasons, a rule of thumb in boundary elements method for the

classical Helmholtz equation consists in imposing that the mean edge is at least eight to ten times smaller than the wavelength of the source. In our software, we first generate the mesh and then apply the Prandtl–Glauert transformation. Therefore, in this test case, the mesh is at most extended by a factor $\gamma_\infty \approx 1.048$. Moreover, the integral operators are computed at the transformed wavenumber $\hat{k}_\infty = \gamma_\infty k_\infty$, resulting on a wavelength of approximately 0.21 m. We then verify that, for Mesh 1, the mean edge of the transformed mesh is eight times smaller than the wavelength. As a consequence, the three other meshes do not satisfy the rule of thumb, but will be used as comparison supports and in numerical experiments requiring a large number of resolutions. In this test case, the extension of the edges of the mesh and the modification of the wavenumber induced by the Prandtl–Glauert transformation are mild, but can become very large as M_∞ is close to 1. The first effect can be canceled by computing the mesh on the geometry already modified by the Prandtl–Glauert transformation.

From Table 3.1, for fine meshes, the number of basis functions used to discretize the unknown p for the variational formulation (3.47) takes a smaller part in the total number of basis functions than for coarse meshes. Therefore, the relative complexity added to (3.36) by the third equation of (3.45) decreases with the total number of unknowns, which is an interesting property when it comes to industrial test cases. Figure 3.2 displays Mesh 1 and the rescaled velocity $\boldsymbol{M}_0$ of the potential flow.

In what follows, a frequency $\mathtt{f}$ is called resonant if $-\hat{k}_\infty^2 = -\frac{4\pi^2\mathtt{f}^2}{\gamma_\infty^2} \in \Lambda$, where Λ is the set of Dirichlet eigenvalues for the Laplacian on $\mathbb{R}^3\backslash\overline{\Omega^+}$. The set Λ depends on the geometric shape of the coupling surface Γ_∞, which slightly changes after each discretization.

	Mesh 1	Mesh 2	Mesh 3	Mesh 4
number of volumic dofs Φ	1796	687	194	79
number of surfacic $\mathbb{P}_0$ dofs λ	808	510	270	148
number of surfacic $\mathbb{P}_1$ dofs p	406	257	137	76
proportion of dofs p in the total number of dofs	11.9%	15.0%	18.6%	20.0%
smallest edge (mm)	7.09	8.78	15.71	19.18
mean edge (mm)	22.64	32.20	49.78	66.46
largest edge (mm)	56.87	70.62	103.59	112.71

Table 3.1. Characteristics of the four considered meshes.

3.5.1 Comparison of pressure fields

As seen in Theorem 3.7, the unstable formulation (3.36) is not well-posed at resonant frequencies. First, a prospective study to identify a resonant frequency for each of the four meshes is carried out by monitoring the condition number of the matrices produced by the discretized version of the unstable formulation (3.36). A resonant frequency for Mesh 1, Mesh 2, Mesh 3, and Mesh 4 is identified around 1509.849 Hz, 1513.431 Hz, 1521.015 Hz, and 1535.704 Hz respectively.

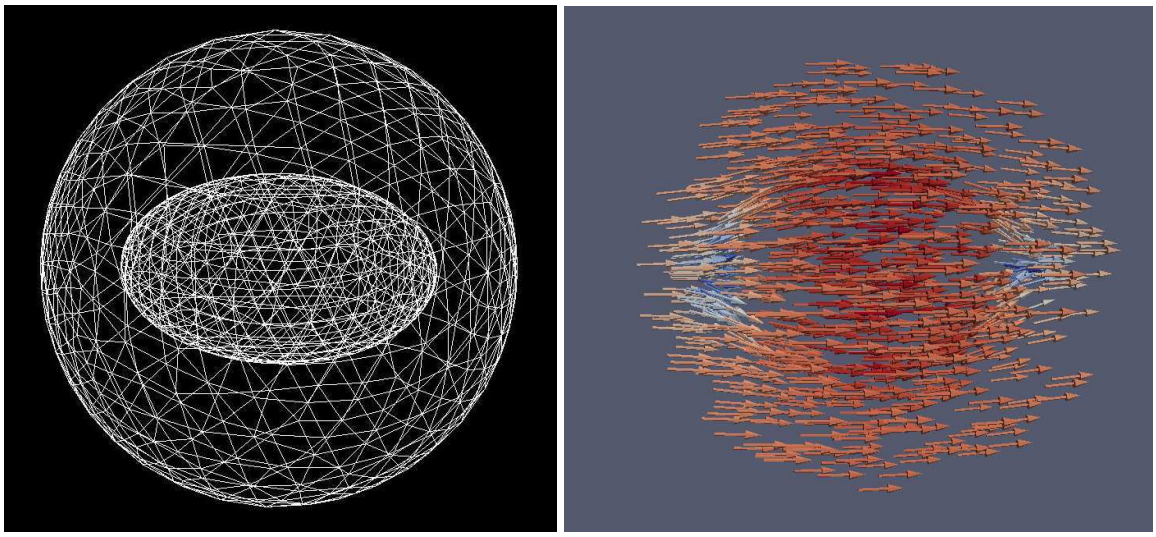

Fig. 3.2. Left: representation of Mesh 1, Right: potential flow around the ellipsoid.

The convergence of the iterative solver is monitored by requiring that the Euclidian norm of the relative residual is smaller than 10^{-6}. Additional tests indicate that the discretized solution to the stable formulation does not change much below this value of the relative residual. For Mesh 1, away from a resonance, say at 1500 Hz, the scattered pressure fields computed with the unstable and stable formulations are very similar. This holds as well for the total pressure fields, see Figure 3.3. At the resonant frequency 1509.849 Hz, the unstable formulation (3.36) yields pressure maps quite different from the ones at 1500 Hz, whereas the stable formulation (3.45) yields pressure maps very similar to the ones at 1500 Hz, see Figure 3.4. From Proposition 3.5, at a resonant frequency, the solutions to (3.36) that differ from the solution to (3.45) should not affect the scattered pressure field in the exterior domain. Actually, the distortion of the scattered field with the unstable formulation (3.36) is the result of the significant magnification of discretization and numerical errors by the ill-conditioning of the linear system approximating (3.36).

3.5.2 Auxiliary variable p

In Figure 3.5, the left plot indicates that with Mesh 1, the magnitude of p is around 0.5% of the scattered pressure. The right plot shows the behaviour of the magnitude of p (measured as $\|p\|_{L^\infty(\Gamma_\infty)}$) with respect to the stopping criterion of the iterative solver for the four meshes. The finer the mesh, the smaller the auxiliary variable p, which is consistent with the fact that the p-component of the solution to (3.45) vanishes (see Proposition 3.9).

3.5.3 Comparison of condition numbers

Figure 3.6 presents the condition numbers of the matrices resulting from the formulations (3.36) and (3.45) with respect to the frequency. In the left plot, the curves are centered at the resonant frequencies. The finer the mesh, the higher the condition number explodes. The width

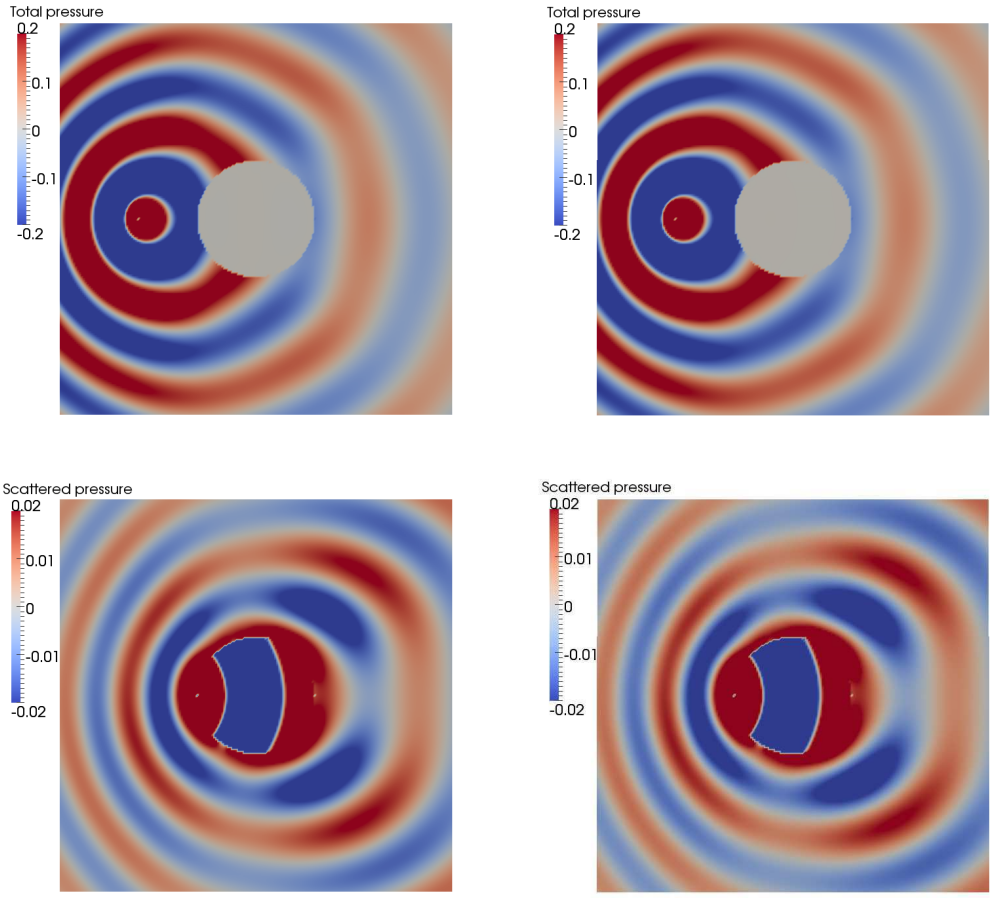

Fig. 3.3. Mesh 1, 1500 Hz. Top : real part of the total pressure; left: unstable formulation (3.36), right: stable formulation (3.45). Bottom : real part of the scattered pressure; left: unstable formulation (3.36), right: stable formulation (3.45). At this non-resonant frequency, both formulations yield similar results.

of the peak at the resonance does not appear to depend on the mesh. In the right plot, a larger bandwidth is considered with Mesh 2. Owing to the frequency sampling (every 5 Hz), some resonances may be missed and the local maxima may not be accurately reached (in particular, from the left plot, the local maximum of 7.2 for $\log(\mathrm{cond}(M))$ at 1513.431 Hz is very underestimated). The stable formulation (3.45) produces somewhat larger condition numbers for the large majority of the frequencies, but, unlike the unstable formulation (3.36), it presents no resonance. Moreover, from the Weyl formula, the number of resonant frequencies smaller than $\mathtt{f}$ increases as $\mathtt{f}^{\frac{3}{2}}$, making the need for a stable formulation even more important for simulations at higher frequencies.

3.5.4 Convergence

To further study the impact of the ill-conditioning of the unstable formulation (3.36) on the computed solution, the preconditioning is not used in what follows. First, the value of the

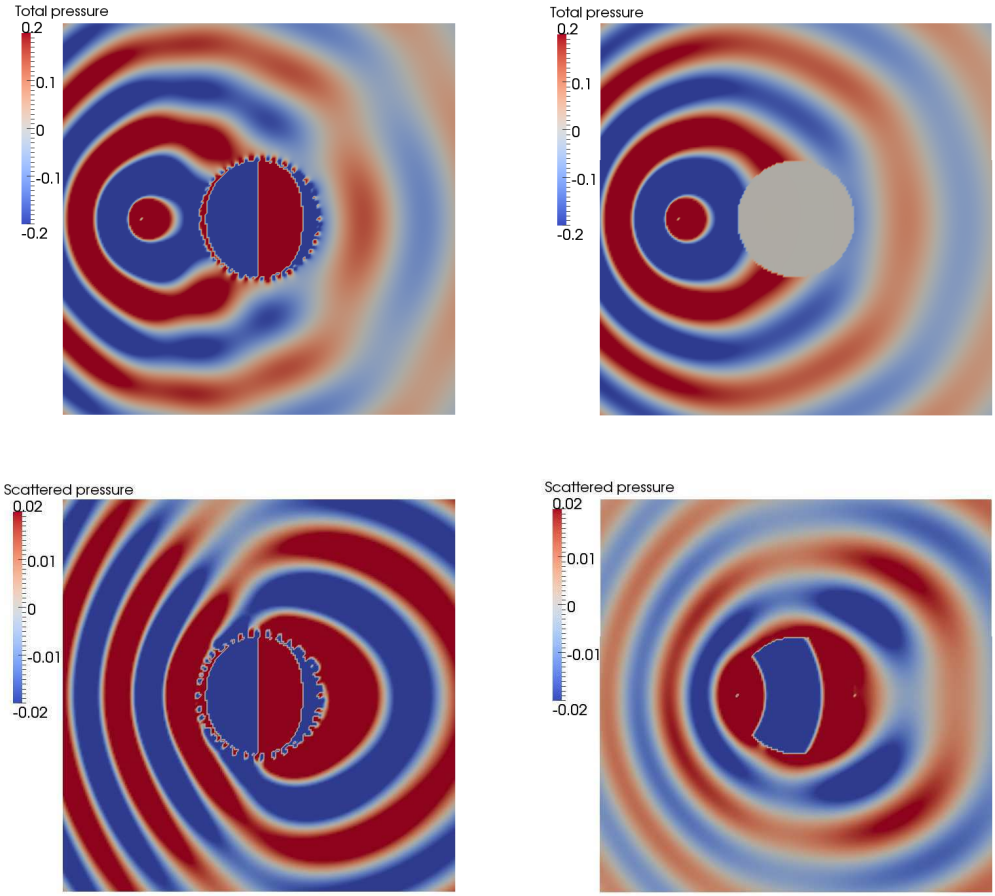

Fig. 3.4. Mesh 1, 1509.849 Hz. Top : real part of the total pressure; left: unstable formulation (3.36), right: stable formulation (3.45). Bottom : real part of the scattered pressure; left: unstable formulation (3.36), right: stable formulation (3.45). At this resonant frequency, the two formulations yield different results.

acoustic pressure on a network of 10000 points located further than 0.5 m from the center of the sphere (therefore in Ω^+) is computed using the stable formulation (3.45) with Mesh 1 at the resonant frequency 1509.849 Hz. This computed acoustic pressure is called the accurate pressure. Next, the acoustic pressure on the same network of points is computed for different values of the number of iterations of the solver, using the unstable formulation (3.36) and the stable formulation (3.45) with Mesh 1 at the resonance 1509.849 Hz. The relative difference between the computed pressure and the accurate pressure in Euclidian norm is called the relative error. Figure 3.7 presents the relative residual and the relative error with respect to the number of iterations. With the unstable formulation (3.36), the relative residual decreases irregularly. In particular, it stays constant during around 200 iterations. The relative error decreases, stays constant, rises after 400 iterations, and finally stabilizes at a large value, whereas the relative residual keeps converging to zero. As for every ill-conditioned problem, the relative residual cannot be used to ascertain convergence towards the correct solution. In particular, after 600 iterations, the relative residual is extremely small, while the error is of order one. With the

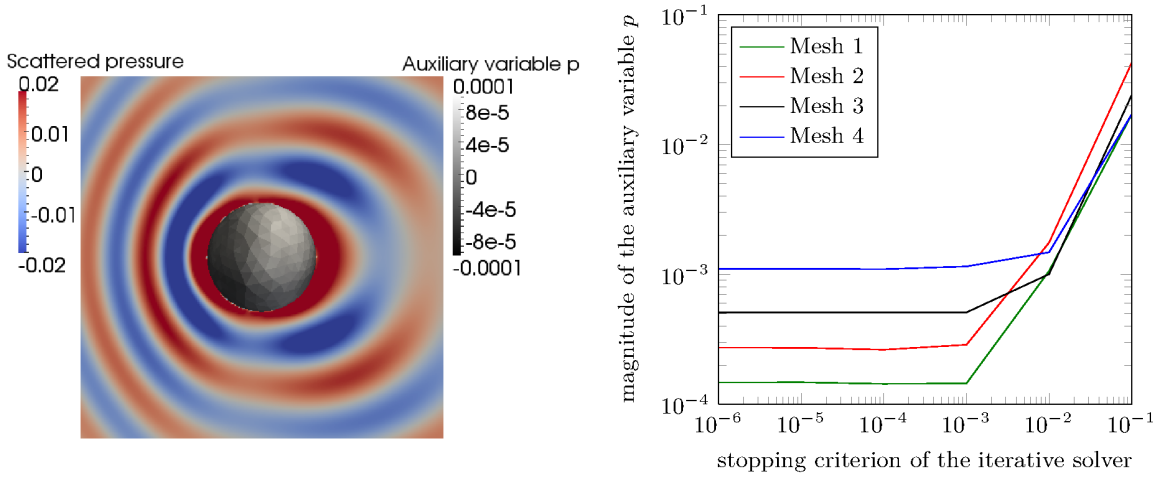

Fig. 3.5. Stable formulation (3.45) at 1500 Hz. Left : real part of the scattered pressure and auxiliary variable p with Mesh 1. Right : Magnitude of the auxiliary variable p as a function of the stopping criterion of the iterative linear solver with all meshes.

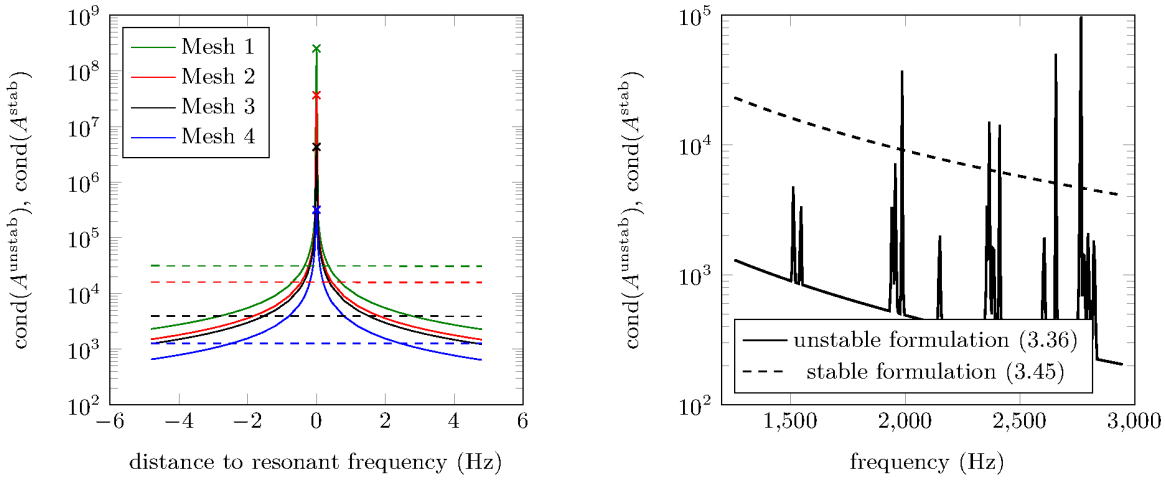

Fig. 3.6. Condition number of the matrix for the unstable formulation (3.36) (solid) and the stable formulation (3.45) (dashed). Left: centered representation around a resonant frequency for the four meshes. Right: larger bandwidth with Mesh 2.

stable formulation (3.45), the relative residual and the relative error decrease regularly, and in the same fashion.

3.5.5 Choice of the coupling parameter η

In the stable formulation (3.45), the choice of the coupling parameter η is expected to have a direct effect of the condition number of the matrix A^{stab}. In Figure 3.8, this condition number is plotted for Mesh 4 and for various values of η. For $\eta = 0$, equations (3.45a)-(3.45b)

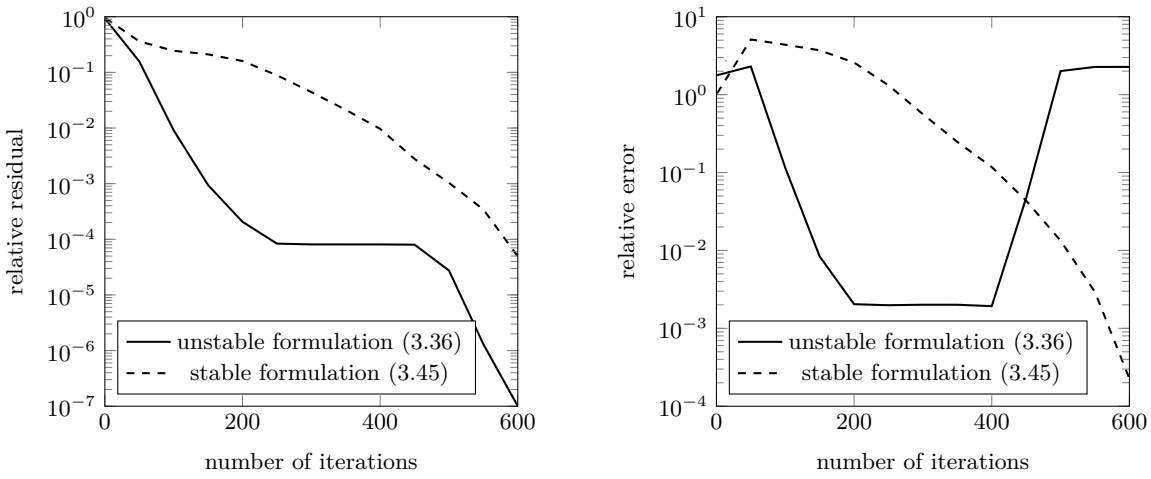

Fig. 3.7. Mesh 1 at resonance 1509.849 Hz; relative residual (left) and relative error (right) with respect to the number of iterations.

are decoupled from (3.45c), and (3.45a)-(3.45b) become (3.36), so that the curve for $\eta = 0.001$ is similar to the curve of the unstable formulation for Mesh 4 in Figure 3.6. The condition number appears to be the smallest for η in the range 1 to 10, and worsens for lower and higher values of η. This motivates the choice $\eta = 1$ made in the above simulations.

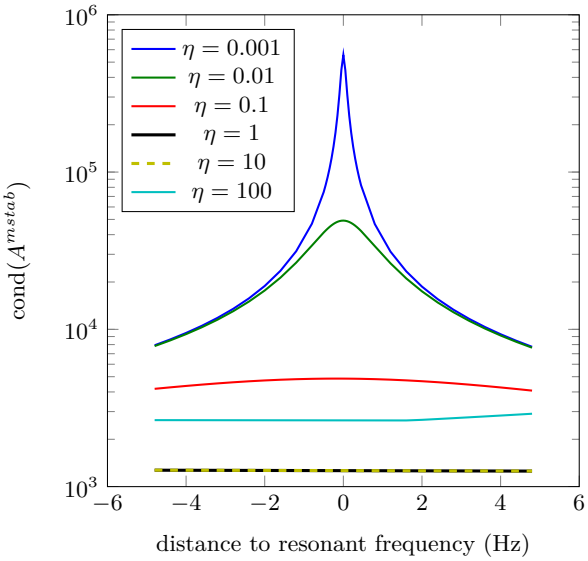

Fig. 3.8. Condition number of the matrix for the stable formulation (3.45) centered around the resonant frequency at 1535.704 Hz for Mesh 4. In this case, the chosen value $\eta = 1$ leads to the minimal condition numbers.

3.6 Conclusion

In this work, we derived two coupled boundary element / finite element methods for the convected Helmholtz equation with non-uniform flow in a bounded domain. The first one leads to an unstable formulation, while the second one leads to a stable formulation. The unstable formulation involves two equations and is well-posed except at some resonant frequencies of the source, while the stable one is unconditionally well-posed, but involves three equations. Even if the unstable formulation admits infinitely many solutions at resonant frequencies, the pressure field resulting from any of these solutions equals the one resulting from the stable formulation. However, our numerical results show that at resonant frequencies, the discretization of the unstable formulation is so ill-conditioned that the pressure field is very different from the one produced by the stable formulation. Moreover, the stable formulation remains tractable within large industrial problems since the relative complexity added by its third equation decreases with the size of the mesh. Its interest is also enhanced by the fact that, at higher frequencies, the density of resonant frequencies is more important.

As long as the uniform flow assumption in the exterior domain is reasonable, more complex flows in the interior domain can be considered, as well as more complex boundary conditions at the surface of the scattering object. These extensions only require to modify the finite element part of the present methodology. An important development for practical simulations is the introduction of modal sources in the interior domain, which is the purpose of [Pr2].

Another interesting extension of this work is the resolution of parametrized aeroacoustic problems, with the frequency of the source as a parameter, using reduced-order models, for instance by means of Proper Generalized Decomposition or Reduced Basis methods. Using the unstable formulation may involve ill-conditioned numerical resolutions if the frequency range of interest contains resonant frequencies, whereas the stable formulation guarantees well-posedness of the procedure. Moreover, the complexity of the online stage of the reduced-order model is not increased by the third equation of the stable formulation.

3.7 Annex: Proof of the mathematical results

As preliminary results, we derive Propositions 3.16 and 3.17. Then, recalling that (3.21) is the original transmission problem, (3.36) the unstable formulation, and (3.45) the stable formulation, the mathematical results are obtained in the following order:

$$\text{Proposition 3.17} \Longrightarrow \begin{cases} \text{Link between (3.21) and (3.45) (Proposition 3.9)} \\ \text{Link between (3.21) and (3.36) (Proposition 3.5)} \end{cases}$$

$$\Downarrow$$

$$\text{Proposition 3.16} \Longrightarrow \begin{cases} \text{Uniqueness for (3.45)} \Longrightarrow \text{Well-posedness of (3.45) (Theorem 3.10)} \\ \qquad\qquad \Longrightarrow \text{Well-posedness of (3.21) (Theorem 3.2)} \\ \text{Conditional uniqueness for (3.36)} \Longrightarrow \text{Conditional well-posedness of (3.36)} \\ \qquad\qquad \text{(Theorem 3.7)} \end{cases}$$

Notice that we do no start with the well-posedness of (3.21) since, to our knowledge, a result directly applicable to it is not available in the literature. Herein, we prove the well-posedness of (3.45) which is equivalent to (3.21).

Proposition 3.16 *Problem* (3.21) *has at most one solution in* $H^1_{\rm loc}(\Omega)$.

Proof. Let $f \in H^1_{\rm loc}(\Omega)$ solve (3.21) with $\varsigma = 0$ (so that $f_{\rm inc} = 0$). From Proposition 3.1, f solves (3.18). Let B be an open ball containing Ω^-. Let $f^t \in H(\Delta, B)$. Using Green's first identity,

$$\begin{aligned} 0 &= \int_{\Omega\cap B} \left(-rk^2\beta\overline{f} - ikr\boldsymbol{V}\cdot\boldsymbol{\nabla}\overline{f} - \boldsymbol{\nabla}\cdot(irk\overline{f}\boldsymbol{V} + (r\Xi\boldsymbol{\nabla}\overline{f})\right) f^t \\ &= \int_{\Omega\cap B} r\Xi\boldsymbol{\nabla}\overline{f}\cdot\boldsymbol{\nabla}f^t - rk^2\beta\overline{f}f^t - ikr\boldsymbol{V}\cdot(\boldsymbol{\nabla}\overline{f}f^t - \boldsymbol{\nabla}f^t\overline{f}) - \left(\gamma^-_{1,\partial B}f, \gamma^-_{0,\partial B}f^t\right)_{\partial B}, \end{aligned} \tag{3.65}$$

where $\gamma^-_{0,\partial B}$ and $\gamma^-_{1,\partial B}$ are the Dirichlet and Neumann traces on ∂B from B. Taking $f^t = f$ yields

$$\left(\gamma^-_{1,\partial B}f, \gamma^-_{0,\partial B}f\right)_{\partial B} = \int_{\Omega\cap B} r\Xi\boldsymbol{\nabla}\overline{f}\cdot\boldsymbol{\nabla}f - rk^2\beta\overline{f}f - 2kr\boldsymbol{V}\cdot\left(\mathrm{Im}\boldsymbol{\nabla}\overline{f}f\right), \tag{3.66}$$

so that $\mathrm{Im}\left(\gamma^-_{1,\partial B}f, \gamma^-_{0,\partial B}f\right)_{\partial B} = 0$. Using Rellich Lemma (see [76, Lemma 9.9]), since $f \in H^1_{\rm loc}(\mathbb{R}^3\backslash\overline{B})$ solves the classical Helmholtz equation in $\mathbb{R}^3\backslash\overline{B}$ and satisfies the Sommerfeld radiation condition, as well as $\mathrm{Im}\left(\gamma^-_{1,\partial B}f, \gamma^-_{0,\partial B}f\right)_{\partial B} \geq 0$, it is inferred that $f|_{\mathbb{R}^3\backslash\overline{B}} \equiv 0$. Equation (3.18a) with $\varsigma = 0$ can be written

$$L(f) := \left(rk^2\beta + \boldsymbol{\nabla}\cdot(irk\boldsymbol{V})\right) f + 2irk\boldsymbol{V}\cdot\boldsymbol{\nabla}f + \boldsymbol{\nabla}\cdot(r\Xi\boldsymbol{\nabla}f) = 0 \quad \text{in } \Omega. \tag{3.67}$$

From [45, Theorem 1.1], since $r\Xi$ is uniformly elliptic with Lipschitz continuous coefficients, and $rk^2\beta + \boldsymbol{\nabla}\cdot(irk\boldsymbol{V})$ and $2irk\boldsymbol{V}$ have bounded coefficients, the differential operator L satisfies the strong unique continuation property in Ω. Hence, $f|_{\mathbb{R}^3\backslash\overline{B}} \equiv 0$ implies that the only $H^1_{\rm loc}(\Omega)$ solution of (3.18) with $\varsigma = 0$ is $f \equiv 0$ in Ω. The assertion follows from Proposition 3.1. ♢

Proposition 3.17 *Let* (Φ, λ, p) *solve* (3.45). *There holds* $p = 0$, $\lambda = \gamma_1^-\Phi$, *and*

$$rk^2\beta\Phi^- + irk\boldsymbol{V}\cdot\boldsymbol{\nabla}\Phi^- + \boldsymbol{\nabla}\cdot(irk\Phi^-\boldsymbol{V} + r\Xi\boldsymbol{\nabla}\Phi^-) = 0 \qquad \text{in } L^2(\Omega^-), \tag{3.68a}$$

$$\gamma^-_{n,\Gamma}\left(irk\Phi\boldsymbol{V} + r\Xi\boldsymbol{\nabla}\Phi\right) = 0 \qquad \text{in } H^{-\frac{1}{2}}(\Gamma), \tag{3.68b}$$

$$N(\gamma_0^-\Phi) + \left(\tilde{D} + \frac{1}{2}I\right)(\lambda) = \gamma_1 f_{\rm inc} \qquad \text{in } H^{-\frac{1}{2}}(\Gamma_\infty), \tag{3.68c}$$

$$\left(D - \frac{1}{2}I\right)(\gamma_0^-\Phi) - S(\lambda) = -\gamma_0 f_{\rm inc} \qquad \text{in } H^{\frac{1}{2}}(\Gamma_\infty). \tag{3.68d}$$

Proof. Set $\boldsymbol{\sigma} := irk\Phi\boldsymbol{V} + r\Xi\boldsymbol{\nabla}\Phi$. Owing to (3.45a) and recalling the definition (3.30) of the sesquilinear form $\mathcal{V}$, there holds, $\forall\Phi^t \in H^1(\Omega^-)$,

$$\int_{\Omega^-}\overline{\boldsymbol{\sigma}}\cdot\boldsymbol{\nabla}\Phi^t = \int_{\Omega^-} rk^2\beta\overline{\Phi}\Phi^t + i\int_{\Omega^-} rk\boldsymbol{V}\cdot\boldsymbol{\nabla}\overline{\Phi}\Phi^t - \left(N(\gamma_0^-\Phi) + \left(\tilde{D} - \frac{1}{2}I\right)(\lambda) - \gamma_1 f_{\rm inc}, \gamma_0^-\Phi^t\right)_{\Gamma_\infty}. \tag{3.69}$$

Restricting the test function Φ^t to $C_c^\infty(\Omega^-)$ shows that (3.68a) holds in $L^2(\Omega^-)$. Moreover, since

$$\int_{\Omega^-} \overline{\boldsymbol{\sigma}} \cdot \boldsymbol{\nabla} \Phi^t = -\int_{\Omega^-} (\boldsymbol{\nabla} \cdot \overline{\boldsymbol{\sigma}})\, \Phi^t + \left(\gamma_n^- \boldsymbol{\sigma}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\gamma_{n,\Gamma}^- \boldsymbol{\sigma}, \gamma_{0,\Gamma}^- \Phi^t\right)_{\Gamma}, \tag{3.70}$$

and recalling that $\gamma_n^- \boldsymbol{\sigma} = \gamma_1^- \Phi$ on Γ_∞, owing to the property of the convective flow at Γ_∞,

$$\left(N(\gamma_0^- \Phi) + \left(\tilde{D} - \frac{1}{2} I\right)(\lambda) - \gamma_1 f_{\text{inc}}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\gamma_1^- \Phi, \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\gamma_{n,\Gamma}^- \boldsymbol{\sigma}, \gamma_{0,\Gamma}^- \Phi^t\right)_{\Gamma} = 0. \tag{3.71}$$

Then, by the surjectivity of the trace operators γ_0^- from $H^1(\Omega^-)$ onto $H^{\frac{1}{2}}(\Gamma_\infty)$ and $\gamma_{0,\Gamma}^-$ from $H^1(\Omega^-)$ onto $H^{\frac{1}{2}}(\Gamma)$ (see [93, Theorem 2.6.11]), it is deduced from (3.71) that (3.68b) holds, and from (3.71) and (3.45b) respectively that

$$N(\gamma_0^- \Phi) + \left(\tilde{D} - \frac{1}{2} I\right)(\lambda) + \gamma_1^- \Phi = \gamma_1 f_{\text{inc}}, \tag{3.72a}$$

$$\left(D - \frac{1}{2} I\right)(\gamma_0^- \Phi) - S(\lambda) - i\eta p = -\gamma_0 f_{\text{inc}}. \tag{3.72b}$$

Owing to (3.45c) and (3.72a), it is inferred that $\delta_{\Gamma_\infty}(p, p^t) = (\lambda - \gamma_1^- \Phi, p^t)_{\Gamma_\infty}$ for all $p^t \in H^1(\Omega^-)$. From the definition (3.38) of M, there holds $p = M(\lambda - \gamma_1^- \Phi)$. Let $x := \lambda - \gamma_1^- \Phi$. Then, from (3.26), (3.72) can be written

$$\begin{pmatrix} \frac{1}{2} I - D & S \\ N & \frac{1}{2} I + \tilde{D} \end{pmatrix} \begin{pmatrix} \gamma_0^- \Phi \\ \lambda \end{pmatrix} = \begin{pmatrix} -i\eta M x + \gamma_0 f_{\text{inc}} \\ x + \gamma_1 f_{\text{inc}} \end{pmatrix}, \tag{3.73}$$

so that $(-i\eta M x + \gamma_0 f_{\text{inc}}, x + \gamma_1 f_{\text{inc}})$ belongs to the range of the block operator defined on the left-hand side of (3.73). Under this condition, from [54, Theorem 4.1], citing [106], a radiating piecewise Helmholtz solution u such that $\gamma_0^- u = -i\eta M x + \gamma_0 f_{\text{inc}}$ and $\gamma_1^- u = x + \gamma_1 f_{\text{inc}}$ can be constructed. Consider v such that $v|_{\Omega^+} = 0$ and $v|_{\mathbb{R}^3 \setminus \overline{\Omega^+}} = u|_{\Omega^-}$, and w such that $w|_{\Omega^+} = 0$ and $w|_{\mathbb{R}^3 \setminus \overline{\Omega^+}} = f_{\text{inc}}$. Since v and w are radiating piecewise Helmholtz solutions, $\hat{u} := v - w$ is also a radiating piecewise Helmholtz solution. Since $[\gamma_0 \hat{u}] = i\eta M x$ and $[\gamma_1 \hat{u}] = -x$, (3.26) implies

$$\begin{pmatrix} \frac{1}{2} I - D & S \\ N & \frac{1}{2} I + \tilde{D} \end{pmatrix} \begin{pmatrix} i\eta M x \\ -x \end{pmatrix} = \begin{pmatrix} i\eta M x \\ -x \end{pmatrix}. \tag{3.74}$$

Consider now $\tilde{u} := \mathcal{S}(x) + \mathcal{D}(i\eta M x)$. From [58, p. 113], the single-layer and double-layer potentials are radiating piecewise Helmholtz solutions. In particular, $\tilde{u}|_{\mathbb{R}^3 \setminus \overline{\Omega^+}}$ solves the Helmholtz equation in $\mathbb{R}^3 \setminus \overline{\Omega^+}$, therefore

$$\left(\gamma_1^- \tilde{u}^-, \gamma_0^- \tilde{u}^-\right)_{\Gamma_\infty} = \int_{\mathbb{R}^3 \setminus \overline{\Omega^+}} \{|\boldsymbol{\nabla} \tilde{u}|^2 - \hat{k}_\infty^2 |\tilde{u}|^2\} \in \mathbb{R}. \tag{3.75}$$

From the trace relations (3.25), there holds $\gamma_0^- \tilde{u}^- = \gamma_0^-(\mathcal{S}(x) + \mathcal{D}(i\eta M x)) = Sx + i\eta \left(D - \frac{1}{2}\right) M x$. Then, using the first line of (3.74), $\gamma_0^- \tilde{u}^- = -i\eta M x$. Likewise, $\gamma_1^- \tilde{u}^- = x$. From (3.75), $-i\eta\, (x, Mx)_{\Gamma_\infty} \in \mathbb{R}$. However, since $(x, Mx)_{\Gamma_\infty} = \|Mx\|^2_{H^1(\Gamma_\infty)} \in \mathbb{R}$ and $\text{Re}(\eta) \neq 0$, there holds

$(x, Mx)_{\Gamma_\infty} = 0$. Therefore $x = 0$, and $p = Mx = 0$, leading to $\lambda = \gamma_1^- \Phi$. Finally, (3.68c)-(3.68d) are directly obtained from (3.72a)-(3.72b) using $\lambda = \gamma_1^- \Phi$ and $p = 0$. ♢

Proof. (of Proposition 3.9) The first part of the proposition is clear from the results of Section 3.3. Let (Φ, λ, p) solve (3.45). Since the single-layer and double-layer potentials are radiating piecewise Helmholtz solutions, $\mathcal{R}(\Phi, \lambda)$ satisfies (3.21b) and (3.21f). From Proposition 3.17, $p = 0$, $\lambda = \gamma_1^- \Phi$, and (3.68) holds. (3.21a) and (3.21c) are just (3.68a) and (3.68b). Then, using the definition of $\mathcal{R}$ given in Proposition 3.5 and the trace identities (3.25), the exterior Dirichlet trace of $\mathcal{R}(\Phi, \lambda)$ is $\gamma_0^+ \mathcal{R}(\Phi, \lambda) = \gamma_0^+(-\mathcal{S}(\lambda) + \mathcal{D}(\gamma_0^- \Phi) + f_{\text{inc}}) = -S(\lambda) + (D + \frac{1}{2} I)(\gamma_0^- \Phi) + \gamma_0 f_{\text{inc}}$. Using (3.68d), there holds $\gamma_0^+ \mathcal{R}(\Phi, \lambda) = \gamma_0^- \Phi$, which is $\gamma_0^- \mathcal{R}(\Phi, \lambda)$, from which we infer that the first transmission condition (3.21d) holds. The second transmission condition (3.21e) is obtained in the same fashion from the exterior Neumann trace of $\mathcal{R}(\Phi, \lambda)$, (3.68c), and using the fact that $\lambda = \gamma_1^- \Phi$. ♢

Proposition 3.18 *Problem* (3.45) *has at most one solution.*

Proof. Let (Φ, λ, p) solve (3.45) with $\gamma_0 f_{\text{inc}} = 0$ and $\gamma_1 f_{\text{inc}} = 0$. From Proposition 3.17, $p = 0$ and $\lambda = \gamma_1^- \Phi$. From Proposition 3.9, $\mathcal{R}(\Phi, \lambda)$ solves (3.21). From the mapping properties of $\mathcal{S}$ and $\mathcal{D}$ (Section 3.3.2), $\mathcal{R}(\Phi, \lambda) \in H^1_{\text{loc}}(\Omega^+ \cup \Omega^-)$. Then, from the transmission conditions (3.21d) and (3.21e), $\mathcal{R}(\Phi, \lambda) \in H^1_{\text{loc}}(\Omega)$. Then, Proposition 3.16 implies that $\mathcal{R}(\Phi, \lambda) = 0$ in Ω. As a result, there holds $\Phi = \mathcal{R}(\Phi, \lambda)|_{\Omega^-} = 0$, and $\lambda = \gamma_1^- \Phi = 0$. ♢

Proof. (of Theorem 3.10) Consider the two sesquilinear forms a_1 and a_2 on $\mathbb{H} \times \mathbb{H}$ such that

$$
\begin{aligned}
a_1\left((\Phi, \lambda, p), \left(\Phi^t, \lambda^t, p^t\right)\right) &:= \int_{\Omega^-} r \Xi \boldsymbol{\nabla} \overline{\Phi} \cdot \boldsymbol{\nabla} \Phi^t + \left(N^0(\gamma_0^- \Phi), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(S^0(\lambda), \lambda^t\right)_{\Gamma_\infty} + \delta_{\Gamma_\infty}(p, p^t) \\
&+ \left(\left(\tilde{D}^0 - \frac{1}{2} I\right)(\lambda), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} - \left(\left(D^0 - \frac{1}{2} I\right)(\gamma_0^- \Phi), \lambda^t\right)_{\Gamma_\infty}, \\
a_2\left((\Phi, \lambda, p), \left(\Phi^t, \lambda^t, p^t\right)\right) &:= -\int_{\Omega^-} r k^2 \beta \overline{\Phi} \Phi^t + i \int_{\Omega^-} r k \boldsymbol{V} \cdot \left(\overline{\Phi} \boldsymbol{\nabla} \Phi^t - \Phi^t \boldsymbol{\nabla} \overline{\Phi}\right) + \left(\left(N - N^0\right)(\gamma_0^- \Phi), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} \\
&+ \left(\left(S - S^0, \lambda^t\right)(\lambda)\right)_{\Gamma_\infty} + \left(\left(\tilde{D} - \tilde{D}^0\right)(\lambda), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} - \left(\left(D - D^0, \lambda^t\right)(\gamma_0^- \Phi)\right)_{\Gamma_\infty} - i\overline{\eta}\left(p, \lambda^t\right)_{\Gamma_\infty} \\
&- \left(N(\gamma_0^- \Phi), p^t\right)_{\Gamma_\infty} - \left(\left(\tilde{D} + \frac{1}{2} I\right)(\lambda), p^t\right)_{\Gamma_\infty},
\end{aligned}
\tag{3.76}
$$

where S^0, D^0, $\tilde{D}^0$ and N^0 are the boundary integral operators S, D, $\tilde{D}$ and N for $\hat{k}_\infty = 0$. Consider the linear form b on $\mathbb{H}$ such that

$$
b\left(\Phi^t, \lambda^t, p^t\right) := \left(\gamma_1 f_{\text{inc}}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\gamma_0 f_{\text{inc}}, \lambda^t\right)_{\Gamma_\infty} - \left(\gamma_1 f_{\text{inc}}, p^t\right)_{\Gamma_\infty}. \tag{3.77}
$$

Problem (3.45) can then be written: Find $(\Phi, \lambda, p) \in \mathbb{H}$, such that $\forall \left(\Phi^t, \lambda^t, p^t\right) \in \mathbb{H}$,

$$
a\left((\Phi, \lambda, p), \left(\Phi^t, \lambda^t, p^t\right)\right) = b\left(\Phi^t, \lambda^t, p^t\right), \tag{3.78}
$$

where $a := a_1 + a_2$. We show the well-posedness of (3.78) by proving successively that (i) a_1 and a_2 are bounded on $\mathbb{H} \times \mathbb{H}$ and b is bounded on $\mathbb{H}$, (ii) a_1 is $\mathbb{H}$-coercive, (iii) the linear map

associated with a_2 is compact from $\mathbb{H}$ into $\mathbb{H}$. With the uniqueness of the solution (Proposition 3.18), the assertion follows from the Fredholm alternative.

(i) From Section 3.2.4, $\Xi\boldsymbol{\nabla}\overline{\Phi}\cdot\boldsymbol{\nabla}\Phi^t \leq \frac{1+M_0^2}{1-M_\infty^2}\|\boldsymbol{\nabla}\overline{\Phi}\|\|\boldsymbol{\nabla}\Phi^t\|$. Then,

$$\left|\int_{\Omega^-} r\Xi\boldsymbol{\nabla}\overline{\Phi}\cdot\boldsymbol{\nabla}\Phi^t\right| \leq \frac{1}{1-M_\infty^2}\|r\|_{L^\infty(\Omega^-)}\|1+M_0^2\|_{L^\infty(\Omega^-)}\|\Phi\|_{H^1(\Omega^-)}||\Phi^t||_{H^1(\Omega^-)}. \tag{3.79}$$

The other volumic integrals are simply controlled by

$$\begin{aligned}\left|\int_{\Omega^-} rk\boldsymbol{V}\cdot\left(\overline{\Phi}\boldsymbol{\nabla}\Phi^t-\Phi^t\boldsymbol{\nabla}\overline{\Phi}\right)\right| &\leq 2\|rk\|_{L^\infty(\Omega^-)}\|\boldsymbol{V}\|_{L^\infty(\Omega^-)^3}\|\Phi\|_{H^1(\Omega^-)}\|\Phi^t\|_{H^1(\Omega^-)},\\ \left|\int_{\Omega} rk^2\beta\overline{\Phi}\Phi^t\right| &\leq \|rk^2\beta\|_{L^\infty(\Omega^-)}\|\Phi\|_{H^1(\Omega^-)}\|\Phi^t\|_{H^1(\Omega^-)}.\end{aligned} \tag{3.80}$$

From [76, Theorem 6.11], all the involved integral operators are bounded in their natural trace spaces. The boundedness constant of an operator A is denoted by C_A, and the continuity constant of the interior Dirichlet trace operator is denoted by $C_{\gamma_0^-}$: $\|\gamma_0^-\Phi\|_{H^{\frac{1}{2}}(\Gamma_\infty)} \leq C_{\gamma_0^-}\|\Phi\|_{H^1(\Omega^-)}$. Moreover, since $H^1(\Gamma_\infty)\subset H^{\frac{1}{2}}(\Gamma_\infty)$, there exists a constant $C_{\Gamma_\infty}>0$ such that $\|p\|_{H^{\frac{1}{2}}(\Gamma_\infty)} \leq C_{\Gamma_\infty}\|p\|_{H^1(\Gamma_\infty)}$. These inequalities lead to

$$\begin{aligned}\left|a_1\left((\Phi,\lambda,p),\left(\Phi^t,\lambda^t,p^t\right)\right)\right| &\leq 2\left[1+\frac{1}{1-M_\infty^2}\|r\|_{L^\infty(\Omega^-)}\|1+M_0^2\|_{L^\infty(\Omega^-)}+C_{S^0}\right.\\ &\left.+C_{\gamma_0^-}\left(1+C_{D^0}+C_{\tilde{D}^0}\right)+C_{\gamma_0^-}^2C_{N^0}\right]\|(\Phi,\lambda,p)\|_{\mathbb{H}}\left\|\left(\Phi^t,\lambda^t,p^t\right)\right\|_{\mathbb{H}},\\ \left|a_2\left((\Phi,\lambda,p),\left(\Phi^t,\lambda^t,p^t\right)\right)\right| &\leq 2\left[\|rk^2\beta\|_{L^\infty(\Omega^-)}+2\|rk\|_{L^\infty(\Omega^-)}\|\boldsymbol{V}\|_{L^\infty(\Omega^-)^3}+C_{\Gamma_\infty}\left(\frac{1}{2}+|\eta|+C_{\tilde{D}}\right)+C_{S^0}\right.\\ &\left.+C_S+C_{\gamma_0^-}\left(C_{\Gamma_\infty}C_N+C_{D^0}+C_D+C_{\tilde{D}^0}+C_{\tilde{D}}\right)+C_{\gamma_0^-}^2\left(C_{N^0}+C_N\right)\right]\|(\Phi,\lambda,p)\|_{\mathbb{H}}\left\|\left(\Phi^t,\lambda^t,p^t\right)\right\|_{\mathbb{H}},\\ \left|b\left(\Phi^t,\lambda^t,p^t\right)\right| &\leq \sqrt{2}\left((C_{\gamma_0^-}+C_{\Gamma_\infty})\|\gamma_1 f_{\text{inc}}\|_{H^{-\frac{1}{2}}(\Gamma_\infty)}+\|\gamma_0 f_{\text{inc}}\|_{H^{\frac{1}{2}}(\Gamma_\infty)}\right)\left\|\left(\Phi^t,\lambda^t,p^t\right)\right\|_{\mathbb{H}}.\end{aligned}$$

(ii) Unlike D and $\tilde{D}$, the operators D^0 and $\tilde{D}^0$ are real-valued. They are therefore adjoint, so that

$$\left(\left(\tilde{D}^0-\frac{1}{2}I\right)(\lambda),\gamma_0^-\Phi\right)_{\Gamma_\infty}-\left(\left(D^0-\frac{1}{2}I\right)(\gamma_0^-\Phi),\lambda^t\right)_{\Gamma_\infty}\in i\mathbb{R}. \tag{3.81}$$

From [32, Theorem 2], the operators N^0 and S^0 are strongly elliptic in their natural trace spaces. Moreover, from Section 3.2.4, for all $\boldsymbol{U}\in\mathbb{C}^3$, $\overline{\boldsymbol{U}}\cdot\Xi\boldsymbol{U}\geq\left(1-M_0^2\right)\|\boldsymbol{U}\|^2$. Then, there holds

$$\begin{aligned}\mathrm{Re}\left(a_1\left((\lambda,\Phi,p),(\lambda,\Phi,p)\right)\right) &= \int_\Omega r\Xi\boldsymbol{\nabla}\overline{\Phi}\cdot\boldsymbol{\nabla}\Phi+\left(N^0(\gamma_0^-\Phi),\gamma_0^-\Phi\right)_{\Gamma_\infty}+\left(S^0(\lambda),\lambda\right)_{\Gamma_\infty}+\delta_{\Gamma_\infty}(p,p)\\ &\geq \inf_{\Omega^-}\left(r\left(1-M_0^2\right)\right)\|\boldsymbol{\nabla}\Phi\|_{L^2(\Omega^-)}^2+K_{N^0}\|\gamma_0^-\Phi\|_{H^{\frac{1}{2}}(\Gamma_\infty)}^2+K_{S^0}\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma_\infty)}^2+\|p\|_{H^1(\Gamma_\infty)}^2,\end{aligned} \tag{3.82}$$

where the coercivity constant of an operator A is denoted by K_A. From the Petree–Tartar Lemma [42, Lemma A.38], it can be shown that there exists a constant $C_{\Omega^-}>0$ such that $\|\Phi\|_{H^1(\Omega^-)}\leq C_{\Omega^-}\left(\|\boldsymbol{\nabla}\Phi\|_{L^2(\Omega^-)^3}+\|\gamma_0^-\Phi\|_{H^{\frac{1}{2}}(\Gamma_\infty)}\right)$. Therefore,

$$\begin{aligned}\mathrm{Re}\left(a_1\left((\lambda,\Phi,p),(\lambda,\Phi,p)\right)\right) \geq & \frac{1}{2C_{\Omega^-}^2}\min\left(\inf_{\Omega^-}\left(r\left(1-M_0^2\right)\right),K_{N^0}\right)\|\Phi\|_{H^1(\Omega^-)}^2 \\ & + K_{S^0}\|\lambda\|_{H^{-\frac{1}{2}}(\Gamma_\infty)}^2 + \|p\|_{H^1(\Gamma_\infty)}^2 \\ \geq & \min\left(\frac{\inf_{\Omega^-}\left(r\left(1-M_0^2\right)\right)}{2C_{\Omega^-}^2},\frac{K_{N^0}}{2C_{\Omega^-}^2},K_{S^0},1\right)\|\left(\Phi,\lambda,p\right)\|_{\mathbb{H}}^2.\end{aligned} \tag{3.83}$$

(iii) Let V be a Hilbert space, let A be an operator from V to V, and let a be a sesquilinear form such that, for all $u,v\in V$, $a(u,v)=(Au,v)_V$. A classical result states that A is compact if and only if, for all weakly convergent sequences $(u_n),(v_n)\in V^{\mathbb{N}}$ such that $u_n\rightharpoonup u$ and $v_n\rightharpoonup v$, there holds, up to subsequences, $a(u_n,v_n)\to a(u,v)$. Let $(\Phi_{1n},\lambda_{1n},p_{1n})\rightharpoonup(\Phi_1,\lambda_1,p_1)$ and $(\Phi_{2n},\lambda_{2n},p_{1n})\rightharpoonup(\Phi_2,\lambda_2,p_1)$ be two weakly convergent sequences in $\mathbb{H}$. Since the injection of $H^1(\Omega^-)$ into $L^2(\Omega^-)$ is compact, then, up to subsequences, $\Phi_{in}\to\Phi_i$ and $\boldsymbol{\nabla}\Phi_{in}\rightharpoonup\boldsymbol{\nabla}\Phi_i$, $i=1,2$, in $L^2(\Omega^-)$. Therefore, up to subsequences,

$$\begin{aligned}-\int_{\Omega^-} rk^2\beta\overline{\Phi_{1n}}\Phi_{2n} + i\int_{\Omega^-} rk\boldsymbol{V}\cdot\left(\overline{\Phi_{1n}}\boldsymbol{\nabla}\Phi_{2n}-\Phi_{2n}\boldsymbol{\nabla}\overline{\Phi_{1n}}\right) \\ \to -\int_{\Omega^-} rk^2\beta\overline{\Phi_1}\Phi_2 + i\int_{\Omega^-} rk\boldsymbol{V}\cdot\left(\overline{\Phi_1}\boldsymbol{\nabla}\Phi_2-\Phi_2\boldsymbol{\nabla}\overline{\Phi_1}\right).\end{aligned} \tag{3.84}$$

Moreover, from [93, Lemma 3.9.8], $S-S^0$, $N-N^0$, $D-D^0$ and $\tilde{D}-\tilde{D}^0$ are compact operators in their natural trace spaces. Hence, up to subsequences, $\left(S-S^0\right)(\lambda_{1n})\to\left(S-S^0\right)(\lambda_1)$ in $H^{\frac{1}{2}}(\Gamma_\infty)$, and the corresponding results hold for the other boundary integral operators. Since the injection of $H^1(\Gamma_\infty)$ in $H^{\frac{1}{2}}(\Gamma_\infty)$ is compact, then, up to subsequences, $p_{in}\to p_i$ in $H^{\frac{1}{2}}(\Gamma_\infty)$, so that $(\lambda_{2n},p_{1n})_{\Gamma_\infty}\to(\lambda_2,p_1)_{\Gamma_\infty}$. The last two terms converge as well by continuity of N, γ_0^- and $\tilde{D}$, as well as the strong convergence of p_{2n} in $H^{\frac{1}{2}}(\Gamma_\infty)$ up to subsequences. This concludes the proof. ♢

Proof. (of Theorem 3.2) This a direct consequence of Proposition 3.9 and Theorem 3.10. ♢

Proof. (of Proposition 3.5) The first part of the proposition is clear from the results of Section 3.3. Let (Φ,λ) solve (3.36). In the same fashion as in the proof of Proposition 3.17, (3.21a) and (3.21c) hold, as well as

$$\begin{pmatrix}\frac{1}{2}I+D & -S\\ -N & \frac{1}{2}I-\tilde{D}\end{pmatrix}\begin{pmatrix}\gamma_0^-\Phi\\ \lambda\end{pmatrix}=\begin{pmatrix}\gamma_0^-\Phi\\ \gamma_1^-\Phi\end{pmatrix}, \tag{3.85}$$

so that $\left(\gamma_0^-\Phi,\gamma_1^-\Phi\right)$ belongs to the range of the block operator defined on the left-hand side of (3.85). Under this condition, from [54, Theorem 4.1], citing [106], a radiating piecewise Helmholtz solution u such that $\gamma_0^+u=\gamma_0^-\Phi$ and $\gamma_1^+u=\gamma_1^-\Phi$ can be constructed. Consider the function v defined as $v|_{\Omega^+}=u|_{\Omega^+}$ and $v|_{\mathbb{R}^3\setminus\overline{\Omega^+}}=0$. The function v is still a radiating piecewise Helmholtz solution, and its jumps of traces are $[\gamma_0 v]_{\Gamma_\infty}=\gamma_0^-\Phi$ and $[\gamma_1 v]_{\Gamma_\infty}=\gamma_1^-\Phi$. From (3.26), $\left(\frac{1}{2}I-D\right)\gamma_0^-\Phi+S\gamma_1^-\Phi=-\gamma_0^-v^-=0$. Together with the first line of (3.85), it is deduced that $\lambda-\gamma_1^-\Phi\in\mathrm{Ker}(S)=\mathrm{Ker}(\tilde{D}-\frac{1}{2}I)$. Therefore, (3.68c) and (3.68d) hold. Then, since single-layer and double-layer potentials are radiating piecewise Helmholtz solutions, $\mathcal{R}(\Phi,\lambda)$ satisfies (3.21b)

and (3.21f). Finally, taking the exterior traces of $\mathcal{R}(\Phi,\lambda)$, the transmission conditions (3.21d) and (3.21e) are directly obtained from (3.68c), (3.68d) and $\lambda-\gamma_1^-\Phi\in\mathrm{Ker}(S)=\mathrm{Ker}(\tilde{D}-\frac{1}{2}I)$, in the same fashion as in the proof of Proposition 3.9. ♢

Proof. (of Theorem 3.7) Let (Φ,λ) solve (3.36) with $\gamma_0 f_{\mathrm{inc}}=0$ and $\gamma_1 f_{\mathrm{inc}}=0$. From Proposition 3.5, $\mathcal{R}(\Phi,\lambda)$ solves (3.21). As seen in the proof of Proposition 3.18, $\mathcal{R}(\Phi,\lambda)\in H^1_{\mathrm{loc}}(\Omega)$. It is then deduced from Proposition 3.16 that $\mathcal{R}(\Phi,\lambda)=0$ in Ω. As a consequence, $\Phi=\mathcal{R}(\Phi,\lambda)|_{\Omega^-}=0$. From the proof of Proposition 3.5, $\lambda-\gamma_1^-\Phi\in\mathrm{Ker}(S)$. Suppose $-\hat{k}_\infty^2\notin\Lambda$. Then, $\mathrm{Ker}(S)=\{0\}$, leading to $\lambda=\gamma_1^-\Phi=0$, so that problem (3.36) has at most one solution; well-posedness is then obtained using the Fredholm alternative by proceeding similarly to the proof of Theorem 3.10. Suppose $-\hat{k}_\infty^2\in\Lambda$. Let $\lambda^*\in\mathrm{Ker}(S)=\mathrm{Ker}(\tilde{D}-\frac{1}{2}I)$, and f be the solution to (3.21). From Proposition 3.6, $(f^-,\gamma_1 f+\lambda^*)$ solves (3.36). ♢

4

Validation campaign and numerical simulations

The scope of this chapter is to validate the BEM/FEM coupled formulation (3.36), that uses the new transformed convected Helmholtz equation (3.18a). We suppose that the frequency of the source is a nonresonant frequency for the considered problem (more precisely, $-\hat{k}_\infty^2 \notin \Lambda$). This can be assessed by verifying that the formula (1.30) does not produce absurd values in $\mathbb{R}^+ \backslash \overline{\Omega^+}$, as it is the case on the left plots of Figure 3.4. All the simulations on this chapter are three-dimensional.

4.1 Validation campaign

The formulation (3.36) is compared to other numerical methods in the following cases:

- flow at rest and uniform properties, comparison with ACTIPOLE with only BEM in Section 4.1.1,
- flow at rest and different properties, comparison with an analytic solution computed by means of Mie series in Section 4.1.2,
- nonuniform flow and nonuniform properties, qualitative comparison with ACTI-HF and ISVR in Section 4.1.4.

Then, industrial test cases are presented in Section 4.2.

4.1.1 Flow at rest and uniform properties: $M = M_\infty = 0$, $\rho = \rho_\infty$, $c = c_\infty$, comparison with ACTIPOLE with only BEM

The code we used to implement the formulations (3.36), ACTIPOLE [33, 34], is a boundary element method code designed by EADS-IW and Airbus to solve Helmholtz exterior problems with uniform (and zero) flows. In its original form, ACTIPOLE computes only surface potentials, and then cannot be used with an inhomogeneous Ω^-. This original state of the code can be used as a reference when the fluid characteristics in Ω^- are equal to the one in Ω^+.

We first consider the geometry presented in Figure 4.1.In this figure, the plain cube is solid object perfectly reflecting the incoming acoustic waves (with boundary condition (3.2) on the acoustic potential), whereas the hatched areas correspond to the interior domain Ω^- meshed with 11510 tetrahedra. The boundary of Ω^- and the boundary of the solid object are each composed of 2418 triangles.

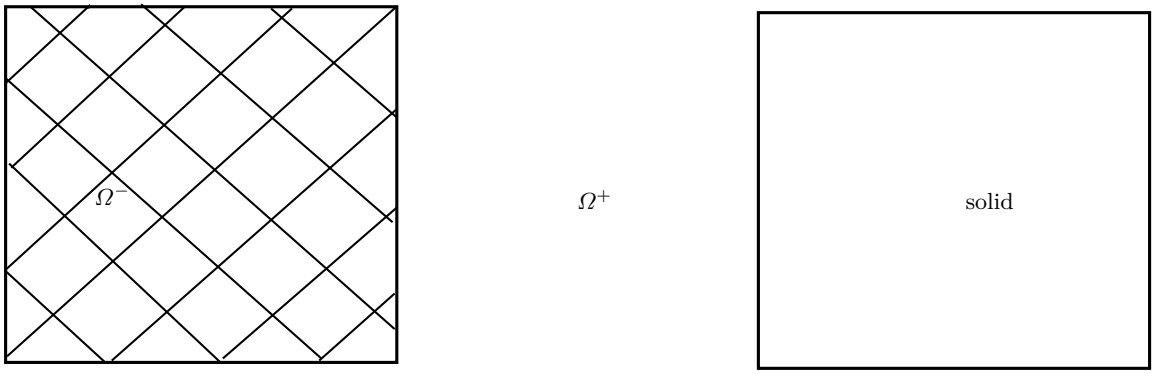

Fig. 4.1. Cross section of the geometry of the test case

We impose $M = M_\infty = 0$, $\rho = \rho_\infty$ and $c = c_\infty$ so that the flow in the interior domain equals the flow in the exterior domain, and the reference version of ACTIPOLE can be used to solve a boundary integral equation at the surface of the solid cube. The pressure fields computed using the formulation (3.36) and the reference version of ACTIPOLE should be very close. We notice that this geometry do not correspond to the one of Figure 3.1, but the formulation (3.36) is easily adapted to the present test case by injecting the boundary condition (3.2) at the boundary of the solid cube.

Comparing for each line of the Figure 4.2 the left and the right pictures, we see that the results of the formulation (3.36) are similar to the ones of the reference version of ACTIPOLE. Notice that the post treatment $\mathcal{R}(\Phi, \lambda)$ of the solution to (3.36) provides the total acoustic pressure. To recover the scattered pressure, one need to subtract the incoming field f_{inc}.

A quantitative comparison is presented in Table 4.1, where we consider a network of 100 points uniformly distributed on a part of a hyperplane located in Ω^+. The difference comes from the fact that with the formulation (3.36), the pressure in the interior domain is computed on a mesh, introducing approximation errors, while with the reference version of ACTIPOLE, the pressure in this domain is computed using the representation formula (1.30). We notice with additional computation that the difference decreases when the mesh size decreases, providing us with the first validation argument.

Number of measure points	10x10
Points defining the part of hyperplane	(-0.5, -0.5, 0.7) (-0.5, 0.5, 0.7) (0.5, -0.5, 0.7)
Relative difference on the scattered pressure	1.08 %

Table 4.1. Relative difference on the scattered pressure in Euclidian norm, between the formulation (3.36) and the reference version of ACTIPOLE.

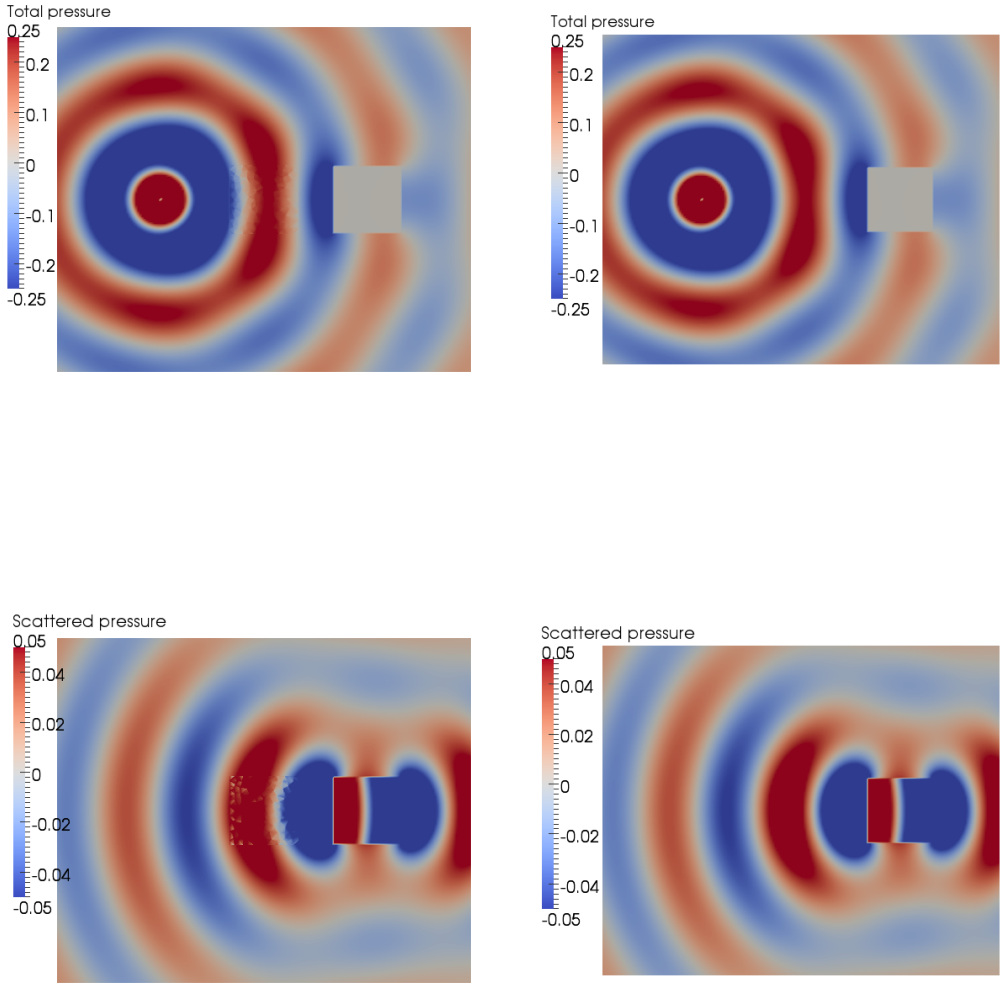

Fig. 4.2. Pressure fields comparison. Left: formulation (3.36), right: reference version of ACTIPOLE. Top: total pressure, bottom: scattered pressure. Notice that the poor quality of the scattered pressure field in Ω^- is due to technical issues of the visualization software: the total pressure field has to be converted from cell data to point data, inducing an averaging of the field over each tetrahedron, before subtracting the incoming field.

4.1.2 Flow at rest and uniform properties: $M = M_\infty = 0$, $\rho = \rho_\infty$, $c = 2c_\infty$, comparison with an analytic solution computed by means of Mie series

Suppose that there is no diffracting object, and that $M = M_\infty = 0$, $\rho = \rho_\infty$, $c = 2c_\infty$. In this case, the pressure field is perturbed when crossing Γ_∞ by the difference of the speed of sound between Ω^- et Ω^+. It is possible to get, for a simple geometry, an approximation of the pressure field in Ω^+ in the form of a Mie series.

The test case is composed of a ball of radius 1, meshed with tetrahedra and centered at the origin, such that $(\rho_0,\, c_0,\, M_0) = (1.2\ kg.m^{-3},\, 680\ m.s^{-1},\, 0)$, embedded in the exterior domain where $(\rho_\infty,\, c_\infty,\, M_\infty) = (1.2\ kg.m^{-3},\, 340\ m.s^{-1},\, 0)$, see Figure 4.3. The source is an acoustic monopole of magnitude 1, and the visualization point is located at $(0, 1.7, 0)$.

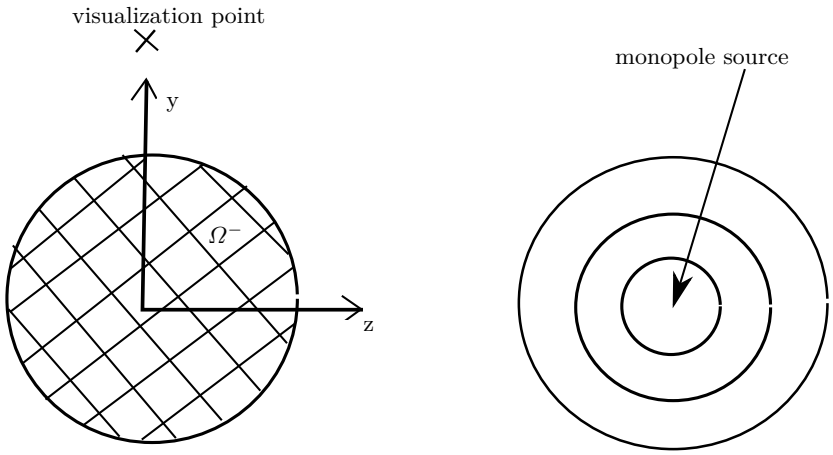

Fig. 4.3. Geometry of the test case

The mesh we used respects the convention of the length of the longest edge being of the order of one tenth of the wavelength of the source for frequencies up to 85 Hz. The solution using Mie series was provided by the company IMACS. We also consider the pressure field in the case $c = c_\infty$ (therefore in the absence of the interior domain Ω^-), whose real part is given by $\frac{\cos(kr)}{4\pi r}$, where r is the distance between the source, are refer to it as pressure "in the absence of ball".

In Figure 4.4, the curve of the pressure computed using formulation (3.36) and the one provided by IMACS are very similar, and these two curves are significantly different from the one correspond to the case in the absence of ball. In Figure 4.5 are represented the difference between the formulation (3.36) and the case in the absence of ball on the one hand, and the difference between the solution computed by IMACS and the case in the absence of ball on the other hand. The relative difference is less than 0.25% for $f < 85\ Hz$ and less than 5% up to 500 Hz.

The difference between the solution using the formulation (3.36) and the solution provided by IMACS is much smaller than the difference of each of these solutions with the reference. Therefore, the perturbation induced by the change in the speed of sound computed by our code

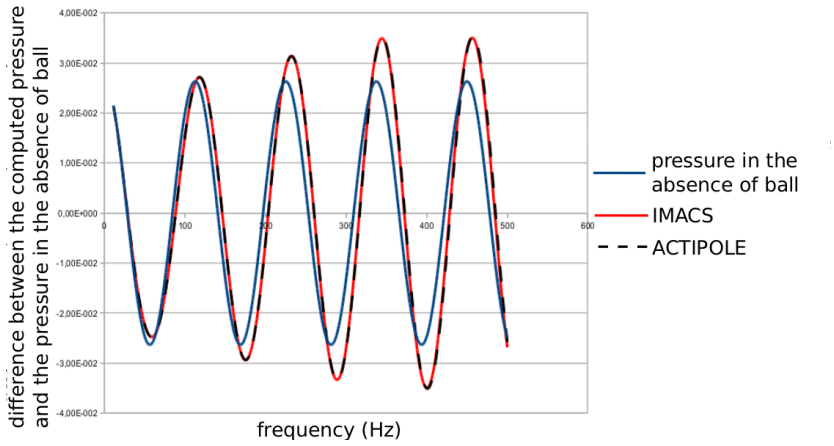

Fig. 4.4. Real part of the pressure measured at the visualization point with respect to the frequency - red: IMACS, dashed black: formulation (3.36), blue: reference. The first two curves superimpose.

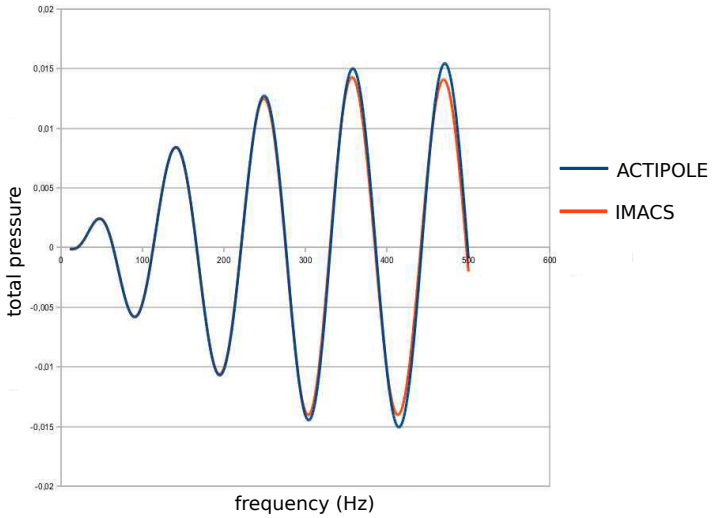

Fig. 4.5. Real part of the difference between the pressure computed at the visualization point and the reference with respect to the frequency of the source - blue: formulation (3.36), red: IMACS.

is validated by comparison with the solution by IMACS. We have verified that this conclusion also holds when the visualization point is located in the interior domain Ω^-.

4.1.3 Uniform flow and properties: $M = M_\infty = 0.5$, $\rho = \rho_\infty$, $c = c_\infty$, comparison with ACTIPOLE with only BEM

The test case is the same as in Section 4.1.1, except that there is now an equal nonzero uniform flow in Ω^- and Ω^+, such that $M = M_\infty = 0.5$. The Mach number being nonzero, the Prandtl–Glauert transformation is used for both the formulation (3.36) and the reference version of ACTIPOLE.

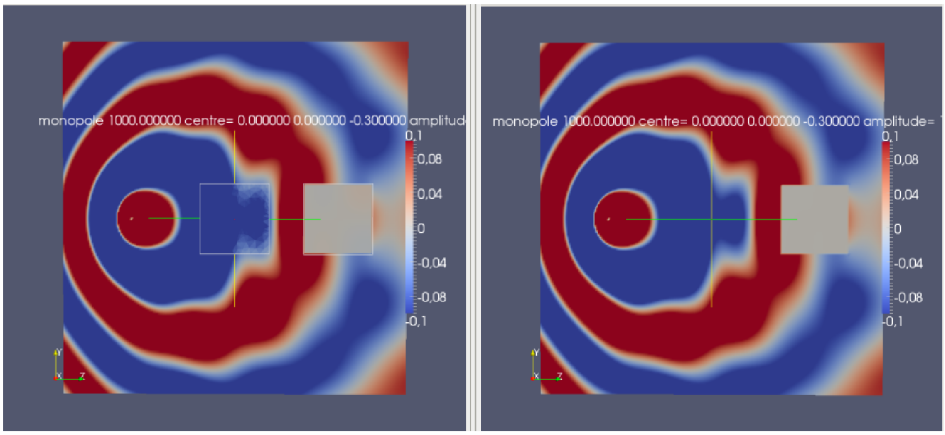

Fig. 4.6. Comparison of the total pressure fields. Left: formulation (3.36), right: reference version of ACTIPOLE.

The total pressure fields are presented in Figure 4.6. The deformation of the waves on the right part of each picture is due to the mean flow going from the left to the right. The pressure fields computed by the formulation (3.36) and the reference version of ACTIPOLE are very similar. The relative difference on the scattered pressure fields in Euclidian norm on a network of 21x21 points located on a part of a hyperplane defined by the points (-1.0, -1.0, 0.5), (-1.0, -1.0, 0.5) et (-1.0, -1.0, 0.5) is 3.5%.

4.1.4 Nonuniform flow and nonuniform properties: $M \neq M_\infty = 0$, $\rho \neq \rho_\infty$, $c \neq c_\infty$, qualitative comparison with ACTI-HF and ISVR

This test case was realized in collaboration with Nolwenn Balin (EADS-IW). We consider the study [90, 88]. This test case simulates the crossing of a hot jet stream (modeled by a cylinder) by an acoustic wave generated by a monopole. The geometry is presented in Figure 4.7.

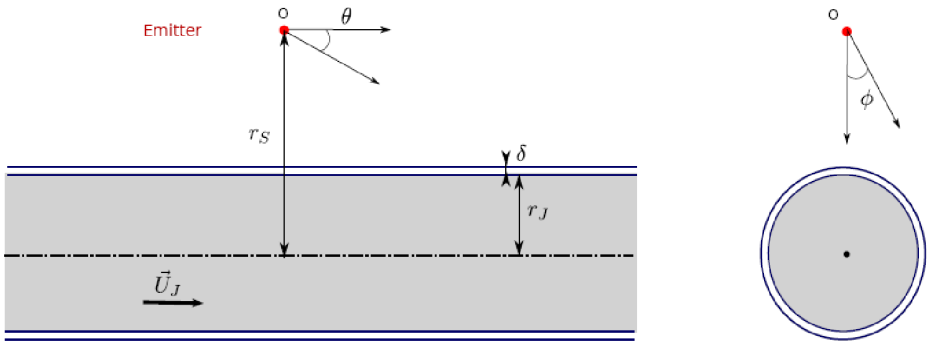

Fig. 4.7. Geometry of the test case

The characteristics of the test case are
– in the exterior domain, $c_{ext} = 340.31\ m.s^{-1}$, $\rho_{ext} = 1.23\ kg.m^{-3}$ et $M_\infty = 0$,
– in the interior domain, $c_{int} = 545.40\ m.s^{-1}$, $\rho_{int} = 0.48\ kg.m^{-3}$ et $M = 0.69$,

– the source is a monopole of magnitude 1 and frequency 4054 Hz located at $r_S = 292.5\ mm$,
– the hot jet stream is modeled by a cylinder of length 2 m and radius $r_J = 117\ mm$ (theoretically, its length is infinite), the thickness of the shear layer is $\delta = 0.14\ mm$.

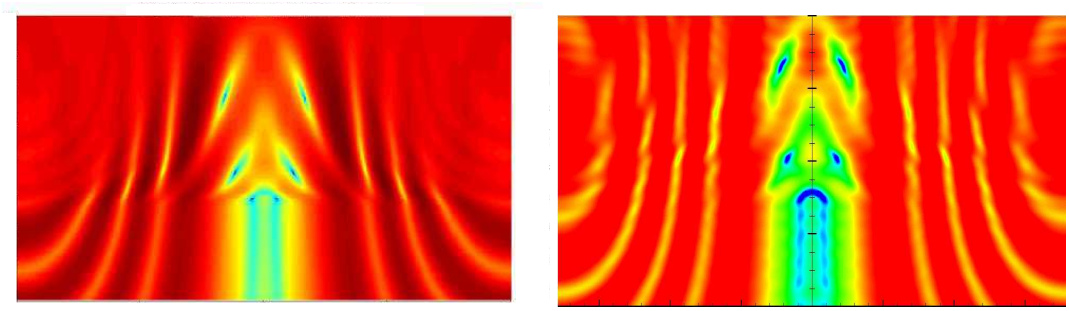

Fig. 4.8. $(Pressure(Jet) - Pressure(No\ Jet))_{db}$ with respect to θ and ϕ defined in Figure 4.7, 30 m away from the source. Left: results obtained in [90, 88], right: results obtained using the formulation (3.36)

Figure 4.8 presents the flied of the difference between the pressure with and without the presence of the jet in dB.

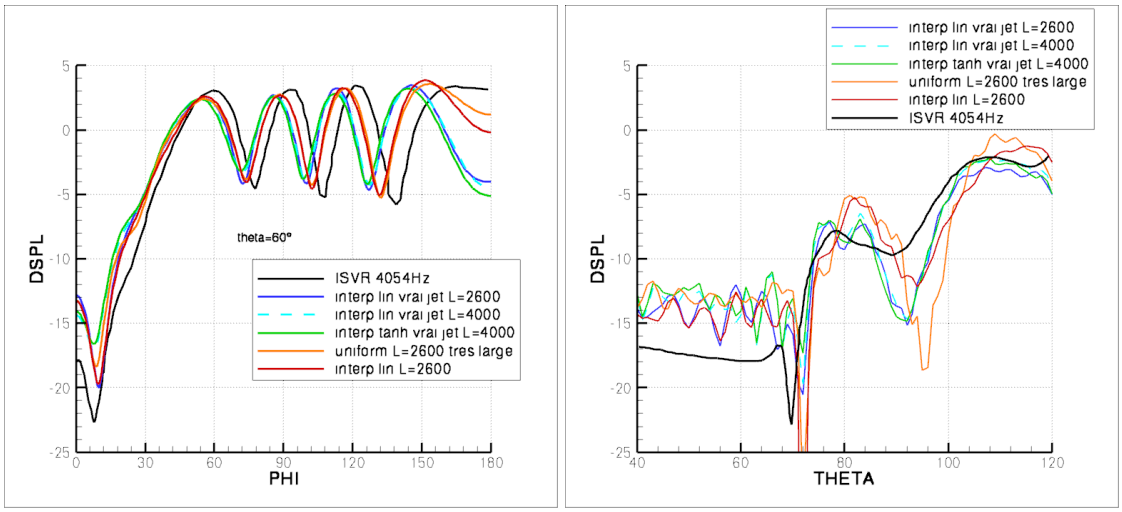

Fig. 4.9. $(Pressure(Jet) - Pressure(No\ Jet))_{db}$, left: cross-section at $\theta = 60°$, right: cross-section at $\phi = 0°$, for several lengths of the cylinder and several velocity profiles in the shear layer

Different velocity profiles have been tried in the shear layer and different lengths of the cylinder have been tried. Two cross section of the flied of the difference between the pressure with and without the presence of the jet in dB, for various combinations of these parameter, are presented in Figure 4.9. In this figure, "interp lin" indicates that the Mach number varies linearly from 0.69 to 0 through the shear layer from the jet the exterior domain, "interp tanh"

indicates that the Mach number varies along a hyperbolic tangent profile, "uniform" indicates that the Mach number equals 0.69 in all the shear layer.

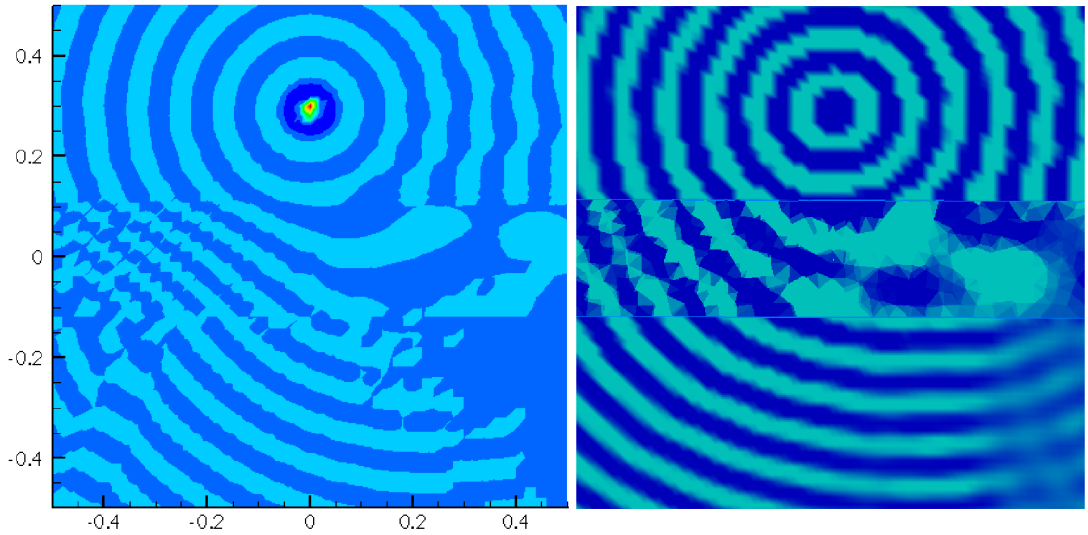

Fig. 4.10. Real part of the total pressure field. Left: results obtained with ACTI-HF, right: results obtained with the formulation (3.36)

Figure 4.10 presents the real part of the total pressure field obtained with ACTI-HF (a code developed by EADS for high frequency computations) and with the formulation (3.36), in the case where the Mach number equals 0.69 in all the shear layer.

Even if the results are not identical between the three methods, the main characteristic elements can be recognized in each case. The differences can be explained by the approximations introduced in each of the methods of resolution. In our method, the cylinder must have a finite length, introducing some errors with respect to the theoretical infinite cylinder case. Besides, some approximation errors result from the finite element and boundary element methods. The validation from this test case is mainly qualitative: since the same behavior is identified using the three techniques, we can conclude that our method provide reasonable results.

4.2 Industrial test cases

The test cases of the section have been realized in collaboration with Nolwenn Balin (EADS-IW). We first present in Section 4.2.1 a test case whose geometry is simple, but the number of unknowns is important. Then, in Section 4.2.2, we present a test case whose geometry is realistic in the aeronautic industry, with a large number of unknowns.

4.2.1 Potential flow around a sphere

The test case is composed of a rigid sphere of radius 0.6 m, located inside a ball of radius 1.2 m meshed with tetrahedra; a potential flow is computed in the interior domain, so that $M_\infty = 0.4$, see Figure 4.11. The source is an acoustic monopole of frequency 1333 Hz located downstream of the object.

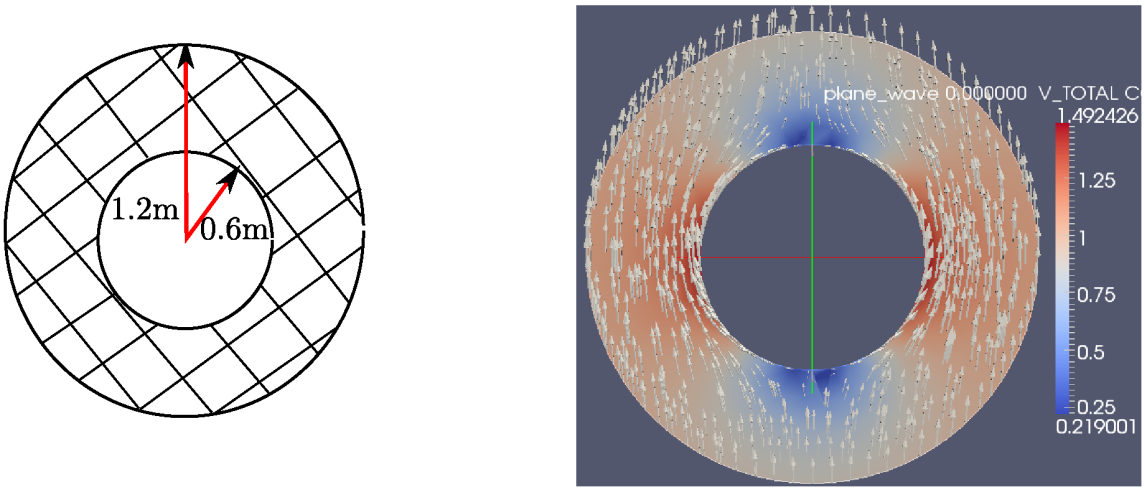

Fig. 4.11. Left: Geometry of the test case, right: potential flow in the interior domain

Figures 4.12 and 4.13 present the scattered and total pressure field, when the mean flow in the interior domain is a uniform flow with Mach number M_∞ and the potential flow presented in Figure 4.11. We notice that the pressure field changes between the two experiments. In particular in Figure 4.12, the magnitude of the pressure in the shadow zone, is strongly underestimated when using a uniform flow in the interior domain even though the source is located downstream the object.

The mesh is composed of 1.5×10^6 tetrahedra, and the problem has 2.5×10^5 volumic unknowns and 10^5 surfacic unknowns. The direct computation lasts 3h on 64 processors, or 1h30 using Fast Multipole Method on 32 processors, with a relative residual of 10^{-3}.

4.2.2 Aircraft turbojet

This test case is presented in [Pr2]. It consists in a simplified engine with modal surfaces orthogonal to $\boldsymbol{e_z}$ to model the upstream and downstream fans (see Figure 4.14). The far field flow is defined by $\boldsymbol{M_\infty} = -0.3\boldsymbol{e_z}$. Three different configurations are considered: a uniform flow defined by $\boldsymbol{M_\infty}$ and potential flows computed such that the Mach number at the upstream modal surface Γ_M is 0.3 and 0.42.

First, the potential flow is computed using an in-house software based on a fixed-point algorithm [87, 40]. The potential flow obtained when $M_M = 0.42$ at the upstream modal surface is plotted on Figure 4.15.

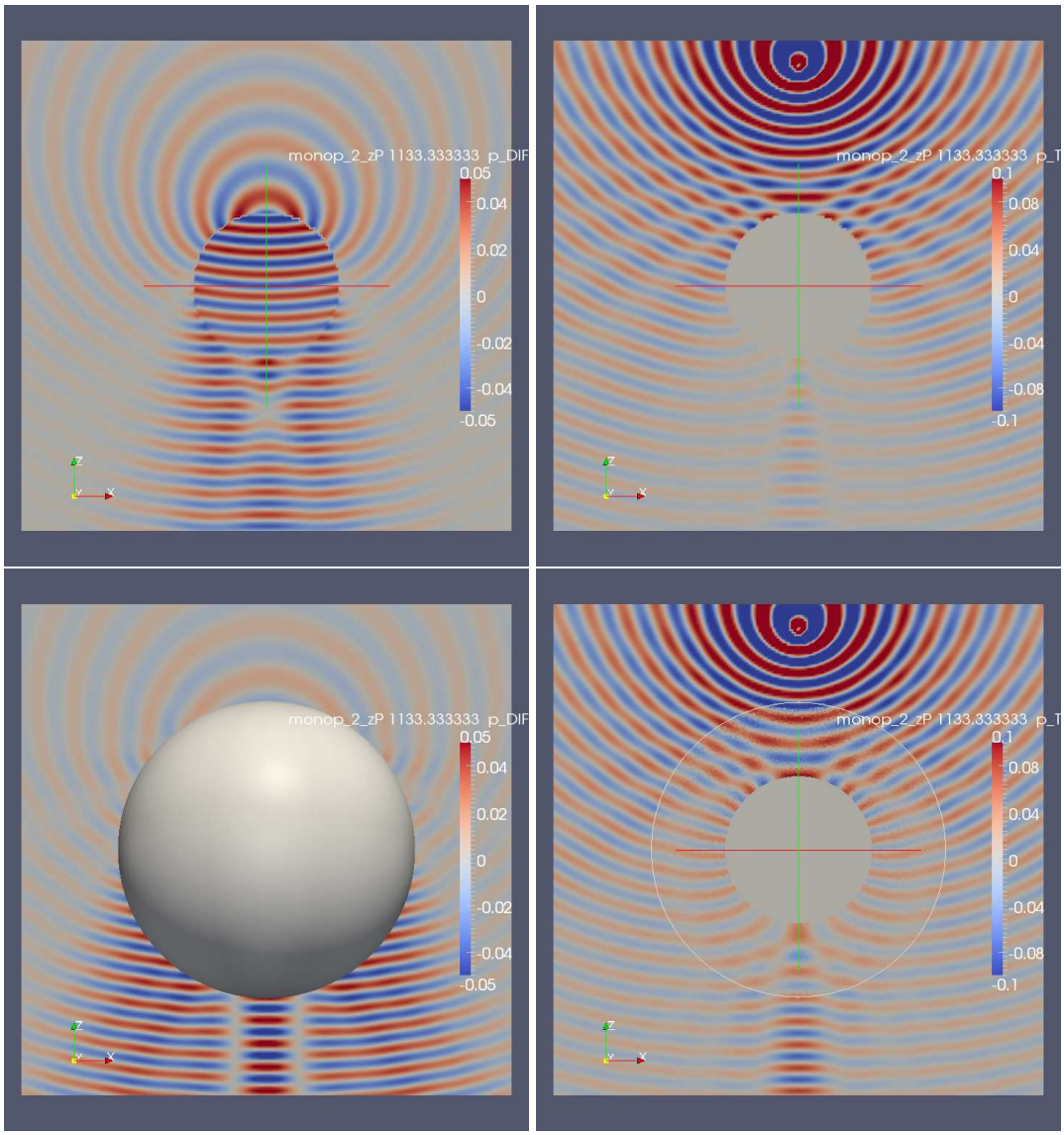

Fig. 4.12. Source located downstream of the object. Top: uniform flow in the interior domain, bottom: potential flow, left: scattered field, right: total field

We now consider the upstream fan modal source model at the frequency of 200 Hz. The mean size of the mesh elements is 83 mm. The model contains 1.2×10^6 degrees of freedom and 11.8×10^6 tetrahedra. The acoustic source consists on an incident field on Γ_M (see Figure 4.14), decomposed on a bases of functions with support on Γ_M called modes. For comparison purposes, the intensity on each mode is set to 100 dB, following Morfey's convention [79].

The pressure obtained in the vicinity of the modal surface is shown on Figures 4.16 and 4.17. For each mode, in the top part of the figure, the pressure is obtained with the uniform flow model and in the bottom part of the figure the pressure is obtained with the potential model

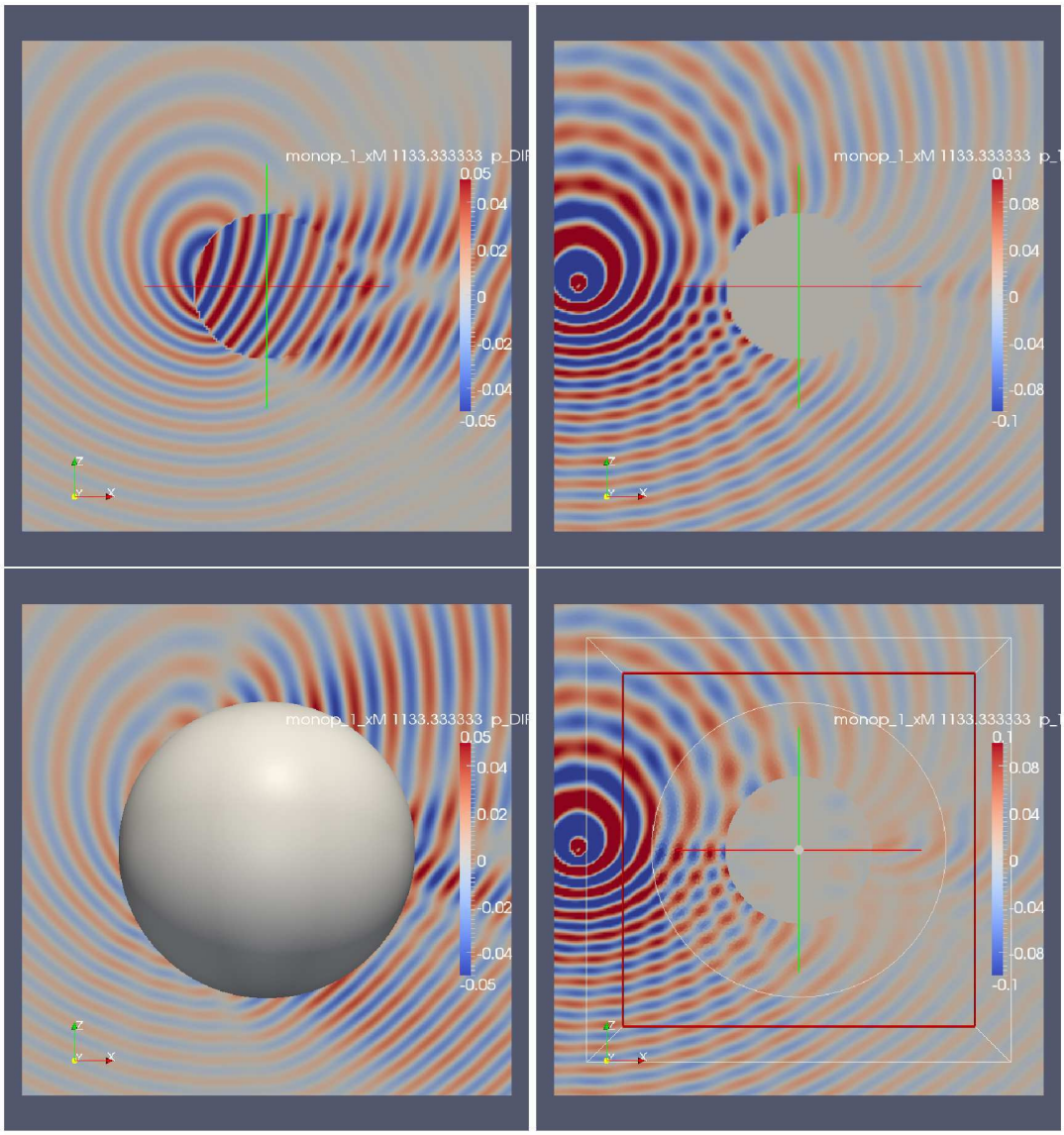

Fig. 4.13. Source located beside the object. Top: uniform flow in the interior domain, bottom: potential flow, left: scattered field, right: total field

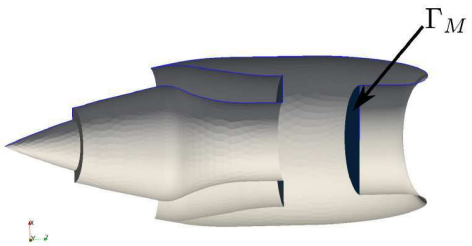

Fig. 4.14. Geometry of the simplified engine.

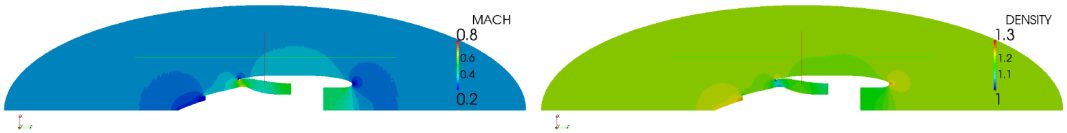

Fig. 4.15. Simplified engine: computed potential flow such that $M_M = 0.42$ and $M_\infty = 0.3$ (*left*: Mach number, *right*: density).

with a Mach number at the modal surface of 0.3 or 0.42. Small variations are observed with the $M_M = 0.3$ condition. The differences are higher when the flow at the modal surface is higher.

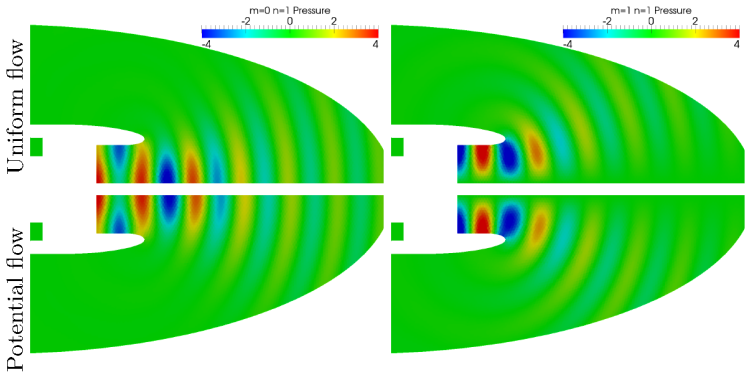

Fig. 4.16. Simplified engine: comparison of the pressure for the uniform flow model and the potential flow model with $M_\infty = M_M = 0.3$.

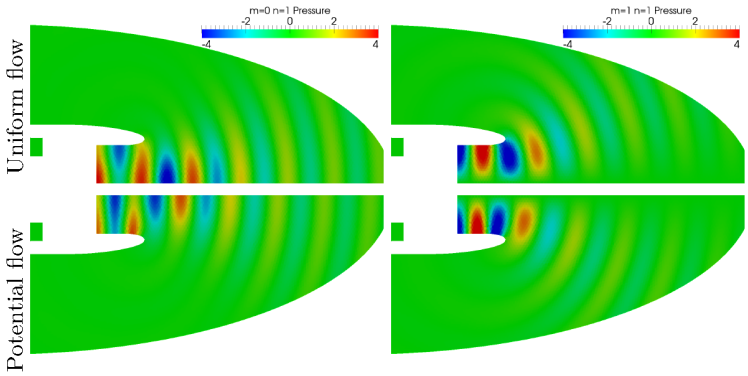

Fig. 4.17. Simplified engine: comparison of the pressure for the uniform flow model ($M_\infty = M_M = 0.3$) and the potential flow model ($M_\infty = 0.3$ and $M_M = 0.42$).

Figure 4.18 shows the pressure in dB obtained on a circle at a distance of 20 m from the center of the modal surface, for different values of M_M. Significant changes in the amplitude are obtained for the different modes by taking into account the potential flow. For instance, for the mode $(1, 1)$ and for the same flow at the modal surface, the amplitude predicted for the

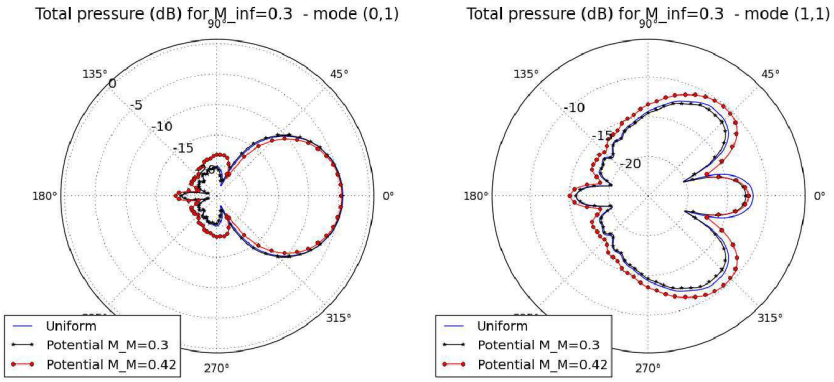

Fig. 4.18. Simplified engine: pressure in dB on a circle at $r = 20$ m for $M_\infty = 0.3$ and some values of M_M (mesh size 75mm).

potential flow is approximately 1 dB lower in the axis direction than the amplitude predicted by the uniform flow model. By increasing the flow through the upstream modal surface, the difference with the uniform flow model is higher and is observed for all the radiation directions. This seams physically reasonable: the sound is convected faster by the jet in the potential case (therefore resulting in larger amplitudes) than in the uniform flow case.

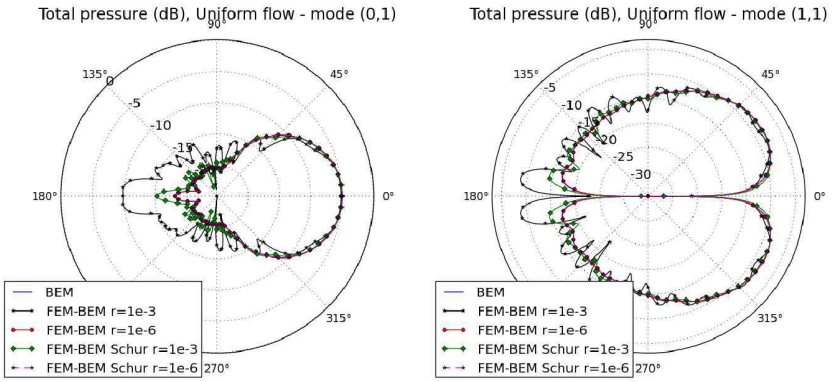

Fig. 4.19. Simplified engine: influence of the convergence criteria and on the pressure in dB on a circle at $r = 20$ m for a uniform flow defined by $M_M = M_\infty = 0.1$.

Figure 4.19 illustrates the influence of the relative residual of the iterative solver on the diffracted pressure field. Results for relative residuals of 10^{-3} and 10^{-6}, and with or without a Schur complement on the volume part of the matrix are presented. Using a Schur complement yields algebraic accuracy on some components of the solution, since the exact inverse of the volume part of the matrix is considered, which is not the case when considering an iterative solver for the whole matrix. From these results, it appears that a convergence with a tolerance of 10^{-3} is not sufficient for a solution without a Schur complement on the volume part of the matrix. In that case, for a mesh containing 4.7 $\times 10^6$ dofs and 25.8 $\times 10^6$ tetrahedra, the computation takes 1.5 h on 160 processors and 231 iterations for the FMM solver without using

the Schur complement and 6.5 h on 120 processors and 204 iterations with the Schur complement, for an achieved residual of 10^{-6}.

Part II

The Reduced Basis Method

5

Accurate and online efficient evaluation of the a posteriori error bound in the reduced basis method

This chapter is based on the article [Ar3]. The last section has been added to explain how we can compute a lower bound of the inf-sup constant in an online-efficient way, when the stability constant is parameter dependent, which is the case in the acoustic problem considered in this chapter.

Summary. The reduced basis method is a model reduction technique yielding substantial savings of computational time when a solution to a parametrized equation has to be computed for many values of the parameter. Certification of the approximation is possible by means of an a posteriori error bound. Under appropriate assumptions, this error bound is computed with an algorithm of complexity independent of the size of the full problem. In practice, the evaluation of the error bound can become very sensitive to round-off errors. We propose herein an explanation of this fact. A first remedy has been proposed in [F. Casenave, Accurate *a posteriori* error evaluation in the reduced basis method. *C. R. Math. Acad. Sci. Paris* **350** (2012) 539–542.]. In this chapter, we improve this remedy by proposing a new approximation of the error bound using the Empirical Interpolation Method (EIM). This method achieves higher levels of accuracy and requires potentially less precomputations than the usual formula. A version of the EIM stabilized with respect to round-off errors is also derived. The method is illustrated on a simple one-dimensional diffusion problem and on a three-dimensional acoustic scattering problem solved by a boundary element method.

5.1 Introduction

In many problems, such as optimization, uncertainty propagation or real-time simulation, one has to evaluate an objective function for a large number of values of some parameters. Evaluating this objective function often implies solving a parametrized partial differential equation for a given parameter value. In an industrial context, one evaluation of the objective function can already be a challenging numerical problem. To keep reasonable computational costs, various model reduction techniques have been developed to speed up computations. We focus on the Reduced Basis (RB) method [71, 72]. This method has been applied to many kinds of problems, including nonlinear problems such as the viscous Burgers equation [104] or the steady incompressible Navier-Stokes equations [103].

As described in Section 5.2, the RB method consists in replacing the sequence $\mathcal{P} \ni \mu \overset{E_\mu}{\mapsto} u_\mu \mapsto Q(u_\mu)$ by the sequence $\mathcal{P} \ni \mu \overset{\hat{E}_\mu}{\mapsto} \hat{u}_\mu \mapsto \hat{Q}(\hat{u}_\mu)$. Here, $\mathcal{P}$ denotes the parameter set, $E_\mu : \mu \mapsto u_\mu$ the model problem, $\hat{E}_\mu : \mu \mapsto \hat{u}_\mu$ its lower-dimensional approximation, $Q(u_\mu)$

the quantity of interest, and $\hat{Q}(\hat{u}_\mu)$ its RB approximation. More specifically, the RB method consists in two steps: (i) A so-called offline stage, where solutions to E_μ for well-chosen values of the parameter μ are computed. During this stage, $\hat{N}$ problems of size N are solved (with $\hat{N} \ll N$), and some quantities related to the $\hat{N}$ solutions are stored, and (ii) a so-called online stage, where the precomputed quantities are used to solve $\hat{E}_\mu$ for many values of μ. In this stage, a certification of the approximation is possible by means of an a posteriori error bound. An important feature in the RB method is the use of an online-efficient a posteriori error bound. The notion of online-efficiency is defined in Section 5.2.4. Moreover, the a posteriori error bound must be as sharp as possible to faithfully represent the error. However, as noticed for example in [86, pp.148-149], standard a posteriori error bounds are subject to round-off errors, especially for the computation of accurate solutions. This difficulty can be encountered in complex industrial applications in the following two cases. First and most importantly, when the stability constant of the underlying bilinear (or sesquilinear) form is very small, the classical formula for the error bound fails to certify the approximation, even at a relatively crude error level, as illustrated in Section 5.5 where the stability constant is about 10^{-6} and the classical error bound stagnates at about 10^{-4}. Second, in some industrial codes, the single-precision format is used to speed up computations, when high precision is not needed. In this case, the classical formula for the error bound fails to deliver values below 10^{-4} for a stability constant of order 1. The purpose of this work is an explanation of these facts and the derivation of a new method to compute the error bound in an accurate and online-efficient way. Additionally, the new formula uses potentially less precomputed quantities than the classical formula.

In Section 5.2, we briefly recall the main ingredients of the RB method, namely (i) the construction of the reduced problem, (ii) the a posterior error bound, (iii) the notion of online-efficiency, and (iv) the offline stage during which the vectors of the reduced basis are constructed. We then explain in Section 5.3 why the classical formula for computing the error bound is ill-conditioned in regard of round-off errors. In Section 5.4, we present our new procedure based on the Empirical Interpolation Method (EIM). A version of the EIM stabilized with respect to round-off errors is also derived, and the various procedures to compute the error bound are compared on a simple one-dimensional diffusion problem. In Section 5.5, we apply this new procedure to a three-dimensional acoustic scattering problem.

5.2 The reduced basis method

5.2.1 The model problem

We suppose that the problem of interest has the following discrete variational form, depending on a parameter μ in a parameter set $\mathcal{P}$: for a finite-dimensional space $\mathcal{V}$ of dimension N (with $N \gg 1$ resulting, e.g., from discretization), find $u_\mu \in \mathcal{V}$ such that

$$E_\mu : a_\mu(u_\mu, v) = b(v), \qquad \forall v \in \mathcal{V}, \tag{5.1}$$

where a_μ is an inf-sup stable bounded sesquilinear form on $\mathcal{V} \times \mathcal{V}$ and b is a continuous linear form on $\mathcal{V}$. We work in complex vector spaces in view of our application to acoustic scattering. In what follows, the complex conjugate of $z \in \mathbb{C}$ is denoted z^*. We define the Riesz isomorphism J from $\mathcal{V}'$ to $\mathcal{V}$ such that for all $l \in \mathcal{V}'$ and all $u \in \mathcal{V}$, $(Jl, u)_\mathcal{V} = l(u)$, where $(\cdot, \cdot)_\mathcal{V}$ denotes

the inner product of $\mathcal{V}$ (antilinear with respect to its second argument) with associated norm $\|\cdot\|_{\mathcal{V}}$. We denote $\beta_\mu := \inf_{u\in\mathcal{V}}\sup_{v\in\mathcal{V}}\frac{|a_\mu(u,v)|}{\|u\|_{\mathcal{V}}\|v\|_{\mathcal{V}}} > 0$ the inf-sup constant of a_μ and $\tilde{\beta}_\mu$ a computable positive lower bound of β_μ. For simplicity, we consider that the linear form b is independent of the parameter μ. The extension to μ-dependent b is straightforward. We refer to the discrete solution u_μ as the "truth solution".

5.2.2 The reduced problem

Suppose that a reduced basis, consisting of $\hat{N}$ solutions u_{μ_i} of E_{μ_i}, $i \in \{1, ..., \hat{N}\}$, has already been constructed. To alleviate the notation, we denote u_i the function u_{μ_i}. How the parameters μ_i are chosen is briefly outlined in Section 5.2.5. Given a parameter value $\mu \in \mathcal{P}$, the reduced problem is then a Galerkin procedure written on the linear space $\hat{\mathcal{V}} = \text{Span}\{u_1, ..., u_{\hat{N}}\} \subset \mathcal{V}$: find $\hat{u}_\mu \in \hat{\mathcal{V}}$ such that

$$\hat{E}_\mu : a_\mu(\hat{u}_\mu, u_j) = b(u_j), \qquad \forall j \in \{1, ..., \hat{N}\}. \tag{5.2}$$

The approximate solution on the reduced basis is written as

$$\hat{u}_\mu = \sum_{i=1}^{\hat{N}} \gamma_i(\mu) u_i. \tag{5.3}$$

Recalling the exact and approximate quantities of interest $Q(u_\mu)$ and $\hat{Q}(\hat{u}_\mu)$, respectively, the quality of the approximation for a given $\mu \in \mathcal{P}$ is quantified by the error measure $\|Q(u_\mu) - \hat{Q}(\hat{u}_\mu)\|$. When we obtain a satisfying error measure with $\hat{N} \ll N$, the RB strategy is successful. Two main cases are generally considered: (i) the so-called general-purpose case, where one is interested in the whole solution: $Q = \hat{Q} = \text{Id}$ and $\|\cdot\| = \|\cdot\|_{\mathcal{V}}$, and (ii) the so-called goal-oriented case, where Q is a linear form on $\mathcal{V}$ and $\|\cdot\| = |\cdot|$. The operator $\hat{Q}$ is consistently built so that $\|Q(u_\mu) - \hat{Q}(\hat{u}_\mu)\|$ vanishes for $\mu = \mu_i$, $i \in \{1, ..., \hat{N}\}$.

5.2.3 A posteriori error bound

In the standard RB method, the a posteriori error bound is a residual-based bound. In what follows, we refer to it simply as error bound. Since this error bound is an upper bound, it provides a way to certify the approximation made by the reduced basis.

Proposition 5.1 (General-purpose case) *The following error bound holds: For all $\mu \in \mathcal{P}$,*

$$\|u_\mu - \hat{u}_\mu\|_{\mathcal{V}} \leq \mathcal{E}_1(\mu) := \tilde{\beta}_\mu^{-1} \|G_\mu \hat{u}_\mu\|_{\mathcal{V}}, \tag{5.4}$$

with G_μ the linear map from $\mathcal{V}$ to $\mathcal{V}$ such that $\mathcal{V} \ni u \mapsto G_\mu u := J\left(a_\mu(u, \cdot) - b\right) \in \mathcal{V}$.

Proof. See [86, Section 4.3.2]. ♢

In the goal-oriented case, one possible approach is to introduce the following dual problem: Find $v_\mu \in \mathcal{V}$ such that

$$E_\mu^d : a_\mu(w, v_\mu) = Q(w), \qquad \forall w \in \mathcal{V}. \tag{5.5}$$

We wrote the dual problem on the same discrete space $\mathcal{V}$, but another space can be considered. A reduced basis procedure is also carried out for the problem E_μ^d, resulting in an approximation $\hat{v}_\mu$ of v_μ. The approximate quantity of interest is then defined as $\hat{Q}(\hat{u}_\mu) := Q(\hat{u}_\mu) - (G_\mu \hat{u}_\mu, \hat{v}_\mu)_\mathcal{V}$, where the second term is the so-called dual-based correction.

Proposition 5.2 (Goal-oriented case) *The following error bound holds: For all $\mu \in \mathcal{P}$,*

$$\left|Q(u) - \hat{Q}(\hat{u}_\mu)\right| \leq \mathcal{E}_1^{\mathrm{go}}(\mu) := \left(\tilde{\beta}_\mu^d\right)^{-1} \|G_\mu \hat{u}_\mu\|_\mathcal{V} \|G_\mu^d \hat{v}_\mu\|_\mathcal{V}, \tag{5.6}$$

where G_μ^d is the linear map from $\mathcal{V}$ to $\mathcal{V}$ such that $\mathcal{V} \ni v \mapsto G_\mu^d u := J\left(a_\mu(\cdot, v) - Q\right) \in \mathcal{V}$ and $\tilde{\beta}_\mu^d$ is a computable lower bound of $\beta_\mu^d = \inf_{u \in \mathcal{V}} \sup_{v \in \mathcal{V}} \frac{|a_\mu(v,u)|}{\|u\|_\mathcal{V} \|v\|_\mathcal{V}}$. Obviously, $\beta_\mu^d = \beta_\mu$ if a_μ is Hermitian.

Proof. See [17, Proposition 23]. ♢

In what follows, we mainly focus on the general-purpose case. Extensions to the goal-oriented case are straightforward.

5.2.4 Online-efficiency of the RB method

The notion of online-efficiency is central to the RB method.

Definition 5.3 *The RB method is said to be online-efficient if in the online stage, (i) the reduced problems can be constructed in complexity independent of N, and (ii) the error bound can be computed in complexity independent of N.*

Definition 5.4 *The sesquilinear form a_μ is said to depend on μ in an affine way if there exist d functions $\alpha_k(\mu) : \mathcal{P} \to \mathbb{C}$ and d μ-independent sesquilinear forms a_k bounded on $\mathcal{V} \times \mathcal{V}$ such that*

$$a_\mu(u, v) = \sum_{k=1}^{d} \alpha_k(\mu) a_k(u, v), \qquad \forall u, v \in \mathcal{V}. \tag{5.7}$$

In what follows, we always assume that the affine decomposition (5.7) holds. This decomposition is instrumental to achieve online-efficiency.

Property 5.5 *If a_μ depends on μ in an affine way, then the RB method is online-efficient.*

Proof. (i) The reduced matrix writes $(\hat{A}_\mu)_{j,i} = a_\mu(u_i, u_j)$ and the reduced right-hand side $(\hat{B})_j = b(u_j)$, for all $1 \leq i, j \leq \hat{N}$. There holds $\hat{A}_\mu = \sum_{k=1}^d \alpha_k(\mu) \hat{A}_k$, where $(\hat{A}_k)_{ij} := a_k(u_i, u_j)$. Therefore, provided the d matrices $\hat{A}_k$ and the vector $\hat{B}$ are precomputed during the offline stage, the reduced problems are constructed in complexity independent of N.

(ii) The operator G_μ inherits the affine dependence of a_μ on μ since, for all $u \in \mathcal{V}$,

$$G_\mu u = -Jb + \sum_{k=1}^{d} \alpha_k(\mu) J a_k(u,\cdot) = G_{00} + \sum_{k=1}^{d} \alpha_k(\mu) G_k u, \tag{5.8}$$

where $G_{00} := -Jb \in \mathcal{V}$ and $G_k u := Ja_k(u,\cdot) \in \mathcal{V}$ for all $k \in \{1,...,d\}$. Using this affine decomposition and recalling (5.3), we infer

$$\mathcal{E}_1(\mu) = \tilde{\beta}_\mu^{-1} \left\| G_{00} + \sum_{i=1}^{\hat{N}} \sum_{k=1}^{d} \alpha_k(\mu)\gamma_i(\mu) G_k u_i \right\|_{\mathcal{V}}. \tag{5.9}$$

The scalar product on which the norm in (5.9) hinges can be expanded to provide another formula for the error bound (see [86, eq.(4.61)]):

$$\begin{aligned} \mathcal{E}_2(\mu) = \tilde{\beta}_\mu^{-1} \Bigg(& (G_{00}, G_{00})_{\mathcal{V}} + 2\mathrm{Re} \sum_{i=1}^{\hat{N}} \sum_{k=1}^{d} \gamma_i(\mu)\alpha_k(\mu)(G_k u_i, G_{00})_{\mathcal{V}} \\ & + \sum_{i,j=1}^{\hat{N}} \sum_{k,l=1}^{d} \gamma_i(\mu)\alpha_k(\mu)\gamma_j^*(\mu)\alpha_l^*(\mu)(G_k u_i, G_l u_j)_{\mathcal{V}} \Bigg)^{\frac{1}{2}}, \end{aligned} \tag{5.10}$$

which is computed in complexity independent of N in the online stage provided that $(G_{00}, G_{00})_{\mathcal{V}}$, $(G_k u_i, G_{00})_{\mathcal{V}}$ and $(G_k u_i, G_l u_j)_{\mathcal{V}}$ are precomputed during the offline stage, and provided that a lower bound $\tilde{\beta}_\mu$ of the stability constant of a_μ is also computed in complexity independent of N (which is possible, for example, by the Successive Constraint Method, see [56, 30]). ♢

An important observation made in [Ar1], and that will be useful below, is that the formula (5.10) defining $\mathcal{E}_2$ can be rewritten in an equivalent way as

$$\mathcal{E}_2(\mu) := \tilde{\beta}_\mu^{-1} \left(\delta^2 + 2\mathrm{Re}(s^t \hat{x}_\mu) + \hat{x}_\mu^{*t} S \hat{x}_\mu \right)^{\frac{1}{2}}, \tag{5.11}$$

where $\delta := \|G_{00}\|_{\mathcal{V}}$, s and $\hat{x}_\mu$ are vectors in $\mathbb{C}^{d\hat{N}}$ with components $s_I := (G_k u_i, G_{00})_{\mathcal{V}}$ and $(\hat{x}_\mu)_I := \alpha_k(\mu)\gamma_i(\mu)$, and S is a matrix in $\mathbb{C}^{d\hat{N}, d\hat{N}}$ with coefficients $S_{I,J} := (G_k u_i, G_l u_j)_{\mathcal{V}}$ (with I and J re-indexing respectively (k,i) and (l,j), for all $1 \le k,l \le d$ and all $1 \le i,j \le \hat{N}$). The t superscript denotes the transposition. The vector s and the matrix S depend on the reduced basis functions $\{u_i\}_{1 \le i \le \hat{N}}$ but are independent of μ, and the vector $\hat{x}_\mu$ depends on the RB approximation $\hat{u}_\mu$ via the coefficients $\gamma_i(\mu)$. Notice that the term between parenthesis on the right-hand side of (5.11) is a multivariate polynomial in $\hat{x}_\mu$ of total degree 2. We would like to stress that $\mathcal{E}_1(\mu) = \mathcal{E}_2(\mu)$ (in infinite precision arithmetic): the indices 1 and 2 are used to denote two different ways to compute the same quantity. In particular, $\mathcal{E}_1(\mu)$ is not online efficient, while $\mathcal{E}_2(\mu)$ is.

5.2.5 The offline stage

Fix a discrete subset of parameters $\mathcal{P}_{\text{trial}} \subset \mathcal{P}$. In the offline stage, the parameters μ_i (from which the reduced basis is constructed) are chosen by a greedy algorithm as elements of $\mathcal{P}_{\text{trial}}$. We denote $\mathcal{P}_{\text{select}}$ the set of these selected parameters; see [86, Section 3.3] for a presentation

of the greedy algorithm. At each step of the algorithm, the new quantities $a_k(u_i, u_j)$ and $b(u_j)$ are computed and stored, as well as the new components of the vector s and of the matrix S to be used in the formula (5.11) for $\mathcal{E}_2$. This task, as that of evaluating G_{00}, typically requires inverting the stiffness matrix in $\mathcal{V}$ by solving, for all $k \in \{1, ..., d\}$ and all $i \in \{1, ..., \hat{N}\}$, the variational problem: find $w_{i,k} \in \mathcal{V}$ such that

$$E_{Gi,k} : (w_{i,k}, v)_{\mathcal{V}} = a_k(u_i, v), \qquad \forall v \in \mathcal{V}. \tag{5.12}$$

Then, $G_k u_i = w_{i,k}$ can be computed. The computation of $(G_k u_i, G_l u_j)_{\mathcal{V}}$ follows from the solutions of $E_{Gi,k}$ and $E_{Gj,l}$. Since the error bounds are evaluated using the formula $\mathcal{E}_2(\mu)$, for all $\mu \in \mathcal{P}_{\text{trial}}$, with the current state of the reduced basis, finding the maximum of the error bound on $\mathcal{P}_{\text{trial}}$ is of complexity independent of N. This allows one to consider very large sets $\mathcal{P}_{\text{trial}}$ without increasing too much the complexity of the whole offline procedure.

5.3 Round-off errors and online certification

In this section, we explain why the online-efficient error bound (5.11) may be sensitive to round-off errors.

5.3.1 Elements of floating-point arithmetic

In a computer, real numbers are represented by a finite number of bits, called floating-point representation. Current architectures are optimized for a format used by a large majority of softwares: IEEE 754 double-precision binary floating-point format. Let x be a real number. The floating point representation of x is denoted by $fl(x)$. When a (nonzero) real number is rounded to the closest floating-point number, the relative error on its floating-point representation is bounded by a number, ϵ, called the machine precision. In double precision, $\epsilon = 5 \times 10^{-16}$ (see [48, Section 1.2]). Let x and y be real numbers. When computing the operation $x + y$, the result returned by the computer can be different from its theoretical value. Whenever the difference is substantial, a loss of significance occurs. A well-known case of loss of significance is when x and y are almost opposite numbers. Suppose that $x = -y$. We denote by $\text{maxfl}(x + y)$ the result that the computer returns when the maximal accumulation of round-off errors occurs when computing the summation. There holds

$$|\text{maxfl}(x + y)| \approx 2\epsilon|x|. \tag{5.13}$$

When implementing an algorithm, one should ensure that each step is free of such a loss of significance. In some cases, simply changing the order of the operations can prevent these situations. As an illustration, consider $x = 1$, $y = 1 + 10^{-7}$, and the operation $x^2 - 2xy + y^2$. This is a sum of terms where the first intermediate result in the sum is 14 orders larger than the result. Therefore, a loss of significance is expected. The relative error of this computation is about 8×10^{-4}. Computing $(x - y)^2$, which is the factorization of the considered operation, leads to a relative error of about 10^{-9}. Thus, the terms of the sum are only 7 orders larger than the results, leading to a less catastrophic loss of significance. In this specific case, the remedy consists in carrying out the sum before the multiplication. In the RB context, the evaluation of the formula $\mathcal{E}_2$ suffers from such a loss of significance, as we now explain.

5.3.2 Validity of the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ for computing the error bound

Consider the two formulae $\mathcal{E}_1$, see (5.9), and $\mathcal{E}_2$, see (5.11), for computing the error bound.

Definition 5.6 *The formula $\mathcal{E}_k$, $k = 1, 2$, is said to be valid for computing the error bound with tolerance* tol *if*

$$\max_{\mu \in \mathcal{P}_{\text{select}}} (\mathcal{E}_k(\mu)) \leq \text{tol}. \tag{5.14}$$

From a theoretical viewpoint, the error $\|u_\mu - \hat{u}_\mu\|_{\mathcal{V}}$ and the residual $G_\mu u_\mu$ vanish for all $\mu \in \mathcal{P}_{\text{select}}$. Hence, any formula for computing the residual-based error bound vanishes as well and therefore is valid with any tolerance. However, the validity of a formula for computing the error bound is to be considered in the presence of some adverse phenomenon introducing errors in the computation, see Figure 5.1. The greedy algorithm in the offline stage stops when $\max_{\mu \in \mathcal{P}_{\text{trial}}} (\mathcal{E}_k(\mu)) < \text{tol}_{\text{RB}}$, where tol_{RB} denotes the maximum acceptable error made by the RB approximation. Therefore, if the minimum tolerance for which an error bound $\mathcal{E}_k$ is valid is larger than tol_{RB}, then the greedy algorithm cannot converge and will keep increasing the set $\mathcal{P}_{\text{select}}$ although the error can be actually very small.

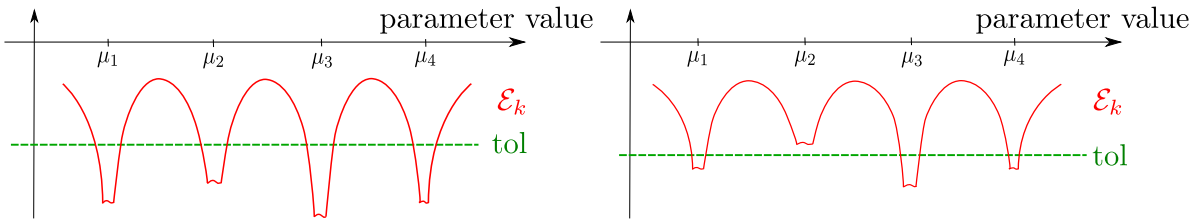

Fig. 5.1. Behavior of the formula $\mathcal{E}_k$ with respect to the parameter value (schematic illustration of Definition 5.6, with $\mathcal{P}_{\text{select}} = \{\mu_1, ..., \mu_4\}$). Left: the formula $\mathcal{E}_k$ is valid for computing the error bound with tolerance tol ; right: the formula is not valid as $\mathcal{E}_k(\mu_2) > \text{tol}$.

We examine the validity of the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ for computing the error bound in the presence of two independent phenomena: round-off errors and approximate reduced basis functions u_i (in the context of inexact linear algebra solvers for E_{μ_i}).

Round-off errors

We investigate the influence of round-off errors when computing the error bounds $\mathcal{E}_1(\mu)$ and $\mathcal{E}_2(\mu)$. As observed at the end of Section 5.3.1, the computation of a polynomial using a factorized form is more accurate than using the developed form, in particular at points close to its roots. Here, $\left(\tilde{\beta}_\mu \mathcal{E}_2(\mu)\right)^2$ is a multivariate polynomial of degree 2 in $\hat{x}_\mu$ computed in a developed form, whereas the scalar product $(G_\mu u_\mu, G_\mu u_\mu)_{\mathcal{V}}$ used in the computation of $\mathcal{E}_1(\mu)$ is not developed.

In this section, we neglect the round-off errors introduced when solving E_μ and $\hat{E}_\mu$, so that the reduced basis functions u_i and the reduced solutions $\hat{u}_\mu$ are considered free of round-off errors. We also suppose that the computable positive lower bound $\tilde{\beta}_\mu$ of the inf-sup constant is computed free of round-off errors, see Remark 5.9.

Proposition 5.7 *Let $\mu \in \mathcal{P}_{\text{select}}$ and let $\text{maxfl}(\tilde{\beta}_\mu \mathcal{E}_{\text{k}}(\mu))$, $k = 1, 2$, denote the evaluation of $\tilde{\beta}_\mu \mathcal{E}_k(\mu)$ when the maximum accumulation of round-off errors occurs. There holds*

$$\begin{aligned} &\text{maxfl}(\tilde{\beta}_\mu \mathcal{E}_1(\mu)) \geq 2\delta\epsilon, \\ &\text{maxfl}(\tilde{\beta}_\mu \mathcal{E}_2(\mu)) \geq 2\delta\sqrt{\epsilon}, \end{aligned} \tag{5.15}$$

where $\delta = \|G_{00}\|_{\mathcal{V}}$ and ϵ is the machine precision.

Proof. Let $\mu \in \mathcal{P}_{\text{select}}$. We present the proof for $\mathcal{E}_1(\mu)$; the proof for $\mathcal{E}_2(\mu)$ is similar. We need to evaluate the right-hand side of (5.9). Let $(\varphi_\rho)_{1 \leq \rho \leq N}$ denote the basis of $\mathcal{V}$, so that, for instance, $G_{00} = \sum_{\rho=1}^{N} (G_{00})_\rho \, \varphi_\rho$. In exact arithmetics, there holds $\mathcal{E}_1(\mu) = 0$, so that $\sum_{i=1}^{\hat{N}} \sum_{k=1}^{d} \gamma_i(\mu)\alpha_k(\mu) \, (G_k u_i)_\rho = -(G_{00})_\rho$ for all $1 \leq \rho \leq N$. As a result, using (5.13), we obtain

$$\left| \text{maxfl}\left((\text{G}_{00})_\rho + \sum_{\text{i}=1}^{\hat{\text{N}}} \sum_{\text{k}=1}^{\text{d}} \gamma_{\text{i}}(\mu)\alpha_{\text{k}}(\mu)(\text{G}_{\text{k}}\text{u}_{\text{i}})_\rho \right) \right| \approx 2|(G_{00})_\rho|\epsilon.$$

Since computing the $\mathcal{V}$-norm on the right-hand side of (5.9) can only increase the round-off errors, we infer the desired lower bound. ♢

Remark 5.8 (Validity of the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$) *We indeed observe in our simulations that the round-off errors on $\mathcal{E}_1$ scale like ϵ, while the round-off errors on $\mathcal{E}_2$ scale like $\sqrt{\epsilon}$ (see Section 5.4.3). The lower bounds in* (5.15) *are actually sharp in term of the scaling in ϵ. Then, if we suppose that the lower bounds are reached in* (5.15)*, the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ are valid for computing the error bound with tolerance* tol *if, respectively,*

$$\begin{aligned} &\text{for } \mathcal{E}_1, \qquad 2\left(\tilde{\beta}_{\min}\right)^{-1} \delta\epsilon \leq \text{tol}, \\ &\text{for } \mathcal{E}_2, \qquad 2\left(\tilde{\beta}_{\min}\right)^{-1} \delta\sqrt{\epsilon} \leq \text{tol}, \end{aligned} \tag{5.16}$$

where $\tilde{\beta}_{\min} = \inf_{\mu \in \mathcal{P}_{\text{select}}} (\tilde{\beta}_\mu)$.

Remark 5.9 (Inf-sup constant) *The computable positive lower bound $\tilde{\beta}_\mu$ of the inf-sup constant suffers from round-off errors as well. However, since it is a multiplicative factor, the quality of its computation does not severely affect the quality of the error bound. Moreover, the value of the inf-sup constant does not depend on the size of the reduced basis, contrary to $\|G_\mu \hat{u}_\mu\|_{\mathcal{V}}$. Therefore, there is no phenomenon susceptible to degrade the accuracy of its computation with the increase of the size of the reduced basis. If the Successive Constraint Method is used, the procedure to compute $\tilde{\beta}_\mu$ is carried out before the greedy algorithm of the RB method.*

Remark 5.10 (Improved floating-point arithmetic) *Increasing the machine precision from ϵ to ϵ^2 (quadruple-precision) for computing the coefficients in* (5.11)*, as well as for evaluating the multivariate polynomial in $\hat{x}_\mu$, is a first solution to recover a good precision with the formula $\mathcal{E}_2$. Moreover, since current architectures are optimized for the double-precision format, changing the floating-point arithmetic can potentially degrade software performance. There are also methods*

allowing one to double the precision of the evaluation of a polynomial while keeping the double-precision format, namely compensated schemes. For instance, the compensated Horner scheme in double-precision [62] doubles the precision and is faster than the full quadruple precision implementation. However, this requires to representing the result of the intermediate operations by two doubles, one for the value in double-precision and another one for the subsequent digits. These strategies are equivalent to quadruple precision (except for the computational savings in evaluating the error bound).

Remark 5.11 (Goal-oriented case, round-off errors) *The same analysis can be carried-out in the goal-oriented case. Let $\mu \in \mathcal{P}_{\text{select}}$. There holds*

$$\begin{aligned} &\operatorname{maxfl}(\tilde{\beta}_\mu^d \mathcal{E}_1^{\text{go}}(\mu)) \geq 2\delta\varsigma\epsilon^2, \\ &\operatorname{maxfl}(\tilde{\beta}_\mu^d \mathcal{E}_2^{\text{go}}(\mu)) \geq 2\delta\varsigma\epsilon, \end{aligned} \tag{5.17}$$

where $\varsigma := \|Q\|_{\mathcal{V}'}$. We indeed observe in our simulations that the round-off errors on $\mathcal{E}_1^{\text{go}}$ scale like ϵ^2, while the round-off errors on $\mathcal{E}_2^{\text{go}}$ scale like ϵ (see Section 5.5). If we suppose that the lower bounds are reached in (5.17)*, then the formulae $\mathcal{E}_1^{\text{go}}$ and $\mathcal{E}_2^{\text{go}}$ are valid for computing the error bound with tolerance* tol *if, respectively,*

$$\begin{aligned} &\text{for } \mathcal{E}_1^{\text{go}}, \qquad 2\left(\tilde{\beta}_{\min}^d\right)^{-1} \delta\varsigma\epsilon^2 \leq \text{tol}, \\ &\text{for } \mathcal{E}_2^{\text{go}}, \qquad 2\left(\tilde{\beta}_{\min}^d\right)^{-1} \delta\varsigma\epsilon \leq \text{tol}, \end{aligned} \tag{5.18}$$

where $\tilde{\beta}_{\min}^d = \inf_{\mu\in\mathcal{P}_{\text{select}}} (\tilde{\beta}_\mu^d)$.

Approximate reduced basis functions

In large-scale simulations, the accuracy of the RB procedure is also limited by the numerical method used for computing the reduced basis functions. We want here to illustrate this fact on a simple example where we suppose that the approximation of the reduced basis functions comes from an iterative solver with prescribed stopping criterion. We recall that for a given value $\mu \in \mathcal{P}_{\text{select}}$, E_μ consists in solving a linear system of size N of the form $A_\mu U_\mu = B$. Thus, for $\mu \in \mathcal{P}_{\text{trial}}$, the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ for the error bound are based on the computation of the residual of E_μ for the reduced solution $\hat{u}_\mu$. Indeed, it is easy to see that $\|G_\mu \hat{u}_\mu\|_{\mathcal{V}} = \|A_\mu \hat{U}_\mu - B\|_{*\mathcal{V}'}$, where for all $\Phi \in \mathbb{C}^N$, $\|\Phi\|_{*\mathcal{V}'} = \sup_{V\in\mathbb{C}^N} \frac{|(V,\Phi)_{\mathbb{C}^N}|}{\|\sum_{i=1}^N V_i\varphi_i\|_{\mathcal{V}}}$, recalling that $(\varphi_\rho)_{1\leq\rho\leq N}$ are the basis functions in $\mathcal{V}$, see [42, §9.1.5].

In this section, we suppose that the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ are free of round-off errors (therefore, for all $\mu \in \mathcal{P}_{\text{trial}}$, $\mathcal{E}_1(\mu) = \mathcal{E}_2(\mu)$), but the problem E_μ is not solved exactly, leading to approximate reduced basis functions such that the residuals do not vanish. Hence, for all $\mu \in \mathcal{P}_{\text{select}}$, $\mathcal{E}_1(\mu) = \mathcal{E}_2(\mu)$ and these error bounds are nonzero owing to inexact linear algebra solves. The reduced problems $\hat{E}_\mu$ are supposed to be solved freely of round-off errors.

Proposition 5.12 (Approximate reduced basis functions) *If the reduced basis functions are computed using an iterative solver with the following stopping criterion on the normalized residual:*

$$\forall \mu \in \mathcal{P}_{\text{trial}}, \qquad \frac{\|A_\mu U_\mu - B\|_{*\mathcal{V}'}}{\|B\|_{*\mathcal{V}'}} \leq \xi, \tag{5.19}$$

then the formulae $\mathcal{E}_1$ *and* $\mathcal{E}_2$ *are valid for computing the error bound with tolerance* tol *if*

$$\tilde{\beta}_{\min}^{-1} \delta \xi \leq \text{tol}. \tag{5.20}$$

Proof. Let $k \in \{1, 2\}$, let $\mu \in \mathcal{P}_{\text{select}}$ and suppose that the stopping criterion (5.19) is satisfied. Then, $\hat{u}_\mu = u_\mu$, but u_μ does not exactly solve E_μ. First, by definition of the $\|\cdot\|_{*\mathcal{V}}$ norm, $\|B\|_{*\mathcal{V}'} = \sup_{V \in \mathbb{C}^N} \frac{\left|b(\sum_{i=1}^N V_i \varphi_i)\right|}{\|\sum_{i=1}^N V_i \varphi_i\|_{\mathcal{V}}} = \|b\|_{\mathcal{V}'} = \|G_{00}\|_{\mathcal{V}} = \delta$. Then, $\|G_\mu \hat{u}_\mu\|_{\mathcal{V}} = \sup_{v \in \mathcal{V}} \frac{(G_\mu \hat{u}_\mu, v)_{\mathcal{V}}}{\|v\|_{\mathcal{V}}} = \sup_{v \in \mathcal{V}} \frac{a_\mu(\hat{u}_\mu, v) - b(v)}{\|v\|_{\mathcal{V}}} = \sup_{V \in \mathbb{C}^N} \frac{(V, A_\mu \hat{U}_\mu - B)_{\mathbb{C}^N}}{\|\sum_{i=1}^N V_i \varphi_i\|_{\mathcal{V}}} = \|A_\mu \hat{U}_\mu - B\|_{*\mathcal{V}'}$. Therefore,

$$\mathcal{E}_k(\mu) = \tilde{\beta}_\mu^{-1} \|G_\mu \hat{u}_\mu\|_{\mathcal{V}} = \tilde{\beta}_\mu^{-1} \|A_\mu \hat{U}_\mu - B\|_{*\mathcal{V}'} = \tilde{\beta}_\mu^{-1} \|A_\mu U_\mu - B\|_{*\mathcal{V}'} \leq \tilde{\beta}_\mu^{-1} \|B\|_{*\mathcal{V}'} \xi = \tilde{\beta}_\mu^{-1} \delta \xi \leq \tilde{\beta}_{\min}^{-1} \delta \xi.$$

Hence, if $\tilde{\beta}_{\min}^{-1} \delta \xi \leq \text{tol}$, the validity of $\mathcal{E}_1$ and $\mathcal{E}_2$ follows from Definition 5.6. ♢

Since the $\|\cdot\|_{*\mathcal{V}'}$ norm is hard to compute, the stopping criterion (5.19) uses in practice the Hermitian norm in $\mathbb{C}^N$ or the $\mathcal{V}$-norm of the corresponding functions in $\mathcal{V}$.

Remark 5.13 (Goal-oriented case, approximate reduced basis functions) *The formulae* $\mathcal{E}_1^{\text{go}}$ *and* $\mathcal{E}_2^{\text{go}}$ *are valid for computing the error bound with tolerance* tol *if* $\left(\tilde{\beta}_{\min}^d\right)^{-1} \delta \gamma \xi^2 \leq \text{tol}$.

Synthesis

Taking into account the round-off errors in the computation of the error bound and the stopping criterion of an iterative solver, and supposing that the bounds (5.15) and (5.17) are reached, the formulae $\mathcal{E}_1$ and $\mathcal{E}_2$ are valid for computing the error bound with tolerance tol if, respectively,

$$\begin{aligned} &\text{for } \mathcal{E}_1, \qquad 2\tilde{\beta}_{\min}^{-1} \delta \max(\xi, \epsilon) \leq \text{tol}, \\ &\text{for } \mathcal{E}_2, \qquad 2\tilde{\beta}_{\min}^{-1} \delta \max(\xi, \sqrt{\epsilon}) \leq \text{tol}, \end{aligned} \tag{5.21}$$

and the formulae $\mathcal{E}_1^{\text{go}}$ and $\mathcal{E}_2^{\text{go}}$ are valid for computing the error bound with tolerance tol if, respectively,

$$\begin{aligned} &\text{for } \mathcal{E}_1^{\text{go}}, \qquad 2\left(\tilde{\beta}_{\min}^d\right)^{-1} \delta \gamma \max\left(\xi^2, \epsilon^2\right) \leq \text{tol}, \\ &\text{for } \mathcal{E}_2^{\text{go}}, \qquad 2\left(\tilde{\beta}_{\min}^d\right)^{-1} \delta \gamma \max\left(\xi^2, \epsilon\right) \leq \text{tol}. \end{aligned} \tag{5.22}$$

Focusing on round-off errors, the formula $\mathcal{E}_1$ for computing the error bound is valid for tolerances scaling as ϵ, but is not online-efficient, whereas the formula $\mathcal{E}_2$ is online-efficient but is valid only for (significantly) higher tolerances, namely tolerances scaling as $\sqrt{\epsilon}$.

5.4 New procedures for accurate and online-efficient evaluation of the error bound

In this section, online-efficient methods, that are valid for tolerances scaling as ϵ, are devised to evaluate the error bound.

5.4.1 Procedure 1: rewriting $\mathcal{E}_2$

We first present the procedure proposed in [Ar1]. We consider that a reduced basis of size $\hat{N}$ has been constructed. Let $\sigma := 1 + 2d\hat{N} + (d\hat{N})^2$. For a given $\mu \in \mathcal{P}_{\text{trial}}$ and the resulting $\hat{u}_\mu \in \text{Span}\{u_1, ..., u_{\hat{N}}\}$ solving the reduced problem, we define $\hat{X}(\mu) \in \mathbb{C}^\sigma$ as the vector with components $(1, \hat{x}_{\mu_I}, \hat{x}^*_{\mu_I}, \hat{x}^*_{\mu_I}\hat{x}_{\mu_J})$, where $\hat{x}_{\mu_I} = \alpha_k(\mu)\gamma_i(\mu)$ (we recall that $\gamma_i(\mu)$ are the coefficients of the reduced solution in the reduced basis, see (5.3), and $\alpha_k(\mu)$ the coefficients of the affine decomposition of a_μ in (5.7)), with $1 \leq I, J \leq d\hat{N}$ (with $I = i + \hat{N}(k-1)$ such that $1 \leq i \leq \hat{N}$, $1 \leq k \leq d$, and with $J = j + \hat{N}(l-1)$ such that $1 \leq j \leq \hat{N}$, $1 \leq l \leq d$). We can write the right-hand side of (5.11) as a linear form in $\hat{X}(\mu)$ as follows:

$$\delta^2 + 2\text{Re}(s^t\hat{x}_\mu) + \hat{x}^{*t}_\mu S\hat{x}_\mu = \sum_{p=1}^{\sigma} t_p\hat{X}_p(\mu), \tag{5.23}$$

where t_p is independent of μ (as δ, s, and S are independent of μ) and $\hat{X}_p(\mu)$ is the p-th component of $\hat{X}(\mu)$.

Now, in the offline stage, we take σ values (e.g. random values) $\mu_r \in \mathcal{P}_{\text{trial}}$, $r \in \{1, ..., \sigma\}$, of the parameter μ. Then, we compute the vectors $\hat{X}(\mu_r)$ and the quantities

$$V_r := \sum_{p=1}^{\sigma} t_p\hat{X}_p(\mu_r). \tag{5.24}$$

Finally, we define $T \in \mathbb{C}^{\sigma\times\sigma}$ as the matrix whose columns are formed by the vectors $\hat{X}(\mu_r)$, that is, $T_{pr} = \hat{X}_p(\mu_r)$ for all $1 \leq p, r \leq \sigma$. We assume that T is invertible, which always happens to be the case in our simulations.

Now, suppose that in the online stage we want to evaluate the error bound for the RB solution $\hat{u}_\mu$ computed at a certain parameter $\mu \in \mathcal{P}_{\text{trial}}$. Then, we evaluate the vector $\hat{X}(\mu)$ and solve the linear system

$$T\lambda(\mu) = \hat{X}(\mu), \tag{5.25}$$

yielding $\lambda(\mu) \in \mathbb{C}^\sigma$. We then obtain $\hat{X}(\mu) = \sum_{r=1}^{\sigma} \lambda_r(\mu)\hat{X}(\mu_r)$ and

$$\sum_{p=1}^{\sigma} t_p\hat{X}_p(\mu) = \sum_{p,r=1}^{\sigma} t_p\lambda_r(\mu)\hat{X}_p(\mu_r) = \sum_{r=1}^{\sigma} \lambda_r(\mu)V_r. \tag{5.26}$$

This yields the following new formula for computing the error bound:

$$\mathcal{E}_3(\mu) := \tilde{\beta}_\mu^{-1} \left(\sum_{r=1}^{\sigma} \lambda_r(\mu)V_r \right)^{\frac{1}{2}}, \tag{5.27}$$

where the quantities $V_r = \|G_{\mu_r}\hat{u}_{\mu_r}\|^2_{\mathcal{V}}$ can be precomputed. Thus, computing $\mathcal{E}_3$ requires solving (5.25) and summing the σ precomputed quantities V_r. Since the complexity of this procedure is independent of N, the formula $\mathcal{E}_3$ is online-efficient for computing the error bound. Notice that in the linear combination in (5.27), the V_r are expected to have the same magnitude as the result of the linear combination, preventing this formula from the loss of significance observed for $\mathcal{E}_1$.

Remark 5.14 (Goal-oriented case) *For the goal-oriented case, the procedure is carried out independently on the two multivariate polynomials* $\|G_\mu \hat{u}_\mu\|_{\mathcal{V}}^2$ *and* $\|G_\mu^d \hat{v}_\mu\|_{\mathcal{V}}^2$.

Notice that $\mathcal{E}_1(\mu)$, $\mathcal{E}_2(\mu)$, and $\mathcal{E}_3(\mu)$ are equal in exact arithmetic. As pointed out in [Ar1], the matrix T exhibits in practice large condition numbers, and there is no guarantee that T is actually invertible. We will see in Section 5.5 for a three-dimensional acoustic scattering problem that $\mathcal{E}_3$ can be in practice as ill-behaved as $\mathcal{E}_2$. Moreover, there is no a priori method for selecting the parameters μ_r for which the quantities V_r are precomputed. In the next section, we propose a new procedure that solves these problems.

5.4.2 Procedure 2: improvement on Procedure 1 using the EIM

In the formula $\mathcal{E}_3$, a potentially ill-conditioned problem $T\lambda(\mu) = \hat{X}(\mu)$ is solved in order to exactly represent $\hat{X}(\mu)$ by the linear combination $\sum_{r=1}^{\sigma} \lambda_r(\mu)\hat{X}(\mu_r)$. Following a suggestion by Patera [84], we propose to approximate $\hat{X}(\mu)$ by means of an interpolation procedure. We want to modify the formula $\mathcal{E}_3$ by an interpolation formula relying on a better conditioned linear system. The price to pay is that the new formula $\mathcal{E}_4$ will not be equal to $\mathcal{E}_1$ in exact arithmetic; the interpolation errors are however marginal, as further discussed in Remark 5.20. We also look for a way to choose the parameters μ_r for which the quantities V_r have to be precomputed. We refer to these values for μ_r as "interpolation points", and to the set of these points as $\mathcal{P}_{\text{inter}}$.

Consider the function of two variables $(p, \mu) \mapsto \hat{X}_p(\mu)$, for all $p \in \{1, ..., \sigma\}$ and all $\mu \in \mathcal{P}_{\text{trial}}$. We look for an approximation of this function in the form

$$\forall \mu \in \mathcal{P}_{\text{trial}}, \forall p \in \{1, ..., \sigma\},\ \hat{X}_p(\mu) \approx \sum_{r=1}^{\hat{\sigma}} \lambda_r^{\hat{\sigma}}(\mu)\hat{X}_p(\mu_r), \tag{5.28}$$

for a certain parameter $\hat{\sigma} \leq \sigma$. The empirical interpolation method (EIM) (more precisely the discrete EIM since p is a discrete variable) provides a numerical procedure to construct this approximation and to choose the interpolation points (see [6, 73]).

The EIM is an offline-online procedure. During the offline stage, $\hat{\sigma}$ basis functions are computed, denoted $q_j : \mathcal{P}_{\text{trial}} \ni \mu \mapsto q_j(\mu) \in \mathbb{C}$, for all $j \in \{1, ..., \hat{\sigma}\}$. These basis functions will be used in the online stage to carry out the interpolation. We define $q^{\hat{\sigma}}$ as the vector-valued map $\mathcal{P}_{\text{trial}} \ni \mu \mapsto q^{\hat{\sigma}}(\mu) := (q_j(\mu))_{1 \leq j \leq \hat{\sigma}} \in \mathbb{C}^{\hat{\sigma}}$. During the offline stage, $\hat{\sigma}$ interpolation points $\mu_r \in \mathcal{P}_{\text{trial}}$ are also selected; these points are collected in the set $\mathcal{P}_{\text{inter}}$. Notice that $\mathcal{P}_{\text{select}}$, the set of parameter values selected by the greedy algorithm of the RB method, is different from $\mathcal{P}_{\text{inter}}$. During the online stage, the matrix $B^{\hat{\sigma}} \in \mathbb{C}^{\hat{\sigma},\hat{\sigma}}$, where $B_{ij}^{\hat{\sigma}} = q_i(\mu_j)$, for $1 \leq i, j \leq \hat{\sigma}$, is constructed. Letting $\mu \in \mathcal{P}_{\text{trial}}$, we solve for $\lambda^{\hat{\sigma}}(\mu) \in \mathbb{C}^{\hat{\sigma}}$ such that

$$B^{\hat{\sigma}} \lambda^{\hat{\sigma}}(\mu) = q^{\hat{\sigma}}(\mu), \tag{5.29}$$

and compute the rank-$\hat{\sigma}$ interpolation operators defined as follows.

Definition 5.15 *Let* $1 \leq k \leq \hat{\sigma}$. *The rank-k interpolation operator* I^k *is defined such that*

$$I^k \hat{X}(\mu) := \sum_{r=1}^{k} \lambda_r^k(\mu)\hat{X}(\mu_r), \tag{5.30}$$

where $\lambda^k(\mu) \in \mathbb{C}^k$ *solves*

$$B^k \lambda^k(\mu) = q^k(\mu). \tag{5.31}$$

Equation (5.30) defines an interpolation in the sense that $I^k \hat{X}_{p_r}(\mu) = \hat{X}_{p_r}(\mu)$ for all $1 \leq r \leq k$ and all $\mu \in \mathcal{P}_{\text{trial}}$. The formula $\hat{X}_p(\mu) \approx (I^{\hat{\sigma}} \hat{X})_p(\mu)$, for all $\mu \in \mathcal{P}_{\text{trial}}$ and all $p \in \{1, ..., \sigma\}$, provides the approximate interpolation formula searched for in (5.28).

Definition 5.16 *The residual operator* $\delta^{\hat{\sigma}}$ *is defined by*

$$\delta^{\hat{\sigma}} := \text{Id} - I^{\hat{\sigma}}. \tag{5.32}$$

Algorithm 2 presents the construction of the function $q^{\hat{\sigma}}$ by a greedy algorithm during the offline stage. This EIM algorithm is a variant from the classical one, described in [73]. The differences stand in the definition of the interpolation operator (5.29), the linear system (5.31) to solve during the online calls, and the definition of the B^k matrix. In particular, the present variant leads to the approximation (5.30), which is nonintrusive in the sens that $I^k \hat{X}(\mu)$ is obtained as a linear combination of evaluations of $\hat{X}$ at some parameter values μ_r. The classical EIM can recover such a property, but to the price of an additional change of basis between $q_k(\cdot)$ and $\hat{X}_{p_k}(\cdot)$. However, contrary to the classical EIM, the variant needs the additional change of basis to be able to compute an approximation between learning points, namely for $\mu \in \mathcal{P}_{\text{trial}} \backslash \mathcal{P}$. We refer to Section 7.2 for more details about the differences between the EIM variant considered here and the classical algorithm.

Algorithm 2 Offline stage of the EIM

1. Choose $\hat{\sigma} > 1$ [Number of interpolation points]
2. Set $k := 1$
3. Compute $p_1 := \underset{p \in \{1,\dots,\sigma\}}{\text{argmax}} \|\hat{X}_p(\cdot)\|_{\ell^\infty(\mathcal{P}_{\text{trial}})}$
4. Compute $\mu_1 := \underset{\mu \in \mathcal{P}_{\text{trial}}}{\text{argmax}} |\hat{X}_{p_1}(\mu)|$ and set $\mathcal{P}_{\text{inter}} = \{\mu_1\}$ [First interpolation point]
5. Set $q_1(\cdot) := \dfrac{\hat{X}_{p_1}(\cdot)}{\hat{X}_{p_1}(\mu_1)}$ [First basis function]
6. Set $B^1_{11} := 1$ [Initialize B matrix]
7. **while** $k < \hat{\sigma}$ **do**
8. Compute $p_{k+1} := \underset{p \in \{1,\dots,\sigma\}}{\text{argmax}} \|(\delta^k \hat{X})_p(\cdot)\|_{\ell^\infty(\mathcal{P}_{\text{trial}})}$
9. Compute $\mu_{k+1} := \underset{\mu \in \mathcal{P}_{\text{trial}}}{\text{argmax}} |(\delta^k \hat{X})_{p_{k+1}}(\mu)|$ [$(k+1)$-th interpolation point]
10. Set $\mathcal{P}_{\text{inter}} := \mathcal{P}_{\text{inter}} \cup \{\mu_{k+1}\}$ [Update of $\mathcal{P}_{\text{inter}}$]
11. Set $q_{k+1}(\cdot) := \dfrac{(\delta^k \hat{X})_{p_{k+1}}(\cdot)}{(\delta^k \hat{X})_{p_{k+1}}(\mu_{k+1})}$ [$(k+1)$-th basis function]
12. $B^{k+1}_{ij} := q_i(\mu_j)$, $1 \leq i, j \leq k+1$ [$(k+1)$-th B matrix]
13. $k \leftarrow k+1$ [Increment the size of the interpolation]
14. **end while**

Definition 5.17 *The new formula for computing the error bound is*

$$\mathcal{E}_4(\mu) := \tilde{\beta}_\mu^{-1} \left(\sum_{r=1}^{\hat{\sigma}} \lambda_r^{\hat{\sigma}}(\mu) V_r \right)^{\frac{1}{2}}, \tag{5.33}$$

where $\lambda^{\hat{\sigma}}(\mu)$ *is the solution to* (5.29)*. We recall that* $V_r = \|G_{\mu_r}\hat{u}_{\mu_r}\|^2_{\mathcal{V}}$.

Proposition 5.18 *The computation of the formula* $\mathcal{E}_4$ *is well defined, and this formula is online-efficient.*

Proof. Owing to [73, Theorem 1], the matrix B is upper triangular with diagonal unity. Hence, $\det B = 1$ and B is guaranteed to be invertible. The online procedure of EIM, consisting in solving a linear system defined by the matrix B, is thus well defined. Then, since the EIM procedure in carried out on $\hat{X}_p(\mu)$, for all $p \in \{1, ..., \sigma\}$ and all $\mu \in \mathcal{P}_{\text{trial}}$, all the computations involved are of complexity independent of N, even the offline part of the EIM. Finally, the complexity of the online part of EIM only depends on $\hat{\sigma}$. ♢

Remark 5.19 (Stopping criterion in Algorithm 2) *For ease of presentation, we chose a simple stopping criterion based on an a priori fixed maximum number of interpolation points. In practice, one possibility is to stop the algorithm when the maximal approximation error in the EIM is below a prescribed value, by monitoring the quantity* $(\delta^k \hat{X})_{p_{k+1}}(\mu_{k+1})$.

Remark 5.20 (Interpolation errors) *As already observed,* $\mathcal{E}_4$ *does not equal* $\mathcal{E}_1$ *in exact arithmetics owing to interpolation errors (when* $\hat{\sigma} < \sigma$*). Thus, although Algorithm 2 yields an accurate approximation of* $\hat{X}_p(\mu)$*, a given interpolation error on* $\hat{X}_p(\mu)$ *does not directly translate into a bound on the difference between* $\mathcal{E}_1(\mu)$ *and* $\mathcal{E}_4(\mu)$ *(the latter depending also on* δ*,* s*, and* S*, as well as on* $\tilde{\beta}_\mu$*). We observe in our numerical experiments that these latter errors are lower than the errors incurred in the evaluation of* $\mathcal{E}_2$ *(due to round-off errors) and in the evaluation of* $\mathcal{E}_3$ *(due to the poor conditioning of* T*).*

Remark 5.21 (Non affine dependence) *When the affine dependence assumption is not available (see Definition 5.4), one can look for an approximation of* a_μ *in the following form:*

$$a_\mu(u, v) \approx \sum_{k=1}^{d} \alpha_k(\mu) a_k(u, v), \qquad \forall u, v \in \mathcal{V}. \tag{5.34}$$

In the reduced basis context, this approximation is usually computed using the EIM. We saw that the formula (5.10) *for* $\mathcal{E}_2$ *makes use of this affine decomposition to ensure online efficiency, and therefore does not account for the approximation in the operator. On the contrary, the formulae* (5.4) *for* $\mathcal{E}_1$ *and* (5.27) *for* $\mathcal{E}_3$ *use the exact operator.*

5.4.3 Illustration

Consider as in [Ar1] a one-dimensional linear diffusion problem, namely the boundary value problem $-u'' + \mu u = 1$ on $]0, 1[$ with $u(0) = u(1) = 0$, with parameter $\mu \in \mathcal{P} := [1, 100]$. The analytic solution is

$$u(x) = -\frac{1}{\mu}\left(\cosh\left(\sqrt{\mu}x\right) - 1\right) + \frac{\cosh\left(\sqrt{\mu}\right) - 1}{\mu \sinh\left(\sqrt{\mu}\right)} \sinh\left(\sqrt{\mu}x\right). \tag{5.35}$$

The Lax–Milgram theory is valid, and the coercivity constant is bounded from below by 1 in the H^1-norm. The error bound is given by $\mathcal{E}_1(\mu) = \|G_\mu \hat{u}_\mu\|_{H^1(]0,1[)}$. Lagrange $\mathbb{P}_1$ finite elements are used with uniform mesh cells of length 0.005. The set $\mathcal{P}_{\text{trial}}$ consists of 1000 points uniformly distributed in $\mathcal{P}$. The RB method is carried out until the formula $\mathcal{E}_2$ suffers from round-off errors, which already happens for a reduced basis of size $\hat{N} = 7$ (since $d = 2$, we obtain $\sigma = 225$). A direct solver is used, so that the only adverse phenomenon to compute the error bound are round-off errors.

In Figure 5.2, we see that the classical formula $\mathcal{E}_2$ is not valid for computing the error bound with any tolerance below 10^{-7}, whereas the formulae $\mathcal{E}_1$, $\mathcal{E}_3$ and $\mathcal{E}_4$ are valid with tolerances down to 10^{-14}. The difference is of 7 orders of magnitude ; given that $\sqrt{\epsilon} \approx 10^{-7}$, this is consistent with Remark 5.8 and Section 5.4.1.

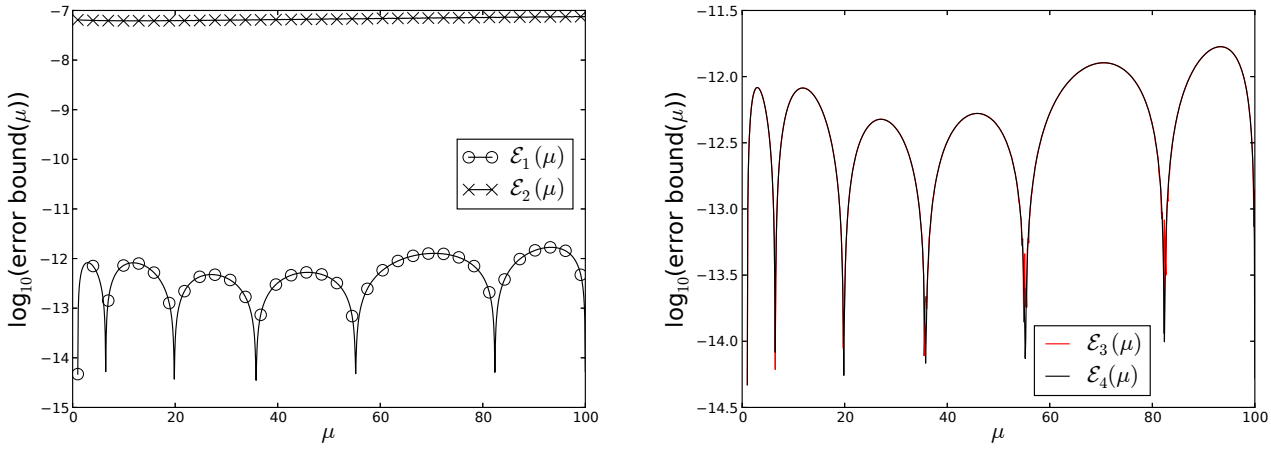

Fig. 5.2. Error bound curves with respect to the parameter. The formula $\mathcal{E}_4$ is computed with $\hat{\sigma} = 23$.

In Figure 5.3, we observe that instabilities occur in the formula $\mathcal{E}_3$, especially for parameter values close to the elements of $\mathcal{P}_{\text{select}}$. This is due to the poor conditioning of the matrix T when solving (5.25). The new formula $\mathcal{E}_4$ based on the EIM is seen to introduce much less numerical errors than $\mathcal{E}_3$.

5.4.4 Procedure 3: improvement of Procedure 2 using a stabilized EIM

In practice, round-off errors are accumulated during the loop in Algorithm 2, and if we keep increasing the number of interpolation points, the coefficients of the matrix B suffer from round-off errors, so that the relation $\det(B) = 1$ no longer holds. Even worse, the matrix B becomes non invertible at some stage. To solve this problem, we now propose a numerical stabilization of EIM based on the following property:

Property 5.22 *There holds*

$$\forall i < j, \ I^j \circ I^i = I^i, \tag{5.36}$$

where the interpolation operators I^j are defined by (5.30).

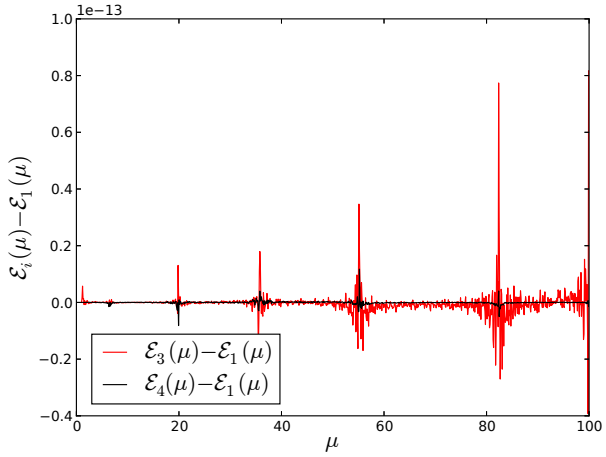

Fig. 5.3. Comparison of the formulae $\mathcal{E}_3$ and $\mathcal{E}_4$, with respect to the formula $\mathcal{E}_1$.

Proof. Using [73, Lemma 1], $I^i\hat{X} \in \mathrm{Span}\,(q_1, ..., q_i)$ and $I^i v = v$ for all $v \in \mathrm{Span}\,(q_1, ..., q_i)$. Therefore, $I^j \circ I^i\hat{X} = I^i\hat{X}$ for all $i < j$. ♢

In our numerical experiments, we observe that, as the number of iterations of the greedy procedure for the EIM grows, the relation (5.36) is no longer verified numerically, due to accumulation of round-off errors. These numerical instabilities can be compensated in the same fashion as the Gram–Schmidt orthonormalization procedure is stabilized (see [50, chapter 5.2.8]). The Gram–Schmidt algorithm transforms a linearly independent family of vectors $\{v_i\}$ into an orthonormal basis $\{u_i\}$. To simplify the presentation, we suppose in what follows that the normalization step is not carried out. Consider the orthogonalization step for the k-th vector. We denote by Π^k the projection operator on $\mathrm{Span}(u_1, ..., u_k)$, and $\delta^k := \mathrm{Id} - \Pi^k$. For the EIM, we suppose that $(k-1)$ interpolation operators I^i, $1 \leq i \leq k-1$, have been constructed, and we wish to construct the k-th interpolation operator I^k. A comparison between the stabilized Gram–Schmidt orthonormalization procedure and the proposed stabilization for the EIM is presented in Table 5.1.

Proposition 5.23 *Let $k \in \mathbb{N}^*$. In exact arithmetic, the following relations hold for the residuals defined in Table 5.1: $\delta^k_{\mathrm{stab}} v = \delta^k v$.*

Proof. We prove by recursion that, for all $i \leq k$, $\delta^{k,i}_{\mathrm{stab}} = \delta^i$. The case $i = 1$ is clear from the definition of the first intermediate residual in Table 5.1. Let $i \leq k$ and suppose that $\delta^{k,i-1}_{\mathrm{stab}} = \mathrm{Id} - I^{i-1}$ for the EIM. There holds

$$\delta^{k,i}_{\mathrm{stab}} = \delta^{k,i-1}_{\mathrm{stab}} - I^i \circ \delta^{k,i-1}_{\mathrm{stab}} = \mathrm{Id} - I^{i-1} - I^i + I^i \circ I^{i-1} = \mathrm{Id} - I^i = \delta^i, \tag{5.37}$$

since $I^i \circ I^{i-1} = I^{i-1}$ owing to Property 5.22. The results follow from the case $i = k$. The same relation is proved likewise for the Gram–Schmidt procedure, for which $\Pi^i \circ \Pi^{i-1} = \Pi^{i-1}$ holds as well. ♢

	stabilized Gram–Schmidt	stabilized EIM
global input	$(v_1, ..., v_{\hat{\sigma}})$ basis of $\mathbb{C}^{\hat{\sigma}}$	$v : \mathcal{P}_{\text{trial}} \to \mathbb{C}^{\hat{\sigma}}$
classical residual at step k	$\delta^k v_k = v_k - \Pi^k v_k$	$(\delta^k v)(\mu) = v(\mu) - (I^k v)(\mu)$
intermediate residuals at step k	$\delta^{k,1}_{\text{stab}} v_k = v_k - \Pi^1 v_k$ $\delta^{k,2}_{\text{stab}} v_k = \delta^{k,1}_{\text{stab}} v_k - \Pi^2 \delta^{k,1}_{\text{stab}} v_k,$ $\vdots$ $\delta^{k,k}_{\text{stab}} v_k = \delta^{k,k-1}_{\text{stab}} v_k - \Pi^k \delta^{k,k-1}_{\text{stab}} v_k$	$(\delta^{k,1}_{\text{stab}} v)(\mu) = v(\mu) - (I^1 v)(\mu)$ $(\delta^{k,2}_{\text{stab}} v)(\mu) = (\delta^{k,1}_{\text{stab}} v)(\mu) - I^2(\delta^{k,1}_{\text{stab}} v)(\mu),$ $\vdots$ $(\delta^{k,k}_{\text{stab}} v)(\mu) = (\delta^{k,k-1}_{\text{stab}} v)(\mu) - I^k(\delta^{k,k-1}_{\text{stab}} v)(\mu)$
stabilized residual at step k	$\delta^k_{\text{stab}} v_k = \delta^{k,k}_{\text{stab}} v_k$	$(\delta^k_{\text{stab}} v)(\mu) = (\delta^{k,k}_{\text{stab}} v)(\mu)$
global output	$(\delta^1_{\text{stab}} v_1, \delta^2_{\text{stab}} v_2, ..., \delta^{\hat{\sigma}}_{\text{stab}} v_{\hat{\sigma}})$ orthogonal basis of $\text{Span}(v_1, ..., v_{\hat{\sigma}})$	$(I^{\hat{\sigma}} v)(\mu)$

Table 5.1. Comparison between stabilized Gram–Schmidt and stabilized EIM.

Definition 5.24 (Stabilized EIM) *The stabilized EIM consists in the same offline procedure as the one described in Section 5.4.2, except that the residuals δ^k are replaced by the stabilized residuals δ^k_{stab} defined in Table 5.1. The online stage is the same as that of the classical EIM.*

The stabilized Gram–Schmidt procedure generates a set of vectors much less polluted by round-off errors (see [16, 46]). By analogy, we expect that the stabilized EIM produces a more accurate interpolation procedure than the classical EIM, that is, much less polluted by round-off errors. This is numerically verified in Figure 5.4, where $\det(B^{\hat{\sigma}})$ and $\text{cond}(B^{\hat{\sigma}})$ are represented as a function of $\hat{\sigma}$. We consider the test case described in Section 5.4.3, where we recall that $\hat{N} = 7$, $d = 2$, and $\sigma = 225$. If the method is stable, then $\det(B^{\hat{\sigma}}) = 1$ should hold throughout the process. Figure 5.4 shows that the stabilized EIM behaves as intended. The classical EIM curve stops since the matrix $B^{\hat{\sigma}}$ becomes noninvertible at some point: a parameter already in $\mathcal{P}_{\text{inter}}$ has been selected by the greedy algorithm. Invertibility can be recovered artificially by ensuring that the new interpolation point is not an element of the current set $\mathcal{P}_{\text{inter}}$. We call this procedure EIM with unique choice. However, this fix is not completely satisfactory, since $\det(B^{\hat{\sigma}}) = 1$ is not satisfied. Moreover, $\text{cond}(B^{\hat{\sigma}})$ is much more ill-behaved with this procedure than with the stabilized EIM.

Remark 5.25 (Computational cost and variant of stabilized EIM) *The computational cost of the stabilized EIM is more than that of the classical EIM, since the stabilized residual requires as many calls to a classical residual as the number of selected interpolation points (i.e. the scaling with $\hat{\sigma}$ is $\hat{\sigma}^2$ for the stabilized EIM as opposed to $\hat{\sigma}$ for the classical EIM). One can think of a cheaper procedure by monitoring $\det(B^{\hat{\sigma}})$ and adding some intermediate residuals $\delta^{k,j}_{\text{stab}}$ until $\det(B^{\hat{\sigma}})$ is close enough to 1.*

5.4.5 Summary

The advantages and drawbacks of the four considered formulae for computing the error bound are summarized in Table 5.2. To estimate the computational complexity of the methods,

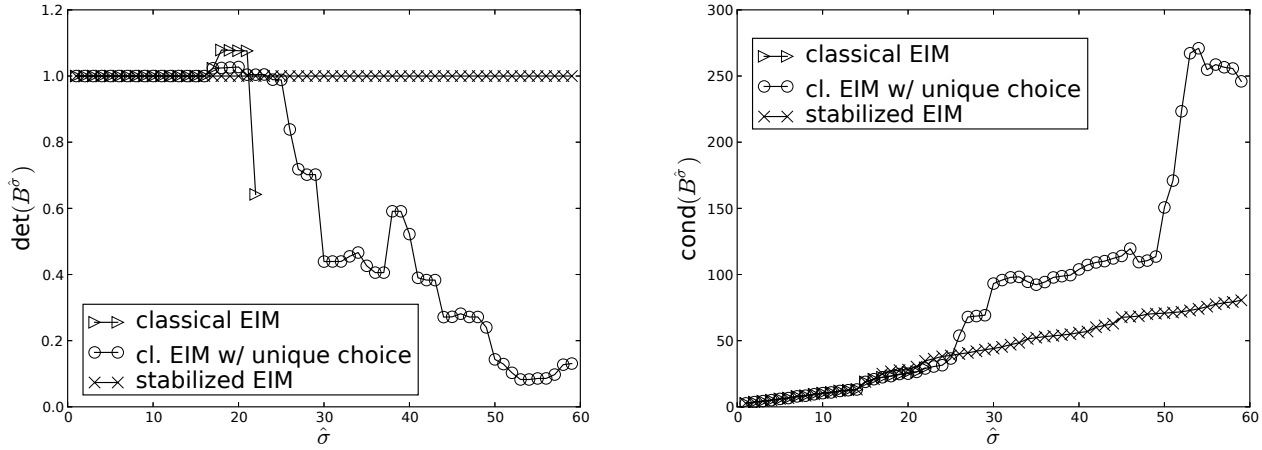

Fig. 5.4. Determinant (left) and condition number (right) of the matrix $B^{\hat{\sigma}}$ as a function of $\hat{\sigma}$, for the classical EIM, the classical EIM with unique choice, and the stabilized EIM. The classical EIM curves stop at 21 interpolation points since $B^{\hat{\sigma}}$ becomes non invertible at 22 points.

we keep only the leading order in operation count. We denote the complexity of the resolution of (5.12) by N_{sol}. The linear systems of size σ, $\hat{\sigma}$, and $\hat{N}$ are supposed to be solved by a direct solver, hence with complexity proportional to σ^3, $\hat{\sigma}^3$, and $\hat{N}^3$, respectively. For the offline stage of $\mathcal{E}_2$ and $\mathcal{E}_3$, we have to evaluate respectively $d\hat{N}+1$ and σ times the functional G_μ, which requires to solve (5.12). For the offline stage of $\mathcal{E}_4$, let M denote the cardinality of $\mathcal{P}_{\text{trial}}$. The k-loop in Algorithm 2 requires at each step to compute a maximum over σ different $\ell^\infty(\mathcal{P}_{\text{trial}})$ norms, and then to solve a linear system of size k, leading to a complexity of $\hat{\sigma}^4\sigma M + \hat{\sigma}N_{\text{sol}}$. If the stabilized EIM is used instead for $\mathcal{E}_4$, each residual evaluation in the k-loop requires solving k linear systems of size 1 to k, leading to a complexity of $\hat{\sigma}^5\sigma M + \hat{\sigma}N_{\text{sol}}$. For the online stage, all the formulae require to solve the problem $\hat{E}_\mu$ of size $\hat{N}$. Moreover, $\mathcal{E}_2$ additionally requires a linear combination of size σ, whereas $\mathcal{E}_3$ and $\mathcal{E}_4$ require to solve a linear system of size σ and $\hat{\sigma}$ respectively. We notice that if $N_{\text{sol}} \gg \hat{\sigma}^4\sigma M$ and $\hat{\sigma} < d\hat{N}+1$, then the offline stage of $\mathcal{E}_4$ with stabilized EIM requires less precomputations than the offline stage of $\mathcal{E}_2$.

Property	$\mathcal{E}_1$	$\mathcal{E}_2$	$\mathcal{E}_3$	$\mathcal{E}_4$
Online efficient	No	Yes	Yes	Yes
Unconditionally well-posed	Yes	Yes	No	Yes
Dependence on ϵ of the observed accuracy	ϵ	$\sqrt{\epsilon}$	ϵ, if well-posed	ϵ
Equals $\mathcal{E}_1$ in exact arithmetics	–	Yes	Yes	Yes, if $\hat{\sigma}=\sigma$ No, if $\hat{\sigma}<\sigma$
Complexity of the offline stage	–	$(d\hat{N}+1)N_{\text{sol}}$	σN_{sol}	$\hat{\sigma}^4\sigma M + \hat{\sigma}N_{\text{sol}}$ with classical EIM $\hat{\sigma}^5\sigma M + \hat{\sigma}N_{\text{sol}}$ with stabilized EIM
Complexity of the online stage	–	$\hat{N}^3+\sigma$	$\hat{N}^3+\sigma^3$	$\hat{N}^3+\hat{\sigma}^3$

Table 5.2. Comparison of the considered formulae for computing the error bound.

5.5 Application to a three-dimensional acoustic scattering problem

5.5.1 Formulation of the problem

We refer to Section 2.2 for more details. We consider a ball $\Omega^- \subset \mathbb{R}^3$ with boundary Γ and $\Omega^+ := \mathbb{R}^3 \backslash \overline{\Omega^-}$, see Figure 5.5. We consider a monopole source located in Ω^+. The surface of the ball is impedant, meaning that any incident wave will be partially absorbed and partially scattered. The proportion of absorbed and scattered parts is quantified by the impedance coefficient μ, which is used in a Robin boundary condition at Γ. We are interested in the computation of the scattered field p_{sc} in Ω^+. We denote p_{inc} the known pressure field created by the source in the absence of the sphere; the total acoustic field in Ω^+ is the sum of p_{inc} and p_{sc}.

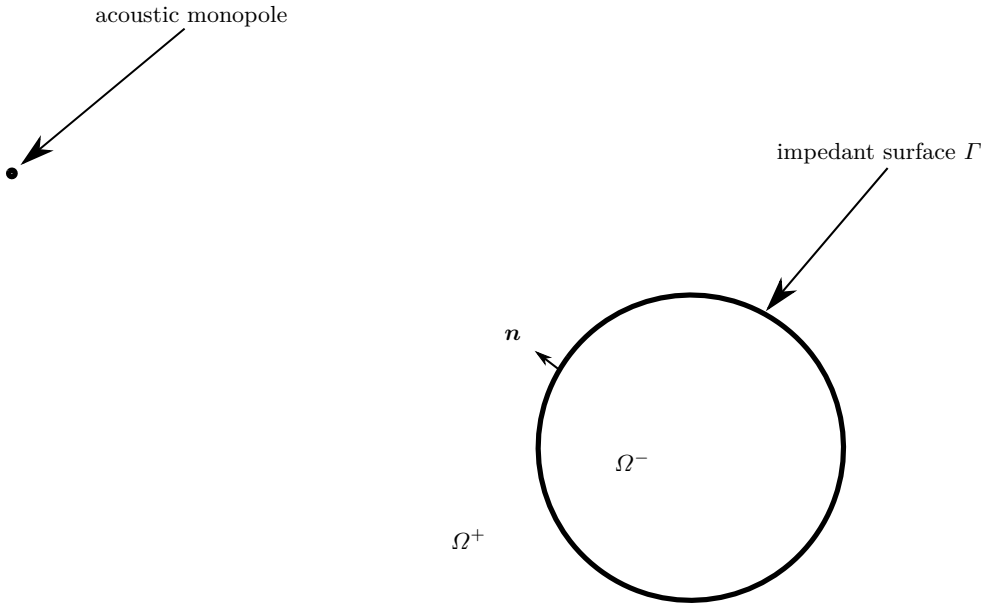

Fig. 5.5. Geometry for the three-dimensional acoustic scattering problem

We define the distribution $v : \Omega^+ \cup \Omega^- \longrightarrow \mathbb{C}$ such that $v_{|\Omega^-} = -p_{\text{inc}}$, $v_{|\Omega^+} = p_{\text{sc}}$. We denote λ and χ the jumps of the Neumann and Dirichlet traces of v across Γ. The Robin boundary condition writes $\lambda + \frac{ik}{\mu}\chi = 0$. Since v solves the homogeneous Helmholtz equation in Ω^+ and in Ω^- and satisfies the Sommerfeld radiation condition at infinity, there holds

$$v = -\mathcal{S}\lambda + \mathcal{D}\chi \quad \text{in } \Omega^+ \cup \Omega^-, \tag{5.38}$$

where $\mathcal{S}$ and $\mathcal{D}$ are respectively the single- and double-layer potentials. Taking the interior Dirichlet and Neumann traces of v in equation (5.38) and injecting the Robin boundary condition, we obtain

$$\begin{bmatrix} N - \frac{ik}{2\mu} I & \tilde{D} \\ D & -S - \frac{i\mu}{2k} I \end{bmatrix} \begin{bmatrix} \chi \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma_1^- p_{\text{inc}} \\ -\gamma_0^- p_{\text{inc}} \end{bmatrix}, \tag{5.39}$$

where k is the wave number of the monopole source, N, $\tilde{D}$, D and S are classical boundary integral operators (see [93]), and $\gamma_0^- p_{\text{inc}}$ and $\gamma_1^- p_{\text{inc}}$ are respectively the interior Dirichlet and Neumann traces of the known function p_{inc}. The software we are using, ACTIPOLE (see [34, 33]), deals with the block system defined in (5.39), which presents the advantage of being invertible for all frequencies of the source when the surface Γ is Lipschitz, see Thoerem 2.2. We denote A_μ the block operator defined by the left-hand side of (5.39). From [55, 76, 93], we infer that A_μ is a bounded bijective operator from $H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ into $H^{-\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$. The variational form is as follows: find $(\chi, \lambda) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ such that for all $(\hat{\chi}, \hat{\lambda}) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$,

$$\begin{cases} \left(N\chi - \dfrac{ik}{2\mu}\chi, \hat{\chi}\right) + \left(\tilde{D}\lambda, \hat{\chi}\right) = (\gamma_1 p_{\text{inc}}, \hat{\chi}), \\ \left\langle \hat{\lambda}, D\chi \right\rangle - \left\langle \hat{\lambda}, S\lambda + \dfrac{i\mu}{2k}\lambda \right\rangle = -\left\langle \hat{\lambda}, \gamma_0 p_{\text{inc}} \right\rangle, \end{cases} \tag{5.40}$$

where $(\cdot, \cdot)$ denotes the $H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma)$ duality product and $\langle \cdot, \cdot \rangle$ denotes the $L^2(\Gamma)$ inner product.

Let $\mathcal{M}$ be a shape-regular triangular mesh of Γ with meshsize h, and let V_h^1 and V_h^0 be respectively the spaces spanned by continuous piecewise affine polynomials on $\mathcal{M}$ and piecewise constant polynomials on $\mathcal{M}$. Let $(\phi_i)_{1 \le i \le P}$ and $(\psi_j)_{1 \le j \le P'}$ be the usual bases of V_h^1 and V_h^0 of size P and P', respectively. The product space $V_h^1 \times V_h^0$ is a conforming approximation of $H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$. The discrete problem is derived from a Galerkin procedure on $V_h^1 \times V_h^0$ using the boundary element method (BEM). The obtained discrete approximation of the problem (7.42) is inf-sup stable for h small enough, see Proposition 2.4. A direct solver is used, in double-precision format.

5.5.2 Application of the RB method

The RB method has recently been applied to problems solved by means of integral equations in electromagnetism, see [44, 29]. In these works, the classical a posteriori error bounds were used. We are here interested in the application of our improved a posteriori error bounds to such problems. We take as parameter for the RB method the value of the impedance μ, which is supposed here to be a positive real number. To recover an affine dependence on the parameter μ, we write the BEM matrix in the form $A_\mu = a_1(\mu)A_1 + a_2(\mu)A_2 + a_3(\mu)A_3$, so that $d = 3$ in the affine decomposition (5.7) with $a_1(\mu) = 1$, $a_2(\mu) = \frac{1}{\mu}$ and $a_3(\mu) = \mu$. Specifically,

$$A_1 = \left[\begin{array}{c|c} \left(N\phi_i, \phi_j\right)_{\substack{1 \le i \le P \\ 1 \le j \le P}} & \left(\tilde{D}\psi_j, \phi_i\right)_{\substack{1 \le i \le P \\ 1 \le j \le P'}} \\ \hline \left\langle D\phi_j, \psi_i \right\rangle_{\substack{1 \le i \le P' \\ 1 \le j \le P}} & \left\langle -S\psi_i, \psi_j \right\rangle_{\substack{1 \le i \le P' \\ 1 \le j \le P'}} \end{array}\right], \tag{5.41}$$

$$A_2 = \left[\begin{array}{c|c} -\frac{ik}{2}\left(\phi_i,\phi_j\right)_{\substack{1\le i\le P\\ 1\le j\le P}} & (0)_{\substack{1\le i\le P\\ 1\le j\le P'}} \\ \hline (0)_{\substack{1\le i\le P'\\ 1\le j\le P}} & (0)_{\substack{1\le i\le P'\\ 1\le j\le P'}} \end{array}\right], \quad A_3 = \left[\begin{array}{c|c} (0)_{\substack{1\le i\le P\\ 1\le j\le P}} & (0)_{\substack{1\le i\le P\\ 1\le j\le P'}} \\ \hline (0)_{\substack{1\le i\le P'\\ 1\le j\le P'}} & -\frac{i}{2k}\left\langle\psi_i,\psi_j\right\rangle_{\substack{1\le i\le P'\\ 1\le j\le P'}} \end{array}\right]. \tag{5.42}$$

In the general-purpose RB, the quantity of interest is the pair of potentials (χ,λ) on Γ. For the goal-oriented case, we consider the value of the pressure at a given point in Ω^+. If this point is far enough from Γ, approximations can be made in the representation formula for the pressure. This is the far-field approximation, which consists in a linear form Q acting on the solution pair (χ,λ) as

$$Q(\chi,\lambda) = \begin{pmatrix} -ik\dfrac{e^{-ik\|x\|_2}}{4\pi\|x\|_2}\left(e^{-iky\cdot\frac{x}{\|x\|_2}}\dfrac{x}{\|x\|_2}\cdot n(y),\chi(y)\right) \\ ik\dfrac{e^{-ik\|x\|_2}}{4\pi\|x\|_2}\displaystyle\int_\Gamma\left(e^{-iky\cdot\frac{x}{\|x\|_2}},\lambda(y)\right) \end{pmatrix} \in \mathbb{C}^2. \tag{5.43}$$

For simplicity, we take the Euclidian norm of vectors in $\mathbb{C}^{P+P'}$ instead of the $H^{\frac{1}{2}}(\Gamma)\times L^2(\Gamma)$ norms of the reconstructed functions in (5.9) and (5.10). This way, the Riesz isomorphism J is simply the identity. Therefore, the computation of the terms $G_\mu u_\mu$, as well as that of the terms $G_k u_i$, does not require to invert the stiffness matrix as in (5.12). The Successive Constraint Method (see Section 5.6) is used to compute a lower bound of the inf-sup constant, which is around 10^{-6} in the present examples.

We define two test cases: (i) one impedant sphere ($d=3$), with $N=584$ and $\mu\in\mathcal{P}:=[0.9,1.1]$, (ii) two impedant spheres ($d=5$), with $N=1561$ and $\mu\in\mathcal{P}:=[0.99,1.01]^2$. We present visualizations of the scattered pressure field, at a random value of the parameter μ, for test case (i) with $\#\mathcal{P}_{\text{trial}}=100$ and $\hat{N}=10$ in Figure 5.6 and for test case (ii) with $\#\mathcal{P}_{\text{trial}}=225$ and $\hat{N}=10$ in Figure 5.7.

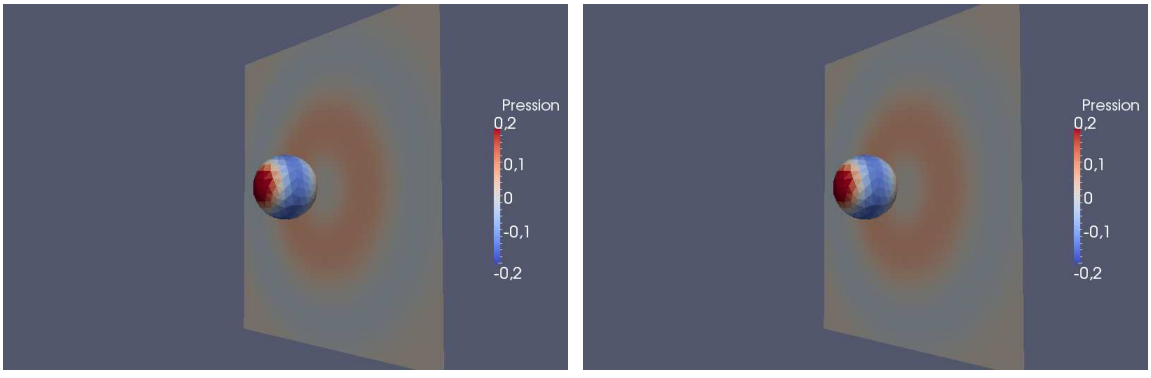

Fig. 5.6. Real part of the pressure field for the BEM solution (left) and the RB solution (right), with a basis of size 10. The difference between the two fields is less than 10^{-15} in infinity norm.

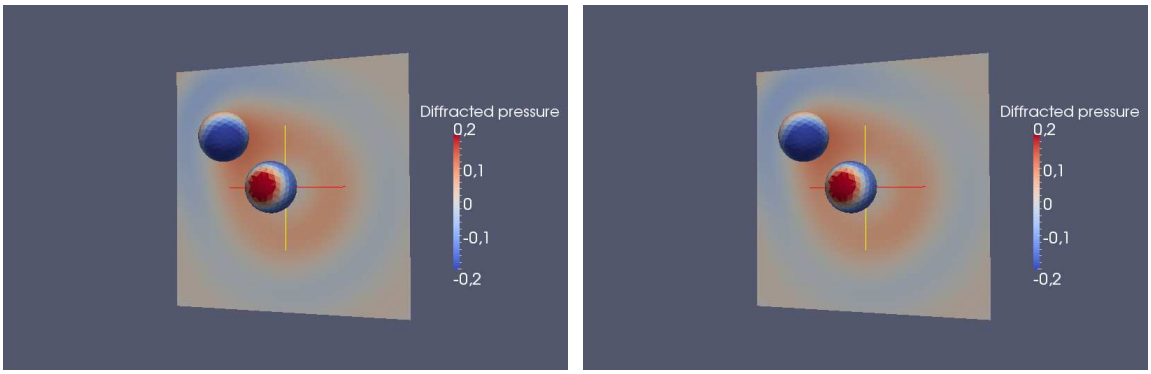

Fig. 5.7. Real part of the pressure field for the BEM solution (left) and the RB solution (right), with a basis of size 10. The difference between the two fields is less than 10^{-15} in infinity norm.

5.5.3 Error bound curves

We present the error bound curves for test case (i) with a general-purpose RB, $\#\mathcal{P}_{\text{trial}} = 100$, $(\hat{N}, \hat{\sigma}, \sigma) = (2, 7, 49), (3, 10, 100), (4, 20, 169)$, and $(5, 30, 256)$ in Figure 5.8 and for test case (ii) with a goal-oriented RB, $\#\mathcal{P}_{\text{trial}} = 225$, $\hat{N} = 8$, $\hat{\sigma} = 60$, and $\sigma = 1681$ in Figure 5.9.

In test case (i), the classical formula $\mathcal{E}_2$ exhibits quite poor performances, since it cannot compute values below 10^{-4}. This is explained by the values of the inf-sup constant which are around 10^{-6}. Furthermore, in agreement with Remark 5.8, the lowest computable values of $\mathcal{E}_1$ and $\mathcal{E}_2$ differ by 8 orders of magnitude. In test case (ii), the behavior of formula $\mathcal{E}_3$ is quite poor, and we do not observe the level of accuracy we observed so far for $\mathcal{E}_3$. Here, the matrix T defined in (5.25) is so ill-conditioned that the numerical errors introduced by its resolution are larger than the ones introduced by the formula $\mathcal{E}_2$. Furthermore, the formula $\mathcal{E}_4$ exhibits, as before, a very good performance. We see in Figure 5.9 that $\underset{\mu \in \mathcal{P}_{\text{select}}}{\text{argmax}}\,(\mathcal{E}_4(\mu)) = (1, 1)$ and $\mathcal{E}_4(1, 1) \approx 10^{-16}$; therefore, the formula $\mathcal{E}_4$ with $\hat{\sigma} = 60$ is valid for computing the error bound in Algorithm 1 with $\text{tol} = 10^{-16}$.

The behavior of $\mathcal{E}_4$ when $\hat{\sigma}$ increases is investigated in Figure 5.10 for test case (i). We consider the values $\hat{\sigma} = 14, 30, 40$ and 50. These four values lead to the same local maxima, and increasing $\hat{\sigma}$ allows the formula $\mathcal{E}_4$ to be valid for smaller tolerances (respectively 5×10^{-8}, 10^{-8}, 8×10^{-9} and 2×10^{-9}). Another interesting observation comes from considering the fourth plot in Figure 5.8 and the first plot in Figure 5.10: the classical formula $\mathcal{E}_2$ requires 16 offline resolutions of (5.12) and stagnates at 10^{-4} while the formula $\mathcal{E}_4$ with $\hat{\sigma} = 14$ only requires 14 offline resolutions of (5.12) and is valid for tolerances down to 5×10^{-8}. This shows that at least in some regimes, the new formula $\mathcal{E}_4$ is valid for lower tolerances than the classical formula $\mathcal{E}_2$, and requires less precomputations. However, contrary to $\mathcal{E}_2$, using $\mathcal{E}_4$ requires that all the quantities V_r defined in (5.24) be recomputed when adding a new vector to the reduced basis.

Conclusion

In this work, we have extended the ideas of [Ar1] by proposing a more stable numerical procedure, using the empirical interpolation method, to represent the a posteriori error bound in

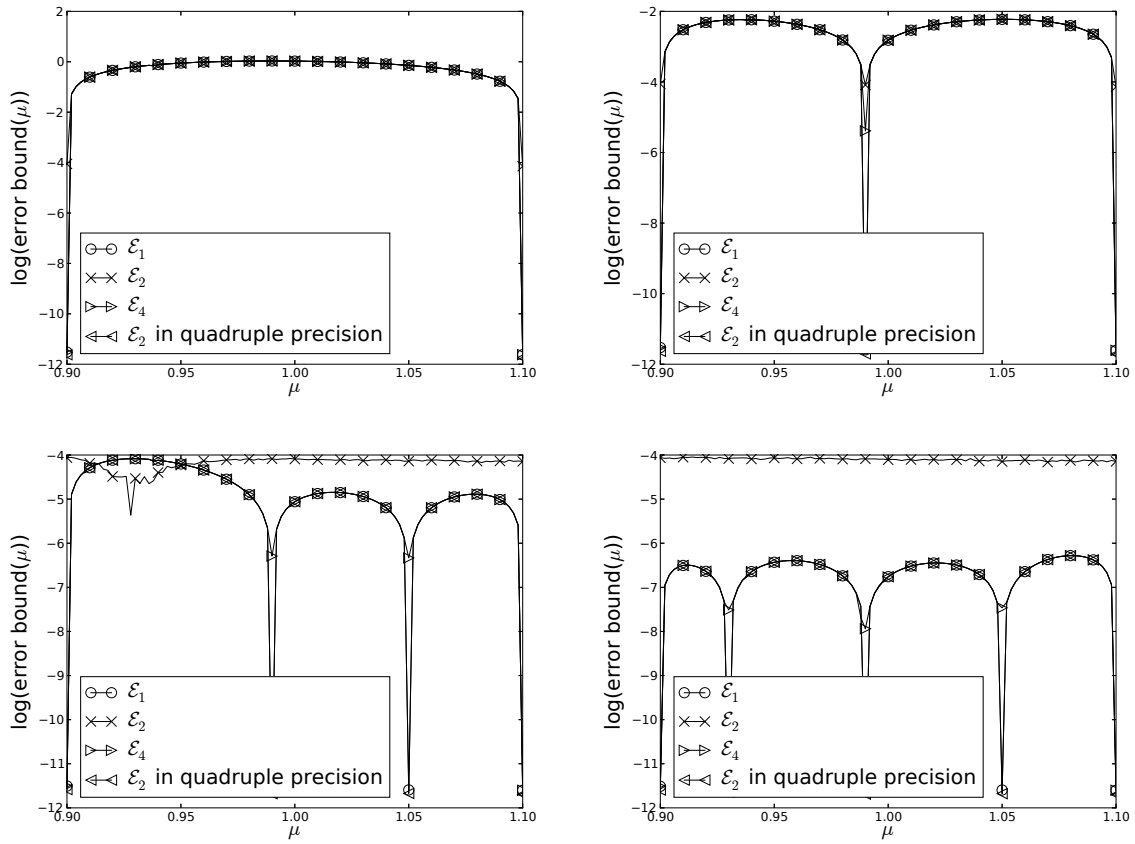

Fig. 5.8. Error bound curves with respect to the impedance coefficient, with $\hat{N}$ equal to 2, 3, 4, and 5 (from left to right and top to bottom). The curve for $\mathcal{E}_2$ computed in quadruple precision superimposes to $\mathcal{E}_1$.

the reduced basis method as a linear combination of its values at given parameter values, called interpolation points. Moreover, the proposed method provides a way of choosing the interpolation points, and yields better accuracy levels than the classical a posteriori error bound and than the procedure proposed in [Ar1]. Besides, our new procedure may require less precomputations than the classical a posteriori error bound. The new error bound derived herein can be of particular interest in three situations: (i) when the stability constant of the original problem is very small (this is the case in many practical problems), (ii) when very accurate solutions are needed, (iii) when considering a nonlinear problem (for which, in some cases, no error bound is possible until a very tight tolerance is reached, see [107]).

5.6 The Successive Constraint Method

The Successive Constraint Method (SCM) enables the computation of a lower bound of the inf-sup constant using an optimization problem, by means of an offline-online procedure. It has been introduced in [56]. We can also find detailed presentations in [30]. A complex-valued operator version has been proposed in [29].

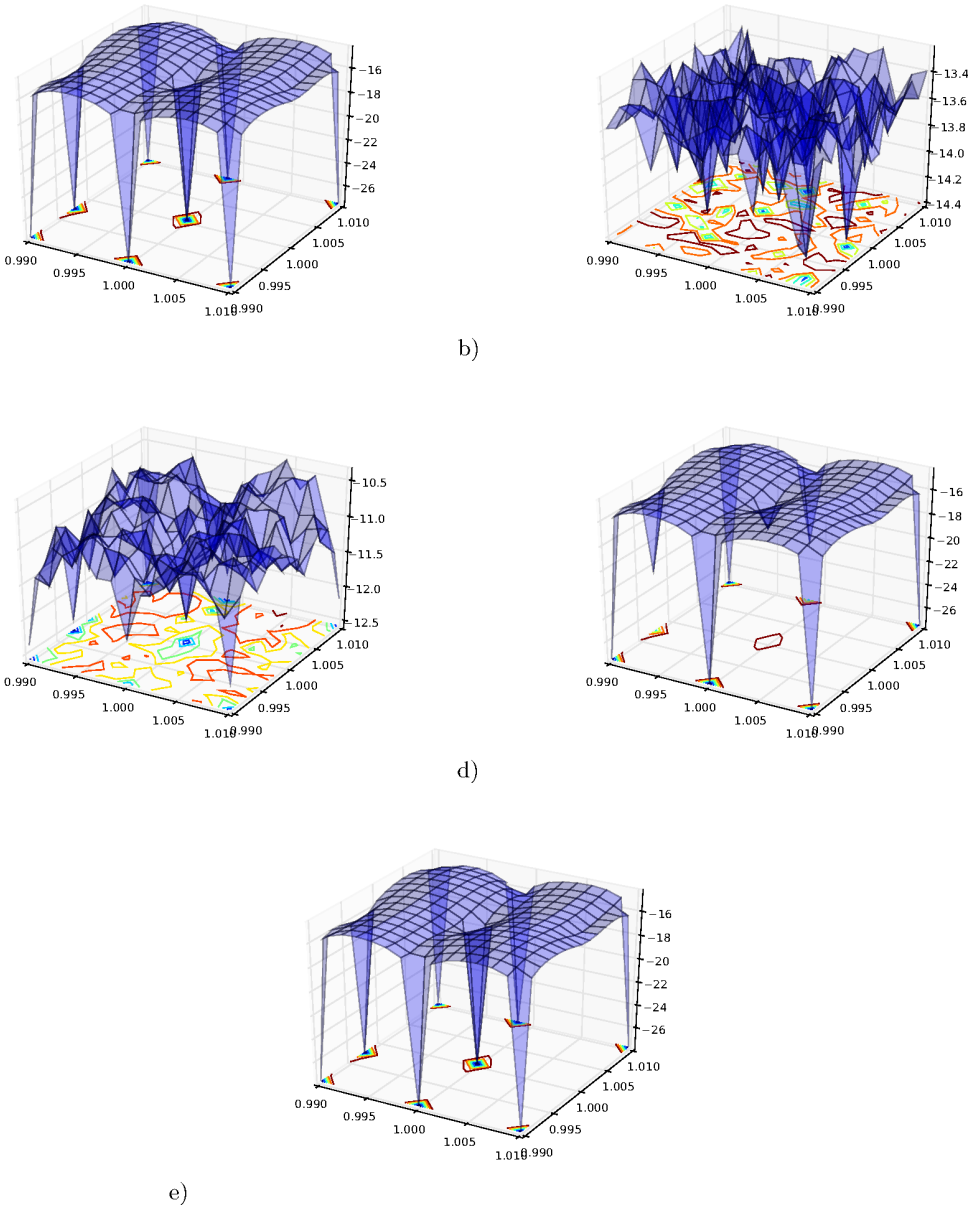

Fig. 5.9. Error bound curves (logarithmic scale) as a function of the impedance coefficients: a) $\mathcal{E}_1$, b) $\mathcal{E}_2$, c) $\mathcal{E}_3$, d) $\mathcal{E}_4$, and e) $\mathcal{E}_2$ computed in quadruple precision.

5.6.1 Principle

The inf-sup constant is defined as

$$\beta_\mu = \inf_{u\in\mathcal{V}} \sup_{v\in\mathcal{V}} \frac{a_\mu(u,v)}{\|u\|_\mathcal{V}\|v\|_\mathcal{V}}. \tag{5.44}$$

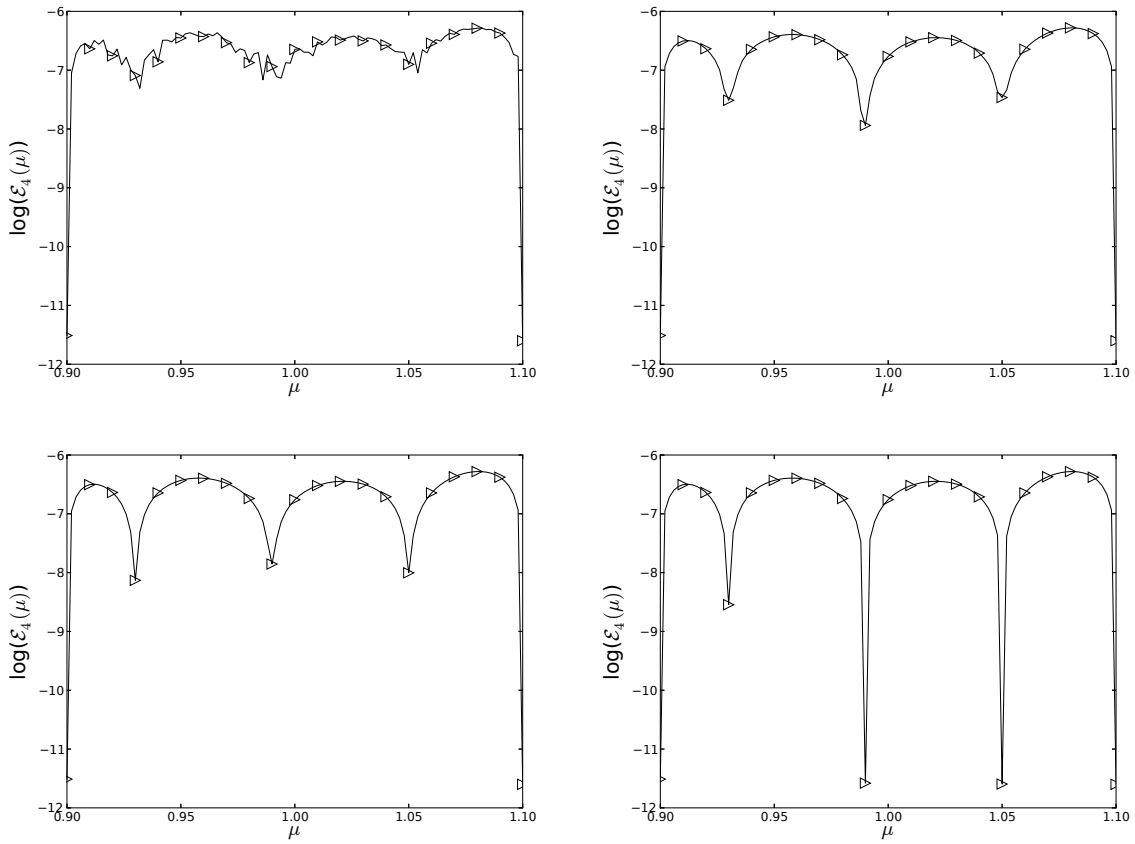

Fig. 5.10. Error bound curve for $\mathcal{E}_4$ with respect to the impedance coefficient, with $\hat{N} = 5$ and $\hat{\sigma}$ equal to 14, 30, 40, and 50 (from left to right and top to bottom).

Consider $T_\mu : \mathcal{V} \to \mathcal{V}$ such that, for all $\mu \in \mathcal{P}$ and all $u \in \mathcal{V}$, $(T_\mu u, v)_\mathcal{V} = a_\mu(u, v)\ \forall v \in \mathcal{V}$. It is shown in [95] that

$$\beta_\mu^2 = \inf_{v \in \mathcal{V}} \frac{(T_\mu v, T_\mu v)_\mathcal{V}}{\|v\|_\mathcal{V}^2}. \tag{5.45}$$

Thus, computing β_μ requires to solve an eigenvalue problem. Since T_μ naturally inherits the affine dependence on μ from a_μ, we inject its affine decomposition in (5.45), where we can identify the functions z_q and the sesquilinear forms $\hat{a}_q$, $1 \leq q \leq Q$, such that

$$\beta_\mu^2 = \inf_{v \in \mathcal{V}} \sum_{q=1}^{Q} z_q(\mu) \frac{\hat{a}_q(v, v)}{\|v\|_\mathcal{V}^2}. \tag{5.46}$$

It is shown in [44] that such a decomposition can be obtained in the case of complex-valued operators with z_q being real-valued functions of μ and $\hat{a}_q$ being Hermitian forms on $\mathcal{V} \times \mathcal{V}$. Then, β_μ can be evaluated by solving a constrained minimization problem in $\mathbb{R}^Q$. Defining

$$\begin{aligned} I : \mathcal{P} \times \mathbb{R}^Q &\to \mathbb{R} \\ (\mu, y) &\mapsto \sum_{q=1}^{Q} z_q(\mu) y_q, \end{aligned} \tag{5.47}$$

there holds

$$\beta_\mu^2 = \min_{y \in \mathcal{Y}} I(\mu, y), \tag{5.48}$$

where

$$\mathcal{Y} = \left\{ y = (y_1, ...y_Q) \in \mathbb{R}^Q \,\middle|\, \exists v \in \mathcal{V} \text{ such that } y_q = \frac{\hat{a}_q(v,v)}{\|v\|_{\mathcal{V}}^2},\ 1 \leqslant q \leqslant Q \right\}. \tag{5.49}$$

Finally, we observe that solving the problem (5.48) on a subset and a superset of $\mathcal{Y}$, we can obtain respectively an upper bound and a lower bound of β_μ^2 as follows:

$$\beta_{\mu\text{LB}}^2 := \min_{y \in \mathcal{Y}_{\text{LB}}} I(\mu, y) \leqslant \beta_\mu^2 \leqslant \beta_{\mu\text{UB}}^2 := \min_{y \in \mathcal{Y}_{\text{UB}}} I(\mu, y), \tag{5.50}$$

where $\mathcal{Y}_{\text{UB}} \subset \mathcal{Y} \subset \mathcal{Y}_{\text{LB}}$.

5.6.2 Algorithm

The SCM consists in a greedy procedure to construct nested spaces verifying

$$\mathcal{Y}_{\text{UB}}^1 \subset \mathcal{Y}_{\text{UB}}^2 \subset ... \subset \mathcal{Y}_{\text{UB}}^n \subset ... \subset \mathcal{Y} \subset ... \subset \mathcal{Y}_{\text{LB}}^n(\mu) \subset ... \subset \mathcal{Y}_{\text{LB}}^2(\mu) \subset \mathcal{Y}_{\text{LB}}^1(\mu), \tag{5.51}$$

where the supersets $\mathcal{Y}_{\text{LB}}^n(\mu)$ defining lower bounds are μ-dependent, whereas the subsets $\mathcal{Y}_{\text{UB}}^n$ defining upper bounds are μ-independent. Before presenting the algorithm to construct these subsets and supersets, we need some preliminary definitions.

Definition 5.26 *(Eigenvector associated to the smallest eigenvalue) We denote by $y(\mu)$ the vector of coordinates*

$$y(\mu)_q = \frac{\hat{a}_q(w,w)}{\|w\|_{\mathcal{V}}^2}, \qquad 1 \leq q \leq Q, \tag{5.52}$$

where w is the generalized eigenvector corresponding to the smallest generalized eigenvalue $\lambda_{\min}$ of the following generalized eigenproblem of unknowns (λ, v):

$$(T_\mu v, T_\mu v)_{\mathcal{V}} = \lambda \|v\|_{\mathcal{V}}^2. \tag{5.53}$$

The inf-sup constant for the parameter μ is then known: $\beta_\mu := \sqrt{\lambda_{\min}}$.

Notice that Definition 5.26 requires solving a large eigenvalue problem. In some cases, there exist efficient algorithms to compute the eigenpair associated to the smallest eigenvalue. For instance, the Locally Optimal Block Preconditioned Conjugate Gradient Method [60] is an iterative algorithm suitable for symmetric positive definite matrices.

Suppose that we are at step n of the offline stage of the SCM: n parameter values μ are selected from $\mathcal{P}_{\text{trial}}$ and stored in a set denoted C_n. For all $\mu' \in C_n$, the smallest eigenvalue

of Problem (5.53) is computed and stored (therefore $\beta^2_{\mu'}$ is known), and the corresponding eigenvector is computed and stored in $\mathcal{Y}^n_{\rm UB}$. Hence, $\mathcal{Y}^n_{\rm UB}$ is a set of vectors from $\mathbb{R}^Q$, and is of cardinal n. At the end of each step of the offline stage of the SCM, a lower bound of β^2_μ is computed for all $\mu \in \mathcal{P}_{\rm trial}$. At each step, this lower bound becomes sharper. In particular at step n, the lower bound computed at step $n-1$ is available for all $\mu \in \mathcal{P}_{\rm trial}$, and is denoted $\left(\beta^2_\mu\right)^{n-1}_{\rm LB}$. Let $\mu \in \mathcal{P}_{\rm trial}$, the nested superset at step n is defined as:

$$\mathcal{Y}^n_{\rm LB}(\mu) := \left\{ y \in \mathcal{B} \middle| I(\mu', y) \geqslant \beta^2_{\mu'} \ \forall \mu' \in \mathtt{P}_{M_1}(\mu, C_n) \text{ and } I(\mu', y) \geqslant \left(\beta^2_{\mu'}\right)^{n-1}_{\rm LB} \ \forall \mu' \in \mathtt{P}_{M_2}(\mu, \mathcal{P}_{\rm trial}) \right\},$$

where $\mathcal{B} = \prod_{q=1}^Q \left[\sigma_q^-, \sigma_q^+\right]$ with $\sigma_q^- = \inf_{v \in \mathcal{V}} \frac{\hat{a}_q(v,v)}{\|v\|^2_{\mathcal{V}}}$, $\sigma_q^+ = \sup_{v \in \mathcal{V}} \frac{\hat{a}_q(v,v)}{\|v\|^2_{\mathcal{V}}}$, $1 \leq q \leq Q$; M_1 and M_2 are positive integers, and, for any subset E of $\mathcal{P}_{\rm trial}$,

$$\mathtt{P}_M(\mu, E) := \begin{cases} M \text{ closest points to } \mu \text{ in } E & \text{if } \mathrm{card}(E) > M, \\ E & \text{if } \mathrm{card}(E) \leqslant M. \end{cases} \tag{5.54}$$

Since $y \to I(\mu, y)$ is linear, $\mathcal{Y}^n_{\rm LB}(\mu)$ is the intersection of $2Q + M_1 + M_2$ half-spaces of $\mathbb{R}^Q$: $2Q$ half-spaces for $y \in \mathcal{B}$, M_1 half-spaces for $I(\mu', y) \geqslant \beta^2_{\mu'}$ $\forall \mu' \in \mathtt{P}_{M_1}(\mu, C_n)$, and M_2 half-spaces for $I(\mu', y) \geqslant \left(\beta^2_{\mu'}\right)^{n-1}_{\rm LB}$ $\forall \mu' \in \mathtt{P}_{M_2}(\mu, \mathcal{P}_{\rm trial})$. Consider now the problem

$$\min_{y \in \mathcal{Y}^n_{\rm LB}(\mu)} I(\mu, y). \tag{5.55}$$

The problem (5.55) belongs to the class of "linear program" optimization problems with Q variables and $2Q + M_1 + M_2$ inequality constraints. Notice that since $\left(\beta^2_{\mu'}\right)^{n-1}_{\rm LB}$ $\forall \mu' \in \mathtt{P}_{M_2}(\mu, \mathcal{P}_{\rm trial})$ and $\beta^2_{\mu'}$ $\forall \mu' \in \mathtt{P}_{M_1}(\mu, C_n)$ have been computed in previous steps of the offline stage of the SCM, the problem (5.55) is of complexity independent of N.

Definition 5.27 *(Error measure) The convergence of the offline stage of the SCM is controlled by monitoring $\eta^n(\mu)$ defined by:*

$$\eta^n(\mu) := 1 - \frac{\min_{y \in \mathcal{Y}^n_{LB}(\mu)} I(\mu, y)}{\min_{y \in \mathcal{Y}^n_{UB}} I(\mu, y)}. \tag{5.56}$$

The offline construction of the nested spaces is described in Algorithm 3.

Once the offline stage of the SCM is done, the online problem consists in solving (5.55), which provides a lower bound of the inf-sup constant, where the sharpness is controlled by $\eta^n(\mu)$ over $\mathcal{P}_{\rm trial}$.

Remark 5.28 *(Choice of M_1 and M_2) The positive integers M_1 and M_2 hide a tradeoff in computational complexity: if M_1 and M_2 are large, the bounds $\left(\beta^2_{\mu'}\right)^{n-1}_{LB}$ are sharp and the offline of the SCM is supposed to converge fast, but the optimization problem* (5.55) *contains more constraints, and therefore is more complex to solve. If M_1 and M_2 are smaller, the number of steps of the SCM is larger, but the optimization problem* (5.55) *is faster to solve. In particular, the complexity of the online problem of the SCM depends on M_1 and M_2.*

Algorithm 3 Offline stage of the SCM algorithm

1. Compute $\sigma_q^- = \inf_{v\in\mathcal{V}} \frac{\hat{a}_q(v,v)}{\|v\|_{\mathcal{V}}^2}$ and $\sigma_q^+ = \sup_{v\in\mathcal{V}} \frac{\hat{a}_q(v,v)}{\|v\|_{\mathcal{V}}^2}$, $1 \leq q \leq Q$, and set $\mathcal{B} = \prod_{q=1}^{Q} \left[\sigma_q^-, \sigma_q^+\right]$
2. Set $n = 1$ and fix $M_1, M_2 \in \mathbb{N}^*$
3. Choose $\mu_1 \in \mathcal{P}_{\text{trial}}$ randomly and set $C_1 := \{\mu_1\}$
4. Compute $y_1 = y(\mu_1)$ and β_{μ_1} using Definition 5.26, and set $\mathcal{Y}_{\text{UB}}^1 := \{y_1\}$
5. Compute $\eta^1(\mu)$ for all $\mu \in \mathcal{P}_{\text{trial}}$ using (5.56), with $\mathcal{Y}_{\text{LB}}^1(\mu) := \left\{ y \in \mathcal{B} \,\middle|\, I(\mu_1, y) \geqslant \beta_{\mu_1}^2 \text{ and } I(\mu', y) \geqslant 0 \ \forall \mu' \in \mathtt{P}_{M_2}(\mu, \mathcal{P}_{\text{trial}}) \right\}$
6. **while** $\max_{\mu\in\mathcal{P}_{\text{trial}}} \eta^n(\mu) \geq$ tolerance **do**
7. Set $\mu_{n+1} = \operatorname{argmax}_{\mu\in\mathcal{P}_{\text{trial}}} \eta^n(\mu)$ and $C_{n+1} := \{\mu_{n+1}\} \cup C_n$
8. Compute $y_{n+1} = y(\mu_{n+1})$ and $\beta_{\mu_{n+1}}$ using Definition 5.26, and set $\mathcal{Y}_{\text{UB}}^{n+1} := \{y^{n+1}\} \cup \mathcal{Y}_{\text{UB}}^n$
9. Compute $\eta^{n+1}(\mu)$ for all $\mu \in \mathcal{P}_{\text{trial}}$ using (5.56), with $\mathcal{Y}_{\text{LB}}^{n+1}(\mu) := \left\{ y \in \mathcal{B} \,\middle|\, I(\mu', y) \geqslant \beta_{\mu'}^2 \ \forall \mu' \in \mathtt{P}_{M_1}(\mu, C_{n+1}) \text{ and } I(\mu', y) \geqslant \left(\beta_{\mu'}^2\right)_{\text{LB}}^n \ \forall \mu' \in \mathtt{P}_{M_2}(\mu, \mathcal{P}_{\text{trial}}) \right\}$
10. $n \leftarrow n + 1$
11. **end while**

6

A nonintrusive EIM to approximate linear systems with nonlinear parameter dependence

This chapter is based on the preprint [Pr1].

Summary. We consider a family of linear systems $A_\mu \alpha = C$ with system matrix A_μ depending on a parameter μ and for simplicity parameter-independent right-hand side C. These linear systems typically result from the finite-dimensional approximation of a parameter-dependent boundary-value problem. We derive a procedure based on the Empirical Interpolation Method to obtain a separated representation of the system matrix in the form $A_\mu \approx \sum_m \beta_m(\mu) A_{\mu_m}$ for some selected values of the parameter. Such a separated representation is in particular useful in the Reduced Basis Method. The procedure is called nonintrusive since it only requires to access the matrices A_{μ_m}. As such, it offers a crucial advantage over existing approaches that instead derive separated representations requiring to enter the code at the level of assembly. Numerical examples illustrate the performance of our new procedure on a simple one-dimensional boundary-value problem and on three-dimensional acoustic scattering problems solved by a boundary element method.

6.1 Introduction

In industrial projects, decisions are often taken after a series of complex computations using computer codes of various origins. To simplify the overall computation, surrogate models can be used to replace some parts of the computation. Some of these surrogates are constructed using only a series of input/output couples. With some hypotheses on the input, confidence intervals can be derived, see e.g. [98] for the kriging method. When additional knowledge on the underlying mathematical formulation is available, model reduction methods can be used. For instance, the Reduced Basis Method (RBM) enables fast resolutions on a basis of precomputed solutions, rather than on a finite element basis (see [72] for a detailed presentation and [15] for some convergence results). We consider a family of linear systems $A_\mu \alpha = C$ of order n, where n is large. For simplicity, we assume that the right-hand side is independent of the parameter μ.

The RBM consists first in an offline stage, where a reduced basis of $\hat{n}$ functions u_j, $1 \leq j \leq \hat{n}$ are computed using a greedy algorithm. These functions are solution of the original problem for some values $(\mu_j)_{1 \leq j \leq \hat{n}}$ of the parameter μ, which are selected using a greedy algorithm. The functions u_j thus write $u_j(x) = \sum_{i=1}^{n} \alpha_i(\mu_j)\theta_i(x)$, where $(\theta_i)_{1 \leq i \leq n}$ is the finite element basis and the vector $\alpha(\mu_j) = (\alpha_i(\mu_j))_{1 \leq i \leq n}$ is such that $A_{\mu_j}\alpha(\mu_j) = C$. In practice, the dimension of the reduced basis is much smaller than the dimension of the finite element basis: $\hat{n} \ll n$. Denote by U the rectangular matrix of size $n \times \hat{n}$ such that $(U)_{i,j} = \alpha_i(\mu_j)$. Second, in the online stage,

for a given value of the parameter μ, a reduced problem is constructed as $\hat{A}_\mu \hat{\alpha}(\mu) = \hat{C}$, where $\hat{A}_\mu = U^t A_\mu U$ and $\hat{C} = U^t C$. Solving this reduced problem for a certain value of μ leads to the approximate solution $\hat{u}_\mu(x) = \sum_{j=1}^{\hat{n}} \hat{\alpha}_j(\mu) u_j(x)$.

To efficiently construct the online problems, a separated representation (also known as an *affine decomposition* in the RBM literature) of the matrix assembled by the code is needed in the form

$$A_\mu \approx \sum_{m=1}^{d} \gamma_m(\mu) A_m, \tag{6.1}$$

so that

$$\hat{A}_\mu \approx \sum_{m=1}^{d} \gamma_m(\mu) U^t A_m U, \tag{6.2}$$

where the matrices $U^t A_m U$ are of small size $\hat{n} \times \hat{n}$ and can be precomputed during the offline stage. The separated representation (6.1) thus enables online problems to be constructed in complexity independent of n, as long as the functions $\mu \mapsto \gamma_m(\mu)$ are also computed in complexity independent of n. Standard techniques (see [73]) to obtain the separated representation (6.1) require in general nontrivial modifications of the assembling routines of the computational code in order to access separately various terms of the variational formulation at hand (See Remark 6.4 below for more details).

The present work provides a step forward in this context, since a procedure that yields a separated representation of A_μ in the form

$$A_\mu \approx \sum_{m=1}^{d} \beta_m(\mu) A_{\mu_m} \tag{6.3}$$

is derived, where $(\mu_m)_{1 \le m \le d}$ are some selected values of the parameter. Since the separated representation (6.3) only uses the complete system matrix at the selected parameter values, this representation requires no implementation effort in the assembly routines of the computational code under the (mild) assumptions that we can indeed access the system matrix A_μ and that we can identify the functional dependencies on μ in the variational formulation under consideration (see below for more details). For this reason, the procedure is called nonintrusive.

In Section 6.2, we present the approximation problems investigated in this work, first a simple introductory example and then problems with a more complex parameter dependence. In Section 6.3, we briefly recall the Empirical Interpolation Method. In Section 6.4, we present our nonintrusive procedure for the introductory example and test it on a one-dimensional boundary-value problem. The procedure is extended to more complex parameter dependence in Section 6.5 where it is also applied to two three-dimensional scattering problems. Some conclusions are drawn in Section 6.6 where, in particular, we observe that our procedure can be extended to the approximation of other quantities.

6.2 The approximation problem

We first present an introductory example. Let $\mathcal{V}$ be a Hilbert space and consider the following weak formulation: Find $u \in \mathcal{V}$ such that for all $u^t \in \mathcal{V}$,

$$\int_\Omega g(\mu,x)\nabla u(x)\cdot\nabla u^t(x)dx+\int_\Omega \mu u(x)u^t(x)dx=b(u^t), \tag{6.4}$$

where Ω is the domain of computation, μ a parameter belonging to a given parameter set $\mathcal{P}$, $g(\mu,x)$ a given function defined on $\mathcal{P}\times\Omega$ and b a bounded linear form on $\mathcal{V}$. Consider now a conforming n-dimensional approximation of the space $\mathcal{V}$ denoted by $\mathcal{V}_h$ (the subscript h refers to an underlying mesh), and a basis of $\mathcal{V}_h$ denoted by $(\theta_i)_{1\le i\le n}$. The finite element approximation of (7.42) requires the computation of the matrix A_μ of size $n\times n$ with entries

$$(A_\mu)_{i,j}:=\left(\int_\Omega g(\mu,x)\nabla\theta_j(x)\cdot\nabla\theta_i(x)dx+\mu\int_\Omega\theta_j(x)\theta_i(x)dx\right)_{i,j}. \tag{6.5}$$

The notation A_μ is adopted to stress the fact that the matrix A_μ depends on the value of the parameter μ. The problem solved by the computational code is

$$A_\mu\alpha=C, \tag{6.6}$$

where $(C)_i=b(\theta_i)$ for all $1\le i\le n$, and where an approximation of the solution u to (7.42) is obtained in the form $u(x)\approx\sum_{i=1}^n\alpha_i\theta_i(x)$.

Let

$$\left(A_\mu^1\right)_{i,j}:=\left(\int_\Omega g(\mu,x)\nabla\theta_j(x)\cdot\nabla\theta_i(x)dx\right)_{i,j}\ \text{and}\ \left(A^0\right)_{i,j}:=\left(\int_\Omega\theta_j(x)\theta_i(x)dx\right)_{i,j} \tag{6.7}$$

so that

$$A_\mu=A_\mu^1+\mu A^0. \tag{6.8}$$

Definition 6.1 (Intrusivity) *A procedure leading to a separated representation of A_μ in the general form* (6.1) *is called*

- *intrusive if it requires to implement new integral terms,*
- *weakly-intrusive if it only requires to precompute independently A_μ^1 for some values of μ and A_0,*
- *nonintrusive if it only requires to precompute A_μ for some values of μ.*

The term "weakly-intrusive" comes from the fact that the user has to enter the routines of the code and to insert switches at the right places to save the terms in A_μ^1 independently from the terms in A^0. In the context of industrial codes, this is not always possible. The notion of nonintrusivity in Definition 6.1 is different from the notion of black-box, which requires only the computation of input / output couples. Our purpose is to develop a nonintrusive procedure leading to the separated representation (6.3) of A_μ.

The above example can be generalized to a class of engineering problems requiring to compute a large, parameter-dependent matrix A_μ for many values of the parameter μ where A_μ is of the form

$$A_\mu=\sum_{\varrho=1}^R A_\mu^\varrho+\sum_{s=1}^S\psi_s(\mu)A^s, \tag{6.9}$$

where A_μ^ϱ are matrices that require to integrate some functions $g^\varrho(\mu,x)$ over Ω, ψ^s are given functions of μ and A^s are μ-independent matrices resulting from some integration over Ω. The introductory example corresponds to $R=1$, $S=1$, and $\psi_1(\mu)=\mu$. To simplify the presentation of the main ideas, we consider the setting of (6.8) in Sections 6.3 and 6.4 and return to the more general setting of (6.9) in Section 6.5.

6.3 Empirical Interpolation Method

The Empirical Interpolation Method (EIM) is a procedure to approximate two-variable functions. In particular, it can be used to approximate the two-variable function $g(\mu, x)$, for all $\mu \in \mathcal{P}$ and all $x \in \Omega$. Denote by EIM^g this particular procedure. EIM^g leads to an interpolation operator $I^g_{d^g}$ such that

$$\left(I^g_{d^g} g\right)(\mu, x) \approx g(\mu, x), \qquad \forall \mu \in \mathcal{P},\ \forall x \in \Omega, \tag{6.10}$$

where d^g is the number of interpolation points (called *magic points* in the context of RBM, see [73]). EIM^g is composed of two stages: (i) an offline stage, where a matrix B^g of size $d^g \times d^g$, a set of d^g x-dependent basis functions $\{q^g_k\}_{1\le k\le d^g}$, a set of d^g points $\{x_k\}_{1\le k\le d^g}$ in Ω, and a set a d^g parameter values $\{\mu_k\}_{1\le k\le d^g}$ in $\mathcal{P}$ are constructed, (ii) an online stage, where the quantities computed in the offline stage are used to carry out the approximation (6.10) (see Section 6.4.2 for more details on the offline / online stages for the whole procedure).

The offline stage of EIM^g is detailed in Algorithm 6. This variant corresponds to the classical EIM, described in [73]. In the loop on k in Algorithm 6, the residual operator δ^g_k is defined by $\delta^g_k = \mathrm{Id} - I^g_k$, where the interpolation operator I^g_k is such that

$$\left(I^g_k g\right)(\mu, x) := \sum_{m=1}^{k} \lambda^g_m(\mu) q^g_m(x), \tag{6.11}$$

and for a given $\mu \in \mathcal{P}$, the $\lambda^g_m(\mu)$'s are defined by

$$\sum_{m=1}^{k} B^g_{l,m} \lambda^g_m(\mu) = g(\mu, x^g_l), \qquad \forall 1 \le l \le k. \tag{6.12}$$

After d^g iterations, the interpolation formula (7.7) leads to the following approximation for A_μ:

$$A_\mu \approx \sum_{m=1}^{d^g} \lambda^g_m(\mu) M_m + \mu A^0, \tag{6.13}$$

where $(M_m)_{i,j} = \int_\Omega q^g_m(x) \boldsymbol{\nabla}\theta_j(x) \cdot \boldsymbol{\nabla}\theta_i(x) dx$. This representation of A_μ is of the form (6.1).

Property 6.2 (Interpolation) $\forall x \in \Omega,\ \forall\ 1 \le m \le d^g,\ \left(I^g_{d^g} g\right)(\mu^g_m, x) = g(\mu^g_m, x)$.

Proof. See [73, Lemma 1]. ♢

Property 6.2 means that, at the parameter values $\left(\mu^g_k\right)_{1\le k\le d^g}$ selected by EIM^g, the approximation (6.13) is exact since $\sum_{m=1}^{d^g} \lambda^g_m(\mu^g_k) M_m = A^1_{\mu^g_k}$ for all $1 \le k \le d^g$. Since $\mathrm{Vect}_{1\le k\le d^g}\left(q^g_k(x)\right) = \mathrm{Vect}_{1\le k\le d^g}\left(g(\mu^g_k, x)\right)$ holds in Algorithm 6, the functions $q^g_k(x)$ can be expressed in terms of the functions $g(\mu^g_k, x)$ in the following form: there exist $\gamma_{l,k}$, $1 \le l \le k \le d^g$ such that $q^g_k(x) = \sum_{l=1}^{d^g} \gamma_{l,k} g(\mu^g_l, x)$, for all $1 \le k \le d^g$. Letting $(\lambda^g_m(\mu))_{1\le m\le d^g}$ solve (7.8) for $k = d^g$, we obtain after exchanging the summations

$$(I^g_{d^g} g)(\mu, x) = \sum_{m=1}^{d^g} \left(\sum_{l=1}^{d^g} \gamma_{m,l} \lambda^g_l(\mu) \right) g(\mu^g_m, x). \tag{6.14}$$

Algorithm 4 Offline stage of EIM^g

1. Choose $d^g > 1$ [Number of interpolation points]
2. Set $k := 1$
3. Compute $\mu_1^g := \underset{\mu\in\mathcal{P}}{\operatorname{argmax}} \|g(\mu,\cdot)\|_{L^\infty(\Omega)}$
4. Compute $x_1^g := \underset{x\in\Omega}{\operatorname{argmax}} |g(\mu_1^g, x)|$ [First interpolation point]
5. Set $q_1^g(\cdot) := \dfrac{g(\mu_1^g,\cdot)}{g(\mu_1^g, x_1^g)}$ [First basis function]
6. Set $B_{1,1}^g := 1$ [Initialize B^g matrix]
7. **while** $k \leq d^g$ **do**
8. Compute $\mu_{k+1}^g := \underset{\mu\in\mathcal{P}}{\operatorname{argmax}} \|(\delta_k^g g)(\mu,\cdot)\|_{L^\infty(\Omega)}$
9. Compute $x_{k+1}^g := \underset{x\in\Omega}{\operatorname{argmax}} |(\delta_k^g g)(\mu_{k+1}^g, x)|$ [$(k+1)$-th interpolation point]
10. Set $q_{k+1}^g(\cdot) := \dfrac{(\delta_k^g g)(\mu_{k+1}^g,\cdot)}{(\delta_k^g g)(\mu_{k+1}^g, x_{k+1}^g)}$ [$(k+1)$-th basis function]
11. Set $B_{i,k+1}^g := q_{k+1}^g(x_i^g)$, for all $1 \leq i \leq k+1$ [Increment matrix B^g]
12. $k \leftarrow k+1$ [Increment the size of the interpolation]
13. **end while**

Define $\eta_m^g(\mu) := \displaystyle\sum_{l=1}^{d^g} \gamma_{m,l} \lambda_l^g(\mu)$. The following property then holds:

Property 6.3 (Weak-intrusivity) EIM^g *leads to a weakly-intrusive procedure, since the resulting approximation of* A_μ *can be written*

$$A_\mu \approx \sum_{m=1}^{d^g} \eta_m^g(\mu) A^1_{\mu_m^g} + \mu A^0. \tag{6.15}$$

Remark 6.4 (Comparison with the standard EIM procedure in the RBM literature) *If the considered variational formulation contains only one term, the above procedure was already proposed in the RBM literature as a nonintrusive method to obtain a separated representation of the linear system under consideration, see [73]. For instance in* (6.15)*, if* $A^0 = 0$*, then* $A^1_{\mu_m^g} = A_{\mu_m^g}$*. In the general setting of* (6.9)*, this corresponds to* $R = 1$ *and* $S = 0$*. In any other case, the classical EIM needs to access independently matrices associated to each term of the variational formulation and thus cannot deliver a separated representation solely based on the* A_μ *matrices.*

6.4 The nonintrusive procedure

6.4.1 Description of the procedure

Denote by $G^g(\mu)$ the vector-valued function with d^g components such that $G_m^g(\mu) = g(\mu, x_m^g)$, for all $1 \leq m \leq d^g$. Then, from (7.8), $\lambda^g(\mu) = (\lambda_m^g(\mu))_{1\leq m\leq d^g}$ can be concisely written as $\lambda^g(\mu) = (B^g)^{-1} G^g(\mu)$. Notice that the computation of $\lambda^g(\mu)$ only requires the matrix B^g and the set of points $\{x_m^g\}_{1\leq m\leq d^g}$. Let $(z_p(\mu))_{1\leq p\leq d_{\max}}$ with $d_{\max} := d^g + 1$, be such that

$$z_p(\mu) := \begin{cases} \lambda_p^g(\mu) & 1 \le p \le d^g, \\ \mu & p = d^g + 1. \end{cases} \tag{6.16}$$

Recalling the notation $(M_m)_{i,j} := \int_\Omega q_m^g(x) \boldsymbol{\nabla}\theta_j(x) \cdot \boldsymbol{\nabla}\theta_i(x) dx$ for all $1 \le m \le d^g$, we infer from (6.13) that

$$A_\mu \approx \sum_{m=1}^{d^g} \lambda_m^g(\mu) M_m + \mu A^0 = \sum_{p=1}^{d_{\max}} z_p(\mu) T_p, \tag{6.17}$$

where the matrices

$$T_p := \begin{cases} M_p & 1 \le p \le d^g, \\ A^0 & p = d^g + 1 = d_{\max}, \end{cases} \tag{6.18}$$

are independent of μ. Note that $d_{\max}$ is the number of matrices to precompute and store when using the approximation (6.13).

The key idea is now to apply a second EIM to approximate $z_p(\mu)$, where z is seen as a function depending on the two variables p and μ. The EIM procedure to approximate $z_p(\mu)$ is denoted by EIM^z and its offline stage is detailed in Algorithm 5. This implementation correspond to the variant from the classical EIM (see Algorithm 2). As explained in Section 5.4.2, this variant leads to a nonintrusive without resorting to an additional change of basis. We refer to Section 7.2 for more details about the differences between the EIM variant considered here and the classical algorithm. The number of interpolation points is denoted by $d^z \le d_{\max}$. In the loop on k in Algorithm 5, the residual operator δ_k^z is defined by $\delta_k^z = \mathrm{Id} - I_k^z$, where

$$\left(I_k^z z\right)_p(\mu) := \sum_{m=1}^{k} \beta_m^z(\mu) z_p(\mu_m^z), \tag{6.19}$$

and

$$\sum_{m=1}^{k} B_{m,l}^z \beta_m^z(\mu) = q_l^z(\mu), \qquad 1 \le l \le k. \tag{6.20}$$

Owing to the interpolation property, there holds $(I_{d^z}^z z)_{p_k^z}(\mu) = z_{p_k^z}(\mu)$ for all $1 \le k \le d^z$ and all $\mu \in \mathcal{P}$. If $d^z = d_{\max}$, all the indices p are selected in Algorithm 5 and $(I_{d_{\max}}^z z)_p(\mu) = z_p(\mu)$ for all $1 \le p \le d_{\max}$ and all $\mu \in \mathcal{P}$. Observe that we can stop EIM^z before $d^z = d_{\max}$ interpolation matrices have been computed, see Sections 6.5.2 and 6.5.3 for some illustrations.

Injecting the approximation (6.19) with $k = d^z$ into the right-hand side of (6.17) with $z_p(\mu)$ replaced by $(I_{d^z}^z z)_p(\mu)$ yields

$$A_\mu \approx \sum_{p=1}^{d_{\max}} T_p \sum_{m=1}^{d^z} \beta_m^z(\mu) z_p(\mu_m^z) = \sum_{m=1}^{d^z} \beta_m^z(\mu) \sum_{p=1}^{d_{\max}} T_p z_p(\mu_m^z) \approx \sum_{m=1}^{d^z} \beta_m^z(\mu) A_{\mu_m^z}, \tag{6.21}$$

where $\beta_m^z(\mu)$ is obtained from (6.20). The right-hand side of (6.21) is the desired separated representation of A_μ that can be built in a nonintrusive way.

6.4.2 Practical implementation

To compute the L^∞-norms and determine the argmax in Algorithms 6 and 5, it is convenient to consider finite subsets of $\mathcal{P}$ and Ω, denoted respectively by $\mathcal{P}_{\text{trial}}$ and Ω_{trial}. This

Algorithm 5 Offline stage of EIM^z

1. Choose $d^z > 1$ [Number of interpolation points]
2. Set $k := 1$
3. Compute $p_1^z := \underset{1 \le p \le d^g+1}{\operatorname{argmax}} \|(z)_p(\cdot)\|_{L^\infty(\mathcal{P})}$
4. Compute $\mu_1^z := \underset{\mu \in \mathcal{P}}{\operatorname{argmax}} |(z)_{p_1^z}(\mu)|$ [First interpolation point]
5. Set $q_1^z(\cdot) := \dfrac{(z)_{p_1^z}(\cdot)}{(z)_{p_1^z}(\mu_1^z)}$ [First basis function]
6. Set $B_{1,1}^z := 1$ [Initialize B^z matrix]
7. **while** $k \le d^z$ **do**
8. Compute $p_{k+1}^z := \underset{1 \le p \le d^g+1}{\operatorname{argmax}} \|(\delta_k^z z)_p(\cdot)\|_{L^\infty(\mathcal{P})}$,
9. Compute $\mu_{k+1}^z := \underset{\mu \in \mathcal{P}}{\operatorname{argmax}} |(\delta_k^z z)_{p_{k+1}^z}(\mu)|$ [$(k+1)$-th interpolation point]
10. Set $q_{k+1}^z(\cdot) := \dfrac{(\delta_k^z z)_{p_{k+1}^z}(\cdot)}{(\delta_k^z z)_{p_{k+1}^z}(\mu_{k+1}^z)}$ [$(k+1)$-th basis function]
11. $B_{i,k+1}^z := q_{k+1}^z(\mu_i^z)$, for all $1 \le i \le k+1$ [Increment matrix B^z]
12. $k \leftarrow k+1$ [Increment the size of the interpolation]
13. **end while**

becomes necessary when, for instance, the function $g(\mu, x)$ is not known analytically, but only for some elements of $\mathcal{P}$ and Ω. It seems natural to take for Ω_{trial} the set of Gauss points on which the quadrature formulae to compute the integrals in (7.42) are defined. However, this supposes to know and manipulate the set of the Gauss points associated with the mesh. Since the functions q^g defined in Algorithm 6 are only used to construct the matrix B^g and are not directly integrated with respect to x to carry out the interpolation (6.21), it is possible to write the procedure with any set Ω_{trial} sampling the geometry. Such an approach is considered in the numerical example of Section 6.5.3. More generally, the sets $\mathcal{P}_{\text{trial}}$ and Ω_{trial} should be fine enough to capture all the phenomena, but not too fine to limit the overall computational cost. The numerical examples of Section 6.5 indicate that high accuracy can be obtained with simple choices for $\mathcal{P}_{\text{trial}}$ and Ω_{trial} on nontrivial cases.

In addition to the two sets $\mathcal{P}_{\text{trial}}$ and Ω_{trial}, the number of interpolation points d^g and d^z for each EIM have to be chosen. The choice we made is to stop the two EIM's when respectively $(\delta_k^g g)(\mu_{k+1}^g, x_{k+1}^g)$ and $(\delta_k^z z)_{p_{k+1}^z}(\mu_{k+1}^z)$ have reached a prescribed threshold, typically set at the level of the machine precision.

Finally, we specify the offline and online stages of our procedure when used within the RBM. EIM^g and the offline stage of EIM^z are part of the offline stage of the RBM. During the online stage of the RBM, the reduced matrix is constructed as

$$\hat{A}_\mu \approx \sum_{m=1}^{d^z} \beta_m^z(\mu) U^t A_{\mu_m^z} U, \tag{6.22}$$

so that only the online stage of EIM^z (i.e., the resolution of (6.20)) is needed.

6.4.3 Illustration

As a first illustration, we consider the following boundary-value problem:

$$-\frac{d}{dx}\left(\exp(\mu x)\frac{du}{dx}(x)\right)+\mu u(x)=1 \qquad \text{in } \Omega:=(-3,3), \tag{6.23}$$

with the following Dirichlet boundary condition $u(-3)=u(3)=0$. The weak form reads: Find $u\in H_0^1(\Omega)$ such that for all $u^t\in H_0^1(\Omega)$,

$$a_\mu(u,u^t)=\int_\Omega u^t(x)dx, \tag{6.24}$$

with

$$a_\mu(u,u^t):=\int_\Omega \exp(\mu x)\frac{du}{dx}(x)\frac{du^t}{dx}(x)dx+\int_\Omega \mu u(x)u^t(x)dx. \tag{6.25}$$

First-order continuous Lagrange finite elements are used, with a three-point quadrature formula in each mesh cell. The mesh is uniform with $h_x=0.015$. Ω_{trial} is taken to be the set of Gauss points on the obtained mesh, and $\mathcal{P}_{\text{trial}}=\{1,1+h_\mu,1+2h_\mu,...,3\}$ with $h_\mu=0.005$. To derive the separated approximation (6.21), EIM^g is first applied to $g(\mu,x):=\exp(\mu x)$. Then, the vector-valued function $z(\mu)$ is constructed using (6.16). The quality of the whole procedure is measured, for various values of d^g and $d^z=d^g+1$ using two error criteria: (i) the relative Frobenius norm error on the matrix A_μ and (ii) the relative $L^2(\Omega)$-norm error on the solution, see Figures 6.1 and 6.2.

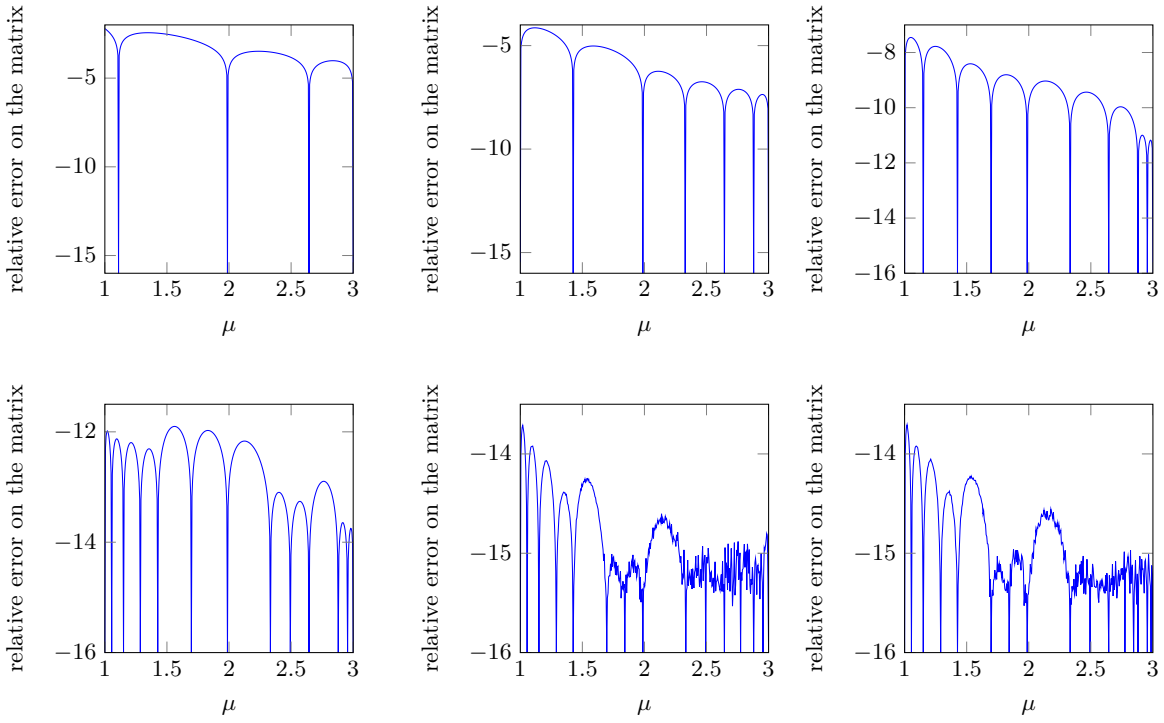

Fig. 6.1. Log_{10} of the relative Frobenius norm error on the matrix A_μ for $d^g=3,6,9,12,14,16$, and $d^z=d^g+1$.

We conclude from this first test case that the present method allows for a very good approximation of the matrix and the solution.

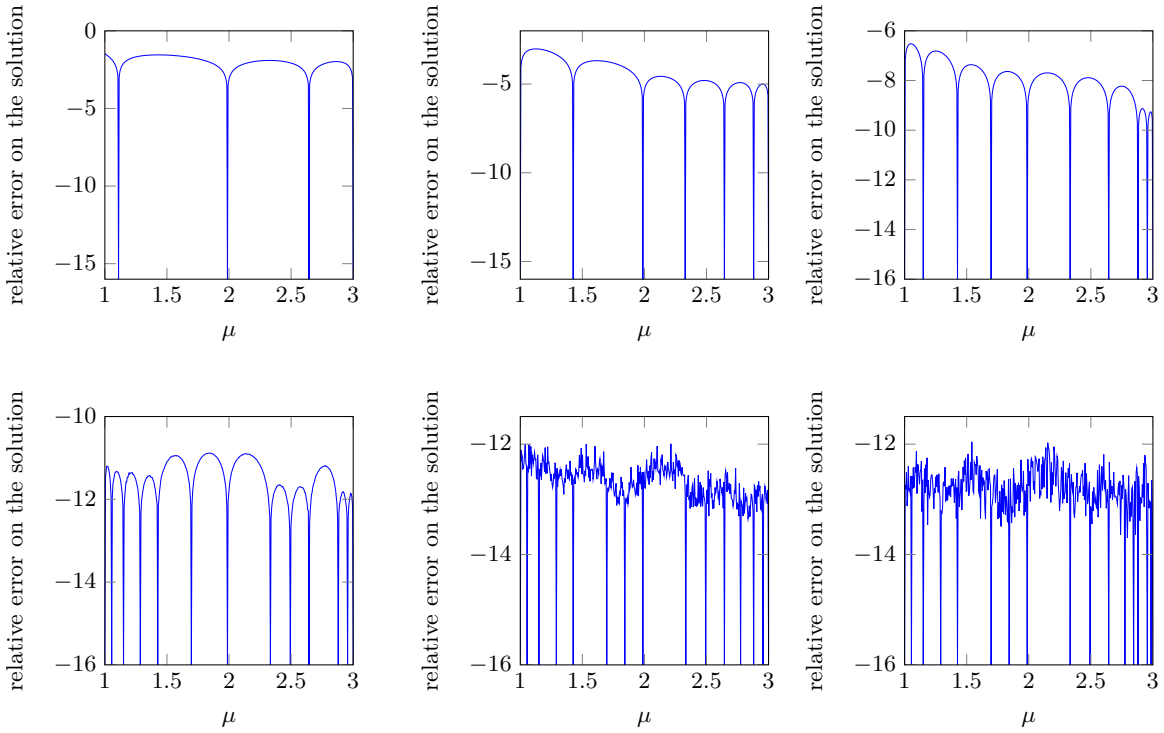

Fig. 6.2. Log_{10} of the relative $L^2(\Omega)$-norm error on the solution for $d^g = 3, 6, 9, 12, 14, 16$, and $d^z = d^g + 1$.

6.5 Extension to more general parameter dependence

The goal of this section is to show how to extend the nonintrusive procedure described in Section 6.4 to more complex parameter dependence. We illustrate the procedure on an industrial test case, namely a frequency-dependent three-dimensional aeroacoustic scattering problem.

6.5.1 Generalization of the nonintrusive procedure

Recall the general form of the matrix A_μ to approximate:

$$A_\mu = \sum_{\varrho=1}^{R} A_\mu^\varrho + \sum_{s=1}^{S} \psi_s(\mu) A^s, \tag{6.26}$$

where A_μ^ϱ are matrices that require to integrate some functions $g^\varrho(\mu, x)$ over Ω, ψ^s are given functions of μ and A^s are μ-independent matrices resulting from some integration over Ω. EIM^g is applied independently to each $g^\varrho(\mu, x)$, for all $1 \le \varrho \le R$, where the number of interpolation points, respectively $(d^g)^\varrho$, may differ from one EIM^g to the other. These procedures lead to the construction of the functions $(\lambda_m^g)^\varrho(\mu)$, for all $1 \le \varrho \le R$, all $1 \le m \le (d^g)^\varrho$, and all $\mu \in \mathcal{P}_{\text{trial}}$, using (7.8). Then, define the functions $(z_p(\mu))_{1 \le p \le d_{\max}}$ with $d_{\max} := \sum_{\varrho=1}^{R} (d^g)^\varrho + S$ such that

$$z_p(\mu) := \begin{cases} (\lambda_m^g)^1(\mu), & 1 \le p \le (d^g)^1, \quad m = p, \\ \vdots & \\ (\lambda_m^g)^R(\mu), & 1 + \sum_{\varrho=1}^{R-1} (d^g)^\varrho \le p \le \sum_{\varrho=1}^{R} (d^g)^\varrho, \quad m = p - \sum_{\varrho=1}^{R-1} (d^g)^\varrho, \\ \psi_1(\mu), & p = \sum_{\varrho=1}^{R} (d^g)^\varrho + 1, \\ \vdots & \\ \psi_S(\mu), & p = \sum_{\varrho=1}^{R} (d^g)^\varrho + S, \end{cases} \tag{6.27}$$

and let EIM^z be applied to $z_p(\mu)$, with d^z interpolation points, such that $d^z \le d_{\max} = \sum_{\varrho=1}^{R} (d^g)^\varrho + S$, to obtain an approximation of A_μ in the same form as (6.21). Note that $d_{\max}$ is the number of matrices to precompute and store when using the approximation (6.13), while the number of matrices to precompute and store when using (6.21) is d^z; in our numerical examples (see below), accurate representations of A_μ are already achieved for d^z smaller than $d_{\max}$. Notice also that in total, there are $(R+1)$ EIM procedures to be applied.

6.5.2 Sound-hard scattering in the air at rest

The problem of interest is the sound-hard scattering of an acoustic monopole source of wave number μ by an aircraft (whose boundary is denoted by Γ) in the air at rest, in the time-harmonic case. To simulate the noise created by one of the engines, the monopole is located under the left wing of the plane. This is a classical Helmholtz exterior problem, for which one possible weak formulation is: Find $u \in H^{\frac{1}{2}}(\Gamma)$ such that for all $u^t \in H^{\frac{1}{2}}(\Gamma)$,

$$a_\mu(u, u^t) = \int_\Gamma f_{\mathrm{inc}_\mu}(x) u^t(x) dx, \tag{6.28}$$

where

$$\begin{aligned} a_\mu(u, u^t) := & \frac{1}{4\pi} \int_\Gamma \int_\Gamma \frac{\exp(i\mu|x-y|)}{|x-y|} \left(\overrightarrow{\mathrm{curl}}_\Gamma u(x) \cdot \overrightarrow{\mathrm{curl}}_\Gamma u^t(y) \right) dxdy \\ & - \frac{\mu^2}{4\pi} \int_\Gamma \int_\Gamma \frac{\exp(i\mu|x-y|)}{|x-y|} u(x) u^t(y) \left(\overrightarrow{n_x} \cdot \overrightarrow{n_y} \right) dxdy, \end{aligned} \tag{6.29}$$

where $\overrightarrow{\mathrm{curl}}_\Gamma$ denotes the surfacic curl on Γ, $\overrightarrow{n_x}$ the unit normal vector on Γ pointing towards the medium of propagation, and f_{inc_μ} is the incident acoustic field created by the source. We refer to [80, Section 3.4] for details on the derivation of (6.28), and justifications on the well-posedness of the integral in (6.29). The parameter of interest is the wave number μ of the acoustic monopole source. The Boundary Element Method (BEM) is used to approximate problem (6.28). This leads to a dense μ-dependent matrix $(A_\mu)_{i,j} = a_\mu(\theta_j, \theta_i)$, where $(\theta_i)_{1 \le i \le n}$ denote the basis functions of the considered finite element space on Γ. Two different meshes, on which the matrices are assembled, are considered, see Table 6.1 and Figure 6.3. The in-house code ACTIPOLE developed by EADS-IW and Airbus [33, 34] is used. This test case is a challenging benchmark for

two reasons. First, the Green kernel $G_\mu(x,y) := \frac{\exp(i\mu|x-y|)}{4\pi|x-y|}$ oscillates at a frequency proportional to the parameter of interest μ, and, secondly, the obtained matrices are dense and complex-valued. Mesh 2 leads to a very large matrix and cannot be stored in an average desktop computer RAM. The tests on Mesh 1 have been computed on a simple laptop with 4 Go of RAM, whereas the tests on Mesh 2 have been computed on CCRT's Curie supercomputer [1].

	Mesh 1	Mesh 2
number of triangles	7,886	40,576
number of vertices	3,945	20,290
smallest edge (mm)	6.53	6.53
mean edge (mm)	437.52	192.92
largest edge (mm)	718.99	389.29
number of complex nonzero coefficients per matrix	1.56×10^7	4.12×10^8
memory usage to store one matrix in binary format (Go)	0.23	6.5

Table 6.1. Characteristics of the two considered meshes.

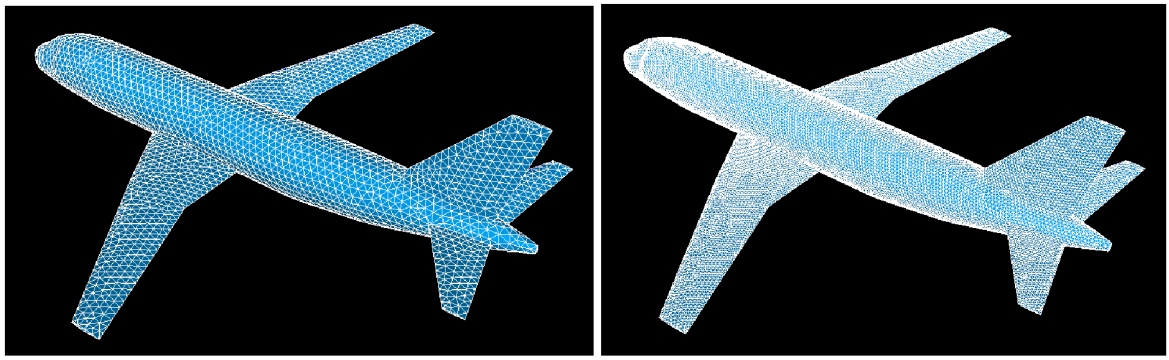

Fig. 6.3. Airbus A319: Mesh 1 and Mesh 2.

To derive the approximation (6.21) for A_μ, we carry out EIM^g to approximate

$$g(\mu, r) := \exp(i\mu r), \quad r = |x-y|, \quad x, y \in \Gamma. \tag{6.30}$$

We choose $\mu \in \mathcal{P}_{\text{trial}} := \{0.005, 0.01, ..., 2.5\}$, a set of 1000 values for the wave number, so that the highest wave number for the source corresponds to a wavelength 5 times larger than the mean edge of Mesh 1. A natural choice for the discrete set of values for x and y is the set of Gauss points associated with the considered mesh, on which the quadrature formulae used to compute the integrals (6.29) are defined. The associated discrete set of values for $r = |x-y|$ is roughly proportional to the square of the number of Gauss points, and equals 7.8×10^6 for Mesh 1. To reduce the computational cost, a subsample of 10^5 values for r, that has a very close density to the one obtained from the set of Gauss points, is chosen, see Figure 6.4.

Once EIM^g has been carried out, we can write

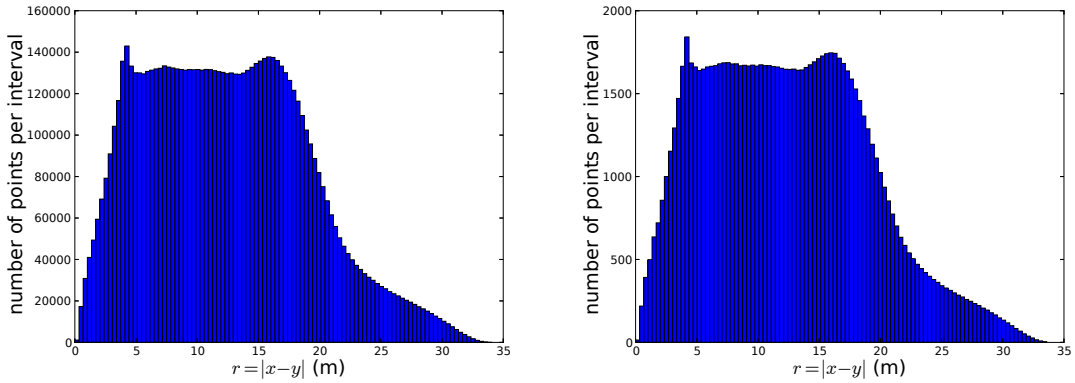

Fig. 6.4. Histograms: discrete values of $r = |x - y|$ over the Gauss points from Mesh 1 (left), and chosen set of size 10^5 (right).

$$A_\mu \approx \left(1 + \mu^2\right) \sum_{m=1}^{d^g} \lambda_m^g(\mu) M_m,$$

where the matrices M_m have been defined in Section 6.4.1, so that the approximation (6.21) can be written using

$$z_p(\mu) := \begin{cases} \lambda_m^g(\mu), & 1 \le m \le d^g, \quad p = m, \\ \mu^2 \lambda_m^g(\mu), & 1 \le m \le d^g, \quad p = m + d^g. \end{cases} \tag{6.31}$$

Note that we exploited the links in the functional dependence on μ for the two terms on the right-hand side of (6.29) to carry out only one EIM^g procedure.

EIM^g and EIM^z are carried out with respectively $d^g = 30$ and $d^z = 32$ interpolation points (notice that $d_{\max} = 60$). To check the accuracy of the approximation, we compute the relative Frobenius norm error on the matrix A_μ and the relative Euclidian norm error on the acoustic pressure computed using the approximate matrix, on a network of 400 points located behind the aircraft. Figure 6.5 presents the results on Mesh 1. In this figure, the relative differences are computed on 100 values of μ, namely one tenth of the considered parameter values, explaining why only 7 minima are achieved on the left plot. On the right plot concerning the acoustic pressure behind the aircraft, a large number of values are at the level of machine precision. Note that the right-hand side of (6.28) also depends on the parameter μ. To compute the right plot of Figure 6.5, we computed the exact values of this right-hand.

Figure 6.6 shows the solution to the problem on Mesh 1 and the relative difference of the solution using the exact matrix and its approximation for $\mu = 2.47$.

The simulation is repeated on Mesh 2, with $d^g = 50$ and $d^z = 50$. A twice as large frequency interval is considered since Mesh 2 has a better spatial resolution than Mesh 1. Figure 6.7 shows the relative Frobenius norm error on the matrix A_μ, confirming the accuracy of the approximation.

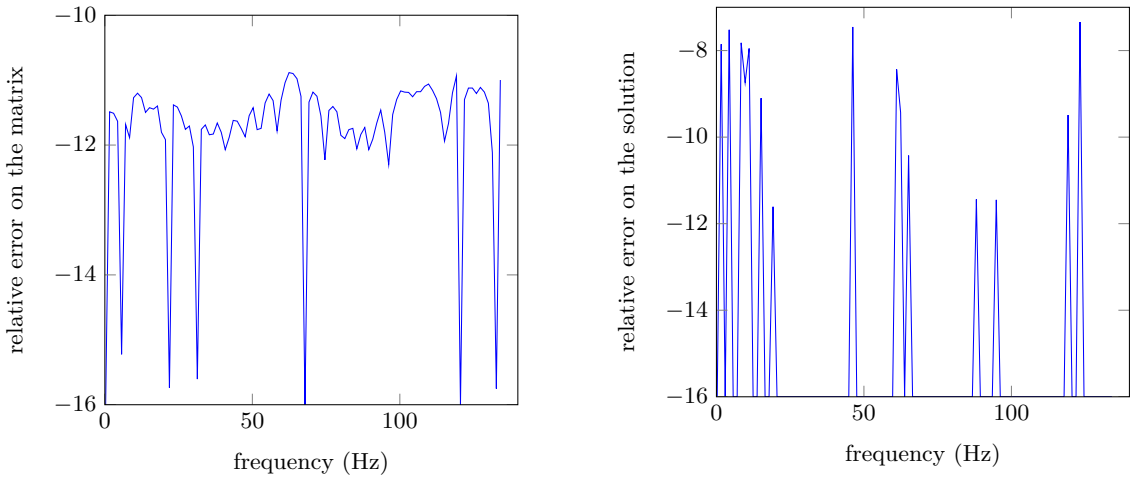

Fig. 6.5. Mesh 1, $\log_{10}$ of the relative error on the Frobenius norm of the matrix A_μ (left), and on the acoustic pressure computed using the approximate matrix on a network of 400 points located behind the plane in Euclidian norm (right), with $d^g = 30$ and $d^z = 32$.

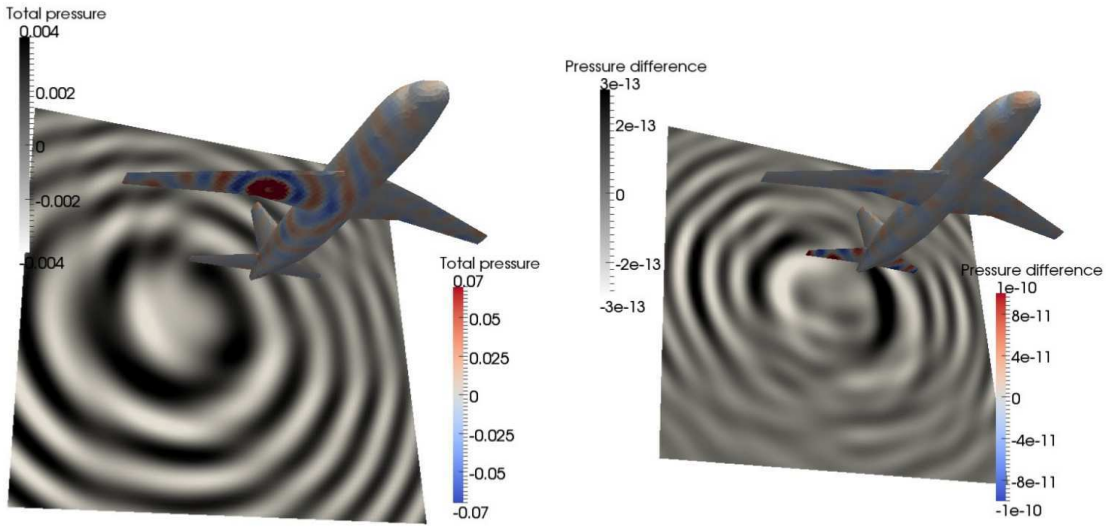

Fig. 6.6. Mesh 1: total acoustic field on the plane and on the network of points (left), and difference between the exact and approximate solution (right), for $\mu = 2.47$.

6.5.3 Sound-hard scattering in a non-uniform flow

Consider an ellipsoid with major axis directed along the z-axis. This object is included inside a larger ball, see Figure 7.3. The external border of the ball after discretization is denoted by Γ_∞. The complement of the ellipsoid in the ball is denoted by Ω^-. A potential flow is precomputed around the ellipsoid and inside the ball, such that the flow is uniform outside the ball, of Mach number 0.3 and directed along the z-axis. The flow is fixed, and does not depend on the parameter μ. An acoustic monopole source lies upstream of the ball, on the z-axis as well. The parameter is again the wave number of the monopole source.

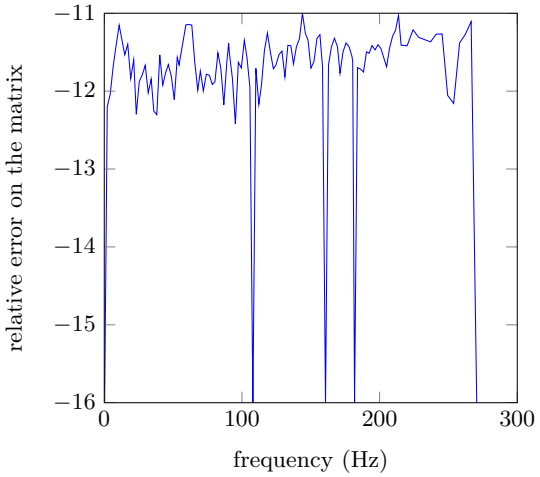

Fig. 6.7. Mesh 2, $\log_{10}$ of the relative error on the Frobenius norm of the matrix A_μ, with $d^g = 50$ and $d^z = 50$.

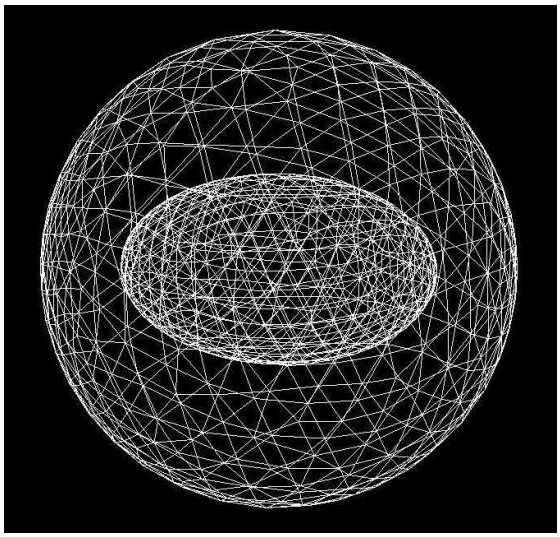

Fig. 6.8. Representation of the mesh.

The considered formulation is a coupled Finite Element Method (FEM) - BEM formulation (3.45) derived in Section 3.3.4. It consists in (i) applying a change of variable to transform the convected Helmholtz equation into the classical Helmholtz equation outside the ball, in order to apply a standard BEM, and (ii) stabilizing the formulation to avoid resonant frequencies associated with the eigenvalues of the Laplacian inside the ball of border Γ_∞. The formulation depends on the wave number of the source in a complex way, but we will see in our numerical tests that our nonintrusive procedure provides an accurate approximation of the resulting matrix as a linear combination of a few snapshots of the complete matrix at some wave numbers of the source.

Consider the product space $\mathbb{H} := H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma_\infty) \times H^1(\Gamma_\infty)$ with inner product $\left((\Phi, \lambda, p), (\Phi^t, \lambda^t, p^t)\right)_{\mathbb{H}} := (\Phi, \Phi^t)_{H^1(\Omega^-)} + (\lambda, \lambda^t)_{H^{-\frac{1}{2}}(\Gamma_\infty)} + (p, p^t)_{H^1(\Gamma_\infty)}$. The weak formulation is: Find $(\Phi, \lambda, p) \in \mathbb{H}$ such that $\forall\, (\Phi^t, \lambda^t, p^t) \in \mathbb{H}$,

$$\mathcal{V}_\mu(\Phi, \Phi^t) + \left(N_\mu(\gamma_0^- \Phi), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} + \left(\left(\tilde{D}_\mu - \frac{1}{2} I\right)(\lambda), \gamma_0^- \Phi^t\right)_{\Gamma_\infty} = \left(\gamma_1 f_{\mathrm{inc}\mu}, \gamma_0^- \Phi^t\right)_{\Gamma_\infty}, \quad (6.32a)$$

$$\left(\lambda^t, \left(D_\mu - \frac{1}{2} I\right)(\gamma_0^- \Phi)\right)_{\Gamma_\infty} - \left(\lambda^t, S_\mu(\lambda)\right)_{\Gamma_\infty} - i\left(\lambda^t, p\right)_{\Gamma_\infty} = -\left(\lambda^t, \gamma_0 f_{\mathrm{inc}\mu}\right)_{\Gamma_\infty}, \quad (6.32b)$$

$$\left(N_\mu(\gamma_0^- \Phi), p^t\right)_{\Gamma_\infty} + \left(\left(\tilde{D}_\mu + \frac{1}{2} I\right)(\lambda), p^t\right)_{\Gamma_\infty} - \delta_{\Gamma_\infty}(p, p^t) = \left(\gamma_1 f_{\mathrm{inc}\mu}, p^t\right)_{\Gamma_\infty}, \quad (6.32c)$$

where $(\cdot,\cdot)_{\Gamma_\infty}$ denotes the extension of the $L^2(\Gamma_\infty)$-inner product to the duality pairing on $H^{-\frac{1}{2}}(\Gamma_\infty) \times H^{\frac{1}{2}}(\Gamma_\infty)$, and where

$$\delta_{\Gamma_\infty}(p, q) := (\boldsymbol{\nabla}_{\Gamma_\infty} p, \boldsymbol{\nabla}_{\Gamma_\infty} q)_{\Gamma_\infty} + (p, q)_{\Gamma_\infty}, \quad (6.33)$$

with $\boldsymbol{\nabla}_{\Gamma_\infty}$ the surfacic gradient on Γ_∞, and

$$\mathcal{V}_\mu(\Phi, \Phi^t) := \int_{\Omega^-} \Xi \boldsymbol{\nabla}\overline{\Phi} \cdot \boldsymbol{\nabla}\Phi^t - \mu^2 \int_{\Omega^-} \beta \overline{\Phi}\Phi^t + i\mu \int_{\Omega^-} \boldsymbol{V} \cdot \left(\overline{\Phi}\boldsymbol{\nabla}\Phi^t - \Phi^t \boldsymbol{\nabla}\overline{\Phi}\right), \quad (6.34)$$

where $\beta := r\left(\left(\varsigma + \gamma_\infty^2 P\right)^2 - \gamma_\infty^4 M_\infty^2\right)$, $\boldsymbol{V} := r\left(\left(\varsigma + \gamma_\infty^2 P\right)\mathcal{N}\boldsymbol{M} - \gamma_\infty^3 \boldsymbol{M}_\infty\right)$, $\Xi := r\mathcal{N}\mathcal{O}\mathcal{N}$ with $r := \frac{\rho}{\rho_\infty}$, $\varsigma := \frac{c_\infty}{c}$, $\gamma_\infty := \frac{1}{\sqrt{1-M_\infty^2}}$, $P := \boldsymbol{M} \cdot \boldsymbol{M}_\infty$, $\mathcal{N} := I + C_\infty \boldsymbol{M}_\infty \boldsymbol{M}_\infty^T$, $\mathcal{O} := I - \boldsymbol{M}\boldsymbol{M}^T$, and $C_\infty := \frac{\gamma_\infty - 1}{M_\infty^2}$. In the above notation, the subscript ∞ is used for quantities outside the ball, ρ is the density of the flow, c is the speed of sound when the flow is at rest and $\boldsymbol{M} = \frac{\boldsymbol{v}}{c}$, where $\boldsymbol{v}$ is the velocity of the flow. The operators γ_0 and γ_1 are Dirichlet and Neumann traces on the coupling surface Γ_∞. The operators N_μ, D_μ, $\tilde{D}_\mu$, and S_μ are boundary integral operators, expressed in terms of the Green kernel $G_\mu(x, y) = \frac{\exp(i\mu|x-y|)}{4\pi|x-y|}$ associated with the Helmholtz equation at wave number μ.

The next step is to identify the dependencies in μ in the formulation (7.47). It turns out that the functions of μ involved in the integrals of the formulation (7.47) are μ, μ^2, $\exp(i\mu r)$, $\mu\exp(i\mu r)$, $\mu^2\exp(i\mu r)$, and $\mu\left(\frac{2i\pi\mu}{c}\mu - 1\right)\exp(i\mu r)$. As in the previous test case, EIM^g is carried out to approximate the function $g(\mu, r) = \exp(i\mu r)$, $r = |x - y|$, $x, y \in \Gamma_\infty$. We choose $\mu \in \mathcal{P}_{\mathrm{trial}} := \{10, 10.03, ..., 40\}$, a set of 1000 values for the wave number, so that the highest wave number of the source corresponds to a wavelength 5 times larger than the mean edge of the mesh. This time, instead of considering a subset of $|x - y|$ where x and y are the Gauss points associated with the mesh, we take $r \in \{0, h, ..., Nh\}$, where $N = 10000$ and $h = \frac{D}{N}$, D being the diameter of the sphere Γ_∞. With this choice, we no longer need to know the position of the Gauss points, but simply the diameter of the geometry of the test case.

Then, the functions $(\lambda_m^g(\mu))_{1 \le m \le d^g}, \mu \in \mathcal{P}_{\mathrm{trial}}$, are computed using (7.8), and the functions $z_p(\mu)$, $1 \le p \le d_{\max} := 3d^g + 3$, are defined by

$$z_p(\mu) := \begin{cases} \lambda_m^g(\mu), & 1 \le p \le d^g, \quad m = p, \\ \mu\lambda_m^g(\mu), & d^g + 1 \le p \le 2d^g, \quad m = p - d^g, \\ \mu^2\lambda_m^g(\mu), & 2d^g + 1 \le p \le 3d^g, \quad m = p - 2d^g, \\ 1, & p = 3d^g + 1, \\ \mu, & p = 3d^g + 2, \\ \mu^2, & p = 3d^g + 3. \end{cases} \tag{6.35}$$

EIM^g and EIM^z are carried out with respectively 17 and 20 interpolation points (notice that $d_{\max} = 71$).

Figure 6.9 shows the relative Frobenius norm error on the matrix A_μ and the relative Euclidian norm error on the acoustic pressure computed using the approximate matrix on a network of 400 points located behind the scattering ellipsoid. In this test case, an excellent accuracy is obtained with only 20 precomputed matrices.

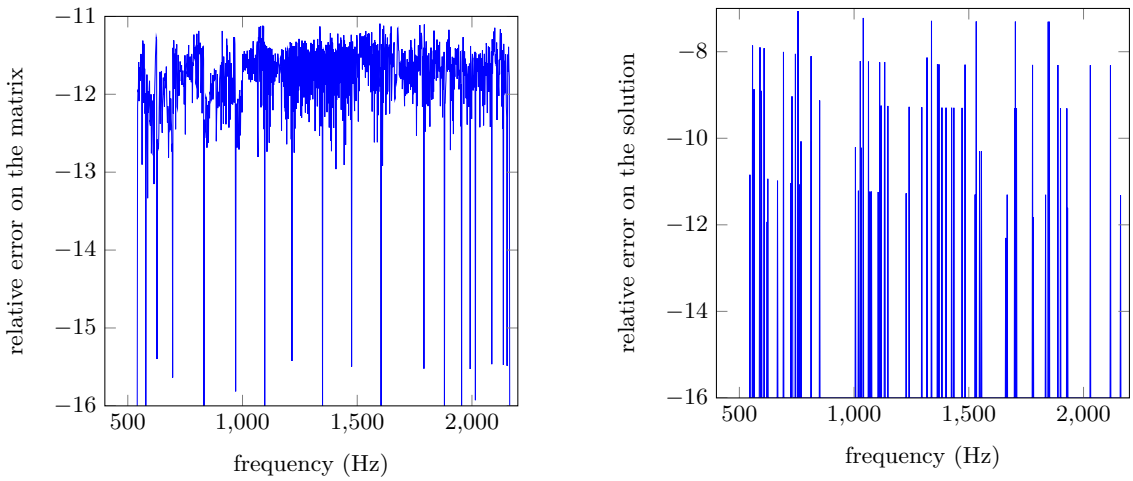

Fig. 6.9. Log_{10} of the relative error on the Frobenius norm of the matrix A_μ (left) and on the acoustic pressure computed using the approximate matrix computed using (6.21) on a network of 400 points located behind the object in Euclidian norm (right), with $d^g = 17$ and $d^z = 20$.

6.6 Outlook

The method described herein provides an efficient nonintrusive approximation of parameter-dependent linear systems, provided that the considered code can return the assembled matrix and that the corresponding weak formulation is known. The method offers a crucial practical advantage over existing methods since it avoids significant implementation efforts. In the present work, the choice has been made to approximate the whole matrix A_μ assembled by the code, but the procedure applies in the same way to the approximation of any linear functional l of the matrix A_μ, whereby

$$l(A_\mu) \approx \sum_{m=1}^{d^z} \beta_m^z(\mu) l(A_{\mu_m^z}), \tag{6.36}$$

where the storage of $A_{\mu_m^z}$ for all $1 \leq m \leq d^z$ is replaced by the storage of $l(A_{\mu_m^z})$ for all $1 \leq m \leq d^z$, which may be much lighter in terms of memory usage. The efficient construction of the reduced matrix $\hat{A}_\mu$ in the RBM corresponds to $l(A_\mu) = U^t A_\mu U$, as explained in the introduction.

Finally, we observe that in the case where the right-hand side C of the problem (6.6) is also dependent on the parameter μ (then written C_μ), the same procedure can be applied to derive a separated representation of C_μ.

7

A nonintrusive Reduced Basis Method applied to aeroacoustic simulations

This chapter is based on the preprint [Pr3]

Summary. The Reduced Basis Method can be exploited in an efficient way only if the so-called affine dependence assumption on the operator and right-hand side of the considered problem with respect to the parameters is satisfied. When it is not, the Empirical Interpolation Method is usually used to recover this assumption approximately. In both cases, one must be able to access and modify the assembly routines of the corresponding computational code, leading to an intrusive procedure. In this work, we derive variants of the EIM algorithm and explain how they can be used to provide a nonintrusive procedure to compute affine decompositions of the operator and of the right-hand side of the problem while being computable for all values of the parameter (and not only at training points). We explain how this algorithm can be effectively applied in various contexts. We present various examples of aeroacoustic problems solved by integral equations using the reduced basis method and show how this algorithm can benefit from the linear algebra tools available in the considered code.

7.1 Introduction

In many problems such as optimization, uncertainty propagation and real-time simulations, one needs to solve a (complex) parametrized problem for many values of the parameters. Among the various available methods to reduce the computational cost, the Reduced Basis Method (RBM) has received increased interest over the last decade (see [72, 73, 85] for a detailed presentation and [36] for some convergence results). Consider the following problem: Find $u_\mu \in \mathcal{V}$ such that

$$a_\mu(u_\mu, v) = c_\mu(v), \qquad \forall v \in \mathcal{V}, \tag{7.1}$$

where $\mu \in \mathcal{P}$ is the parameter, a_μ is a sesquilinear form, c_μ is a linear form, and $\mathcal{V}$ is a finite-dimensional functional space of size n, where n is typically very large. Since the linear problem (7.1) is written on a finite-dimensional space, we can consider the following matrix form:

$$A_\mu U_\mu = C_\mu, \tag{7.2}$$

where $A_\mu \in \mathbb{C}^{n\times n}$ and $C_\mu \in \mathbb{C}^n$. We refer to the solutions to (7.1) as truth solutions.

The RBM allows one to compute very fast an approximation of the truth solution u_μ by means of an offline/online procedure. The online stage is a Galerkin procedure written on a basis of truth solutions u_{μ_j}, $1 \leq j \leq \hat{n} \ll n$, rather than on a basis of $\mathcal{V}$. The parameter

values μ_j are selected by a greedy algorithm in the offline stage, where the functions u_{μ_j} of the reduced basis are also precomputed. Denote by U the rectangular matrix of size $n \times \hat{n}$ such that $(U)_{i,j} = \gamma_i(\mu_j)$, where $\gamma_i(\mu_j)$, $1 \leq i \leq n$, are the coefficients of u_{μ_j} on the basis of $\mathcal{V}$. Then, the RBM approximation is computed by solving the reduced problem $\hat{A}_\mu \hat{\gamma}(\mu) = \hat{C}_\mu$, where $\hat{A}_\mu = U^t A_\mu U$ and $\hat{C}_\mu = U^t C_\mu$, so that $\hat{u}_\mu(x) := \sum_{j=1}^{\hat{n}} \hat{\gamma}_j(\mu) u_{\mu_j}(x) \approx u_\mu(x)$.

The efficiency of the RBM hinges on the assumption of an affine dependence of the operator and the right-hand side with respect to the parameter. This assumption states that

$$A_\mu = \sum_{i=1}^{d} \alpha_i(\mu) A_i. \tag{7.3}$$

We only discuss the case of the operator A_μ, the right-hand side C_μ being treated in the same way. Owing to the separated representation (7.3), the assembly of the reduced problems and the computation of the a posteriori error bound are performed in complexity independent of n. When affine dependence does not hold, the Empirical Interpolation Method (EIM) can be used to recover it approximately. In any case, replacing A_μ by the right-hand side of (7.3) so as to assemble the reduced problems in complexity independent of n requires in general nontrivial modifications of the assembling routines of the computational code since various terms of the variational formulation at hand corresponding to the matrices A_i in (7.3) have to be accessed separately.

When considering the approximation of A_μ and C_μ by the EIM in a Reduced Basis context, some requirements are crucial in practical applications: (i) the assembly of the online problems must be of complexity independent of n to ensure online-efficiency, (ii) the approximation must be computable for all $\mu \in \mathcal{P}$, and not only on some training points in $\mathcal{P}$, (iii) the procedure should be nonintrusive in the sense that it should not require to assemble new terms such as the matrices A_i, but just the matrices A_μ for selected values of the parameter μ. In this work, we derive various variants for the classical EIM algorithm and discuss their advantages, drawbacks, and properties. We show that some of the derived approximation procedures meet the requirements (i) and (ii), but not (iii). Meeting requirement (iii) hinges on deriving a separated representation of A_μ in the form

$$A_\mu \approx \sum_{m=1}^{r} \beta_m(\mu) A_{\mu_m}, \tag{7.4}$$

where μ_m, $1 \leq m \leq r$, are some selected values of the parameter (which are different from the parameter values μ_j, $1 \leq j \leq \hat{n}$, selected by the greedy algorithm in the offline stage of the RBM). In this work, we derive such nonintrusive approximation procedures, to the price of an additional EIM approximation of which we can control the accuracy. One of these was first considered in Chapter 6; we explain that among all the possible choices, this approximation property is optimal in terms of computational savings and, at a certain limit, interpolant with respect to the parameter.

In Section 7.2, we briefly recall the classical EIM algorithm, and present some new variants that are useful in the present context. Then, procedures to approximate A_μ meeting the requirements (i), (ii), and (iii) are derived in Section 7.3. Finally, numerical simulations are presented on aeroacoustic problems solved by integral equations in Section 7.4, where the use of the nonintrusive formulae is crucial.

7.2 Classical EIM and variants

Consider a function $g(\mu, x)$ defined over $\mathcal{P} \times \Omega$ for two sets $\mathcal{P}$ and Ω. We look for an approximation of this function in a separated form with respect to μ and x. There are different possible ways to achieve such an approximation using EIM-like algorithms. An EIM algorithm consists in an offline stage, where some quantities are precomputed within a greedy procedure, and an online stage (where the approximation is computed making use of these precomputed quantities).

First, we recall the classical EIM as defined in [73]. We denote the offline stage of this algorithm by EIM$^{\mathrm{S1}}$, S1 refering to *Slice 1*, since the obtained approximation is interpolant with respect to the first variable (see Proposition 7.2 below). Fix an integer $d > 1$ (the total number of interpolation points). For all $1 \leq k < d$, the rank-k approximation operator is defined as

$$\left(I_k^{\mathrm{S1}} g\right)(\mu, x) := \sum_{m=1}^{k} \lambda_m^{\mathrm{S1}}(\mu) q_m^{\mathrm{S1}}(x), \tag{7.5}$$

where the functions $\lambda_m^{\mathrm{S1}}(\mu)$, $1 \leq m \leq k$, solve the linear system

$$\sum_{m=1}^{k} B_{l,m}^{\mathrm{S1}} \lambda_m^{\mathrm{S1}}(\mu) = g(\mu, x_l^{\mathrm{S1}}), \qquad \forall 1 \leq l \leq k. \tag{7.6}$$

The notation λ always refers to the solution of the online problem, which can be a function of μ or of x in what follows. The functions $q_m^{\mathrm{S1}}(\cdot)$ and the matrices $B^{\mathrm{S1}} \in \mathbb{R}^{k \times k}$ are constructed as described in Algorithm 6, where $\delta_k^{\mathrm{S1}} = \mathrm{Id} - I_k^{\mathrm{S1}}$.

Note that this algorithm constructs a set of points $\{x_l^{\mathrm{S1}}\}_{1 \leq l \leq d}$ in Ω used in (7.6), and also a set of points $\{\mu_l^{\mathrm{S1}}\}_{1 \leq l \leq d}$ in $\mathcal{P}$.

Algorithm 6 Offline stage EIM$^{\mathrm{S1}}$

1. Choose $d > 1$ [Number of interpolation points]
2. Set $k := 1$
3. Compute $\mu_1^{\mathrm{S1}} := \underset{\mu \in \mathcal{P}}{\mathrm{argmax}} \|g(\mu, \cdot)\|_{L^\infty(\Omega)}$
4. Compute $x_1^{\mathrm{S1}} := \underset{x \in \Omega}{\mathrm{argmax}} |g(\mu_1^{\mathrm{S1}}, x)|$ [First interpolation point]
5. Set $q_1^{\mathrm{S1}}(\cdot) := \dfrac{g(\mu_1^{\mathrm{S1}}, \cdot)}{g(\mu_1^{\mathrm{S1}}, x_1^{\mathrm{S1}})}$ [First basis function]
6. Set $B_{1,1}^{\mathrm{S1}} := 1$ [Initialize matrix B^{S1}]
7. **while** $k < d$ **do**
8. Compute $\mu_{k+1}^{\mathrm{S1}} := \underset{\mu \in \mathcal{P}}{\mathrm{argmax}} \|(\delta_k^{\mathrm{S1}} g)(\mu, \cdot)\|_{L^\infty(\Omega)}$
9. Compute $x_{k+1}^{\mathrm{S1}} := \underset{x \in \Omega}{\mathrm{argmax}} |(\delta_k^{\mathrm{S1}} g)(\mu_{k+1}^{\mathrm{S1}}, x)|$ [$(k+1)$-th interpolation point]
10. Set $q_{k+1}^{\mathrm{S1}}(\cdot) := \dfrac{(\delta_k^{\mathrm{S1}} g)(\mu_{k+1}^{\mathrm{S1}}, \cdot)}{(\delta_k^{\mathrm{S1}} g)(\mu_{k+1}^{\mathrm{S1}}, x_{k+1}^{\mathrm{S1}})}$ [$(k+1)$-th basis function]
11. Set $B_{i,k+1}^{\mathrm{S1}} := q_{k+1}^{\mathrm{S1}}(x_i^{\mathrm{S1}})$, for all $1 \leq i \leq k+1$ [Increment matrix B^{S1}]
12. $k \leftarrow k + 1$ [Increment the size of the decomposition]
13. **end while**

The online stage of the classical EIM$^{\mathrm{S1}}$ amounts to (7.5)-(7.6) for $k = d$. This yields

$$\left(I_d^{\mathrm{S1O1}} g\right)(\mu, x) := \sum_{m=1}^{d} \lambda_m^{\mathrm{S1O1}}(\mu) q_m^{\mathrm{S1}}(x), \tag{7.7}$$

where the functions $\lambda_m^{\mathrm{S1O1}}(\mu)$, $1 \leq m \leq d$, solve the linear system

$$\sum_{m=1}^{d} B_{l,m}^{\mathrm{S1}} \lambda_m^{\mathrm{S1O1}}(\mu) = g(\mu, x_l^{\mathrm{S1}}), \qquad \forall 1 \leq l \leq d. \tag{7.8}$$

In (7.7)-(7.8), the exponent O1 refers to *Online problem 1.*

Consider now the following online problem, denoted by O2 for *Online problem 2.* Like O1, O2 is based on quantities precomputed in Algorithm 6:

$$\left(I_d^{\mathrm{S1O2}} g\right)(\mu, x) := \sum_{m=1}^{d} \lambda_m^{\mathrm{S1O2}}(x) g(\mu, x_m^{\mathrm{S1}}), \tag{7.9}$$

where the functions $\lambda_m^{\mathrm{S1O2}}(x)$, $1 \leq m \leq d$, solve the linear system

$$\sum_{m=1}^{d} (B^{\mathrm{S1}})_{l,m}^{t} \lambda_m^{\mathrm{S1O2}}(x) = q_l^{\mathrm{S1}}(x), \qquad \forall 1 \leq l \leq d. \tag{7.10}$$

Proposition 7.1 *For all $x \in \Omega$ and all $\mu \in \mathcal{P}$,*

$$(I_d^{\mathrm{S1O1}} g)(\mu, x) = (I_d^{\mathrm{S1O2}} g)(\mu, x). \tag{7.11}$$

Proof. The expression $\lambda_m^{\mathrm{S1O1}}(\mu) = \sum_{l=1}^{d} (B^{\mathrm{S1}})_{m,l}^{-1} g(\mu, x_l^{\mathrm{S1}})$, $1 \leq m \leq d$, holds from (7.8), leading to $(I_d^{\mathrm{S1O1}} g)(\mu, x) = \sum_{m=1}^{d} \sum_{l=1}^{d} (B^{\mathrm{S1}})_{m,l}^{-1} g(\mu, x_l^{\mathrm{S1}}) q_m^{\mathrm{S1}}(x)$. In the same fashion, the expression $\lambda_l^{\mathrm{S1O2}}(x) = \sum_{m=1}^{d} (B^{\mathrm{S1}})_{m,l}^{-1} q_m^{\mathrm{S1}}(x)$, $1 \leq l \leq d$, holds from (7.10), leading to $(I_d^{\mathrm{S1O2}} g)(\mu, x) = \sum_{l=1}^{d} \sum_{m=1}^{d} (B^{\mathrm{S1}})_{m,l}^{-1} q_m^{\mathrm{S1}}(x) g(\mu, x_l^{\mathrm{S1}})$. We recognize the expression of $(I_d^{\mathrm{S1O1}} g)(\mu, x)$ by switching the two dummy indices l and m. □ ♢

Although the approximations $I_d^{\mathrm{S1O1}} g$ and $I_d^{\mathrm{S1O2}} g$ are equal and both based on Algorithm 6 for the offline stage, they rely on different online problems (7.8) and (7.10), which induces different properties when considering the approximation of certain quantities based on $g(\mu, x)$, see Section 7.3.

A variant of Algorithm 6 is obtained by switching the roles of x and μ in the offline stage. We denote this variant by $\mathrm{EIM}^{\mathrm{S2}}$, S2 refering to *Slice 2*, since the obtained approximation is interpolant with respect to the second variable (see Proposition 7.3 below). Fix an integer $d > 1$ (the total number of interpolation points). Then, for all $1 \leq k < d$, the rank-k approximation operator is defined as

$$\left(I_k^{\mathrm{S2}} g\right)(\mu, x) := \sum_{m=1}^{k} \lambda_m^{\mathrm{S2}}(x) q_m^{\mathrm{S2}}(\mu), \tag{7.12}$$

where the functions $\lambda_m^{\mathrm{S2}}(x)$, $1 \leq m \leq k$, solve the linear system

$$\sum_{m=1}^{k} B_{l,m}^{\mathrm{S2}} \lambda_m^{\mathrm{S2}}(x) = g(\mu_l^{\mathrm{S2}}, x), \qquad \forall 1 \leq l \leq k. \tag{7.13}$$

The functions $q_m^{\mathrm{S2}}(\cdot)$ and the matrices $B^{\mathrm{S2}} \in \mathbb{R}^{k\times k}$ are constructed during the offline stage, described in Algorithm 7, where $\delta_k^{\mathrm{S2}} = \mathrm{Id} - I_k^{\mathrm{S2}}$.

Note that this algorithm constructs a set of points $\{\mu_l^{\mathrm{S2}}\}_{1\le l\le d}$ in $\mathcal{P}$ used in (7.13), and also a set of points $\{x_l^{\mathrm{S2}}\}_{1\le l\le d}$ in Ω.

Algorithm 7 Offline stage EIM$^{\mathrm{S2}}$

1. Choose $d > 1$ [Number of interpolation points]
2. Set $k := 1$
3. Compute $x_1^{\mathrm{S2}} := \underset{x\in\Omega}{\mathrm{argmax}} \|g(\cdot, x)\|_{L^\infty(\mathcal{P})}$
4. Compute $\mu_1^{\mathrm{S2}} := \underset{\mu\in\mathcal{P}}{\mathrm{argmax}} |g(\mu, x_1^{\mathrm{S2}})|$ [First interpolation point]
5. Set $q_1^{\mathrm{S2}}(\cdot) := \dfrac{g(\cdot, x_1^{\mathrm{S2}})}{g(\mu_1^{\mathrm{S2}}, x_1^{\mathrm{S2}})}$ [First basis function]
6. Set $B_{1,1}^{\mathrm{S2}} := 1$ [Initialize matrix B^{S2}]
7. **while** $k < d$ **do**
8. Compute $x_{k+1}^{\mathrm{S2}} := \underset{x\in\Omega}{\mathrm{argmax}} \|(\delta_k^{\mathrm{S2}} g)(\cdot, x)\|_{L^\infty(\mathcal{P})}$
9. Compute $\mu_{k+1}^{\mathrm{S2}} := \underset{\mu\in\mathcal{P}}{\mathrm{argmax}} |(\delta_k^{\mathrm{S2}} g)(\mu, x_{k+1}^{\mathrm{S2}})|$ [$(k+1)$-th interpolation point]
10. Set $q_{k+1}^{\mathrm{S2}}(\cdot) := \dfrac{(\delta_k^{\mathrm{S2}} g)(\cdot, x_{k+1}^{\mathrm{S2}})}{(\delta_k^{\mathrm{S2}} g)(\mu_{k+1}^{\mathrm{S2}}, x_{k+1}^{\mathrm{S2}})}$ [$(k+1)$-th basis function]
11. Set $B_{i,k+1}^{\mathrm{S2}} := q_{k+1}^{\mathrm{S2}}(\mu_i^{\mathrm{S2}})$, for all $1 \le i \le k+1$ [Increment matrix B^{S2}]
12. $k \leftarrow k+1$ [Increment the size of the decomposition]
13. **end while**

The online stage O1 of EIM$^{\mathrm{S2}}$ is computed by solving (7.12)-(7.13) for $k = d$. This yields

$$\left(I_d^{\mathrm{S2O1}} g\right)(\mu, x) := \sum_{m=1}^{d} \lambda_m^{\mathrm{S2O1}}(x) q_m^{\mathrm{S2}}(\mu), \tag{7.14}$$

where the functions $\lambda_m^{\mathrm{S2O1}}(x)$, $1 \le m \le d$, solve the linear system

$$\sum_{m=1}^{d} B_{l,m}^{\mathrm{S2}} \lambda_m^{\mathrm{S2O1}}(x) = g(\mu_l^{\mathrm{S2}}, x), \qquad \forall 1 \le l \le d. \tag{7.15}$$

The online stage O2 is defined as

$$\left(I_d^{\mathrm{S2O2}} g\right)(\mu, x) := \sum_{m=1}^{d} \lambda_m^{\mathrm{S2O2}}(\mu) g(\mu_m^{\mathrm{S2}}, x), \tag{7.16}$$

where the functions $\lambda_m^{\mathrm{S2O2}}(\mu)$, $1 \le m \le d$, solve the linear system

$$\sum_{m=1}^{d} (B^{\mathrm{S2}})_{l,m}^t \lambda_m^{\mathrm{S2O2}}(\mu) = q_l^{\mathrm{S2}}(\mu), \quad \forall 1 \le l \le d. \tag{7.17}$$

Similarly to Proposition 7.1, we infer that for all $x \in \Omega$ and all $\mu \in \mathcal{P}$, $(I_d^{\mathrm{S2O1}} g)(\mu, x) = (I_d^{\mathrm{S2O2}} g)(\mu, x)$. Since the roles of x and μ are not symmetric, the algorithms EIM$^{\mathrm{S1}}$ and EIM$^{\mathrm{S2}}$

	Online problem 1	Online problem 2
Slice 1	S1O1	S1O2
Slice 2	S2O1	S2O2

Table 7.1. Notation for the four possible approximation procedures

lead to different approximations of the function g: in general, $I_d^{\mathrm{S1O1}} = I_d^{\mathrm{S1O2}} \neq I_d^{\mathrm{S2O1}} = I_d^{\mathrm{S2O2}}$. We refer to Table 7.1 for notation of the four possible approximation procedures.

The classical interpolation property from [73, Lemma 1] (corresponding to the choice S1O1) and Proposition 7.1 yield the following propositions.

Proposition 7.2 (Interpolation with S1) *The approximation procedures S1O1 and S1O2 are interpolant with respect to the first variable: for all $x \in \Omega$,*

$$(I_d^{\mathrm{S1O1}} g)(\mu_m^{\mathrm{S1}}, x) = (I_d^{\mathrm{S1O2}} g)(\mu_m^{\mathrm{S1}}, x) = g(\mu_m^{\mathrm{S1}}, x), \quad \forall 1 \leq m \leq d. \tag{7.18}$$

Proposition 7.3 (Interpolation with S2) *The approximation procedures S2O1 and S2O2 are interpolant with respect to the second variable: for all $\mu \in \mathcal{P}$,*

$$(I_d^{\mathrm{S2O1}} g)(\mu, x_m^{\mathrm{S2}}) = (I_d^{\mathrm{S2O2}} g)(\mu, x_m^{\mathrm{S2}}) = g(\mu, x_m^{\mathrm{S2}}), \quad \forall 1 \leq m \leq d. \tag{7.19}$$

7.3 Nonintrusive procedure

The goal of this section is to obtain an approximation of the following objects:

$$Q_t(\mu) = \sum_{s=1}^{\varsigma} \int_\Omega g_s(\mu, x) \Psi_{s,t}(x) dx, \qquad \forall 1 \leq t \leq N, \tag{7.20}$$

where $\varsigma \geq 2$ while N is supposed to be large, using an offline-online procedure. We want the procedure to be robust with respect to N. This means that EIM algorithms can only be carried out to approximate the functions $(\mu, x) \mapsto g_s(\mu, x)$ and not the functions $(\mu, x) \mapsto g_s(\mu, x)\Psi_{s,t}(x)$. The index t refers to basis functions or couples of basis functions when evaluating the entries of the vector C_μ and the matrix A_μ in (7.2), see Section 7.4 for various examples.

Applying the four approximation formulae from Table 7.1 to the ς functions $g_s(\mu, x)$ leads to the construction of ς sets of points x, points μ, matrices B and vector-valued functions $q(\cdot)$. We denote these quantities with an additional index s; for instance, $\mathrm{EIM}^{\mathrm{S1}}$ carried out on $g_s(\mu, x)$ leads to the construction of the vector-valued functions $q_s^{\mathrm{S1}}(\cdot)$, of components $q_{s,m}^{\mathrm{S1}} : x \mapsto q_{s,m}^{\mathrm{S1}}(x)$, for all $1 \leq m \leq d$. For simplicity, we assume that each EIM algorithm stops at the same rank d. This leads to the following approximations for $Q_t(\mu)$:

– S1O1:

$$(I_d^{\mathrm{S1O1}} Q_t)(\mu) := \sum_{s=1}^{\varsigma} \sum_{m=1}^{d} \lambda_{s,m}^{\mathrm{S1O1}}(\mu) \int_\Omega q_{s,m}^{\mathrm{S1}}(x) \Psi_{s,t}(x) dx, \tag{7.21}$$

where $\lambda_{s,m}^{\mathrm{S1O1}}(\mu)$ solves (7.8),

– S1O2:

$$(I_d^{\mathrm{S1O2}} Q_t)(\mu) := \sum_{s=1}^{\varsigma} \sum_{m=1}^{d} g_s(\mu, x_{s,m}^{\mathrm{S1}}) \int_\Omega \lambda_{s,m}^{\mathrm{S1O2}}(x) \Psi_{s,t}(x) dx, \tag{7.22}$$

where $\lambda_{s,m}^{\mathrm{S1O2}}(x)$ solves (7.10),

– S2O1:

$$(I_d^{\mathrm{S2O1}} Q_t)(\mu) := \sum_{s=1}^{\varsigma} \sum_{m=1}^{d} q_{s,m}^{\mathrm{S2}}(\mu) \int_\Omega \lambda_{s,m}^{\mathrm{S2O1}}(x) \Psi_{s,t}(x) dx, \tag{7.23}$$

where $\lambda_{s,m}^{\mathrm{S2O1}}(x)$ solves (7.15),

– S2O2:

$$(I_d^{\mathrm{S2O2}} Q_t)(\mu) := \sum_{s=1}^{\varsigma} \sum_{m=1}^{d} \lambda_{s,m}^{\mathrm{S2O2}}(\mu) \int_\Omega g_s(\mu_{s,m}^{\mathrm{S2}}, x) \Psi_{s,t}(x) dx, \tag{7.24}$$

where $\lambda_{s,m}^{\mathrm{S2O2}}(\mu)$ solves (7.17).

We require that the approximation formula for $Q_t(\mu)$ is (i) online-efficient (in the sense that integrations over Ω are not allowed during the online calls), (ii) computable for all $\mu \in \mathcal{P}$, and not only on some training points in $\mathcal{P}$, and (iii) nonintrusive (in the sense that the only allowed integration over Ω in the offline stage of the procedure is the quantity $Q_t(\mu)$ itself). We first moderate the requirement of nonintrusivity by that of weak-intrusivity saying that the only integrations over Ω in the offline stage of the procedure are $\int_\Omega g_s(\mu, x) \Psi_{s,t}(x) dx$, $1 \leq s \leq \varsigma$, for all $\mu \in \mathcal{P}$.

We show in Section 7.3.1 that S1O1 and S2O2 lead to online-efficient procedures, while S1O2 and S2O1 do not. This suggests to discard the approximation procedures S1O2 and S2O1. In Section 7.3.2, we show that S2O2 is weakly intrusive while S1O1 is not, and that S1O1 allows to compute the approximation for all $\mu \in \mathcal{P}$ while S2O2 only allows the computation at the training points. Then, we show in Section 7.3.3 that to the price of an additional linear system to be solved online, S1O1 can be turned to be weakly intrusive as well and that to the same price, the approximation using S2O2 can be computed for all $\mu \in \mathcal{P}$ as well. However, none of the two procedures is nonintrusive. Finally, we show in Section 7.3.4 that following an idea from Chapter 6 and to the price of an additional EIM approximation of which we can control the accuracy, we can devise a nonintrusive procedure for both S1O1 ans S2O2.

7.3.1 Online-efficient procedures

We notice from (7.22) and (7.23) that S1O2 and S2O1 need to integrate solutions to the online problems over Ω. Thus, these procedures are not online-efficient. On the contrary, we notice from (7.21) and (7.24) that for S1O1 and S2O2, the integrals over Ω can be precomputed once and for all during the offline stage and reused in the online call. Therefore, the procedures S1O1 and S2O2 are online-efficient.

7.3.2 Computation between training points and weak intrusivity

In practice, the EIM algorithms are carried out on finite-dimensional subsets $\Omega_{\mathrm{trial}} \subset \Omega$ and $\mathcal{P}_{\mathrm{trial}} \subset \mathcal{P}$. This means that the functions $q_{s,m}^{\mathrm{S1}}(\cdot)$ and $q_{s,m}^{\mathrm{S2}}(\cdot)$ are constructed respectively

over Ω_{trial} and $\mathcal{P}_{\text{trial}}$. We suppose that the functions $g_s(\mu, x)$ are known for all $x \in \Omega$ and all $\mu \in \mathcal{P}$. For instance, for S1O1, $\lambda_{s,m}^{\text{S1O1}}(\mu)$ such that $\sum_{m=1}^{d}(B_s^{\text{S1}})_{l,m}\lambda_{s,m}^{\text{S1O1}}(\mu) = g_s(\mu, x_{s,l}^{\text{S1}})$ can be computed for all $\mu \in \mathcal{P}$, but the approximation $(I_d^{\text{S1O1}} g_s)(\mu, x) = \sum_{m=1}^{d} \lambda_{s,m}^{\text{S1O1}}(\mu) q_{s,m}^{\text{S1}}(x)$ can be computed only for $x \in \Omega_{\text{trial}}$, due to the evaluation of $q_{s,m}^{\text{S1}}(x)$. As a result, the approximation of $g_s(\mu, x)$ is available on

- $\mathcal{P} \times \Omega_{\text{trial}}$ for S1O1,
- $\mathcal{P}_{\text{trial}} \times \Omega$ for S2O2.

In practice, the integrations over Ω are computed numerically on a mesh, using for instance quadrature formulae. The values of $g_s(\mu, x)$ are then needed only on the Gauss points corresponding to the quadrature: we can take the set of these Gauss points as Ω_{trial}, so that S1O1 provides an approximation computable at all the needed points. Instead, S2O2 provides an approximation computable only at the training points in $\mathcal{P}_{\text{trial}}$ and not for all $\mu \in \mathcal{P}$.

Concerning intrusivity, we see from (7.21) that S1O1 is not weakly intrusive since the integration of the functions $q_{s,m}^{\text{S1}}(\cdot)$ over Ω is required. On the contrary, we see from (7.24) that S2O2 is weakly intrusive.

To sum up, the two online-efficient procedures enjoy the following exclusive properties:

- S1O1: intrusive approximation, computable for all $\mu \in \mathcal{P}$,
- S2O2: weakly intrusive approximation, computable only for $\mu \in \mathcal{P}_{\text{trial}}$.

7.3.3 Modification of the online problems

Consider the offline stage EIM^{S1} described in Algorithm 6. By construction, it is clear that for all $1 \le s \le \varsigma$, $\underset{1\le m\le d}{\text{Vect}}\left(q_{s,m}^{\text{S1}}(\cdot)\right) = \underset{1\le m\le d}{\text{Vect}}\left(g_s(\mu_m^{\text{S1}}, \cdot)\right)$. Therefore, there exists a matrix $\Gamma_s^{\text{S1}} \in \mathbb{R}^{d\times d}$ such that, for all $1 \le l \le d$,

$$\sum_{m=1}^{d} (\Gamma_s^{\text{S1}})_{l,m} q_{s,m}^{\text{S1}}(x) = g_s(\mu_{s,l}^{\text{S1}}, x), \qquad \forall x \in \Omega_{\text{trial}}. \tag{7.25}$$

Lemma 7.4 *The matrix Γ_s^{S1} can be constructed recursively in the loop in k of Algorithm 6 in the following way:*

- $k = 1$*:*

$$(\Gamma_s^{\text{S1}})_{1,1} = g_s(\mu_{s,1}^{\text{S1}}, x_{s,1}^{\text{S1}}), \tag{7.26}$$

- $k \to k+1$*:*

$$\begin{aligned} (\Gamma_s^{\text{S1}})_{k+1,k+1} &= (\delta_k^{\text{S1}} g_s)(\mu_{s,k+1}^{\text{S1}}, x_{s,k+1}^{\text{S1}}), \\ (\Gamma_s^{\text{S1}})_{l,k+1} &= 0, && \forall 1 \le l \le k, \\ (\Gamma_s^{\text{S1}})_{k+1,l} &= \kappa_{s,l}^{\text{S1}}, && \forall 1 \le l \le k, \end{aligned} \tag{7.27}$$

where the vector κ_s^{S1} is such that $\sum_{m=1}^{k}(B_s^{\text{S1}})_{l,m}\kappa_{s,m}^{\text{S1}} = g_s(\mu_{s,k+1}^{\text{S1}}, x_{s,l}^{\text{S1}})$, for all $1 \le l \le k$.

Proof. The case $k = 1$ results from line 5 of Algorithm 6. Suppose that the assertion holds at rank k. Using the definition (7.25) of Γ_s^{S1} at rank $(k+1)$, for all $1 \le l \le k$ and all $x \in \Omega_{\text{trial}}$, there holds $(\Gamma_s^{\text{S1}})_{l,k+1} q_{s,k+1}^{\text{S1}}(x) + \sum_{m=1}^{k}(\Gamma_s^{\text{S1}})_{l,m} q_{s,m}^{\text{S1}}(x) = g_s(\mu_{s,l}^{\text{S1}}, x)$. Using the same definition at rank k leads to $(\Gamma_s^{\text{S1}})_{l,k+1} = 0$ for all $1 \le l \le k$. Then, using the same definition for $l = k+1$, it

is inferred that $(\Gamma_s^{\mathrm{S1}})_{k+1,k+1} q_{s,k+1}^{\mathrm{S1}}(x) + \sum_{m=1}^{k} (\Gamma_s^{\mathrm{S1}})_{k+1,m} q_{s,m}^{\mathrm{S1}}(x) = g_s(\mu_{s,k+1}^{\mathrm{S1}}, x)$. Using line 10 of Algorithm 6, we identify $(\Gamma_s^{\mathrm{S1}})_{k+1,k+1} = (\delta_{s,k}^{\mathrm{S1}} g)(\mu_{s,k+1}^{\mathrm{S1}}, x_{s,k+1}^{\mathrm{S1}})$ and $\sum_{m=1}^{k} (\Gamma_s^{\mathrm{S1}})_{k+1,m} q_{s,m}^{\mathrm{S1}}(x) = (I_k^{\mathrm{S1}} g_s)(\mu_{s,k+1}^{\mathrm{S1}}, x)$. From (7.7)-(7.8), there holds

$$\sum_{m=1}^{k} (\Gamma_s^{\mathrm{S1}})_{k+1,m} q_{s,m}^{\mathrm{S1}}(x) = \sum_{l=1}^{k} \sum_{m=1}^{k} (B_s^{\mathrm{S1}})_{m,l}^{-1} q_{s,m}^{\mathrm{S1}}(x) g(\mu_{s,k+1}^{\mathrm{S1}}, x_{s,l}^{\mathrm{S1}}). \tag{7.28}$$

Therefore, $(\Gamma_s^{\mathrm{S1}})_{k+1,m} = \sum_{l=1}^{k} (B_s^{\mathrm{S1}})_{m,l}^{-1} g_s(\mu_{s,k+1}^{\mathrm{S1}}, x_{s,l}^{\mathrm{S1}})$, finishing the proof. □ ♢

Proposition 7.5 *To the price of an additional online problem, the procedure S1O1 is weakly intrusive.*

Proof. From Lemma 7.4, $q_{s,m}^{\mathrm{S1}}(x) = \sum_{l=1}^{d} (\Gamma_s^{\mathrm{S1}})_{m,l}^{-1} g_s(\mu_{s,l}^{\mathrm{S1}}, x)$ holds for all $1 \leq m \leq d$, so that (7.21) can be written

$$(I_d^{\mathrm{S1O1}} Q_t)(\mu) = \sum_{s=1}^{\varsigma} \sum_{m=1}^{d} \lambda_{s,m}^{\mathrm{S1O1}}(\mu) \sigma_{m,s,t}, \tag{7.29}$$

where the coefficients $\sigma_{m,s,t}$, for all $1 \leq m \leq d$, solve the linear system

$$\sum_{l=1}^{d} (\Gamma_s^{\mathrm{S1}})_{m,l} \sigma_{s,l,t} = \int_{\Omega} g_s(\mu_{s,m}^{\mathrm{S1}}, x) \Psi_{s,t}(x) dx, \tag{7.30}$$

and $\lambda_{s,m}^{\mathrm{S1O1}}(\mu)$ solves (7.8). □ ♢

Consider now the offline stage $\mathrm{EIM}^{\mathrm{S2}}$ described in Algorithm 7. There exists a matrix $\Gamma_s^{\mathrm{S2}} \in \mathbb{R}^{d \times d}$ such that

$$\sum_{m=1}^{d} (\Gamma_s^{\mathrm{S2}})_{l,m} q_{s,m}^{\mathrm{S2}}(\mu) = g_s(\mu, x_{s,l}^{\mathrm{S2}}), \qquad \forall \mu \in \mathcal{P}_{\mathrm{trial}}. \tag{7.31}$$

Notice that even if the function $q_{s,m}^{\mathrm{S2}}(\cdot)$ is constructed on $\mathcal{P}_{\mathrm{trial}}$, the above expression allows one to extend the definition of $q_{s,m}^{\mathrm{S2}}(\cdot)$ to $\mathcal{P}$. In the same fashion as Lemma 7.4, we obtain the following result.

Lemma 7.6 *The matrix Γ_s^{S2} can be constructed recursively in the loop in k of Algorithm 7 in the following way:*

– *$k = 1$:*

$$(\Gamma_s^{\mathrm{S2}})_{1,1} = g_s(\mu_{s,1}^{\mathrm{S2}}, x_{s,1}^{\mathrm{S2}}), \tag{7.32}$$

– *$k \to k+1$:*

$$\begin{aligned} (\Gamma_s^{\mathrm{S2}})_{k+1,k+1} &= (\delta_k^{\mathrm{S2}} g_s)(\mu_{s,k+1}^{\mathrm{S2}}, x_{s,k+1}^{\mathrm{S2}}), \\ (\Gamma_s^{\mathrm{S2}})_{l,k+1} &= 0, && \forall 1 \leq l \leq k, \\ (\Gamma_s^{\mathrm{S2}})_{k+1,l} &= \kappa_{s,l}^{\mathrm{S2}}, && \forall 1 \leq l \leq k, \end{aligned} \tag{7.33}$$

where the vector κ_s^{S2} is such that $\sum_{m=1}^{k} (B_s^{\mathrm{S2}})_{m,l} \kappa_{s,m}^{\mathrm{S2}} = g_s(\mu_{s,l}^{\mathrm{S2}}, x_{s,k+1}^{\mathrm{S2}})$, for all $1 \leq l \leq k$.

Proposition 7.7 *To the price of an additional online problem, the approximation of $Q_t(\mu)$ using the procedure S2O2 can be computed between training points.*

Proof. Let $\mu \in \mathcal{P}$. Compute first $q^{\mathrm{S2}}_{s,m}(\mu)$ solving (7.31), and compute then $\lambda^{\mathrm{S2O2}}_{s,m}(\mu)$ solving (7.17). Then, $(I^{\mathrm{S2O2}}_d Q_t)(\mu) = \sum_{s=1}^{\varsigma}\sum_{m=1}^{d}\lambda^{\mathrm{S2O2}}_{s,m}(\mu)\int_\Omega g_s(\mu^{\mathrm{S2}}_{s,m},x)\Psi_{s,k}(x)dx$ is computable for all $\mu\in\mathcal{P}$. □ ♢

To sum up, with the above-discussed modifications of the online problems, the two online-efficient procedures S1O1 and S2O2 are now weakly intrusive and computable for all $\mu \in \mathcal{P}$. Unfortunately, none of the two procedures is nonintrusive in the sense defined above.

7.3.4 The nonintrusive procedures

The observation made in Chapter 6 is that the expressions (7.21) and (7.24) are linear forms in a vector $z \in \mathbb{R}^{\varsigma d}$, whose components, denoted by $z_p(\mu)$, $1 \le p \le \varsigma d$, contain all the μ-dependencies. Notice that this is still the case when considering the additional online problems of Section 7.3.3. For instance, consider the procedure S1O1 and define

$$z_p(\mu) := \begin{cases} \lambda^{\mathrm{S1O1}}_{1,m}(\mu), & 1 \le p \le d, & m = p, \\ \lambda^{\mathrm{S1O1}}_{2,m}(\mu), & 1+d \le p \le 2d, & m = p-d, \\ & \vdots & \\ \lambda^{\mathrm{S1O1}}_{\varsigma,m}(\mu), & 1+(\varsigma-1)d \le p \le \varsigma d, & m = p-(\varsigma-1)d. \end{cases} \tag{7.34}$$

$$\mathcal{Q}_{t,p} := \begin{cases} \displaystyle\int_\Omega q^{\mathrm{S1}}_{1,m}(x)\Psi_{1,t}(x)dx, & 1 \le p \le d, & m = p, \\ \displaystyle\int_\Omega q^{\mathrm{S1}}_{2,m}(x)\Psi_{2,t}(x)dx, & 1+d \le p \le 2d, & m = p-d, \\ & \vdots & \\ \displaystyle\int_\Omega q^{\mathrm{S1}}_{\varsigma,m}(x)\Psi_{\varsigma,t}(x)dx, & 1+(\varsigma-1)d \le p \le \varsigma d, & m = p-(\varsigma-1)d. \end{cases} \tag{7.35}$$

Then,

$$(I^{\mathrm{S1O1}}_d Q_t)(\mu) = \sum_{p=1}^{\varsigma d} z_p(\mu)\mathcal{Q}_{t,p}. \tag{7.36}$$

The idea consists in applying another EIM to $z_p(\mu)$ seen as the two-variable function $(\mu,p)\mapsto z_p(\mu)$. Here again, the four approximation procedures S1O1(z), S1O2(z), S2O1(z), and S2S2(z) are possible for the approximation of $z_p(\mu)$, where now p plays the role that x played in Section 7.2, and where we indicate specifically in the notation that these procedures are related to the approximation of $z_p(\mu)$. We consider an approximation of $z_p(\mu)$ using either the procedure S1O1(z) or the procedure S2O2(z) (see Remark 7.8 below). We then inject the approximation of $z_p(\mu)$ in the right-hand side of (7.36). This leads to

– S1O1(z):

$$(I^{\mathrm{S1O1}}_d Q_t)(\mu) \approx \sum_{m=1}^{d^z} \beta^{\mathrm{S1O1}(z)}_m(\mu)\sum_{p=1}^{\varsigma d} q^{\mathrm{S1}(z)}_m(p)\mathcal{Q}_{t,p}, \tag{7.37}$$

where $\beta_m^{\mathrm{S1O1}(z)}(\mu)$ solves

$$\sum_{m=1}^{d^z} B_{l,m}^{\mathrm{S1}(z)} \beta_m^{\mathrm{S1O1}(z)}(\mu) = z_{p_l^{\mathrm{S1}(z)}}(\mu), \qquad \forall 1 \leq l \leq d^z, \tag{7.38}$$

– S2O2(z):

$$(I_d^{\mathrm{S2O2}} Q_t)(\mu) \approx \sum_{m=1}^{d^z} \beta_m^{\mathrm{S2O2}(z)}(\mu) \sum_{p=1}^{\varsigma d} z_p(\mu_m^{\mathrm{S2}(z)}) \mathcal{Q}_{t,p}, \tag{7.39}$$

where $\beta_m^{\mathrm{S2O2}(z)}(\mu)$ solves

$$\sum_{m=1}^{d^z} (B^{\mathrm{S2}(z)})_{l,m}^t \beta_m^{\mathrm{S2O2}(z)}(\mu) = q_l^{\mathrm{S2}(z)}(\mu), \qquad \forall 1 \leq l \leq d^z. \tag{7.40}$$

In the above equations, we now denote the solution to the online problem of the EIM procedures to approximate $z_p(\mu)$ by β. We keep the same notation for the constructed matrices B and vector-valued functions $q_m(\cdot)$, and the selected points μ_m, while we introduced in (7.38) the indices p_m selected by the EIM procedures to approximate $z_p(\mu)$.

With S2O2(z), (7.39) can be rewritten as

$$Q_t(\mu) \approx \sum_{m=1}^{d^z} \beta_m^{\mathrm{S2O2}(z)}(\mu) Q_t(\mu_m^{\mathrm{S2}(z)}), \tag{7.41}$$

which provides a nonintrusive procedure for the approximation of $Q_t(\mu)$. As explained in Section 7.3.3, we need to solve an additional online problem to retrieve this approximation for all $\mu \in \mathcal{P}$. In the same fashion, we can prove that with S1O1(z), we can derive from (7.37), to the price of an additional online problem, another nonintrusive formula for the approximation of $Q_t(\mu)$, which is directly available for all $\mu \in \mathcal{P}$.

Remark 7.8 *Like before, we want to be able to precompute as many terms as possible in the offline stage. To be able to precompute summations of size* ςd *in* (7.37) *and* (7.39)*, we can discard S1O2(z) and S2O1(z). Note however that in general,* ςd *is not as large as* N*.*

We can repeat this work for the case where the EIMs on $g_s(\mu, x)$ are performed with S2O2, and achieve the same kind of conclusions. Actually, we have four possible choices for the complete procedure at our disposal, namely the product of the choices S1O1 and S2O2 for the approximation of $g_s(\mu, x)$, times the choices S1O1(z) and S2O2(z) for the approximation of $z_p(\mu)$. The overall choice is driven by (approximate) interpolation properties we wish our formula to exhibit and by avoiding additional online problems if a certain choice permits it. Other performance considerations can be considered, for instance if $\#\mathcal{P}_{\mathrm{trial}} \gg \#\Omega_{\mathrm{trial}}$, one may prefer to construct functions q defined on Ω_{trial} rather than on $\mathcal{P}_{\mathrm{trial}}$. In our simulations, we chose S1O1 for the EIM on $g_s(\mu, x)$ and S2O2(z) for the EIM on $z_p(\mu)$. For the EIM on $g_s(\mu, x)$, S1O1 enables the $\lambda_{s,m}^{\mathrm{S1O1}}(\mu)$ involved in the online problems to be computed for all $\mu \in \mathcal{P}$, but the weak intrusivity allowed by S2O2 is not useful for the approximation of $g_s(\mu, x)$. For the EIM on $z_p(\mu)$, S2O2(z) provides (7.41), a nonintrusive procedure for the approximation of $Q_t(\mu)$, while S1O1(z) allows to compute the approximation of $Q_t(\mu)$ with no additional online problem. When

nonintrusivity is crucial for practical purposes, the additional online problem needed for S1O1(z) to be nonintrusive is always required, while the additional online problem needed for S2O2(z) to be computed is only needed for $\mu \in \mathcal{P} \backslash \mathcal{P}_{\text{trial}}$. Hence, the choice S1O1/S2O2(z) requires no additional problem in the offline stage (when the approximation operators are evaluated), and a single additional online problem for computing the approximation for $\mu \in \mathcal{P} \backslash \mathcal{P}_{\text{trial}}$. Table 7.2 presents the number of additional online problems to be solved when considering an offline call for the approximation operators (i.e. for $\mu \in \mathcal{P}_{\text{trial}}$) and an online call (i.e. for $\mu \in \mathcal{P} \backslash \mathcal{P}_{\text{trial}}$). Among the two choices that minimize the computation cost, S1O1/S2O2(z) was chosen since it provides an approximation for $Q_t(\mu)$ that is interpolant with respect to μ at the limit $d^z = \varsigma d$.

	S1O1/S1O1(z)	S1O1/S2O2(z)	S2O2/S1O1(z)	S2O2/S2O2(z)
Offline stage	1	0	1	0
Online stage	1	1	1	1

Table 7.2. Number of additional online problems for each evaluation of the approximation operator in the offline and online stages, for the four EIM combinations, when a nonintrusive procedure is required. Offline stage refers to an approximation for $\mu \in \mathcal{P}_{\text{trial}}$ and online stage refers to an approximation for $\mu \in \mathcal{P} \backslash \mathcal{P}_{\text{trial}}$

7.4 Nonintrusive RBM for aeroacoustic problems

In this section, we use the nonintrusive formula (7.41) for the approximation of the matrix and the right-hand side of discrete variational formulations arising in aeroacoustic problems modelled by the Helmholtz equation or the convected Helmholtz equation. The finite element method (FEM) and the boundary element method (BEM) are used to obtain the matrix and right-hand side of the problem. Both quantities are of the form $Q_t(\mu)$ as defined in (7.20). For the matrix, the index t in $\Psi_{s,t}(x)$ refers to the product of two finite element basis functions, while for the right-hand side, the index t refers to the finite element functions themselves. Using the nonintrusive approximation (7.41), we only need to compute matrix-vector products involving A_μ and scalar products to precompute in the offline stage all the quantities needed to construct efficiently the reduced problem and compute the error bound in the online stage.

7.4.1 Implementation of the RBM

For simplicity, we take the Euclidian norm of the discretized vectors in the computation of the a posteriori error bound in the RBM, to avoid dealing with the computation of dual norms. When considering reduced basis strategies for the (convected) Helmholtz equation approximated by BEM with the frequency as a parameter, the approximate affine decomposition (7.41) has a quite large number of terms. When applying the Successive Constraints Methods (SCM, see [56]) as an online-efficient procedure for computing the inf-sup constant, we have to solve constrained linear optimization problems, with a number of constraints proportional to the square of the number of terms in the decomposition (7.41). For simplicity, we do not look for an online-efficient way to compute the inf-sup constant. In Sections 7.4.2 and 7.4.3, we compute a single value of this constant (for centered values of the parameters) and use it for any error bound evaluation.

Even if the inf-sup constant depends on the parameters, its values are not expected to exhibit significant variations since the considered formulations do not feature any resonant frequency. In Section 7.4.4, the test case has much more unknowns that those from the two previous sections, and we do not compute the inf-sup constant so as to control the offline computational cost. The power iteration method (see [77]) associated with the inverse matrix can be used to approximately compute the smallest eigenvalue. This would imply to solve many eigenvalue problems associated with the inverse operator, and therefore does not appear reasonable for industrial test cases. Dealing further with this issue goes beyond the present scope.

7.4.2 An optimization problem for an impedant object in the air at rest

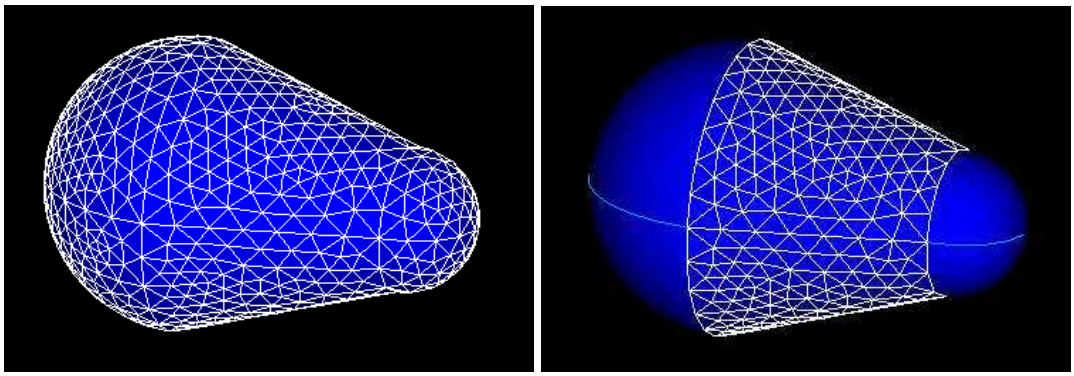

Fig. 7.1. Test case 1. Left: mesh for test case 1. Right: impedant surface Γ_2

Consider the object whose mesh is represented in the left panel of Figure 7.1. The surface of this object, denoted by Γ, consists of three zones denoted by Γ_1, Γ_2 and Γ_3 respectively. The surface Γ_2 is represented in the right panel of Figure 7.1. On each of these zones, a Robin boundary condition is enforced with a specific impedance coefficient μ_i for $i \in \{1,2,3\}$. Thus, the impedance coefficient on Γ, denoted by μ_Γ, is piecewise constant and takes the form $\mu_\Gamma(x) = \mu_1 \mathbb{1}_{\Gamma_1}(x) + \mu_2 \mathbb{1}_{\Gamma_2}(x) + \mu_3 \mathbb{1}_{\Gamma_3}(x)$, for all $x \in \Gamma$, where $\mathbb{1}_{\Gamma_i}$, $i \in \{1,2,3\}$, are characteristic functions. The source is a plane wave whose wave vector is supported by the axis of symmetry of the object, creating an incident acoustic pressure field denoted by p_{inc}. The variational formulation of the problem is as follows: Find $(\chi, \lambda) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$ such that for all $(\hat{\chi}, \hat{\lambda}) \in H^{\frac{1}{2}}(\Gamma) \times L^2(\Gamma)$,

$$\begin{cases} \left(N_\mu \chi - \dfrac{i\mu}{2\mu_s}\chi, \hat{\chi}\right)_\Gamma + \left(\tilde{D}_\mu \lambda, \hat{\chi}\right)_\Gamma = (\gamma_1 p_{\text{inc}}, \hat{\chi})_\Gamma \,, \\ \left(\hat{\lambda}, D_\mu \chi\right)_\Gamma - \left(\hat{\lambda}, S_\mu \lambda + \dfrac{i\mu_s}{2\mu}\lambda\right)_\Gamma = -\left(\hat{\lambda}, \gamma_0 p_{\text{inc}}\right)_\Gamma \,, \end{cases} \tag{7.42}$$

where $(\cdot,\cdot)_\Gamma$ denotes the extension of the $L^2(\Gamma)$-inner product to the duality pairing on $H^{\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma)$, $\mu = \frac{\omega}{c}$, with ω the pulsation of the source and c the speed of sound in the air at rest and where $\mathring{u}\gamma_0$ and γ_1 respectively denote the Dirichlet and Neumann traces on Γ. The operators N_μ, D_μ, $\tilde{D}_\mu$, and S_μ are boundary integral operators, expressed in terms of the Green kernel

$G_\mu(x,y) = \frac{\exp(i\mu|x-y|)}{4\pi|x-y|}$ associated with the Helmholtz equation at wave number μ. The pressure field around the object is then obtained by applying a representation formula to (χ, λ), the solution to (7.42). We refer to Chapter 2 for more details on this formulation and its well-posedness. The considered finite-dimensional approximation of (7.42) has 2240 unknowns.

The parameters of the problem are the frequency of the source $fr = \frac{\omega}{2\pi}$, and the impedance coefficient of each of the three zones composing the surface of the object. The frequency varies from 487 to 1082 Hz, and each impedance coefficient varies from 1 to 5. The quantity of interest is the far-field acoustic pressure along the axis of symmetry of the object, but in the opposite direction of the source. A goal-oriented RBM is carried out to select a basis of $\hat{n} = 20$ truth solutions using the nonintrusive formula (7.41) to approximate the matrix, the right-hand side of the direct problem, and the right-hand side of the adjoint problem needed to evaluate the quantity of interest. For the matrix, the approximation procedure S1O1 is applied to

$$g(\mu, r) := \exp(i\mu r), \ r = |x-y|, \ x, y \in \Gamma, \tag{7.43}$$

and the procedure S2O2(z) is applied to

$$z_p(\mu, \mu_1, \mu_2, \mu_3) := \begin{cases} \lambda_m^{\mathrm{S1O1}}(\mu), & 1 \le m \le d, \quad p = m, \\ \mu\lambda_m^{\mathrm{S1O1}}(\mu), & 1 \le m \le d, \quad p = m+d, \\ \mu^2\lambda_m^{\mathrm{S1O1}}(\mu), & 1 \le m \le d, \quad p = m+2d, \\ \dfrac{\mu}{\mu_1}, & p = 3d+1, \\ \dfrac{\mu_1}{\mu}, & p = 3d+2, \\ \dfrac{\mu}{\mu_2}, & p = 3d+3, \\ \dfrac{\mu_2}{\mu}, & p = 3d+4, \\ \dfrac{\mu}{\mu_3}, & p = 3d+5, \\ \dfrac{\mu_3}{\mu}, & p = 3d+6. \end{cases} \tag{7.44}$$

For the approximation formula of the right-hand side of the direct and dual problems, the procedure S1O1 is applied to

$$g(\mu, x) := \exp(i\mu \boldsymbol{d} \cdot \boldsymbol{x}), \ x \in \Gamma, \tag{7.45}$$

where $\boldsymbol{d}$ is respectively the direction of the incoming plane wave and the direction of measure of the far-field; and the procedure S2O2(z) is applied to

$$z_m(\mu) := \lambda_m^{\mathrm{S1O1}}(\mu), \qquad 1 \le m \le d, \quad p = m. \tag{7.46}$$

The EIM algorithms are carried out with $d = 13$ and $d^z = 20$ for the matrix, and $d = 13$ and $d^z = 13$ for right-hand side of the direct and dual problems. Over the considered parameter values, the relative error for the three nonintrusive formulae is of the order of 10^{-12} (in Frobenius norm for the matrix and Euclidian norm for the vectors). The maximum error bound (over a

discretization $\mathcal{P}_{\text{trial}}$) is of the order of 10^{-6}, the online stage takes 2.8×10^{-3} s to compute a reduced solution and the error bound, while the full direct problem is solved in about 30 s in parallel on 4 processors, which corresponds to an acceleration factor of 10^4.

To formulate the optimization problem, we consider as an illustration a set of values fr_i, $1 \leq i \leq p$, for the frequency of the source and we denote by $J_i(\mu_1, \mu_2, \mu_3)$, the quantity of interest computed for the frequency fr_i of the source and depending on the three impedance coefficients. Consider the following cost function: $(\mu_1, \mu_2, \mu_3) \mapsto \mathcal{J}(\mu_1, \mu_2, \mu_3) := \sum_{i=1}^{p} \alpha_i J_i(\mu_1, \mu_2, \mu_3) + h(\mu_1, \mu_2, \mu_3)$. The goal of the study is to find the values of the impedance coefficients that minimize the cost function. With such a cost function, we can minimize the far-field acoustic pressure scattered by the object, taking into account that some frequencies are more harmful than others for the human ear (through the weights α_i), and that some treatments of the object surface to modify the impedance coefficients are more expensive than others (through the function h). To illustrate, we choose $p = 20$, $\alpha_i = 2$ for $1 \leq i \leq 7$, $\alpha_i = 1$ for $8 \leq i \leq 13$ and $\alpha_i = 3$ for $14 \leq i \leq 20$, and $h(\mu_1, \mu_2, \mu_3) = \frac{1}{6}(0.2\mu_1^{-0.5} + 0.3\mu_1^{-0.8} + 0.5\mu_1^{-1}) - 8$. The cost function is computed for 1000 values of the impedance coefficients (each coefficient being sampled by 10 values). Notice that for each evaluation of the cost function, we need to compute the solution of the aeroacoustic problem for 20 values of the frequency. Using the online stage, the minimum of the cost function over this sample of impedance coefficients is 0.366, reached for $(\mu_1, \mu_2, \mu_3) = (2.8, 1, 1.9)$, and is found in less than 24 s.

Figure 7.2 shows a screenshot of a java applet computing the quantity of interest at 50 values of the frequency, and at values of the impedance coefficients selected by the user.

7.4.3 An uncertainty quantification problem for an object surrounded by a potential flow

Consider an ellipsoid with major axis directed along the z-axis. This object is included inside a larger ball, see Figure 7.3. The external border of the ball after discretization is denoted by Γ_∞. The complement of the ellipsoid in the ball is denoted by Ω^-. A potential flow is precomputed around the ellipsoid and inside the ball, such that the flow is uniform outside the ball, of Mach number 0.3 and directed along the z-axis. An acoustic monopole source lies upstream of the object, on the z-axis as well.

The considered formulation is a coupled BEM-FEM formulation. It consists in (i) applying a change of variables to transform the convected Helmholtz equation into the classical Helmholtz equation outside the ball, in order to apply a standard BEM on Γ_∞, and (ii) stabilizing the formulation to avoid resonant frequencies associated with the eigenvalues of the Laplacian inside the ball of boundary Γ_∞. Consider the product space $\mathbb{H} := H^1(\Omega^-) \times H^{-\frac{1}{2}}(\Gamma_\infty) \times H^1(\Gamma_\infty)$ with inner product $\left((\Phi, \lambda, p), (\Phi^t, \lambda^t, p^t)\right)_{\mathbb{H}} := (\Phi, \Phi^t)_{H^1(\Omega^-)} + (\lambda, \lambda^t)_{H^{-\frac{1}{2}}(\Gamma_\infty)} + (p, p^t)_{H^1(\Gamma_\infty)}$. The weak formulation is: Find $(\Phi, \lambda, p) \in \mathbb{H}$ such that $\forall (\Phi^t, \lambda^t, p^t) \in \mathbb{H}$,

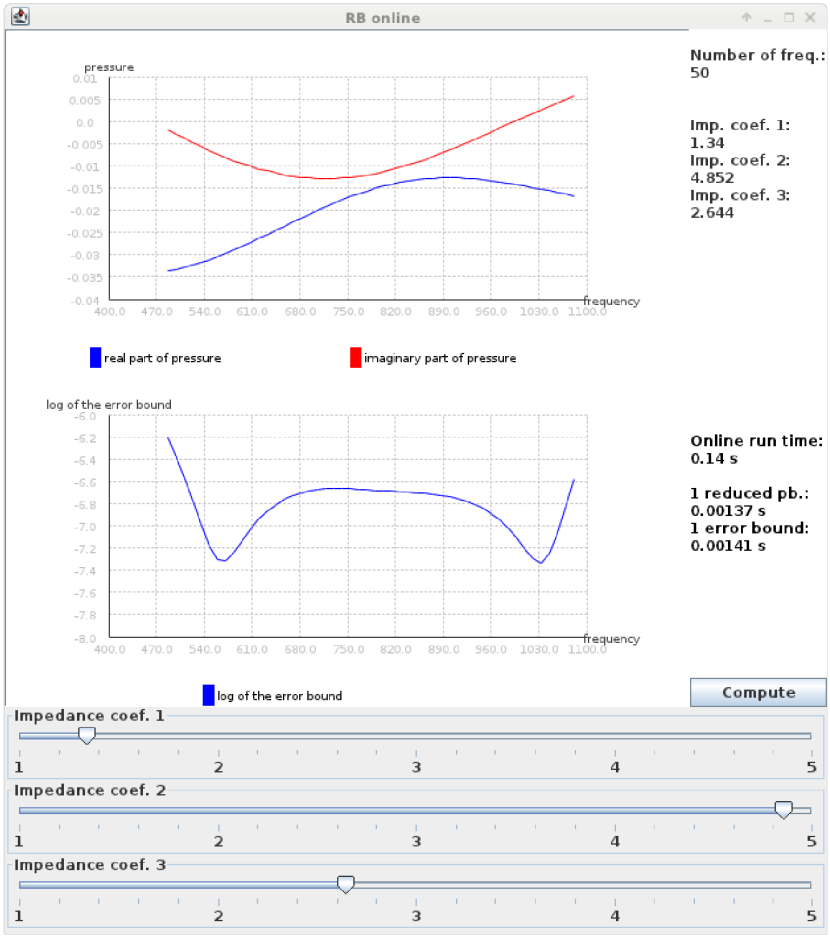

Fig. 7.2. Java applet for the online stage of the RBM for test case 1. Top panel: real part and imaginary part of the far-field pressure for 50 values of the frequency. Middle panel: error bound. Bottom panel: selection of the impedance coefficients

$$\mathcal{V}_\mu(\Phi,\Phi^t)+\left(N_\mu(\gamma_0^-\Phi),\gamma_0^-\Phi^t\right)_{\Gamma_\infty}+\left(\left(\tilde{D}_\mu-\frac{1}{2}I\right)(\lambda),\gamma_0^-\Phi^t\right)_{\Gamma_\infty}=\left(\gamma_1 f_{\mathrm{inc}_\mu},\gamma_0^-\Phi^t\right)_{\Gamma_\infty}, \tag{7.47a}$$

$$\left(\lambda^t,\left(D_\mu-\frac{1}{2}I\right)(\gamma_0^-\Phi)\right)_{\Gamma_\infty}-\left(\lambda^t,S_\mu(\lambda)\right)_{\Gamma_\infty}-i\left(\lambda^t,p\right)_{\Gamma_\infty}=-\left(\lambda^t,\gamma_0 f_{\mathrm{inc}_\mu}\right)_{\Gamma_\infty}, \tag{7.47b}$$

$$\left(N_\mu(\gamma_0^-\Phi),p^t\right)_{\Gamma_\infty}+\left(\left(\tilde{D}_\mu+\frac{1}{2}I\right)(\lambda),p^t\right)_{\Gamma_\infty}-\delta_{\Gamma_\infty}(p,p^t)=\left(\gamma_1 f_{\mathrm{inc}_\mu},p^t\right)_{\Gamma_\infty}, \tag{7.47c}$$

where $(\cdot,\cdot)_{\Gamma_\infty}$ denotes the extension of the $L^2(\Gamma_\infty)$-inner product to the duality pairing on $H^{-\frac{1}{2}}(\Gamma_\infty)\times H^{\frac{1}{2}}(\Gamma_\infty)$, and where

$$\delta_{\Gamma_\infty}(p,q):=(\nabla_{\Gamma_\infty}p,\nabla_{\Gamma_\infty}q)_{\Gamma_\infty}+(p,q)_{\Gamma_\infty}, \tag{7.48}$$

with ∇_{Γ_∞} the surfacic gradient on Γ_∞, and

$$\mathcal{V}_\mu(\Phi,\Phi^t):=\int_{\Omega^-}\Xi\nabla\overline{\Phi}\cdot\nabla\Phi^t-\mu^2\int_{\Omega^-}\beta\overline{\Phi}\Phi^t+i\mu\int_{\Omega^-}\boldsymbol{V}\cdot\left(\overline{\Phi}\nabla\Phi^t-\Phi^t\nabla\overline{\Phi}\right), \tag{7.49}$$

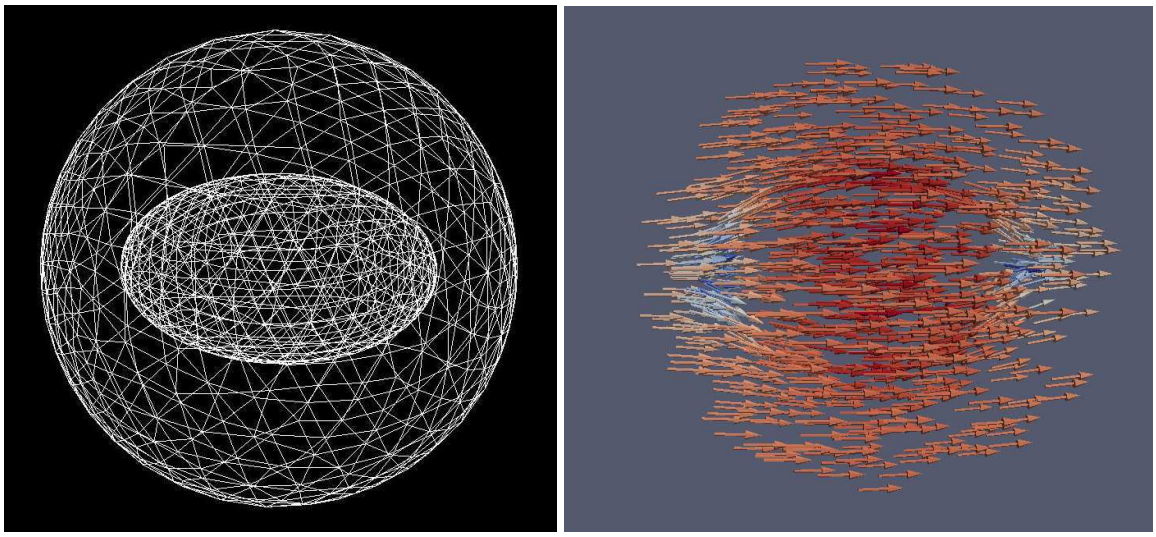

Fig. 7.3. Test case 2. Left: representation of the mesh. Right: potential flow around the ellipsoid

where $\beta := r\left(\left(\varsigma+\gamma_\infty^2 P\right)^2-\gamma_\infty^4 M_\infty^2\right)$, $\boldsymbol{V} := r\left(\left(\varsigma+\gamma_\infty^2 P\right)\mathcal{N}\boldsymbol{M}-\gamma_\infty^3\boldsymbol{M}_\infty\right)$, $\Xi := r\mathcal{N}\mathcal{O}\mathcal{N}$ with $r := \frac{\rho}{\rho_\infty}$, $\varsigma := \frac{c_\infty}{c}$, $\gamma_\infty := \frac{1}{\sqrt{1-M_\infty^2}}$, $P := \boldsymbol{M}\cdot\boldsymbol{M}_\infty$, $\mathcal{N} := I + C_\infty \boldsymbol{M}_\infty \boldsymbol{M}_\infty^T$, $\mathcal{O} := I - \boldsymbol{M}\boldsymbol{M}^T$, and $C_\infty := \frac{\gamma_\infty - 1}{M_\infty^2}$. In the above notation, the subscript ∞ is used for quantities outside the ball, ρ is the density of the flow, c is the speed of sound when the flow is at rest and $\boldsymbol{M} = \frac{\boldsymbol{v}}{c}$, where $\boldsymbol{v}$ is the velocity of the flow. The operators γ_0 and γ_1 are Dirichlet and Neumann traces on the coupling surface Γ_∞. We refer to Chapter 3 for more details on this formulation and its well-posedness. The considered finite-dimensional approximation of (7.47) has 1711 unknowns.

The potential flow, represented in the right panel of Figure 7.3, is part of the data for the problem. We perturb this flow uniformly in space. Although the boundary condition on the solid surface Γ and transmission condition on Γ_∞ are violated as soon as the perturbation of the flow is nonzero, the present study can be viewed as a first step towards quantifying uncertainties on the potential flow and their impact on a quantity of interest. The flow perturbation takes the form $\delta\boldsymbol{M} = \mu_1 \boldsymbol{e}_x + \mu_2 \boldsymbol{e}_y + \mu_3 \boldsymbol{e}_z$. The quantity of interest is the acoustic pressure at a point located on the axis of symmetry, downstream of the object. The parameters of the problem are the frequency of the source, and the magnitude of the uniform perturbations of the potential flow in each Cartesian direction. The frequency varies from 487 to 1082 Hz, and the magnitude of the uniform perturbations of the flow varies from 0 to 0.1. Denote by μ the wavenumber of the source (so that the frequency of the source is $\frac{340\mu}{2\pi}$ in the air at rest), and by μ_1, μ_2, μ_3 the three magnitudes of the perturbation. A goal-oriented reduced basis method is carried out to select a basis of $\hat{n} = 20$ truth solutions using the nonintrusive formula (7.41) to approximate the matrix of the problem, the right-hand side of the direct problem, and the right-hand side of the adjoint problem corresponding to our quantity of interest. For the matrix, the approximation procedure S1O1 is applied to

$$g(\mu, r) := \exp\left(i\mu r\right), \; r = |x-y|, \; x, y \in \Gamma_\infty, \tag{7.50}$$

and the procedure S2O2(z) is applied to

$$z_p(\mu,\mu_1,\mu_2,\mu_3) := \begin{cases} \lambda_m^{\mathrm{S1O1}}(\mu), & 1\le m\le d, \quad p=m, \\ \mu\lambda_m^{\mathrm{S1O1}}(\mu), & 1\le m\le d, \quad p=m+d, \\ \mu^2\lambda_m^{\mathrm{S1O1}}(\mu), & 1\le m\le d, \quad p=m+2d, \\ 1, & p=3d+1, \\ \mu, & p=3d+2, \\ \mu^2, & p=3d+3, \\ \mu^2\mu_3, & p=3d+4, \\ \mu^2\mu_3^2, & p=3d+5, \\ \mu\mu_i, & 1\le i\le 3, \quad p=3d+5+i, \\ \mu\mu_i\mu_3, & 1\le i\le 3, \quad p=3d+8+i, \\ \mu_i\mu_j, & 1\le i,j\le 3, \quad p=3d+11+i+3(j-1), \end{cases} \tag{7.51}$$

where these parameter dependencies have been identified upon injecting $\boldsymbol{M} \to \boldsymbol{M} + \delta\boldsymbol{M}$ in (7.47), while using that $\boldsymbol{M}_\infty$ is collinear to $\boldsymbol{e}_z$. For the right-hand side of the direct and dual problems, the approximation procedure S1O1 is applied to

$$g(\mu,x) := \exp\left(i\mu|x-x_0|\right),\ x\in\Gamma_\infty, \tag{7.52}$$

where x_0 is respectively the position of the source and the point where the quantity of interest is computed, and the approximation procedure S2O2(z) is applied to

$$z_p(\mu) := \begin{cases} \lambda_m^{\mathrm{S1O1}}(\mu), & 1\le m\le d, \quad p=m \\ \mu\lambda_m^{\mathrm{S1O1}}(\mu), & 1\le m\le d, \quad p=m+d. \end{cases} \tag{7.53}$$

The EIM algorithms are carried out with $d=13$ and $d^z=25$ for the matrix, and $d=13$ and $d^z=18$ for the right-hand side of the direct and dual problems. Over the considered parameter values, the relative error for the three nonintrusive formulae is of the order of 10^{-12} (in Frobenius norm for the matrix and Euclidian norm for the vectors). The maximum error bound (over a discretization $\mathcal{P}_{\mathrm{trial}}$) is of the order of 10^{-7}, the online stage takes 2.8×10^{-3} s to compute a reduced solution and the error bound, while the full direct problem is solved in about 14 s, which corresponds to an acceleration factor of 5×10^3.

To illustrate, we suppose that the perturbation of the potential flow is modelled by random variables: the law of μ_1 is a truncated Gaussian, that of μ_2 is a uniform law, and that of μ_3 is a truncated log-normal low. The goal is to compute the probability density function of the quantity of interest. Figure 7.4 shows a screenshot of a java applet computing an histogram of the values taken by the quantity of interest, at a frequency selected by the user.

7.4.4 A scalable RBM implementation applied to an industrial test case of an impedant aircraft in the air at rest

In BEM implementations for the Helmholtz equation, the Fast Multipole Method (FMM) allows one to approximately compute matrix-vector products, and then approximately solve linear systems using iterative methods, in complexity scaling with $n\log n$, where n denotes the

Fig. 7.4. Java applet for the online stage of the reduced basis method for test case 2. Top panel: histograms of the real part and imaginary part of the quantity of interest. Bottom panel: histograms of the three components of the perturbation of the flow

number of unknowns [27, 99]. For boundary integral systems, the matrices are dense, and have a priori n^2 nonzero complex coefficients. In this section, we consider a test case where the matrices $A_{\mu_m^{\mathrm{S2}(z)}}$, where $\mu_m^{\mathrm{S2}(z)}$ are the parameter values selected when applying the nonintrusive formula (7.41) to the approximation of A_μ, are so large that they cannot be stored on the hard drive of the computer used for the simulations. Therefore, each time a matrix-vector product is carried out, the matrix is assembled, and the FMM is used.

We consider the same problem as in Section 7.4.2, i.e., the scattering of an incoming acoustic field by an object whose surface has been coated on three zones by three impedant materials. However, the considered scattering object is now an aircraft, see Figure 7.5. Two meshes are considered: one leading to a discrete formulation with 11831 unknowns, the other leading to a discrete formulation with 60866 unknowns. The source is an acoustic monopole, located under the right wing of the plane. The parameters of the problem are the frequency of the source, and the impedance of the three zones composing the surface of the aircraft. The frequency varies from 27 to 135 Hz, and each impedance coefficient varies from 1 to 2. We take 532400 parameter values in $\mathcal{P}_{\mathrm{trial}}$ (400 values for the frequency, and 11 values for each impedance coefficient).

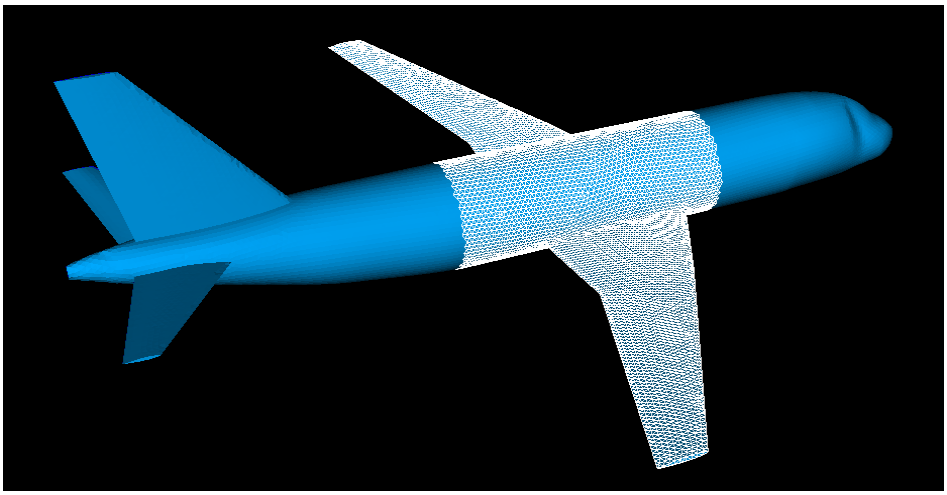

Fig. 7.5. Second impedant surface, with the finest mesh for test case 3

First, the RBM is applied to the problem on the coarser mesh. To recover the affine dependence assumption, we use the nonintrusive approximations detailed in Section 7.4.2, with (7.43)-(7.44) for the matrix decomposition (with now $d = 35$ and $d^z = 50$) and (7.45)-(7.46) for the decomposition of the right-hand side of the problem (with $d = 50$ and $d^z = 60$). With $\hat{n} = 30$ basis vectors selected by the greedy algorithm, the relative error between the direct solution and the reduced solution, in Euclidian norm, at the value of the parameters that maximizes the error bound, is less than 3%. Notice that the procedure is scalable with respect to the number of available CPUs: the matrix-vector products in FMM and the exploration of $\mathcal{P}_{\text{trial}}$ by the greedy algorithm, which are the two steps with high computational complexity (dependent on the number of unknowns and on the cardinality of $\mathcal{P}_{\text{trial}}$), are parallel. Therefore, the procedure is expected to highly benefit from large clusters.

We now consider the finer mesh. Each time a vector U_{μ_j} is added to the reduced basis, we have to compute the $d^z = 50$ matrix-vector products $A_{\mu_m^{\text{S2}(z)}} U_{\mu_j}$, $1 \leq m \leq d^z$, where the $\mu_m^{\text{S2}(z)}$'s are the values of the parameter in the nonintrusive approximation formula (7.41). Therefore, in addition to the resolution of the direct problem, 50 matrices have to be assembled at each step of the greedy algorithm, which is time-consuming. However, once a matrix is constructed, it is relatively cheap to compute many matrix-vector products with the same matrix. Hence, a greedy algorithm is not considered on the finer mesh, but the values of the parameters selected by the greedy algorithm on the coarser mesh are directly used to build the reduced basis. This way, the 50 matrices are constructed once, and the 30 matrix-vector products (corresponding to $\hat{n} = 30$ values of the parameter selected by the greedy algorithm on the coarser mesh) are carried out at once. The simulations have been performed on a laptop with a quadricore CPU, and 4 Go of RAM. The formula (7.41) allows us to directly use the FMM. Without the FMM, this simulation on this computer would have been impossible, since one matrix needs 60 Go to be stored. An approach consisting in computing and storing the 50 matrices of the decomposition would need 3 To of memory.

The online stage takes 1.5×10^{-2} s to compute a reduced solution and the error bound, while the full direct problem is solved in about 40 minutes, which corresponds to an acceleration

factor of 1.6×10^5. The offline stages are computed in about 2 days, and the last step of the greedy algorithm in the offline stage of the RBM with the coarser mesh takes 1 hour. The FMM we used computes matrix-vector products with a relative accuracy of approximately 10^{-3}; therefore, we cannot expect to achieve a much more accurate RB approximation.

The acoustic field in the exterior domain is computed from the solution to (7.42) using a representation formula, which is a linear operation. We consider the acoustic field on an array of 1681 points located behind the aircraft. We can precompute this field using the vectors of the reduced basis as solutions, and the quantity of interest is directly obtained at any parameter values from these precomputed fields and the components of the reduced solutions. Figure 7.6 shows a screenshot of a java applet computing this acoustic field at a set of parameters selected by the user (frequency and impedance coefficients).

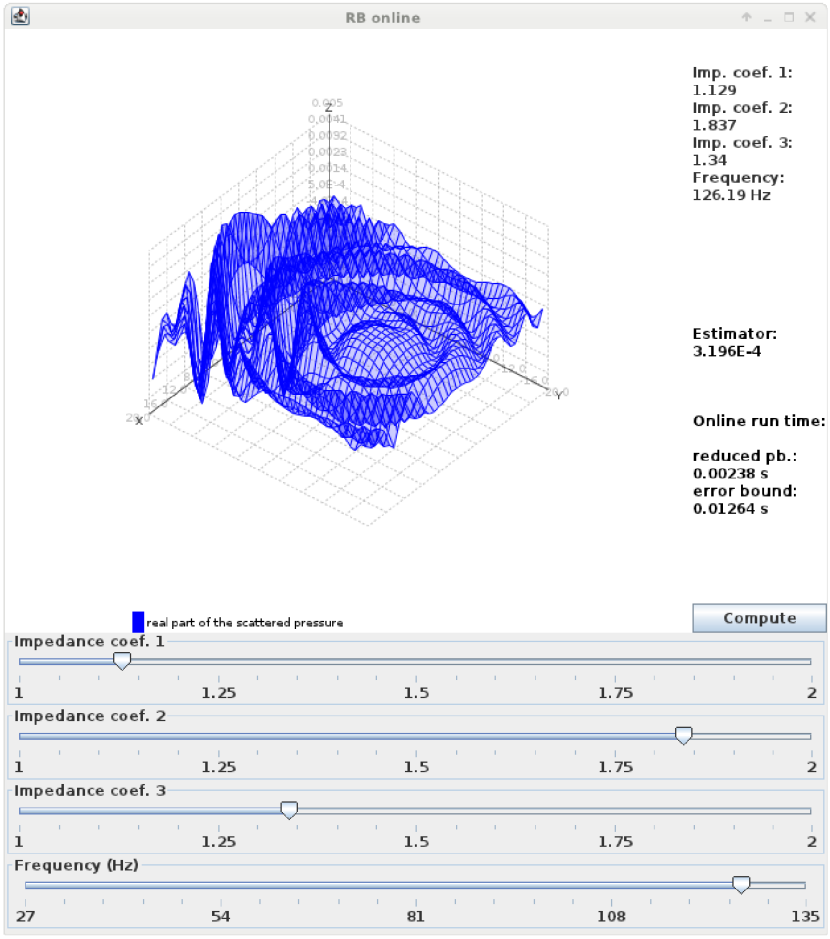

Fig. 7.6. Java applet for the online stage of the reduced basis method test case 3. Top panel: total acoustic pressure field on an array of 1681 points located behind the aircraft. Bottom panel: selection of the impedance coefficients and of the frequency

Consider the following parameter values: frequency = 122.3 Hz, $\mu_1 = 1.21$, $\mu_2 = 1.87$ and $\mu_3 = 1.45$. The error bound is 5.4×10^{-4}, and the relative error between the direct solution

and the reduced solution, in Euclidian norm, is 1%. On the array of 1681 points located behind the aircraft, the relative error for the scattered acoustic field is 1.4%. Figure 7.7 shows the corresponding acoustic pressure fields, and the difference between the reduced basis and direct solutions.

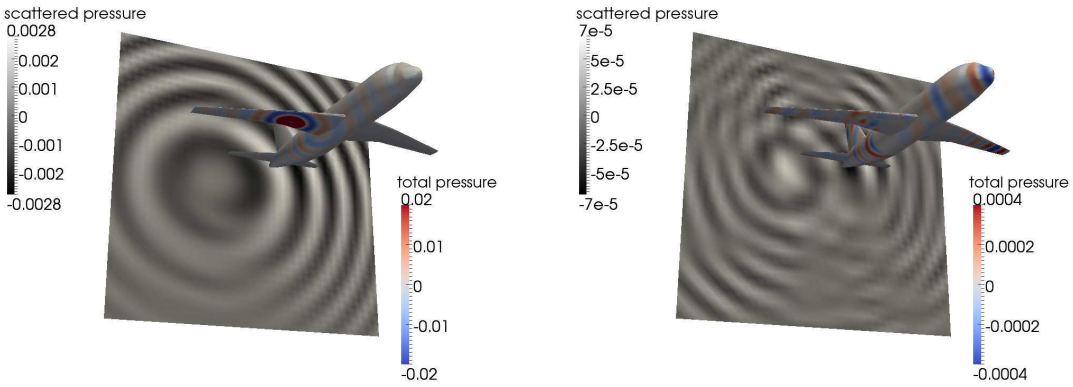

Fig. 7.7. Test case 3. Left: acoustic pressure fields on the aircraft and on an array of 1681 points located behind the aircraft computed solving the direct problem. Right: difference between the reduced basis and the direct solution

7.5 Conclusion

In this work, we derived a nonintrusive procedure for the reduced basis method. Its implementation is relatively simple: it has been successfully and easily applied to the approximation of various matrices and right-hand sides. In particular, this procedure allows for the direct use of advanced linear algebra tools, since we are only dealing with quantities already assembled by the computational code at hand.

8

A multiscale problem in thermal science

This chapter is based on the article [Ar2].

Summary. We consider a multiscale heat problem in civil aviation: determine the temperature field in a plane in flying conditions, with air conditioning. Ventilated electronic components in the bay bring a heat source, introducing a second scale in the problem. First, we present three levels of modeling for the physical phenomena, which are applied to the two sub-problems: the plane and the electronic component. Then, having reduced the complexity of the problem to a linear non-symmetric coercive PDE, we will use the reduced basis method for the electronic component problem.

8.1 Introduction

In the civil aircraft industry, one of the main stakes is fuel efficiency. The use of composite materials, to replace aluminum alloys, enables the manufacturers to lighten the plane while keeping the required mechanical properties. However, these materials present lower thermal conductivity, leading to new air conditioning problems. The goal of the present work is to develop fast tools to compute the temperature in an aircraft cabin in flying conditions, with presence of heat sources: mainly the electronic components.

A closer look will be taken at two problems:
- the passenger comfort in which case the output of interest is the temperature in the cabin;
- the equipment failure in which case the output of interest is the maximum of the temperature in the electronic components.

After presenting three levels of modeling for the physical phenomena, we present numerical simulations of the model providing the best trade-off between physical realism and computational accuracy. Then, a reduced basis approach will be developed for the electronic component problem to further speed up the computations.

8.2 Physical modeling

In both cabin and equipment problems, the flow as well as the temperature have to be computed. We first introduce three models of increasing complexity to solve this physical prob-

lem. The standard continuum model for natural convection phenomenon is the compressible Navier-Stokes system (CNS). This model present both theoretical and numerical difficulties in the sense that the equations of conservation are strongly coupled and nonlinear. We choose to consider a hierarchy of simplifications of (CNS).

8.2.1 A hierarchy of models

Consider a bounded domain $\Omega \subset \mathbb{R}^2$ representing the cabin or the electronic components. The domain Ω is split in two parts:

$$\Omega = \Omega_{\text{solid}} \cup \Omega_{\text{air}},$$

where Ω_{solid} stands for solid structures in the cabin or in the electric component. Thus, the velocity field is considered as non-zero on $\Omega \setminus \Omega_{\text{solid}}$ and extended by 0 on Ω_{solid}. As a first simplification, we will consider the Boussinesq equations, neglecting density variations except in the body force so that the fluid is divergence free. The coupling between the velocity and temperature fields appears in the body force terms in the equations of the fluid and the advective term in the heat equation. This model expresses conservation of momentum and mass of the fluid coupled to the heat equation. The unsteady equations of conservation reads, for a time $t_{\text{S}} > 0$, for all $t \in [0, t_{\text{S}}]$:

$$\begin{cases} \rho_0 \left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} \right) = \rho_0 \left(1 - \dfrac{T - T_0}{T_0} \right) \boldsymbol{g} - \nabla p + \eta \Delta \boldsymbol{u} \text{ in } \Omega_{\text{air}}, \\ \text{div } (\boldsymbol{u}) = 0 \text{ in } \Omega_{\text{air}}, \\ \rho_0 c_p(x) \left(\dfrac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T \right) = \text{div } (\kappa(x) \nabla T) + Q(x) \text{ in } \Omega, \end{cases} \tag{8.1}$$

Here, $\boldsymbol{u}$ denote the velocity field, p the pressure and T the temperature. T_0 is a reference temperature (300 Kelvin), ρ_0 is the air density at temperature T_0, $\boldsymbol{g}$ the gravity constant, η the air dynamic viscosity. Then, c_p and κ are space dependent discontinuous functions and represent the heat capacity and the heat conductivity of the considered medium (*e.g* air, aircraft structure). Eventually, Q is a space dependent function representing a source term in the heat problem. In the following, the function κ, c_p, Q have the form

$$\kappa(x) = \sum_i \kappa_i \mathbf{1}_{\Omega_i}, \quad c_p(x) = \sum_i c_p^i \mathbf{1}_{\Omega_i}, \quad Q(x) = \sum_i Q_i \mathbf{1}_{\Omega_i}, \tag{8.2}$$

where indices i refer to the air and different solid parts depending on the considered problem (aircraft structure, electronic component part,...).
If we assume that the variations of temperature do not modify the velocity field, we can decouple the fluid and heat problems. Moreover, we consider the fluid at steady state thereby neglecting any feedback of the temperature on the convection of the air. The conservation equations writes for a time $t_{\text{S}} > 0$, for all $t \in [0, t_{\text{S}}]$:

$$\begin{cases} \rho_0 \boldsymbol{u}_{\text{NS}} \cdot \nabla \boldsymbol{u}_{\text{NS}} = \rho_0 \boldsymbol{g} - \nabla p + \eta \Delta \boldsymbol{u}_{\text{NS}} \text{ in } \Omega_{\text{air}}, \\ \text{div } (\boldsymbol{u}_{\text{NS}}) = 0 \text{ in } \Omega_{\text{air}}, \\ \rho_0 c_p(x) \left(\dfrac{\partial T}{\partial t} + \boldsymbol{u}_{\text{NS}} \cdot \nabla T \right) = \text{div } (\kappa(x) \nabla T) + Q(x) \text{ in } \Omega. \end{cases} \tag{8.3}$$

The last level of modeling consists, as in the previous case, in taking a stationary regime and in decoupling the fluid and heat problems for the same reasons. The next simplification is to consider a basic model for the fluid equation, namely a potential flow. All the phenomena induced by the viscosity are not captured by this type of model (recirculation zones, boundary layers).

The conservation equations writes:

$$\begin{cases} -\Delta\psi = 0 \text{ in } \Omega_{\text{air}}, \\ \rho_0 c_p(x)\left(\partial_t T + \boldsymbol{u}_{\text{pot}} \cdot \nabla T\right) = \text{div }\left(\kappa(x)\nabla T\right) + Q(x) \text{ in } \Omega. \end{cases} \tag{8.4}$$

Here, $\boldsymbol{u}_{\text{pot}} = \nabla\psi$ is the velocity field associated with the potential ψ. All the three models are supplemented with initial and boundary conditions. A nice consequence for the numerical calculations is that for the two last systems (8.3) and (8.4), a single evaluation of the velocity field is required.

8.2.2 Geometry and boundary conditions

Now, we describe the different boundary conditions that we consider in the numerical experiments. Assume that the boundary of Ω is partitioned as follow:

$$\partial\Omega = \Gamma_{\text{in}} \cup \Gamma_{\text{out}} \cup \Gamma_{\text{wall}}.$$

The portions $\Gamma_{\text{in/out}}$ represent parts of the domain where there are exchange of air (fans and evacuations) and Γ_{wall} are solid adiabatic walls. We enforce non-penetration and no-slip boundary conditions for the fluid flow on the walls:

$$\boldsymbol{u} = 0 \text{ on } \Gamma_{\text{wall}}, \tag{8.5}$$

and an inflow of air is imposed on Γ_{in} through a Dirichlet boundary condition:

$$\begin{cases} \boldsymbol{u} \cdot \boldsymbol{n} = u_{\text{in}} \text{ on } \Gamma_{\text{in}}, \\ \eta\left(\nabla\boldsymbol{u} \cdot \boldsymbol{n}\right) \cdot \boldsymbol{\tau} = 0 \text{ on } \Gamma_{\text{in}}, \end{cases} \tag{8.6}$$

where u_{in} is a scalar function that we will specify for the numerical calculations and τ a tangent vector to Γ_{in}. In the case of systems (8.1) and (8.3), we enforce natural boundary condition, requiring:

$$\eta\left(\nabla\boldsymbol{u} \cdot \boldsymbol{n}\right) = p\boldsymbol{n} \text{ on } \Gamma_{\text{out}}. \tag{8.7}$$

In the case of system (8.4), in order to have a well posed problem, we impose exact conservation of mass enforcing:

$$\boldsymbol{u}_{\text{pot}} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma_{\text{out}}, \tag{8.8}$$

so that the Poisson problem for the potential flow has to be solved with non-homogeneous Neumann boundary conditions. With this setting, the potential is determined up to an additive constant. For the numerical experiments we fix $\int_{\Omega_{\text{air}}} \psi = 0$. This condition ensure the well-posedness of the problem.

For the temperature field, we assume Dirichlet boundary conditions in the inflow part. Since we assume that the wall are adiabatic, the boundary conditions for the temperature reads:

$$\begin{cases} T = T_{\text{in}} \text{ on } \Gamma_{\text{in}}, \\ \nabla T \cdot \boldsymbol{n} = 0 \quad \text{on } \Gamma_{\text{out}} \cup \Gamma_{\text{wall}}. \end{cases} \tag{8.9}$$

8.2.3 Time and space discretization

We now describe the numerical implementation used to solve the previous systems of conservation equations. We describe the numerical algorithm used to solve unsteady Navier-Stokes equations and the heat equation. We use those algorithms in an iterative process to solve the Boussinesq equations. We start by describing the Navier-Stokes solver. The discretization is based on finite elements in space and implicit Euler scheme in time. Let δt be the time step, taken to be constant for simplicity. We denote by $t^n = n\delta t$ the n-th discrete time. We introduce $V_h(\Omega_{\text{air}})^2$ and $M_h(\Omega_{\text{air}})$ the finite elements spaces for velocity and pressure. We define the fluid problem at time t^{n+1} for a temperature T by: given $\boldsymbol{u}^n \in V_h(\Omega_{\text{air}})^2$, we seek $(\boldsymbol{u}^{n+1}, p^{n+1}) \in V_h(\Omega_{\text{air}})^2 \times M_h(\Omega_{\text{air}})$ such that for all $(\boldsymbol{v}, q) \in V_h(\Omega_{\text{air}})^2 \times M_h(\Omega_{\text{air}})$,

$$\text{Fluid}(n+1,T) := \begin{cases} \dfrac{1}{\delta t}\displaystyle\int_{\Omega_{\text{air}}} \rho_0 \boldsymbol{u}^{n+1}\cdot\boldsymbol{v}\,d\boldsymbol{x} + \int_{\Omega_{\text{air}}} \rho_0\left(\boldsymbol{u}^{n+1}\nabla\boldsymbol{u}^{n+1}\right)\cdot\boldsymbol{v}\,d\boldsymbol{x} + \eta\displaystyle\int_{\Omega_{\text{air}}} \boldsymbol{\nabla}\boldsymbol{u}^{n+1} : \boldsymbol{\nabla}\boldsymbol{v}\,d\boldsymbol{x} \\ -\displaystyle\int_{\Omega_{\text{air}}} p^{n+1}\text{div}\,(\boldsymbol{v})\,d\boldsymbol{x} = \dfrac{1}{\delta t}\displaystyle\int_{\Omega_{\text{air}}} \rho_0\boldsymbol{u}^n\cdot\boldsymbol{v}\,d\boldsymbol{x} + \int_{\Omega_{\text{air}}} \rho_0\left(1-\dfrac{T-T_0}{T_0}\right)\boldsymbol{g}\cdot\boldsymbol{v}\,d\boldsymbol{x}, \\ \displaystyle\int_{\Omega_{\text{air}}} q\text{div}\,(\boldsymbol{u}^{n+1})\,d\boldsymbol{x} = 0. \end{cases} \tag{8.10}$$

Let us denote $\tilde{\boldsymbol{u}}$ the extension by zero of $\boldsymbol{u}$ on Ω. We search the temperature in the same finite element space than the velocity components. We define the heat problem at time t^{n+1} for a fluid velocity $\tilde{\boldsymbol{u}}$ by: given $T^n \in M_h(\Omega)$, we seek $T^{n+1} \in M_h(\Omega)$ such that for all $\Theta \in M_h(\Omega)$:

$$\text{Heat}(n+1,\tilde{\boldsymbol{u}}) := \begin{cases} \dfrac{1}{\delta t}\displaystyle\int_{\Omega} \rho_0 c_p(x) T^{n+1}\Theta\,d\boldsymbol{x} + \int_{\Omega} \rho_0 c_p(x)\left(\tilde{\boldsymbol{u}}\cdot\nabla T^{n+1}\right)\Theta\,d\boldsymbol{x} + \\ \displaystyle\int_{\Omega} k(x)\nabla T^{n+1}\cdot\nabla\Theta\,d\boldsymbol{x} = \int_{\Omega} Q(x)\Theta\,d\boldsymbol{x} + \dfrac{1}{\delta t}\int_{\Omega}\rho_0 c_p(x)T^n\Theta\,d\boldsymbol{x}. \end{cases} \tag{8.11}$$

The variational formulation for steady Navier-Stokes equations and equation can be straightforwardly obtained from (8.10) and (8.11). Note that the time scheme for the system (8.10) is nonlinear and we resort to a Newton-Raphson algorithm to solve this nonlinear problem at each time step. In the case of the potential flow, one can resort to a variational formulation of the Poisson equation with Neumann boundary conditions in conjunction with a Lagrange multiplier to ensure that $\int_{\Omega_{\text{air}}}\psi = 0$. For instance, we take the inf-sup stable pair of discrete finite elements spaces $V_h(\Omega_{\text{air}})^2 \times M_h(\Omega_{\text{air}}) = (\mathbb{P}_2)^2 \times \mathbb{P}_1$ for the fluid equations and $M_h(\Omega) = \mathbb{P}_1$ for the heat equation. Let us depict the iterative process allowing to solve the Boussinesq system:

Initialize with $\boldsymbol{u}_0 \in V_h(\Omega_{\text{air}})^2$ and $T_0 \in M_h(\Omega)$; For all $n \geq 0$,
(i) solve Fluid$(n+1, T^n)$ to get $\boldsymbol{u}^{n+1}$;
(ii) solve Heat$(n+1, \boldsymbol{u}^{n+1})$ to get T^{n+1};

All calculations are performed using the finite element solver `FreeFEM++` (see [52]). As a matter of illustration, the computationnal code was tested on a simple configuration: consider a square box, with initial state uniform temperature and fluid at rest. A source term Q for the heat problem is localized in the lower part of the box. As we can see in Figure 8.1, the temperature dependent density in the gravitational term lightens hot air and weighs down cold air. This enable a nonzero velocity field to arise, and a convective dissipation of the heat.

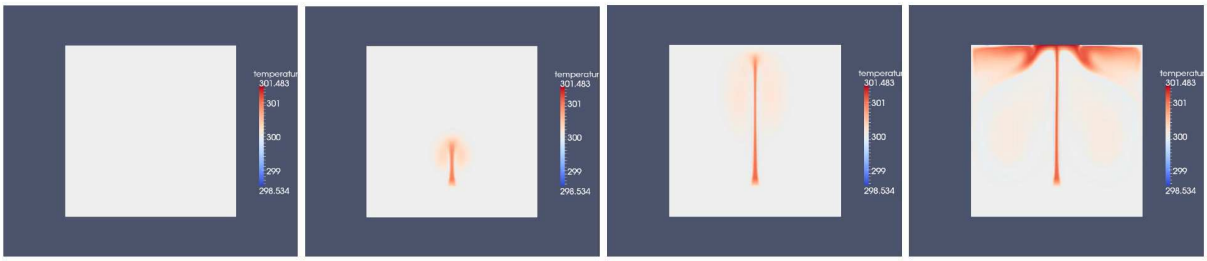

Fig. 8.1. Temperature field at times t=0, 0.5, 1 and 4 s

8.2.4 Numerical results

The plane

Consider a bidimensional cross section of a plane (see Figure 8.2). The upper part, the cabin, presents an inflow of cool air coming from an air conditionning system. Air is leaving the plane through a hole in the lower part, the bay, which contains the electronic components that create heat by Joule effect. A simplified geometry is considered: the goal here is to develop a methodology, not to carry out an accurate industrial simulation. We consider that the heat

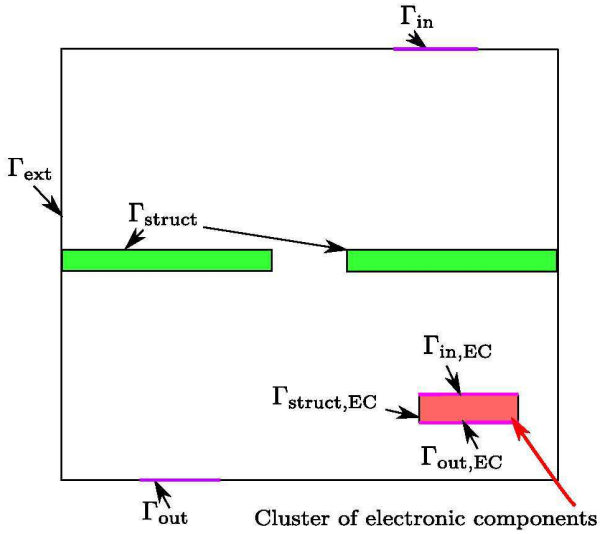

Fig. 8.2. Geometry of the plane test case

produced by the components is completely brought to the bay. Therefore, we can take a constant effective surfacic source term Q in the heat problem. The velocity on $\Gamma_{\text{in,EC}}$ is imposed by the fans of the components. This methodology has been applied for the three different models considered previously.

– Unsteady Boussinesq:
  See Figures 8.3 and 8.4. As expected in such physical situation, the Boussinesq solution does not reach a steady state. The convection effects induced by temperature gradients are the dominant effects, therefore the Boussinesq coupling could not be simplified. Cool

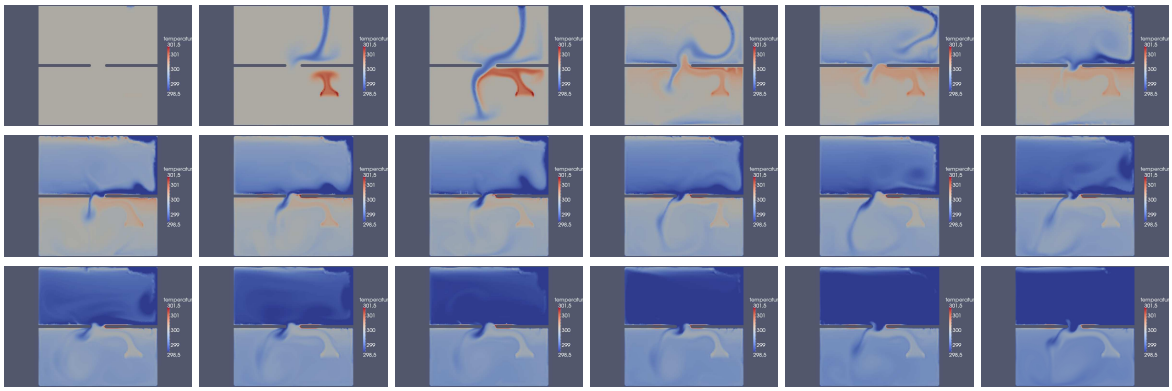

Fig. 8.3. Temperature field in the Boussinesq model, with the cluster at the right of the bay, at times 0, 100, 200,... 1700

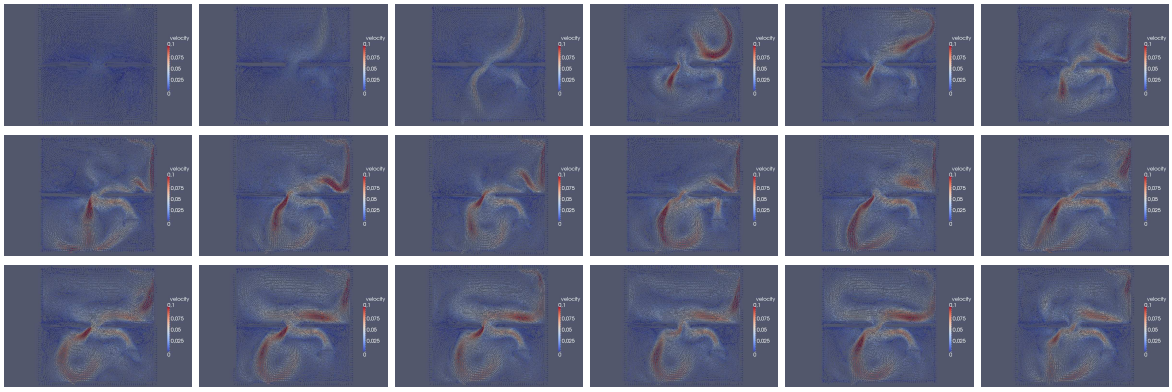

Fig. 8.4. Velocity field in the Boussinesq model, with the cluster at the right of the bay, at times 0, 100, 200,... 1700

air is coming from the upper part and the air in the lower part is being heated up, and the hydrostatic equilibrium is not reached at the end of the simulation.

– Steady decoupled NS/heat: See Figure 8.5. The velocity field has been computed once

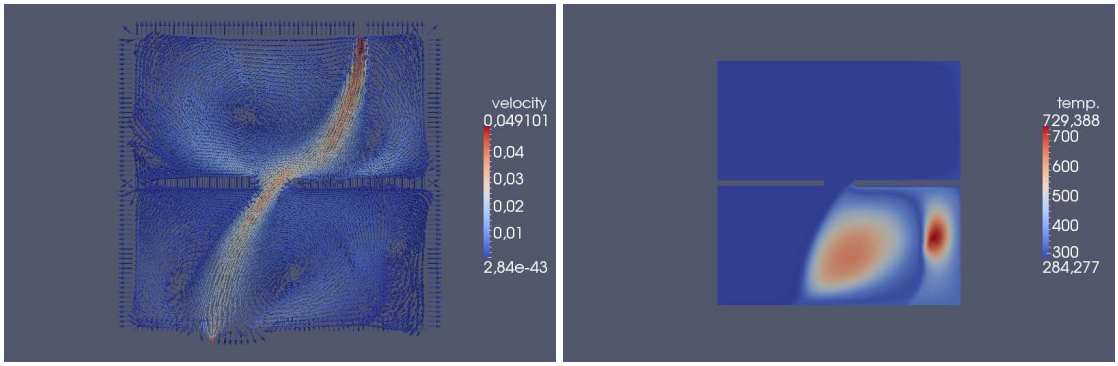

Fig. 8.5. Plane test case in the decoupled NS/heat model. Velocity and temperature fields.

for all, and is assumed not to depend on the temperature. The air viscosity produces recirculation zones. Without Boussinesq effect, the cooling down is only ensured by diffusion and convection from the precomputed velocity field. These two effects are too low, and the temperature reaches very high values.

– Unsteady decoupled potential/heat: See Figure 8.6. The precomputed flow is very simple,

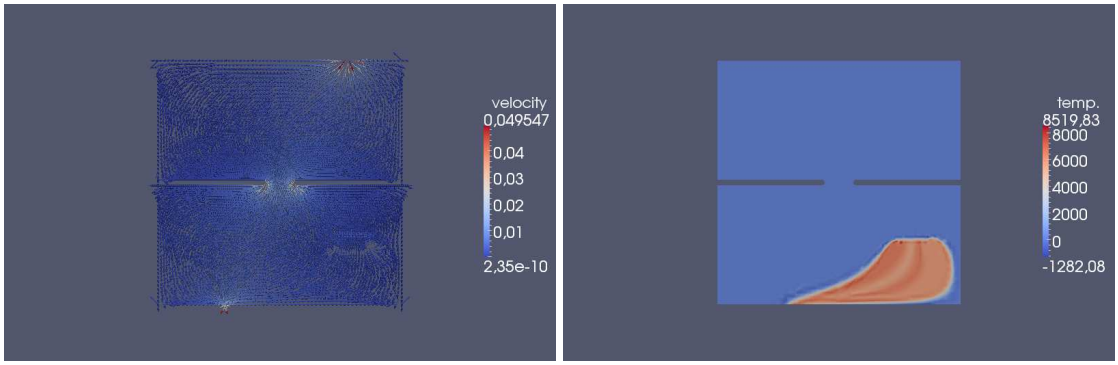

Fig. 8.6. Plane test case in the decoupled potential/heat model. Velocity and temperature fields.

viscosity is neglected. Recirculation zones are not obtained and therefore the cooling of the components (away from the air conditioning main stream) is even worse. Diffusion and convection are way too low, and the temperature diverges.

For this situation, only the unsteady Boussinesq can capture all the physical phenomena, and is therefore the model to be considered. In the two other models, the convection brought by the Boussinesq term is not captured, and the components are not cooled down enough. Note that there do not seem to be any steady state to reach for this problem. We simulated up to 40 min, and the velocity and temperature fields appear to reach a periodic in time behavior.

In the present section, we simply modeled the presence of electronic components by a constant surfacic heat source term. We will now develop a model to simulate the velocity and the temperature inside the electronic component.

The electronic component

Consider a 2D section of an electronic component (see Figure 8.7). The green area represent the support board for the red-colored integrated circuits. The blue zone is filled with air, pushed from the bottom part by a fan, and leaving the box through the top part. The red components will heat up by Joule effect while functioning. Periodic boundary conditions are enforced at Γ_{per}.

In this case, the steady decoupled incompressible Navier-Stokes / heat model gives the same result as the long time Boussinesq (see Figure 8.8). Actually, the Péclet number is large ($\approx 10^4$), indicating that convective effects are dominant. Moreover, the flux is strongly guided inside a channel, which is different from the previous case where air was blown inside a large volume at rest. Forced convection dominates convection induced by local thermal fluxes. The unsteady decoupled potential / heat is not satisfactory: the absence of boundary layer does not

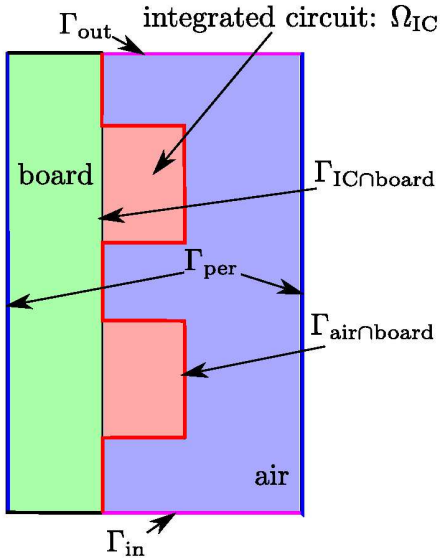

Fig. 8.7. Distorted geometry of the electronic component test case (the height of the component is actually 16 times larger than its width)

allow the air close to the integrated circuits to heat up. In this case, contrary to the previous one, the convection is overestimated in the heating area.

As a conclusion, a decoupling of the fluid and the heat problems is possible. For instance, in order to determine the conductivities and heat capacities of the elements of the electronic component that minimize the temperature, a reduced basis approach can be carried out to drastically reduce the computation time in such a many queries context.

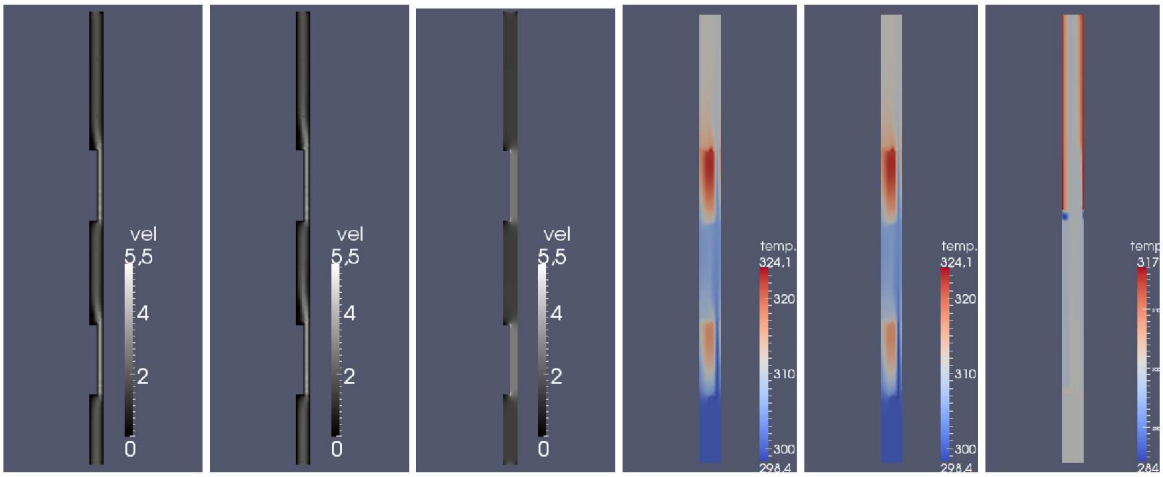

Fig. 8.8. Left: velocity field, right: temperature field, for Boussinesq, NS and potential cases, in the international system of units.

8.3 A reduced basis approach for the electronic component problem

8.3.1 Review of the method

The reduced basis (RB) method aims at reducing the computation time in a precise context: running many times the same calculation with a small change of a set of parameters. The idea is somehow close to modal decomposition in mechanical vibrations: the solution should be well represented by a small set of precomputed solutions (the reduced basis). A wide variety of problems can be tackled by this method (see [72], [19]). We will present it in the context of the present study: the steady decoupled incompressible Navier-Stokes / heat model (8.3) applied to the electronic component. The reduced basis method has been applied to a heat conduction problem by Sen (see [94]), and recently to the unsteady Boussinesq equations by Knezevic, Nguyen and Patera (see[59]). Consider that the fluid equations have been solved once for all, the RB method will be applied to the set of equations:

$$\rho_0 c_p \boldsymbol{u}_{\mathrm{NS}} \cdot \nabla T - \mathrm{div}\ (\kappa(x)\nabla T) = Q(x), \tag{8.12}$$

where $\boldsymbol{u}_{\mathrm{NS}}$ is the result of the preliminary incompressible Navier-Stokes computation, and $Q(x) = q\mathbf{1}_{x\in\Omega_{\mathrm{IC}}}$ (Ω_{IC} stands for integrated circuits and is defined in Figure 8.7). $\boldsymbol{u}_{\mathrm{NS}}$ is extended by $\mathbf{0}$ in the integrated circuits and in the board.

Equation (8.12) can be rewritten as

$$A_\mu T_\mu = f, \tag{8.13}$$

with μ a set of n parameters (for instance $c_{p_{\mathrm{IC}}}$, κ_{board}, etc...) in a n-dimensional subspace $\mathcal{D}$ of $\mathbb{R}^n$. Intervals of variation for the parameters are specified in section 8.3.3. Here, A_μ is the linear operator of the problem, f the source term and T_μ represents the temperature, solution of equation (8.12) with the parameters set μ.

Let us first check that the RB approach is reasonable in the present case, by selecting by hand a finite set of parameter values, computing the corresponding solutions T_i, selecting a problem-related scalar product - $(T,\Theta) = \int_\Omega \nabla T \cdot \nabla\Theta$, and checking the variation of the eigenvalues of the matrix $M_{ij} = (T_i, T_j)$. This is close to finding principal components of the energy operator (like modal basis). On Figure 8.9, we see that the eigenvalues of the gramian matrix M of the problem lose 8 orders of magnitude with a reduced basis of size 81 chosen randomly, and the decreasing is exponential. This indicates that the solutions, for $\mu \in \mathcal{D}$, can be efficiently represented by a linear combination of a small number of functions.

We are interested in the variational formulation of equation (8.13): Find $T_\mu \in M_h$ such that $\forall\Theta \in M_h$

$$a_\mu(T_\mu, \Theta) = l(\Theta), \tag{8.14}$$

where M_h is a finite dimensional subspace of $X := H^1_{\mathrm{per},0}(\Omega) = \{T \in H^1(\Omega) \,|T_{\Gamma_{\mathrm{in}}} = 0, T \text{ periodic at } \Gamma_{\mathrm{per}}\}$, (we can make the Dirichlet condition homogeneous, considering a lifting of T_{in}). We use the Lagrange's $\mathbb{P}_1$ finite elements for this heat equation (see section 8.2.3). Since we have mixed boundary condition with a homogeneous Dirichlet condition, $||.||_{H^1_0(\Omega)}$ is a norm on $H^1_{\mathrm{per},0}(\Omega)$. The considered output is $s_\mu = l(T_\mu) = (f, T_\mu)_{H^1_0(\Omega)} = q\int_{\Omega_{IC}} T_\mu$, a quantity

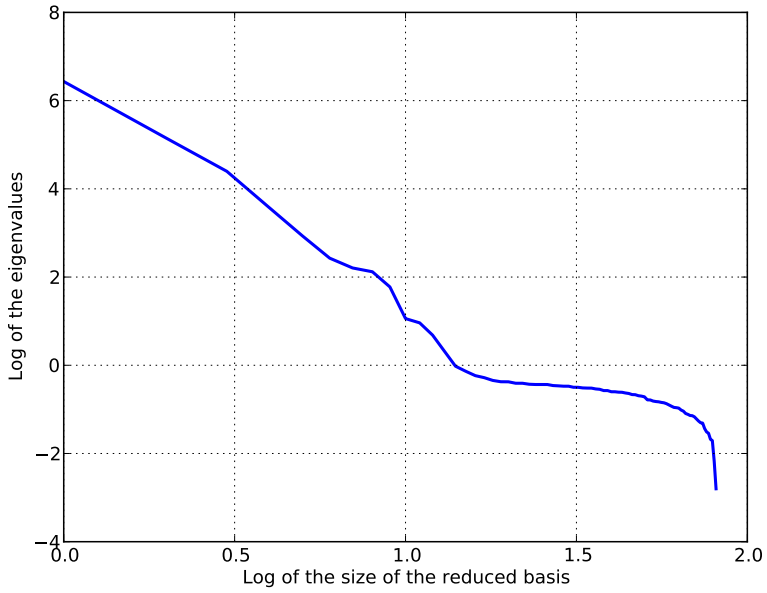

Fig. 8.9. Eigenvalues of the gramian matrix M

proportional to the mean temperature in the integrated circuits. This specific output allows efficient mathematical analysis.

The RB consists in two steps:
- A computationaly heavy offline stage: construct a low dimensional basis, which is a good basis for the high dimensional problem (8.12) for every parameter μ in $\mathcal{D}$,
- Fast online stages: solve light low dimensional problems.

The challenge of this approach is to guaranty that the approximate solution of equation (8.12) is a good one. This is possible thanks to the efficient computation of an a posteriori error estimate for our specific quantity of interest.

8.3.2 Goal-oriented a posteriori error estimate : certified RB

Consider T_μ^{FE} the solution of (8.14) and T_μ^{RB} the current RB approximation of (8.14) (the proper construction of T_μ^{RB} is explained in section 5.2.5). These quantities verify $a_\mu\left(T_\mu^{\mathrm{FE}}, \Theta\right) = l(\Theta),\ \forall \Theta \in M_h$ and $a_\mu\left(T_\mu^{\mathrm{RB}}, \Theta\right) = l(\Theta),\ \forall \Theta \in M_{\mathrm{RB}}$, where M_{RB} is the space spanned by the current reduced basis, and has much lower dimension than M_h.

The quantity of interest is s_μ^{FE}, which is computed using the expensive finite element solution, is approximated by s_μ^{RB}, which is computed using reduced basis approximation. We need an accurate and fastly computable a posteriori error estimate Δ_μ for the approximation error between the RB and the FE solution : $|s_\mu^{\mathrm{RB}} - s_\mu^{\mathrm{FE}}|$. By "fast", we mean that the evaluation of Δ_μ should not require the computation of the FE solution.

The bilinear form a_μ in Equation (8.14) is coercive and non symmetric: given $\boldsymbol{u}_{\mathrm{NS}} \in V_h^2(\Omega_{\mathrm{air}})$ we seek $T \in M_h$ such that for all $\Theta \in M_h$:

$$a_\mu(T,\Theta) = \int_\Omega \rho_0 c_p (\boldsymbol{u}_{\mathrm{NS}} \cdot \nabla\Theta) T + k\nabla\Theta \cdot \nabla T + \frac{1}{2}\int_\Omega \rho_0 c_p \mathrm{div}(\boldsymbol{u}_{\mathrm{NS}}) T\Theta.$$

where the last integral is the Temam term.

The non-symmetry is induced by the convective terme $\boldsymbol{u}_{\mathrm{NS}} \cdot \nabla T$. Recall that the precomputed velocity field $\boldsymbol{u}_{\mathrm{NS}}$ is incompressible and cancels on $\partial\Omega$:

$$\begin{aligned} a_\mu(T,T) &= \int_\Omega \rho_0 c_p (\boldsymbol{u}_{\mathrm{NS}} \cdot \nabla T) T + \int_\Omega k\nabla T \cdot \nabla T + \frac{1}{2}\int_\Omega \rho_0 c_p \mathrm{div}(\boldsymbol{u}_{\mathrm{NS}}) T^2 \\ &= -\frac{1}{2}\int_\Omega \rho_0 c_p \mathrm{div}(\boldsymbol{u}_{\mathrm{NS}}) T^2 + \int_\Omega k\nabla T \cdot \nabla T + \frac{1}{2}\int_\Omega \rho_0 c_p \mathrm{div}(\boldsymbol{u}_{\mathrm{NS}}) T^2. \end{aligned} \tag{8.15}$$

Therefore, $a_\mu(T,T) = \int_\Omega \kappa_\mu \nabla T \cdot \nabla T \geq \min(\kappa_\mu) ||T||^2_{H_0^1(\Omega)}$; the Temam term ensures coercivity for this bilinear form defined on M_h^2 ($T \in \mathbb{P}_1$, therefore $T^2 \notin \mathbb{P}_1$ and $\int_\Omega \mathrm{div}(\boldsymbol{u}_{\mathrm{NS}}) T^2 \neq 0$). This lower bound is quite pessimistic, especially when κ_μ varies a lot in the computational domain. For simplicity, we take this one in the present work. Methods exist to determine sharper constant for the discrete problem, based on solving eigenvalue problems (see [17] remark 16 p.114 for symmetric problems, see [56] for a presentation of the successive constraint method used in nonsymmetric problems).

We introduce the adjoint problem of (8.14): Find $\Psi_\mu^{*\mathrm{FE}} \in M_h$ such that $\forall v \in M_h$

$$a_\mu\left(v, \Psi_\mu^{*\mathrm{FE}}\right) = -l(v). \tag{8.16}$$

We have to construct a reduced basis for this problem as well: define $\Psi_\mu^{*\mathrm{RB}}$ the current RB approximation of (8.16) (verifying $a_\mu\left(v, \Psi_\mu^{*\mathrm{RB}}\right) = -l(v),\ \forall v \in M_{\mathrm{RB}}^*$, the space spanned by the current reduced basis of (8.16)).

Consider the residual for the direct and the adjoint problems: for all $w, v \in M_h$,

$$\begin{cases} g_\mu(w,v) = a_\mu(w,v) - l(v), \\ g_\mu^*(w,v) = a_\mu(w,v) + l(w). \end{cases} \tag{8.17}$$

Make use of the version of Riesz-Fréchet representation theorem applied to continuous bilinear forms: there exists a unique application $G_\mu : M_h \longmapsto M_h$ such that $\forall w, v \in M_h$, $g_\mu(w,v) = (G_\mu w, v)_{H_0^1}$. Define G_μ^* in the same way: $\forall w, v \in M_h$, $g_\mu^*(w,v) = \left(w, G_\mu^* v\right)_{H_0^1}$ (note that $G_\mu T_\mu^{\mathrm{FE}} = 0$ and $G_\mu^* \Psi_\mu^{*\mathrm{FE}} = 0$).

The output RB approximation of s_μ^{FE} is computed as $s_\mu^{\mathrm{RB},*\mathrm{RB}} = l\left(T_\mu^{\mathrm{RB}}\right) + g_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right)$ (see [17] section 4-I-B-c). This specific output has been chosen to ensure the posteriori estimate given in the proposition 8.1. Note that the FE corresponding quantity is $l\left(T_\mu^{\mathrm{FE}}\right) + g_\mu\left(T_\mu^{\mathrm{FE}}, \Psi_\mu^{*\mathrm{FE}}\right) = l\left(T_\mu^{\mathrm{FE}}\right) = s_\mu^{\mathrm{FE}}$, so $g_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right)$ only contains approximation errors introduced by the two reduced basis.

Make use of proposition 23 p.115 of Boyaval's PhD thesis [17] (see also [18] eq.22):

Proposition 8.1

$$|s_\mu^{\mathrm{RB},*\mathrm{RB}} - s_\mu^{\mathrm{FE}}| \leq \Delta_\mu := \frac{||G_\mu T_\mu^{\mathrm{RB}}||_{H_0^1(\Omega)} ||G_\mu^* \Psi_\mu^{*\mathrm{RB}}||_{H_0^1(\Omega)}}{\alpha_{\mathrm{LB},\mu}},$$

where $\alpha_{\mathrm{LB},\mu}$ is a computable lower bound for the coercivity constant α_μ of $a_\mu(.,.)$ (recall that a bound for the continuous case is $\alpha_{\mathrm{LB},\mu} = \min(\kappa_\mu)$).

Proof. Using elements of proof from Boyaval's PhD thesis (propositions 18 and 23):

$$|s_\mu^{\mathrm{RB},*\mathrm{RB}} - s_\mu^{\mathrm{FE}}| = |l\left(T_\mu^{\mathrm{FE}} - T_\mu^{\mathrm{RB}}\right) - g_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right)|. \tag{8.18}$$

Using (8.16), $T_\mu^{\mathrm{FE}} - T_\mu^{\mathrm{RB}} \in M_h \Longrightarrow l\left(T_\mu^{\mathrm{FE}} - T_\mu^{\mathrm{RB}}\right) = -a_\mu\left(T_\mu^{\mathrm{FE}} - T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{FE}}\right)$.
Using (8.14), $\Psi_\mu^{*\mathrm{FE}} \in M_h \Longrightarrow a_\mu\left(T_\mu^{\mathrm{EF}}, \Psi_\mu^{*\mathrm{FE}}\right) = l\left(\Psi_\mu^{*\mathrm{FE}}\right)$.
Then, $|s_\mu^{\mathrm{RB},*\mathrm{RB}} - s_\mu^{\mathrm{FE}}| = |a_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{FE}}\right) - l\left(\Psi_\mu^{*\mathrm{FE}}\right) - g_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right)| = |g_\mu\left(T_\mu^{\mathrm{RB}}, \Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}\right)|$, by definition (8.17).
Using Cauchy-Schwarz inequality, we can write:

$$\left|s_\mu^{\mathrm{RB},*\mathrm{RB}} - s_\mu^{\mathrm{FE}}\right| \leq \left\|G_\mu T_\mu^{\mathrm{RB}}\right\|_{H_0^1} \left\|\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}\right\|_{H_0^1}. \tag{8.19}$$

Then by coercivity, (8.16), (8.17) and the Cauchy-Schwarz inequality:

$$\begin{aligned}
\left\|\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}\right\|_{H_0^1}^2 &\leq \frac{|a_\mu\left(\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}, \Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}\right)|}{\alpha_{\mathrm{LB},\mu}}, \\
&= \frac{|a_\mu\left(\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right) + l\left(\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}\right)|}{\alpha_{\mathrm{LB},\mu}}, \\
&= \frac{|g_\mu^*\left(\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}, \Psi_\mu^{*\mathrm{RB}}\right)|}{\alpha_{\mathrm{LB},\mu}}, \\
&\leq \frac{||G_\mu^* \Psi_\mu^{*\mathrm{RB}}||_{H_0^1}}{\alpha_{\mathrm{LB},\mu}} ||\Psi_\mu^{*\mathrm{FE}} - \Psi_\mu^{*\mathrm{RB}}||_{H_0^1}.
\end{aligned}$$

Pluging this in (8.19), we get the inequality (8.1). ♢

Remark 8.2 *Note that the error estimate Δ_μ does not require the evaluation of the FE solutions, enabling fast computation.*

The inexpensive a posteriori error estimate (8.1) is useful to check rapidly if an online call approximates well the FE reference and to construct iteratively the reduced basis using a greedy algorithm (see section 5.2.5). This algorithm was proposed by Patera, Prud'homme, Rovas and Veroy in [85].

8.3.3 Computation aspects and construction of the basis with a greedy algorithm

Offline stage: precomputation and greedy algorithm

The certified RB method consists in iteratively construct a basis with solutions of the considered problem, computed at particular values of the parameters μ.

Ideally, we would like to choose a tolerance ϵ for the error $|s_\mu^{\mathrm{RB},*\mathrm{RB}} - s_\mu^{FE}|$ and make use of the error estimate Δ_μ to construct a low dimensional basis where the error is guaranteed to be smaller than ϵ for all μ in the parameter space $\mathcal{D}$. Unfortunately, such a result does not exist. One can refer to recent work by Buffa, Maday, Patera, Prud'homme and Turinici [24] for theoretical results on greedy algorithms convergence performances.

In the present setting, we have an affine dependence of the operator on the parameters (thanks to equations (8.2)): for all $w, v \in M_h$,

$$a_\mu(w,v) = a_0(w,v) + \sum_{i=1}^{n} \mu_i a_i(w,v) = \sum_{i=0}^{n} \mu_i a_i(w,v).$$

In the same fashion, the residuals are written: for all $w, v \in M_h$,

$$\begin{cases} g_\mu(w,v) = -l(v) + \sum_{i=0}^{n} \mu_i a_i(w,v), \\ g_\mu^*(w,v) = l(w) + \sum_{i=0}^{n} \mu_i a_i(w,v), \end{cases}$$

where the a_i do not depend on μ.

Define now G_i, G_i^*: $\forall w, v \in M_h$, $\forall i = 1, ...n$, $(G_i w, v)_{H_0^1} = a_i(w,v)$ and $(w, G_i^* v)_{H_0^1} = a_i^*(w,v)$. Take g_0 and g_0^*, solutions of $(g_0, v)_{H_0^1} = -l(v)$ and $(w, g_0^*)_{H_0^1} = l(w)$ (note that in this case, $g_0 = -g_0^*$).

In practice, we only need to compute $G_i w$ and $G_i^* v$ for given vectors $w, v \in M_h$. This is done by solving variational problems. Consider for instance $G_i w \in M_h$: Let $w \in M_h$. Find $u \in M_h$ such that $\forall v \in M_h$:

$$\int_\Omega \nabla u \cdot \nabla v = a_i(w,v),$$

w being fixed, $v \longmapsto a_i(w,v)$ is a continuous linear form, and $G_i w := u$.

The other quantities defined above are computed in the same way. This will speed up the online stage since lots of quantities will be precalculated during the offline stage, letting to the online stage simple low dimension algrebra operations (see section 8.3.3).

Take a finite set $\mathcal{D}_{\mathrm{trial}} \subset \mathcal{D}$. The greedy algorithm is given in Algorithm 8.

Note that we did not explain how to chose the finite set $\mathcal{D}_{\mathrm{trial}}$. In our case, we simply discretized each direction of $\mathcal{D}$ with a constant step: see Figure 8.10. The efficient evaluation of Δ_μ is explained in the next section 8.3.3.

Algorithm 8 Greedy algorithm for the offline stage of the reduced basis method

1. Choose $\mu_1 \in \mathcal{D}_{\text{trial}}$ randomly and remove it from $\mathcal{D}_{\text{trial}}$
2. Compute $T_{\mu_1}^{\text{FE}}$ and $\Psi_{\mu_1}^{*\text{FE}}$, then set $V_1^{\text{RB}} = \text{Span}\left(T_{\mu_1}^{\text{FE}}\right)$ and $V_1^{\text{RB}} = \text{Span}\left(\Psi_{\mu_1}^{*\text{FE}}\right)$
3. Compute $G_i T_{\mu_1}^{\text{FE}}$ and $G_i^* \Psi_{\mu_1}^{*\text{FE}}$, $\forall i \in [0, \dots n]$, as well as g_0 and g_0^*
4. Store $a_i\left(T_{\mu_1}^{\text{FE}}, T_{\mu_1}^{\text{FE}}\right)$, $\left(G_i T_{\mu_1}^{\text{FE}}, G_i T_{\mu_1}^{\text{FE}}\right)_{H_0^1}$, $\left(G_i^* \Psi_{\mu_1}^{*\text{FE}}, G_i^* \Psi_{\mu_1}^{*\text{FE}}\right)_{H_0^1}$, $\left(g_0, G_i T_{\mu_1}^{\text{FE}}\right)_{H_0^1}$ and $\left(g_0^*, G_i^* \Psi_{\mu_1}^{*\text{FE}}\right)_{H_0^1}$ $\forall i \in [0, \dots n]$, as well as $(g_0, g_0)_{H_0^1}$ and $(g_0^*, g_0^*)_{H_0^1}$
5. **while** $\max\{\Delta_{\mu_i}, \mu_i \in \mathcal{D}_{\text{trial}}\} \geq \epsilon$ **do**
6. Solve $\mu_N = \text{argmax}\{\Delta_{\mu_i},\ \mu_i \in \mathcal{D}_{\text{trial}}\}$ and remove it from $\mathcal{D}_{\text{trial}}$
7. Compute $T_{\mu_N}^{\text{FE}}$ and $\Psi_{\mu_N}^{*\text{FE}}$, then set $V_N^{\text{RB}} = \text{Span}\left(T_{\mu_1}^{\text{FE}}, \dots T_{\mu_N}^{\text{FE}}\right)$ and $V_N^{\text{RB}} = \text{Span}\left(\Psi_{\mu_1}^{*\text{FE}}, \dots \Psi_{\mu_N}^{*\text{FE}}\right)$
8. Compute $G_i T_{\mu_N}^{\text{FE}}$ and $G_i^* \Psi_{\mu_N}^{*\text{FE}}$, $\forall 0 \in [1, \dots n]$
9. Store $a_i\left(T_{\mu_k}^{\text{FE}}, T_{\mu_N}^{\text{FE}}\right)$, $\left(G_i T_{\mu_k}^{\text{FE}}, G_j T_{\mu_N}^{\text{FE}}\right)_{H_0^1}$ and $\left(G_i^* \Psi_{\mu_k}^{*\text{FE}}, G_j^* \Psi_{\mu_N}^{*\text{FE}}\right)_{H_0^1}$ $\forall i, j \in [0, \dots n]$, $\forall k \in [1, \dots N]$, as well as $\left(g_0, G_i T_{\mu_N}^{\text{FE}}\right)_{H_0^1}$ and $\left(g_0^*, G_i^* \Psi_{\mu_N}^{*\text{FE}}\right)_{H_0^1}$ $\forall i \in [0, \dots n]$

   Remark: since the problem is not symmetric, we also have to store $a_i\left(T_{\mu_N}^{\text{FE}}, T_{\mu_k}^{\text{FE}}\right)$, $\left(G_i T_{\mu_N}^{\text{FE}}, G_j T_{\mu_k}^{\text{FE}}\right)_{H_0^1}$ and $\left(G_i^* \Psi_{\mu_N}^{*\text{FE}}, G_j^* \Psi_{\mu_k}^{*\text{FE}}\right)_{H_0^1}$ $\forall i, j \in [0, \dots n], \forall k \in [1, \dots N-1]$
10. $N = N + 1$
11. **end while**

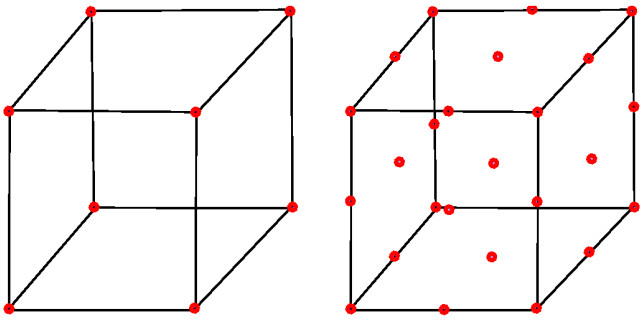

Fig. 8.10. Determination of $\mathcal{D}_{\text{trial}}$ in the case of tridimensional parameter space. Left: two values per direction, right: three values per direction

Online stage

Consider that a basis $(b)_i$ of size d has been constructed (and a basis $(b^*)_i$ for the adjoint problem) , and that we want to efficiently compute the temperature using the RB for a given parameter.

An online call consists in solving the following Galerkin problem: Find $T_\mu^{\text{RB}} \in M^{\text{RB}} := \text{Vect}\,(b_1, b_2, \dots b_d)$ such that $\forall v \in M^{\text{RB}}$

$$a_\mu\left(T_\mu^{\text{RB}}, v\right) = l\,(v)\,.$$

Write $T^{\text{RB}} = \sum_{i=1}^d \alpha_i b_i$, the Galerkin problem becomes: Find $(\alpha_1, \alpha_2, \dots \alpha_d)$ such that $\forall j = 1, \dots d$

$$\sum_{i=1}^{d} \alpha_i a_\mu\,(b_i, b_j) = l\,(b_j)\,.$$

We have $A_\mu := \sum_{k=1}^d \mu_k\,(a_k\,(b_i, b_j))_{ij}$,where all $a_k\,(b_i, b_j)$ have been computed and stored during the offline stage.

Defining the vectors $(\alpha)_j = \alpha_j$, and $(F)_j = l(b_j)$, one then just has to solve the d-dimensional linear system

$$A_\mu \alpha = F,$$

and then compute the RB solution $u_\mu^{\mathrm{RB}} = \sum_{i=1}^{d} \alpha_i b_i$. Note that the α_i contain the dependence of the solution on μ.

In practice, when the size of the reduced basis increases, the matrix A_μ may be close to singular. One can improve this by orthonormalizing the basis using modified or simple Gram-Schmidt (see [17], p.109).

Then, an online call should also contain an evaluation of the a posteriori error estimate Δ_μ, and if it is larger than ϵ, one can compute the FE solution and enrich the basis with this function. This way, the error estimate is guaranteed, but we may need to compute the expensive solution sometimes.

We also use precomputed quantities to get an evaluation of Δ_μ in a complexity independent of the FE problem size.

$$\begin{aligned}
||G_\mu T_\mu^{\mathrm{RB}}||_{H_0^1}^2 &= \left(G_\mu T_\mu^{\mathrm{RB}}, G_\mu T_\mu^{\mathrm{RB}}\right)_{H_0^1} \\
&= \left(g_0 + \sum_i^n \mu_i G_i T_\mu^{\mathrm{RB}}, g_0 + \sum_j^n \mu_i G_j T_\mu^{\mathrm{RB}}\right)_{H_0^1} \\
&= (g_0, g_0)_{H_0^1} + 2\sum_i^n \sum_k^d \mu_i \alpha_k \left(g_0, G_i b_k\right)_{H_0^1} + \sum_{i,j=1}^n \sum_{k,l=1}^d \mu_i \mu_j \alpha_k \alpha_l \left(G_i b_k, G_j b_l\right)_{H_0^1},
\end{aligned}$$

where the affine dependence has been used in the second line and the RB decomposition in the third line. All the scalar products involved have been precomputed and stored during the offline stage. Therefore

$$\Delta_\mu = \frac{||G_\mu T_\mu^{\mathrm{RB}}||_{H_0^1(\Omega)} ||G_\mu^* \Psi_\mu^{*\mathrm{RB}}||_{H_0^1(\Omega)}}{\alpha_{\mathrm{LB},\mu}}$$

is of complexity $O(n^2 d^2)$, which corresponds to a fast computation compared to the FE problem. The faster the estimator is computed, the larger $\mathcal{D}_{\mathrm{trial}}$ we can explore in the greedy algorithm for a given computation time.

Numerical results

The greedy algorithm has been carried out to compute 10 basis functions, with a parameter set of $20,000$ points. The finite elements problem has 9012 degrees of freedom.

We took as parameters: κ_{IC}, κ_{board}, κ_{air} the integrated circuit, board and air heat conductivity and $c_{p_{\mathrm{air}}}$ the air thermal capacity, varying in the following intervals (international system of units):

	κ_{IC}	κ_{board}	κ_{air}	$c_{p_{\text{air}}}$
min	0.5	0.06	0.028	1080
max	5	0.6	0.032	1120

We first check that the growth of the basis reduces projection errors (see Figure 8.11):

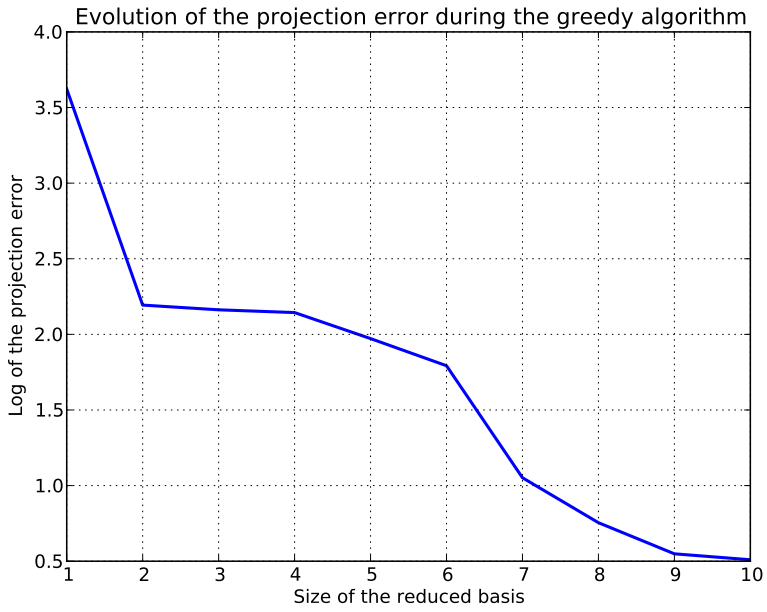

Fig. 8.11. Evolution of $\left\|T_{\hat{\mu}}^{\text{FE}} - T_{\hat{\mu}}^{\text{RB}}\right\|_{H_0^1}$ for $\hat{\mu} \in \mathcal{D}$ taken randomly with the size of the reduced basis

Consider now the evolution of the maximum error estimate Δ_μ for $\mu \in \mathcal{D}_{\text{trial}}$ and the corresponding value of the error $\left|s_\mu^{\text{RB},*\text{RB}} - s_\mu^{\text{FE}}\right|$ (see Figure 8.12). The error estimate and the error are not strictly decreasing. Maybe this is due to the fact that the RB is certified for a scalar quantity different from the projection error represented in Figure 8.11. Recall also that we did not compute a satisfying lower bound for the coercivity constant of the bilinear form a_μ. The error estimate for a random parameter is not strictly decreasing with the size of the basis, but globally tends to zero. Remark that an error of 8×10^{-5} correspond to a mean temperature error of 1 K over the integrated circuits.

Once the basis constructed, we tried an online call taking a random parameter in $\mathcal{D}$ (see Figure 8.13):

– A posteriori error estimate : 7.24×10^{-6}
– Error : 1.20×10^{-6}

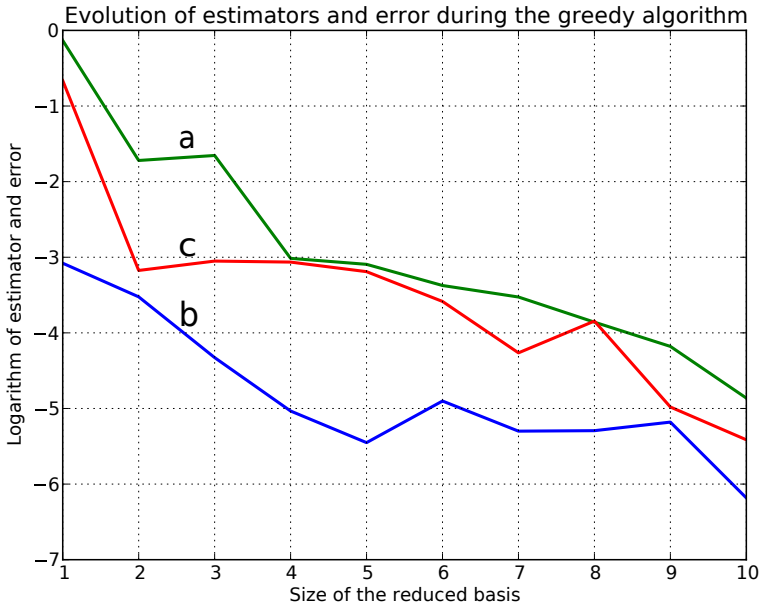

Fig. 8.12. Evolution of the maximum error estimate Δ_μ for $\mu \in \mathcal{D}_{\text{trial}}$ (a), the corresponding value of the error $|s_\mu^{\text{RB},*\text{RB}} - s_\mu^{\text{FE}}|$ (b) and the error estimate $\Delta_{\hat{\mu}}$ for $\hat{\mu} \in \mathcal{D}$ taken randomly (c)

– Integrated circuits temperature RB: 40.99 K; direct computation: 40.96 K
relative difference = 0.074 %

Duration of the different stages:
– offline with greedy: 8min37s
– one finite element resolution: 0.28s
– one online call: 1.4×10^{-6}s

Conclusion

We developped different models for solving the velocity and the temperature of the air under different conditions. The full Boussinesq model yields satisfactory results in the cases we considered, whereas decoupled incompressible Navier-Stokes / heat was satisfactory for the electronic component case (forced convection in a pipe).

We then tested a certified RB method for the heat problem with non homogeneous convection. Our code makes use of an a posteriori error estimate to build a basis iteratively using a greedy algorithm. The affine parametric assumption allows to precompute many terms, and online calls are reduced to add and inverse low dimensional matrices. This enables us to drastically speed up the resolution in the electric component, with a physically satisfying model, and controlled approximation errors.

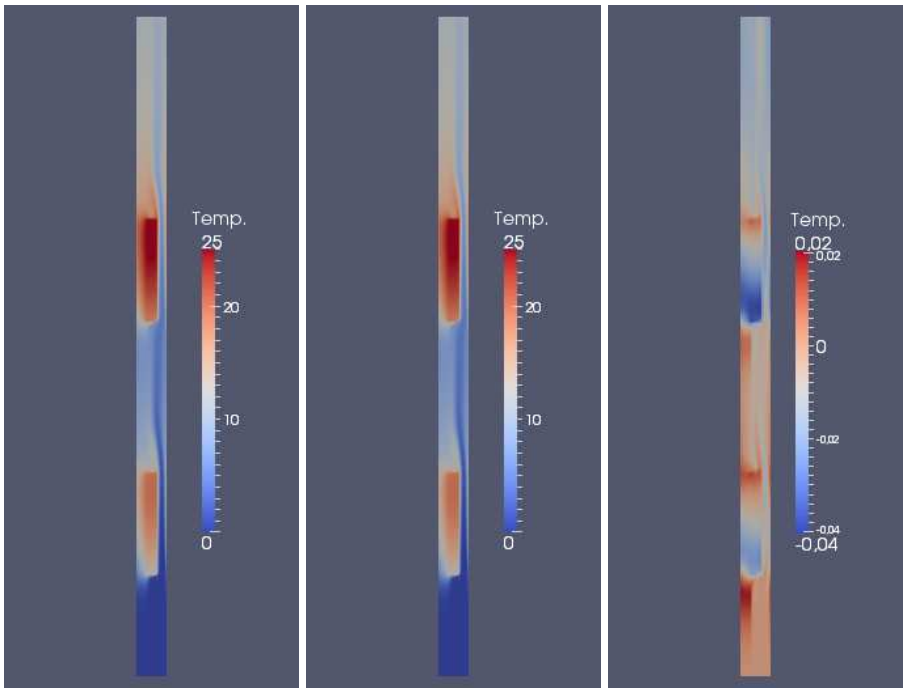

Fig. 8.13. Temperature maps for the electronic component case. Left: RB online call $\boldsymbol{u}_{\mathrm{RB}}$, center: FE calculation $\boldsymbol{u}_{\mathrm{FE}}$, right: difference $\boldsymbol{u}_{\mathrm{FE}} - \boldsymbol{u}_{\mathrm{RB}}$

Annexe

A

A well conditioned kernel interpolation

A.1 Kernel interpolation

Consider a unknown function $\mu \to f(\mu)$, $\mu \in \mathcal{P}$. We want to derive an interpolation formula for f: given interpolation points $\mu_i \in \mathcal{P}$, $1 \leq i \leq d$ where we dispose of the value of $f(\mu_i)$, how can we derive an approximation $\hat{f}(\mu)$ of $f(\mu)$. Consider a given kernel $K(\mu, \mu')$, the Kernel Ridge Regression with kernel K and parameter $\lambda > 0$ is defined by

$$\hat{f}_0(\mu) := \sum_{i=1}^{d} b_i K(\mu_i, \mu), \tag{A.1}$$

where

$$(\hat{K} + \lambda I)b = y, \tag{A.2}$$

with $y_i = f(\mu_i)$, $1 \leq i \leq d$ and $\hat{K}_{ij} = K(\mu_i, \mu_j)$, $1 \leq i, j \leq d$. In (A.1) is used a mapping between the set $\mathcal{P}$ and an reproducing kernel Hilbert space, determined by the kernel K, without having the compute explicitly (we only compute evaluation of the kernel). For these reasons, (A.1) is often referred as kernel trick).

The following proposition is readily seen.

Proposition A.1 (Interpolation) *For all* $1 \leq i \leq d$,

$$\lambda = 0 \Longrightarrow \hat{f}_0(\mu_i) = f(\mu_i). \tag{A.3}$$

The parameter λ allows the linear system (A.2) to be numerically solvable, since in practice, the matrix $\hat{K}$ can have a very large condition number, especially when some learning points are very close to each other. However, with $\lambda \neq 0$, Proposition A.1 is lost in the general case. In what follows, we use the Empirical Interpolation Method (EIM) to improve the numerical behavior of the linear system (A.2) when $\lambda = 0$.

A.2 Empirical interpolation method

Consider a discrete subset of $\mathcal{P}$ denoted $\mathcal{P}_{\text{trial}}$. Suppose that the following EIM approximation of the kernel has been computed:

$$K(\mu, \mu') \approx \sum_{J=1}^{D} \lambda_J(\mu) K(\mu_J, \mu'), \tag{A.4}$$

where $\lambda_J(\mu)$ can be computed for all $\mu \in \mathcal{P}_{\text{trial}}$ using the online stage of the EIM, and $\mu_J \in \mathcal{P}_{\text{trial}}$, $1 \leq J \leq D$ is a set of interpolation points. We suppose in this section that $d = \#\mathcal{P}_{\text{trial}}$, so that, for all $\mu \in \mathcal{P}_{\text{trial}}$,

$$\sum_{i=1}^{d} b_i K(\mu_i, \mu) = f(\mu). \tag{A.5}$$

Inject (A.4) into (A.1) to define:

$$\begin{aligned} \hat{f}_1(\mu) &:= \sum_{i=1}^{d} b_i \sum_{J=1}^{D} \lambda_J(\mu_i) K(\mu_J, \mu) \\ &= \sum_{J=1}^{D} K(\mu_J, \mu) \sum_{i=1}^{d} b_i \lambda_J(\mu_i), \end{aligned} \tag{A.6}$$

where $t_J := \sum_{i=1}^{d} b_i \lambda_J(\mu_i)$ is independent of μ. Then, using the symmetry of K,

$$\begin{aligned} \hat{f}_1(\mu) &= \sum_{J=1}^{D} K(\mu, \mu_J) t_J \\ &\approx \sum_{J=1}^{D} \sum_{J'=1}^{D} \lambda_{J'}(\mu) K(\mu_{J'}, \mu_J) t_J \\ &= \sum_{J'=1}^{D} \lambda_{J'}(\mu) \sum_{J=1}^{D} K(\mu_J, \mu_{J'}) t_J \\ &\approx \sum_{J'=1}^{D} \lambda_{J'}(\mu) \hat{f}_1(\mu_{J'}). \end{aligned} \tag{A.7}$$

Proposition A.2 *For all* $1 \leq J' \leq D$, $\hat{f}_1(\mu_{J'}) = f(\mu_{J'})$.

Proof. Using the interpolation property of the EIM, there holds, for all $\mu \in \mathcal{P}_{\text{trial}}$, for all $1 \leq J' \leq D$,

$$\sum_{J=1}^{D} \lambda_J(\mu) K(\mu_J, \mu_{J'}) = K(\mu, \mu_{J'}). \tag{A.8}$$

Therefore,

$$\begin{aligned} \hat{f}_1(\mu_{J'}) &= \sum_{i=1}^{d} b_i \sum_{J=1}^{D} \lambda_J(\mu_i) K(\mu_J, \mu_{J'}) \\ &= \sum_{i=1}^{d} b_i K(\mu_i, \mu_{J'}). \end{aligned} \tag{A.9}$$

Then, since $\mu_{J'} \in \mathcal{P}_{\text{trial}}$ and $d = \#\mathcal{P}_{\text{trial}}$, the assertion follows from (A.5). ♢

This leads to the new interpolation formula:

$$\hat{f}_2(\mu) := \sum_{J'=1}^{D} \lambda_{J'}(\mu) f(\mu_{J'}). \tag{A.10}$$

In this case, the interpolation is computed by a linear combination of values of f. Providing that the values $\lambda_{J'}(\mu)$ are of order 1, the formula (A.10) is then mush less sensitive to round-off errors than the formula (A.1).

Remark A.3 *We are not constrained by the use of the interpolation points selected by EIM. The same method can be derived on any set of interpolation points μ_J.*

A.3 Simple numerical illustration

We apply the two previously defined interpolation method to the function $\mu \to \sin(10\mu)$, $0 \leq \mu \leq 1$. $\mathcal{P}_{\text{trial}}$ is obtained by a uniform discretization of $(0, 1)$ using 501 points, and the same set of interpolation points is used for both methods, namely the points selected by EIM. The gaussian kernel $K(\mu, \mu') := \exp\left(\frac{(\mu-\mu')^2}{2}\right)$ has been chosen. The results are in Figures A.1-A.2: the numerical pollution due to the ill-conditioning of the matrix $\hat{K}$ increases with the number of interpolation points for the first method. For the second method, no such pollution is observed, and the error at the interpolation points is always zero in all the simulations.

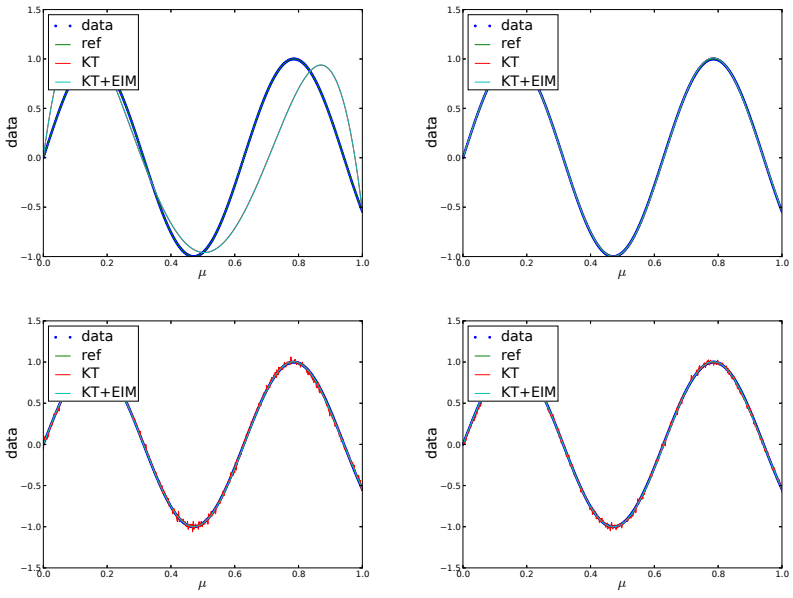

Fig. A.1. Comparison of the interpolation formulae for 5, 10, 11 and 12 interpolation points. KT refers to the kernel trick (A.1), and KT+EIM refers to the new formula (A.10).

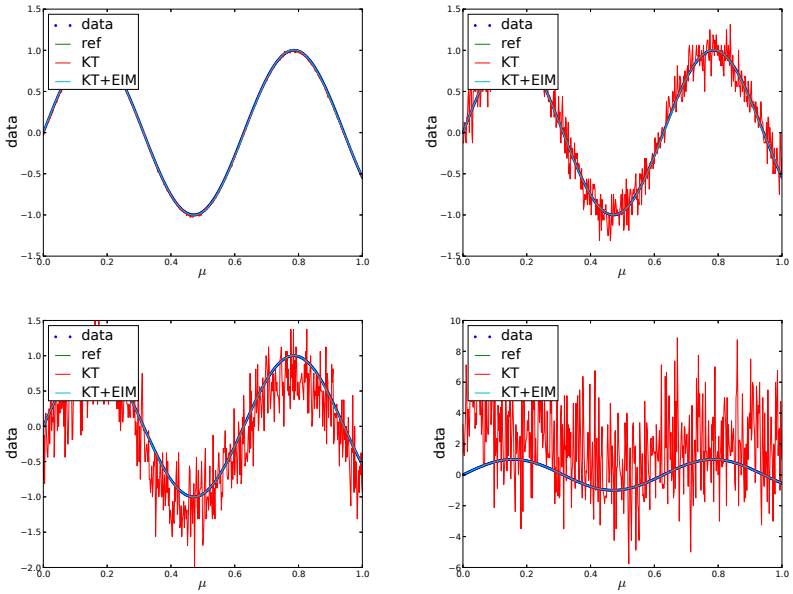

Fig. A.2. Comparison of the interpolation formulae for 13, 14, 15, and 20 interpolation points.

B

Étude d'un modèle d'incertitude non paramétrique

Dans cette annexe, nous nous intéressons à la relation entre le modèle mécanique des plaques constitutives du fuselage d'un avion et la puissance acoustique transmise de l'extérieur de l'avion vers les passagers. Nous considérons la modélisation des plaques mal connue ou aléatoire. Dans cette situation, une première possibilité est de supposer que les coefficients classiques de modélisation (rigidité, vitesse des ondes) sont des variables aléatoires dont nous imposons une loi, basée sur l'expérience ou l'expertise. Une deuxième possibilité, que nous développons dans cette annexe, consiste à modéliser le comportement aléatoire des plaques par une approche probabiliste non paramétrique où les aléas sont quantifiés par un scalaire δ, qui représente une variance de l'opérateur mécanique complet. Enfin, nous considérons une fonction coût dépendant de δ et nous étudions l'optimisation de cette fonction coût sous une contrainte en probabilité de dépassement de seuil de la norme en énergie de la solution du système mécanique (qui correspond à l'énergie acoustique transmise à travers une plaque du fuselage). Dans le cas de plusieurs plaques reliées à leurs bords, on peut s'intéresser au problème d'allocation des incertitudes entre les plaques, sous la même contrainte en probabilité exprimée sur l'énergie acoustique totale transmise à travers le réseau de plaques.

Pour simplifier la présentation, nous considérons une corde vibrante tendue soumise à une force linéique. Ce problème peut être remplacé sans difficulté par le problème d'une plaque en vibration, car les mêmes types de matrices interviennent.

Dans la section B.1, nous introduisons le problème modèle et les notations. Dans la section B.2, nous détaillons la modélisation probabiliste de la matrice du problème obtenu, basée sur les travaux de Soize [96]. Enfin, le problème d'optimisation sous contrainte en probabilité est présenté dans la section B.3 et une illustration numérique est fournie en section B.4. Dans la section B.5, nous donnons les premiers éléments pour étendre le raisonnement au cas de plusieurs cordes reliées à leurs extrémités.

B.1 Problème modèle

Dans le domaine fréquentiel, l'équation de la corde vibrante s'écrit

$$-y'' - \frac{\omega^2}{a^2} y = f, \qquad \text{dans }]0,1[, \tag{B.1}$$

où l'inconnue y est le déplacement vertical, a la vitesse locale de propagation des ondes et ω la fréquence. Nous prenons des conditions aux limites de Dirichlet homogènes en 0 et en 1, et considérons pour simplifier que le terme source vaut $f = 1$. En écrivant une formulation variationnelle sur $H_0^1(]0,1[)$, puis en prenant les éléments finis P1 en dimension 1, constitués des fonctions continues affines par morceaux sur $]0,1[$, nous obtenons

$$\left(K - \omega^2 M\right) Y = F, \tag{B.2}$$

où $]0,1[$ est discrétisé en segments de même taille $h = \frac{1}{N+1}$, avec N la taille des matrices obtenues. On a

$$M = h \begin{pmatrix} \frac{2}{3} & \frac{1}{6} & & & \\ \frac{1}{6} & . & . & & \\ & . & . & . & \\ & & . & . & \frac{1}{6} \\ & & & \frac{1}{6} & \frac{2}{3} \end{pmatrix}, \; K = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & . & . & & \\ & . & . & . & \\ & & . & . & -1 \\ & & & -1 & 2 \end{pmatrix} \text{ et } F = h \begin{pmatrix} 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}. \tag{B.3}$$

Les composantes du vecteur Y sont les composantes de la fonction approchant y dans la base des éléments finis considérés.

La matrice M étant symétrique définie positive, on peut écrire

$$\left(M^{-\frac{1}{2}} K M^{-\frac{1}{2}} - \omega^2\right) M^{\frac{1}{2}} Y = M^{-\frac{1}{2}} F. \tag{B.4}$$

Soit $H = M^{-\frac{1}{2}} K M^{-\frac{1}{2}}$, que l'on diagonalise. La matrice H étant symétrique définie positive, il existe une matrice orthogonale B et une matrice diagonale D telles que $B^t H B = D$. En posant $Q = B^t M^{\frac{1}{2}} Y$, (B.2) devient $HBQ - \omega^2 BQ = M^{-\frac{1}{2}} F$, et en posant $A = -\omega^2 I + D$, il vient

$$AQ = B^t M^{-\frac{1}{2}} F. \tag{B.5}$$

L'introduction de la nouvelle inconnue Q peut s'apparenter à un changement de base. Nous considérons que les composantes de Q sont les coordonnées de la solution de (B.4) dans cette base, appelée base modale. Dans cette base, la matrice de masse vaut l'identité, et la matrice de rigidité est diagonale, et contient les valeurs propres de $M^{-\frac{1}{2}} K M^{-\frac{1}{2}}$. On peut aussi réduire le nombre de degrés de liberté en ne prenant que les modes de basse énergie, c'est-à-dire en réduisant D à une matrice diagonale ne contenant que les n plus petites valeurs propres. Dans ce cas, la matrice de passage B est rectangulaire, et on ne résout qu'une approximation de (B.2).

B.2 Modélisation probabiliste

Nous nous appuyons sur les travaux de Soize [96] pour modéliser la méconnaissance ou les aléas sur la corde en considérant que la matrice A du problème (B.5) est une matrice aléatoire

suivant une loi de Wishart. En invoquant un principe de maximisation d'entropie en se fixant des contraintes a priori (matrices simulées symétriques définies positives, moyenne et dispersion imposées), Soize montre que le problème se réduit (à une constante près) à une loi de Wishart dont les paramètres sont une matrice symétrique définie positive (le paramètre d'échelle), que l'on peut relier à la matrice moyenne, et un entier (le nombre de degrés de liberté), que l'on peut relier à la dispersion de la loi [96, Section 3]. Dans cette section, nous détaillons d'abord comment les paramètres de la loi de Wishart sont reliés aux propriétés de moyenne et dispersion des matrices aléatoires suivant cette loi, puis nous étudions la loi de la norme d'énergie de la solution Q de (B.5) lorsque A suit une loi de Wishart.

B.2.1 La loi de Wishart

Nous supposons que $A \sim W_n\left(\frac{1}{m}A_{\text{moy}}, m\right)$, c'est-à-dire que A suit une loi de Wishart de paramètre d'échelle $\frac{1}{m}A_{\text{moy}}$ et de degrés de liberté $m > n-1$, où n est la taille de la matrice A_{moy}, voir [5] pour des détails sur cette loi. Dans la suite, nous notons l'égalité en loi par $\sim$. Considérons la décomposition de Cholesky $A_{\text{moy}} = L_A L_A^t$, et prenons m vecteurs $\{U_i\}_{1\leq i\leq m} \in \mathbb{R}^n$, dont les coefficients sont des variables aléatoires gaussiennes centrées réduites indépendantes. Les vecteurs $L_A U_i$ suivent la loi normale n-dimensionnelle centrée de matrice de covariance A_{moy}, et $\frac{1}{m}\sum_{i=1}^m (L_A U_i)(L_A U_i)^t$ suit la loi $W_n\left(\frac{1}{m}A_{\text{moy}}, m\right)$. Cela fournit un moyen pratique de générer des réalisations de la loi de Wishart. Considérons $\hat{A} = L_A^{-1} A L_A^{-t}$. Il vient

$$\begin{aligned}\hat{A} &\sim \frac{1}{m} L_A^{-1} \sum_{i=1}^m L_A U_i \left(L_A U_i\right)^t L_A^{-t} \\ &\sim \frac{1}{m}\sum_{i=1}^m U_i U_i^t.\end{aligned} \tag{B.6}$$

La matrice $\sum_{i=1}^m U_i U_i^t$ a ses termes diagonaux qui suivent des lois du χ^2 à m degrés de liberté (d'espérance m) et des termes hors diagonale qui sont des sommes de produits de normales centrées réduites indépendantes (donc d'espérance nulle). Ainsi, $\mathbb{E}\left(\hat{A}\right) = I$ et

$$\mathbb{E}\left(A\right) = L_A \mathbb{E}\left(\hat{A}\right) L_A^t = A_{\text{moy}}, \tag{B.7}$$

ce qui relie simplement le paramètre d'échelle de la loi de Wishart à l'espérance des matrices aléatoires. Il paraît alors naturel de définir une dispersion δ pour la loi de A par la formule

$$\delta = \left(\frac{\mathbb{E}\left(\|A - A_{\text{moy}}\|_F^2\right)}{\mathbb{E}\left(\|A_{\text{moy}}\|_F^2\right)}\right)^{\frac{1}{2}}, \tag{B.8}$$

où $\|\cdot\|_F$ désigne la norme de Frobenius. On a

$$\begin{aligned}\mathbb{E}\left(\|A - A_{\text{moy}}\|_F^2\right) &= \sum_{i,j} \mathbb{E}\left(|A_{ij} - {A_{\text{moy}}}_{ij}|^2\right) \\ &= \sum_{i,j} \text{Var}\left(A_{ij}\right).\end{aligned} \tag{B.9}$$

En repassant par la matrice $\hat{A}$ et en utilisant la formule de la variance de la somme et du produit de variables aléatoires indépendantes, on montre que

$$\mathrm{Var}\left(A_{ij}\right) = \frac{1}{m}\left(\left(A_{\mathrm{moy}}\right)_{ii}\left(A_{\mathrm{moy}}\right)_{jj} + \left(A_{\mathrm{moy}}\right)_{ij}^2\right). \tag{B.10}$$

Ainsi,

$$\mathbb{E}\left(\|A - A_{\mathrm{moy}}\|_F^2\right) = \frac{1}{m}\sum_{i,j}\left({A_{\mathrm{moy}}}_{ij}^2 + {A_{\mathrm{moy}}}_{ii}{A_{\mathrm{moy}}}_{jj}\right) = \frac{1}{m}\left(trA_{\mathrm{moy}}^2 + \left(trA_{\mathrm{moy}}\right)^2\right) \tag{B.11}$$

Comme par ailleurs on a $\|A_{\mathrm{moy}}\|_F^2 = tr\left(A_{\mathrm{moy}}A_{\mathrm{moy}}^t\right) = tr\left(A_{\mathrm{moy}}^2\right)$, il vient

$$\delta^2 = \frac{\mathbb{E}\left(||A - A_{\mathrm{moy}}||_F^2\right)}{||A_{\mathrm{moy}}||_F^2} = \frac{1}{m}\left(1 + \frac{\left(trA_{\mathrm{moy}}\right)^2}{trA_{\mathrm{moy}}^2}\right), \tag{B.12}$$

ce qui permet de relier le degré de liberté de la loi de Wishart à la dispersion des matrices aléatoires par la formule

$$m = \frac{1}{\delta^2}\left(1 + \frac{\left(trA_{\mathrm{moy}}\right)^2}{trA_{\mathrm{moy}}^2}\right). \tag{B.13}$$

Dans la suite, par souci de concision, nous notons $\varsigma := \frac{\left(trA_{\mathrm{moy}}\right)^2}{trA_{\mathrm{moy}}^2}$.

B.2.2 La loi de la norme d'énergie de la solution

Nous supposons que la matrice A du problème (B.5) est une matrice aléatoire de loi $\frac{1}{m}W_n\left(A_{\mathrm{moy}}, m\right)$. La matrice $A_{\mathrm{moy}} = -\omega^2 I + D$ est connue : ω est la pulsation de la source et D est fixée à partir des connaissances a priori du modèle tel que décrit dans la section B.1. L'entier m nous permet de contrôler la dispersion de ces matrices aléatoires autour de A_{moy} par la formule (B.13). La solution Q est notre inconnue sur la base modale, et nous prenons pour variable d'intérêt la norme d'énergie de ce vecteur induite par A :

$$\begin{aligned} V_\delta &= \|Q\|_A^2 \\ &= Q^t A Q \\ &= (A^{-1}B^t M^{-\frac{1}{2}}F)^t A (A^{-1}B^t M^{-\frac{1}{2}}F) \\ &= (B^t M^{-\frac{1}{2}}F)^t A^{-1}(B^t M^{-\frac{1}{2}}F). \end{aligned} \tag{B.14}$$

Comme A_{moy} est diagonale à coefficients positifs, nous pouvons prendre $L_A = diag\left(\sqrt{{A_{\mathrm{moy}}}_{ii}}\right)$. Comme $\hat{A} := {L_A}^{-1} A {L_A}^{-t}$ et en posant $\hat{F} = {L_A}^{-1}B^t M^{-\frac{1}{2}}F$, la quantité d'intérêt s'écrit

$$V_\delta = \hat{F}^t \hat{A}^{-1}\hat{F}. \tag{B.15}$$

Lorsque la taille de la base réduite est $n = 1$, $\hat{A}$ est scalaire et $\hat{A} \sim \frac{c(m)}{m}$, où $c(m)$ est une variable aléatoire suivant une loi du χ^2 à m degrés de liberté, et $V_\delta \sim m||\hat{F}||_2^2 \frac{1}{c(m)}$, où $||\cdot||_2$ est la norme Euclidienne.

Proposition B.1 *En dimension $n > 1$,*

$$V_\delta \sim m||\hat{F}||_2^2 \frac{1}{c(m-n+1)}. \tag{B.16}$$

Preuve. Rappelons que $\hat{A} \sim W_n\left(\frac{1}{m}I, m\right)$. La loi de $\hat{A}$ est invariante par rotation: pour toute matrice de rotation R, $R^t\hat{A}R \sim W_n\left(\frac{1}{m}I, m\right)$. Donc $V_\delta \sim ||\hat{F}||_2^2 e_1^t\hat{A}^{-1}e_1 = ||\hat{F}||_2^2\hat{A}_{11}^{-1}$. Il nous reste à déterminer la loi du premier coefficient de l'inverse d'une matrice de Wishart. Soit $B = \hat{A}^{-1}$. Il est connu, sous le nom de décomposition de Bartlett, que $\hat{A} \sim Z^tZ$, où Z est une matrice triangulaire supérieure telle que tous les coefficients non nuls suivent des lois indépendantes: pour tout $1 \leq i \leq n$, Z_{ii}^2 suit une loi du χ^2 à $n-i+1$ degrés de liberté et tous les coefficients au dessus de la diagonale suivent des loi normales centrées réduites.

$$B_{11} = \left(\hat{A}^{-1}\right)_{11} = e_1^t\hat{A}^{-1}e_1 \sim e_1^tZ^{-1}Z^{-t}e_1 = \left\|Z^{-t}e_1\right\|^2. \tag{B.17}$$

Posons $(x_1, x_2..., x_n) = (Me_1, Me_2..., Me_n)$ et $d = ||x_1 - \sum_{j=2}^n \left(x_1, \frac{x_j}{\|x_j\|}\right)\frac{x_j}{\|x_j\|}||$, la distance entre la première colonne de M et l'espace vectoriel généré par les $n-1$ autres colonnes. Nous remarquons que $(M^{-t}e_1, Me_j) = e_1^tM^{-1}Me_j = \delta_{1j}$. Ainsi, $M^{-t}e_1$ est orthogonal à tous les vecteurs $x_j, j \geq 2$, voir figure B.1. Ainsi,

$$\begin{aligned}\sqrt{B_{11}}d &= \left(M^{-t}e_1, x_1 - \sum_{j=2}^n \left(x_1, \frac{x_j}{\|x_j\|}\right)\frac{x_j}{\|x_j\|}\right) \\ &= \left(M^{-t}e_1, Me_1\right) \\ &= 1.\end{aligned} \tag{B.18}$$

Donc $B_{11} \sim \frac{1}{d^2}.$, D'après le théorème de Cochran, d^2 est celle d'une χ^2 à $m-n+1$ degrés de liberté, ce qui permet de conclure. ♢

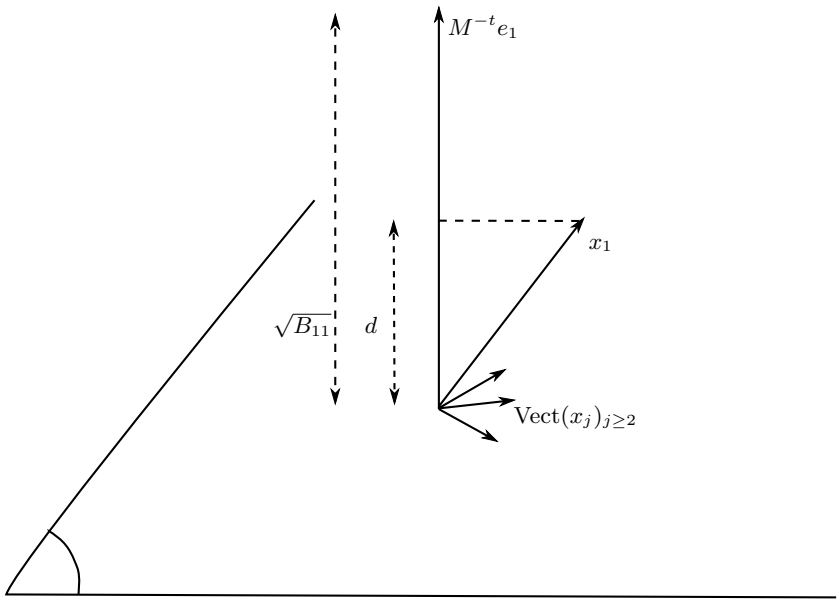

Fig. B.1. Représentation des divers vecteurs considérés

B.3 Le problème d'optimisation sous contraintes en probabilité dans le cas d'un seul objet

Considérons une fonction coût, notée $\mathcal{C}$, dépendant de la dispersion δ de la loi sur A, un seuil S et une probabilité de défaillance $\epsilon \in]0,1[$. Considérons le problème d'optimisation sous contrainte en probabilité suivant :

$$\delta_{\mathrm{opt}} = \mathrm{argmin}\left\{\mathcal{C}\left(\delta\right),\ \text{t.q. } \forall\ 0 < \delta \leq \delta_{\max}, \mathbb{P}\left(V_\delta > S\right) \leq \epsilon\right\}, \tag{B.19}$$

où $\delta_{\max}$ est une dispersion maximale fixée a priori. Nous rappelons que la quantité d'intérêt V_δ est définie par (B.14). Du point de vue d'un processus de production industrielle, ce problème consiste à trouver la dispersion δ_{opt} sur la loi de A telle que le coût de fabrication de la corde soit minimal, tout en s'assurant que pour toutes les dispersions $\delta \leq \delta_{\mathrm{opt}}$, la norme en énergie de la solution ne dépasse pas un seuil donné avec probabilité supérieure à $1-\epsilon$. Nous supposons que le coût est décroissant en δ : définir un processus industriel dont la dispersion des matrices des systèmes produits est faible coûte plus cher qu'un processus où la dispersion autorisée serait plus grande. La solution de (B.19) vérifie :

$$\delta_{\mathrm{opt}} = \max\left\{\delta,\ \text{t.q. } \forall\ 0 < \delta \leq \delta_{\max}, \mathbb{P}\left(V_\delta > S\right) \leq \epsilon\right\}, \tag{B.20}$$

L'intervalle $]0, \delta_{\max}]$ est appelé ensemble des états admissibles du problème (B.19).

Proposition B.2 *Supposons $S > \mathbb{E}(V_{\delta_{\max}})$. Nous disposons de l'inégalité de concentration suivante :*

$$\mathbb{P}\left(V_\delta > S\right) \leq e^{-\phi(\delta,S)}, \tag{B.21}$$

où $\phi\left(\delta, S\right) := \frac{1+\varsigma}{2\delta^2}\frac{||\hat{F}||_2^2}{S} - \frac{1}{2}\left(\frac{1+\varsigma}{\delta^2} - n + 1\right)\left[1 - \log\left(\frac{S}{||\hat{F}||_2^2}\left(1 - \frac{n-1}{1+\varsigma}\delta^2\right)\right)\right]$ *et* $\delta \leq \delta_{\max}$.

Preuve. Remarquons d'abord que $\mathbb{E}(V_{\delta_{\max}}) = \frac{||\hat{F}||_2^2}{1+\frac{1-n}{m_{\min}}}$, où $m_{\min} := \frac{1+\varsigma}{\delta_{\max}^2}$. Ensuite,

$$\begin{aligned}\mathbb{P}\left(V_\delta > S\right) &= \mathbb{P}\left(c(m-n+1) < \theta\right) \text{ où } \theta = \frac{m||\hat{F}||_2^2}{S} \\ &= \mathbb{P}\left(e^{-tc(m-n+1)} > e^{-t\theta}\right), \quad \forall t > 0.\end{aligned} \tag{B.22}$$

Nous utilisons ensuite l'inégalité de Markov :

$$\mathbb{P}\left(V_\delta > S\right) \leq e^{t\theta}\mathbb{E}\left(e^{-tc(m-n+1)}\right). \tag{B.23}$$

Nous reconnaissons la fonction génératrice de $c(m-n+1)$, si bien que

$$\mathbb{P}\left(V_\delta > S\right) \leq (1+2t)^{-\frac{m-n+1}{2}} e^{t\theta} =: g(t), \quad \forall t > 0. \tag{B.24}$$

Nous minimisons la fonction $g(t)$ sur $\mathbb{R}^+$ pour obtenir la meilleure borne possible. Le minimum est unique et atteint en $t_{min} = \frac{1}{2}\left(\frac{m-n+1}{\theta} - 1\right)$. La condition $S > \frac{\left\|\hat{F}\right\|_2^2}{1+\frac{1-n}{m_{\min}}}$ sert à garantir $t_{min} > 0$ pour tous les m considérés. Le résultat est obtenu en injectant la formule pour t_{min} et $m = \frac{1+\varsigma}{\delta^2}$ (cf (B.13)) dans (B.24). ♢

Cette expression n'est pas très pratique à utiliser, notamment la monotonie de ϕ en δ n'est pas évidente, et l'inversion de $\delta \mapsto \phi$ n'est pas directe. Une autre inégalité de concentration est donnée dans la proposition suivante.

Proposition B.3

$$\mathbb{P}\left(V_\delta \geq \frac{m||\hat{F}||_2^2}{m-n+1-t}\right) \leq \exp\left(-\frac{t^2}{4\left(m-n+1\right)}\right), \quad \forall t > 0, \tag{B.25}$$

où nous rappelons que m et δ sont reliés par (B.13).

Preuve. Voir [25] et [64, Section 4.1.]. ♢

Proposition B.4 *Supposons $S > ||\hat{F}||_2^2$ et posons $\alpha := \frac{||\hat{F}||_2^2}{S} \in [0;1[$ et*

$$\delta^* := \sqrt{\frac{1+\varsigma}{\frac{n-1}{1-\alpha} - \frac{2\log\epsilon}{(1-\alpha)^2}\left(1+\sqrt{1-\frac{n-1}{\log\epsilon}\alpha(1-\alpha)}\right)}}. \tag{B.26}$$

L'intervalle $]0,\delta^*[$ est inclus dans l'ensemble des états admissibles de (B.19).

Preuve. Soit t tel que $\frac{m||\hat{F}||_2^2}{m-n+1-t} = S$, c'est-à-dire $t = m(1-\alpha)+1-n$. Pour avoir $t > 0$, il faut que $m > \hat{m} := \frac{n-1}{1-\alpha}$. Nous pouvons montrer que pour que la variance de A^{-1} soit finie, il faut que $m > n+3$. En appliquant la proposition B.3,

$$\forall m > \hat{m}, \quad \mathbb{P}\left(V_\delta > S\right) \leq \exp\left(-\frac{\left(m(1-\alpha)+1-n\right)^2}{4\left(m+1-n\right)}\right) =: \exp\left(-f(m)\right). \tag{B.27}$$

On calcule $f'(m) = \frac{m(1-\alpha)+1-n}{4(m+1-n)^2}\left(m(1-\alpha)+1-n+2\alpha(n-1)\right) = \frac{t}{4(m+1-n)^2}\left(t+2\alpha(n-1)\right)$. Pour tout $m > \hat{m}$, $f'(m) > 0$. Ainsi, l'équation $\exp(-f(m^*)) = \epsilon$ admet une unique solution, qui vérifie $M = m^* - \hat{m}$ tel que

$$M^2 + 4\frac{\log\epsilon}{(1-\alpha)^2}\left(M + \hat{m}\alpha\right) = 0. \tag{B.28}$$

La solution vérifiant $m^* > \hat{m}$ est donnée par $m^* := \frac{n-1}{1-\alpha} - \frac{2\log\epsilon}{(1-\alpha)^2}\left(1 + \sqrt{1 - \frac{n-1}{\log\epsilon}\alpha(1-\alpha)}\right)$, et $\delta^* = \sqrt{\frac{1+\varsigma}{m^*}}$. ♢

L'intervalle $]0, \delta^*[$ est d'autant plus proche de l'ensemble des états admissibles que l'inégalité de concentration de la Proposition B.3 est précise. Nous n'obtenons pas l'ensemble des états admissibles complet, car rien ne nous assure qu'il n'existe pas de δ tel que $\mathbb{P}\left(V_\delta > S\right) < \epsilon < e^{-f(m(\delta))}$. Ce que nous avons démontré jusqu'à présent est représenté dans la figure B.2.

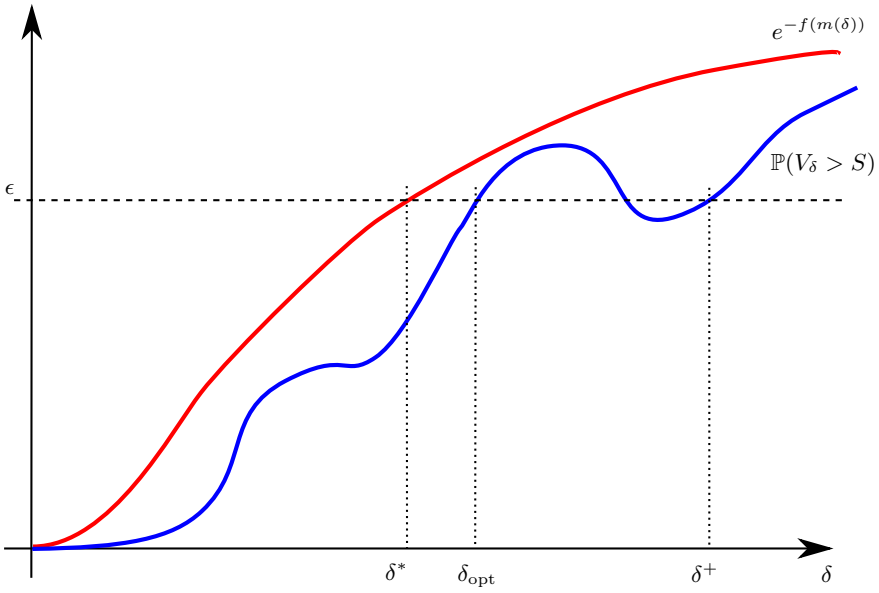

Fig. B.2. Représentation de $\delta \mapsto \mathbb{P}\left(V_\delta > S\right)$ (bleu) et de $\delta \mapsto e^{-f(m(\delta))}$ (rouge) avec les informations de la Proposition B.4

Nous n'avons pour l'instant pas d'information sur la monotonie de $\delta \mapsto \mathbb{P}\left(V_\delta > S\right)$. Dans le cas représenté à la figure B.2, l'équation $\mathbb{P}\left(V_\delta > S\right) = \epsilon$ a plusieurs solutions, notamment δ_{opt} et δ^+, et $\{\delta,\ \text{t.q. } \mathbb{P}\left(V_\delta > S\right) \leq \epsilon\}$ est une union de deux intervalles disjoints. Tout algorithme d'optimisation serait alors inutilisable, car selon l'initialisation, il pourrait renvoyer δ^+, par exemple, qui n'est pas dans l'ensemble des états admissibles de (B.19).

Remarque B.5 *$\mathcal{C}$ étant décroissante, nous avons $\delta_{opt} \geq \delta^*$. Comme $\mathbb{E}\left(V_\delta\right) = \frac{m}{m-n+1}||\hat{F}||_2^2$, alors $\mathbb{E}\left(V_\delta\right) \underset{\delta\to 0}{\to} ||\hat{F}||_2^2$. En supposant $S \geq ||\hat{F}||_2^2$, il vient alors $\mathbb{P}\left(V_\delta > S\right) \underset{\delta\to 0}{\to} 0$.*

Nous énonçons un lemme technique que nous utiliserons par la suite.

Lemme B.6 $\forall \alpha \in \left]0, \frac{1}{2}\right]$,

$$Q_l := \sum_{k=0}^{l-1} \frac{2^k}{k!}\alpha^k + (2\alpha)^l \left(e - \sum_{k=0}^{l-1} \frac{1}{k!} \right) \geq e^{2\alpha}. \tag{B.29}$$

Preuve.

$$\begin{aligned} Q_l - e^{2\alpha} &= \sum_{k=0}^{l-1} \frac{2^k}{k!}\alpha^k + (2\alpha)^l \sum_{k=l}^{+\infty} \frac{1}{k!} - \sum_{k=0}^{+\infty} \frac{2^k}{k!}\alpha^k \\ &= (2\alpha)^l \sum_{k=l}^{+\infty} \frac{1 - (2\alpha)^{k-l}}{k!}. \end{aligned} \tag{B.30}$$

$\forall k \geq l$, $\forall \alpha \in \left]0, \frac{1}{2}\right]$, $1 - (2\alpha)^{k-l} \geq 0$. Donc $Q_l \geq e^{2\alpha}$. ♢

Nous disposons d'un résultat sur la monotonie de $\delta \mapsto \mathbb{P}(V_\delta \geq S)$.

Proposition B.7 *Supposons que l'on se restreigne à des valeurs entières de* $l := \frac{m-n+1}{2}$. *Supposons* $S \geq ||\hat{F}||_2^2$. *Alors la fonction* $\delta \mapsto \mathbb{P}(V_\delta \geq S)$ *est croissante.*

Preuve. Posons $\alpha := \frac{||\hat{F}||_2^2}{2S}$. $S \geq ||\hat{F}||_2^2$ implique $0 \leq \alpha \leq \frac{1}{2}$. Comme $V_\delta \sim m||\hat{F}||_2^2 \frac{1}{c(m-n+1)}$, alors $\mathbb{P}(V_\delta \geq S) = F_{c(m-n+1)}(2m\alpha)$, où $F_{c(m-n+1)}$ désigne la fonction de répartition d'une χ^2 à $m-n+1$ degrés de liberté. Nous avons $F_{c(m-n+1)}(2m\alpha) = \frac{\gamma\left(\frac{m-n+1}{2}, m\alpha\right)}{\Gamma\left(\frac{m-n+1}{2}\right)}$, où $\gamma(\cdot,\cdot)$ est la fonction Gamma incomplète inférieure, et $\Gamma(\cdot)$ est la fonction Gamma. On a $F_{c(m-n+1)}(2m\alpha) = 1 - \frac{\Gamma\left(\frac{m-n+1}{2}, m\alpha\right)}{\Gamma\left(\frac{m-n+1}{2}\right)}$, où $\Gamma(\cdot,\cdot)$ est la fonction Gamma incomplète supérieure. Lorsque $l = \frac{m-n+1}{2}$ est entier, nous avons l'expression :

$$F_{c(m-n+1)}(2m\alpha) = 1 - e^{-(2l+n-1)\alpha} \sum_{i=0}^{l-1} \frac{\left((2l+n-1)\alpha\right)^i}{i!}. \tag{B.31}$$

Il nous reste à montrer que $m \mapsto F_{c(m-n+1)}(2m\alpha)$ est décroissante (car $m = \frac{1+\varsigma}{\delta^2}$), ou, de façon équivalente, que $l \mapsto f_l^\alpha := e^{-(2l+n-1)\alpha} \sum_{i=0}^{l-1} \frac{((2l+n-1)\alpha)^i}{i!}$ est croissante pour tout $\alpha \in \left]0, \frac{1}{2}\right]$. Nous avons

$$f_{l+1}^\alpha - f_l^\alpha = e^{-(2l+n+1)\alpha} \left(\frac{(2l+n+1)^l \alpha^l}{l!} + \sum_{i=0}^{l-1} \frac{\alpha^i}{i!} \left[(2l+n+1)^i - e^{2\alpha}(2l+n-1)^i \right] \right). \tag{B.32}$$

La fonction $l \mapsto f_l^\alpha$ est croissante si et seulement si

$$\Delta f_l^\alpha := \frac{(2l+n+1)^l \alpha^l}{l!} + \sum_{i=0}^{l-1} \frac{\alpha^i}{i!}(2l+n+1)^i - e^{2\alpha} \sum_{i=0}^{l-1} \frac{\alpha^i}{i!}(2l+n-1)^i \geq 0. \tag{B.33}$$

En utilisant le Lemme B.6, il vient

$$
\begin{aligned}
\Delta f_l^\alpha &\geq \frac{(2l+n+1)^l\,\alpha^l}{l!} + \sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l+n+1)^i - \left(\sum_{k=0}^{l-1}\frac{2^k}{k!}\alpha^k + (2\alpha)^l\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\right)\sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l+n-1)^i \\
&= \frac{(2l+n+1)^l\,\alpha^l}{l!} - (2\alpha)^l\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l+n-1)^i \\
&\quad + \sum_{k=0}^{l-1}\sum_{j=0}^{k}\frac{2^j\,(2l+n-1)^{k-j}}{j!\,(k-j)!}\alpha^k - \sum_{k=0}^{l-1}\sum_{i=0}^{l-1}\frac{2^k\,(2l+n-1)^i}{i!k!}\alpha^{i+k} \\
&= \frac{(2l+n+1)^l\,\alpha^l}{l!} - (2\alpha)^l\left[\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l+n-1)^i + \sum_{k=0}^{l-1}\sum_{i=l-k}^{l-1}\frac{2^{k-l}\,(2l+n-1)^i}{i!k!}\alpha^{i+k-l}\right] \\
&\quad + \sum_{k=0}^{l-1}\sum_{j=0}^{k}\frac{2^j\,(2l+n-1)^{k-j}}{j!\,(k-j)!}\alpha^k - \sum_{k=0}^{l-1}\sum_{i=0}^{l-k-1}\frac{2^k\,(2l+n-1)^i}{i!k!}\alpha^{i+k} \\
&= \alpha^l\left\{\frac{(2l+n+1)^l}{l!} - 2^l\left[\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l+n-1)^i + \sum_{k=0}^{l-1}\sum_{i=l-k}^{l-1}\frac{2^{k-l}\,(2l+n-1)^i}{i!k!}\alpha^{i+k-l}\right]\right\} \\
&\quad + \sum_{k=0}^{l-1}\sum_{j=0}^{k}\frac{2^j\,(2l+n-1)^{k-j}}{j!\,(k-j)!}\alpha^k - \sum_{k=0}^{l-1}\sum_{j=k}^{l-1}\frac{2^k\,(2l+n-1)^{j-k}}{(j-k)!k!}\alpha^j.
\end{aligned}
\tag{B.34}
$$

Dans la dernière double somme, $j \geq k$, donc en intervertissant les deux sommations, cette quantité devient $\sum_{j=0}^{l-1}\sum_{k=0}^{j}\frac{2^k(2l-n+1)^{j-k}}{k!(j-k)!}\alpha^j$, et la dernière ligne dans l'expression précédente se simplifie. Nous obtenons

$$
\frac{\Delta f_l^\alpha}{\alpha^l} \geq \frac{(2l+n+1)^l}{l!} - 2^l\left[\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\sum_{i=0}^{l-1}\frac{\alpha^i}{i!}\,(2l-n+1)^i + \sum_{k=0}^{l-1}\sum_{i=l-k}^{l-1}\frac{2^{k-l}\,(2l+n-1)^i}{i!k!}\alpha^{i+k-l}\right].
\tag{B.35}
$$

$\forall l \in \mathbb{N}^*$, $e - \sum_{k=0}^{l-1}\frac{1}{k!} \geq 0$ et $\forall i \in \mathbb{N}$, $-\alpha^i \geq -\frac{1}{2^i}$, donc nous pouvons régler la dépendance en α :

$$
\begin{aligned}
\frac{\Delta f_l^\alpha}{\alpha^l} &\geq \frac{(2l+n+1)^l}{l!} - 2^l\left[\left(e-\sum_{k=0}^{l-1}\frac{1}{k!}\right)\sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!} + \sum_{k=0}^{l-1}\sum_{i=l-k}^{l-1}\frac{(2l+n-1)^i}{2^i i!k!}\right] \\
&= \frac{(2l+n+1)^l}{l!} - 2^l\left[e\sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!} + \sum_{k=0}^{l-1}\frac{1}{k!}\left(\sum_{i=l-k}^{l-1}\frac{(2l+n-1)^i}{2^i i!} - \sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!}\right)\right] \\
&= \frac{(2l+n+1)^l}{l!} - 2^l\left[e\sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!} - \sum_{k=0}^{l-1}\frac{1}{k!}\sum_{i=0}^{l-k-1}\frac{(2l+n-1)^i}{2^i i!}\right] \\
&= \frac{(2l+n+1)^l}{l!} - 2^l\left[e\sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!} - \sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!}\sum_{k=0}^{l-i-1}\frac{1}{k!}\right] \\
&= \frac{(2l+n+1)^l}{l!} - 2^l\sum_{i=0}^{l-1}\frac{(2l+n-1)^i}{2^i i!}\sum_{k=l-i}^{+\infty}\frac{1}{k!}.
\end{aligned}
\tag{B.36}
$$

Donc,

$$\frac{l!}{(2l+n+1)^l}\frac{\Delta f_l^\alpha}{\alpha^l} \geq 1 - l!\sum_{i=0}^{l-1}\frac{2^{l-i}}{i!}\frac{(2l+n-1)^i}{(2l+n+1)^l}\sum_{k=l-i}^{+\infty}\frac{1}{k!}. \tag{B.37}$$

$\forall l \in \mathbb{N}^*$, $\forall i \in \mathbb{N}^*$ tel que $i < l$, la fonction $x \mapsto g(x) := \frac{(2l+x-1)^i}{(2l+x+1)^l}$ est décroissante sur $\mathbb{R}^+$. En effet, $g'(x) = \frac{(2l+x-1)^i}{(2l+x+1)^l}\left(\frac{i}{2l+n-1} - \frac{l}{2l+n+1}\right)$, donc $g'(x) \leq 0 \Leftrightarrow \frac{i+1}{l} - \frac{1}{l} \leq 1 - \frac{1}{l+\frac{x-1}{2}}$. Or, d'après nos hypothèses, $\frac{i+1}{l} \leq 1$ et $-\frac{1}{l} \leq -\frac{1}{l+\frac{x-1}{2}}$. En particulier, $\forall l \in \mathbb{N}^*$, $\forall i \in \mathbb{N}^*$ tel que $i < l$, $\forall n \in \mathbb{N}^*$, $-\frac{(2l+n-1)^i}{(2l+n+1)^l} \geq -\frac{2^{i-l}l^i}{(l+1)^i}$, et nous avons réglé la dépendance en n en écrivant

$$\frac{l!}{(2l+n+1)^l}\frac{\Delta f_l^\alpha}{\alpha^l} \geq 1 - \frac{l!}{(l+1)^l}\sum_{i=0}^{l-1}\frac{l^i}{i!}\sum_{k=l-i}^{+\infty}\frac{1}{k!} := 1 - A_l. \tag{B.38}$$

Nous voulons $\Delta f_l^\alpha \geq 0$. Il nous reste à prouver que $(A_l)_l$ est une suite majorée par 1 pour montrer que $l \mapsto f_l^\alpha$ est croissante $\forall \alpha \in \left]0, \frac{1}{2}\right]$.

$$\begin{aligned}
A_l &= \frac{l^l}{(l+1)^l}\sum_{i=0}^{l-1}\frac{l!l^{i-l}}{i!}\sum_{k=l-i}^{+\infty}\frac{1}{k!} \\
&= \left(1+\frac{1}{l}\right)^{-l}\sum_{i=0}^{l-1}\frac{l!l^{i-l}}{i!}\int_0^1\frac{t^{l-i-1}}{(l-i-1)!}e^{1-t}dt \\
&= \left(1+\frac{1}{l}\right)^{-l} e\int_0^1 e^{-t}\sum_{i=0}^{l-1}\frac{(l-1)!}{i!(l-i-1)!}\left(\frac{t}{l}\right)^{l-i-1}dt \\
&= \left(1+\frac{1}{l}\right)^{-l} e\int_0^1 e^{-t}\left(1+\frac{t}{l}\right)^{l-1}dt \\
&= \left(1+\frac{1}{l}\right)^{-l} e\left(\left[e^{-t}\left(1+\frac{t}{l}\right)^{l}\right]_0^1 - \int_0^1 -e^{-t}\left(1+\frac{t}{l}\right)^{l-1}dt\right) \\
&= 1 + \left(1+\frac{1}{l}\right)^{-l} e\left(I_l - 1\right),
\end{aligned} \tag{B.39}$$

où $I_l := \int_0^1 e^{-t}\left(1+\frac{t}{l}\right)^l dt$. Il nous reste à voir que la suite $(I_l)_l$ est bien majorée par 1 pour que $(A_l)_l$ le soit également. Posons $g(t) = ln\left(e^{-t}\left(1+\frac{1}{l}\right)^l\right)$. On a $g'(t) = -\frac{t}{l}\left(1+\frac{t}{l}\right)^{-1} \leq 0$ pour $t \in \mathbb{R}^+$. Donc $e^{g(t)} \leq e^{g(0)} = 1$. Ainsi, $I_l \leq 1$, et la Proposition B.7 est démontrée. ♢

Corollaire B.8 *La fonction $\delta \mapsto \mathbb{P}(V_\delta \geq S)$ étant croissante, l'ensemble des états admissibles du problème d'optimisation* (B.19) *est un intervalle de $\mathbb{R}^+$. La fonction de coût étant décroissante, δ_{opt} est donné par la solution unique de $\mathbb{P}\left(V_{\delta_{opt}} > S\right) = \epsilon$.*

Nous pouvons désormais imaginer un algorithme de résolution de $\mathbb{P}\left(V_{\delta_{opt}} > S\right) = \epsilon$ initialisé par δ^* donné par le théorème B.4.

Corollaire B.9 *Tout modèle de Wishart pour le système mécanique avec la même matrice moyenne et avec un nombre de degrés de liberté supérieur à $m_{opt} := m(\delta_{opt})$ verra sa solution vérifier la contrainte du problème d'optimisation* (B.19).

B.4 Tests numériques dans le cas d'une corde vibrante

Nous considérons le cas de la corde vibrante décrit dans la Section B.1. Dans un premier temps, nous nous plaçons dans le cas $n = 1$, avec $S = 0.5$, de sorte que $||\hat{F}||_2^2 \approx 0.2045$.

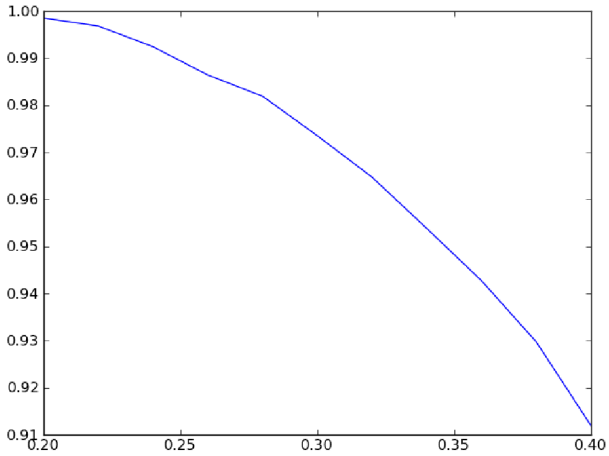

Fig. B.3. $\delta \mapsto \mathbb{P}(V_\delta < S)$ pour $n = 1$

La figure B.3 représente la fonction $\delta \mapsto \mathbb{P}(V_\delta < S) = 1 - \mathbb{P}(V_\delta > S)$. La solution du problème d'optimisation sous contrainte en probabilité pour $\epsilon = 5\%$ est alors $\delta_{opt} \approx 0.34$.

Nous nous plaçons ensuite dans le cas $n = 10$, avec $S = 1.5$, de sorte que $||\hat{F}||_2^2 \approx 1.025$.

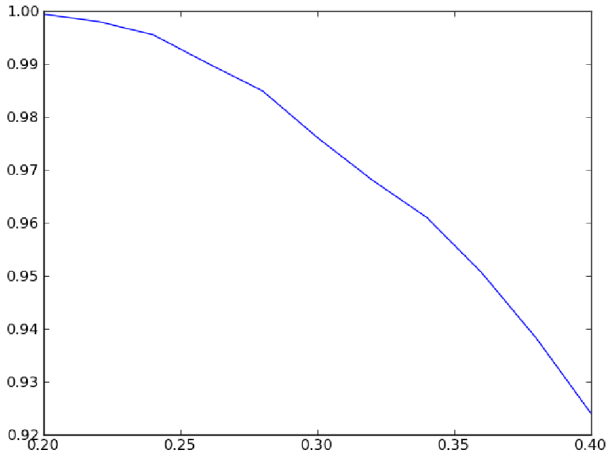

Fig. B.4. $\delta \mapsto \mathbb{P}(V_\delta < S)$ pour $N = 10$

La figure B.4 représente la fonction $\delta \mapsto \mathbb{P}(V_\delta < S)$. La solution du problème d'optimisation sous contrainte en probabilité pour $\epsilon = 5\%$ est alors $\delta_{opt} \approx 0.36$.

B.5 Le problème d'optimisation sous contraintes stochastiques dans le cas de deux objets : décomposition de domaines

On utilise la décomposition de Craig–Bampton, qui permet de coupler deux cordes par une extrémité (ou deux objets mécaniques par leur interface sur les nœeuds du maillage commun), tout en gardant une résolution modale de chacune des deux sous-parties, voir [97] pour la description détaillée. Notons avec un indice i les éléments relatifs à la sous-partie i. Soit $E_i = M_i^{-1}K_i - \omega^2 I$ et $G_i = M_i^{-1}F_i$. Nous obtenons le système

$$\begin{pmatrix} E_i & E_{Ci} \\ {E_C}_i^t & E_{\Sigma i} \end{pmatrix} \begin{pmatrix} Y_i \\ Y_\Sigma \end{pmatrix} = \begin{pmatrix} G_i \\ G_\Sigma \end{pmatrix}, \tag{B.40}$$

où Y_i est l'inconnue déplacement dans les nœuds du maillage intérieurs au sous-domaine i, et Y_Σ est l'inconnue déplacement à l'interface. Considérons la transformation modale de la façon suivante :

$$\begin{pmatrix} Y_i \\ Y_\Sigma \end{pmatrix} = \begin{pmatrix} B_i & S_i \\ 0 & I \end{pmatrix} \begin{pmatrix} Q_i \\ Y_\Sigma \end{pmatrix}, \tag{B.41}$$

où B_i est la matrice de changement de base intervenant dans la diagonalisation de $M_i^{-1}K_i$ (éventuellement rectangulaire si l'on ne garde pas toutes les valeurs propres de $M_i^{-1}K_i$, comme discuté dans la Section B.1), et où $S = -{K_i}_{(1,1)}^t K_{i(1,2)}$, tel que

$$K_i = \begin{pmatrix} K_{i(1,1)} & K_{i(1,2)} \\ {K_i}_{(1,2)}^t & K_{i(2,2)} \end{pmatrix}, \tag{B.42}$$

où le partitionnement des matrices K_i se fait sur les degrés de liberté intérieurs, puis de frontière. En injectant (B.41) dans (B.40) pour les 2 sous-domaines, et en multipliant à gauche par

$$\begin{pmatrix} {B_i}^t & 0 \\ {S_i}^t & I \end{pmatrix}, \tag{B.43}$$

nous obtenons

$$\begin{pmatrix} A_1 & 0 & \Omega_1 \\ 0 & A_2 & \Omega_2 \\ \Omega_1^t & \Omega_2^t & \Delta_1 + \Delta_2 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \\ Y_\Sigma \end{pmatrix} = \begin{pmatrix} B_1^t G_1 \\ B_2^t G_2 \\ S_1^t G_1 + S_2^t G_2 + G_\Sigma \end{pmatrix}, \tag{B.44}$$

où $\Delta_i = S_i^t E_i S_i + E_i S_i + E_{C_i}^t S_i + S_i^t E_{C_i} + E_{\Sigma_i}$, $\Omega_i = {B_1}^t (E_1 S_1 + E_{C_1})$, et $A_i = B_i^t E_i B_i = D_i + \omega^2 I$, avec D_i la matrice diagonale contenant les (n premières) valeurs propres de E_i. Réécrivons (B.44) sous la forme de système :

$$\begin{cases} A_1 Q_1 + \Omega_1 Y_\Sigma = B_1^t G_1, \\ A_2 Q_2 + \Omega_2 Y_\Sigma = B_2^t G_2, \\ \Omega_1^t Q_1 + \Omega_2^t Q_2 + (\Delta_1 + \Delta_2) Y_\Sigma = S_1^t G_1 + S_2^t G_2 + G_\Sigma. \end{cases} \tag{B.45}$$

Nous avons donc

$$\begin{cases} Y_\Sigma = \left[\Omega_1^t A_1^{-1}\Omega_1 + \Omega_2 A_2^{-1}\Omega_2 - (\Delta_1+\Delta_2)\right]^{-1}\left(\Omega_1^t A_1^{-1}B_1^t G_1 + \Omega_2^t A_2^{-1}B_2^t G_2 - S_1^t G_1 - S_2^t G_2 - G_\Sigma\right), \\ Q_1 = A_1^{-1}\left(B_1^t G_1 - \Omega_1 Y_\Sigma\right), \\ Q_2 = A_2^{-1}\left(B_2^t G_2 - \Omega_2 Y_\Sigma\right). \end{cases} \tag{B.46}$$

Supposons maintenant que $A_i \sim W_n\left(\frac{1}{m_i}\left(D_i+\omega^2 I\right), m_i\right)$ et posons $\delta_i = \left(\frac{1}{m_i}\left(1+\frac{\left(tr\left(D_i+\omega^2 I\right)\right)^2}{tr(D_i+\omega^2 I)^2}\right)\right)^{\frac{1}{2}}$. Nous choisissons comme quantité d'intérêt

$$V_{\delta_1,\delta_2} = Y^t Y, \tag{B.47}$$

où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_\Sigma \end{pmatrix} = \begin{pmatrix} B_1 & 0 & S_1 \\ 0 & B_2 & S_2 \\ 0 & 0 & I \end{pmatrix}\begin{pmatrix} Q_1 \\ Q_2 \\ Y_\Sigma \end{pmatrix}, \tag{B.48}$$

avec Q_1, Q_2 et Y_Σ donnés par (B.46).

Nous supposons maintenant que la dispersion sur la loi de A_1 est connue (δ_1 est fixé), et nous définissons le problème d'optimisation sous contraintes en probabilité suivant :

$$\underset{\delta_2}{\operatorname{argmin}}\left(\mathcal{C}\left(\delta_1,\delta_2\right)\right),\ \text{tq}\ \mathbb{P}\left(V_{\delta_1,\delta_2} < S\right) \geq 1-\epsilon. \tag{B.49}$$

Contrairement à la Proposition B.1 dans cas d'une seule corde, nous ne sommes pas en mesure de caractériser simplement la quantité d'intérêt en termes de variables aléatoires classiques. Un résultat équivalent à la Proposition B.7 serait : si $S \geq M = \lim_{\delta_2\to 0}\mathbb{E}\left(V_{\delta_1,\delta_2}\right)$, alors la fonction $\delta_2 \mapsto \mathbb{P}\left(V_{\delta_1,\delta_2} \geq S\right)$ est croissante. Nous vérifions si cette conjecture est raisonnable en représentant à la figure B.5 la fonction $\delta_2 \mapsto \mathbb{P}\left(V_{\delta_1,\delta_2} < S\right)$ avec $\delta_1 = 0.2$, $S = 2.2$, $N_1 = N_2 = 100$ et $n_1 = n_2 = 14$. La monotonie de la fonction est cohérente avec le résultat conjecturé.

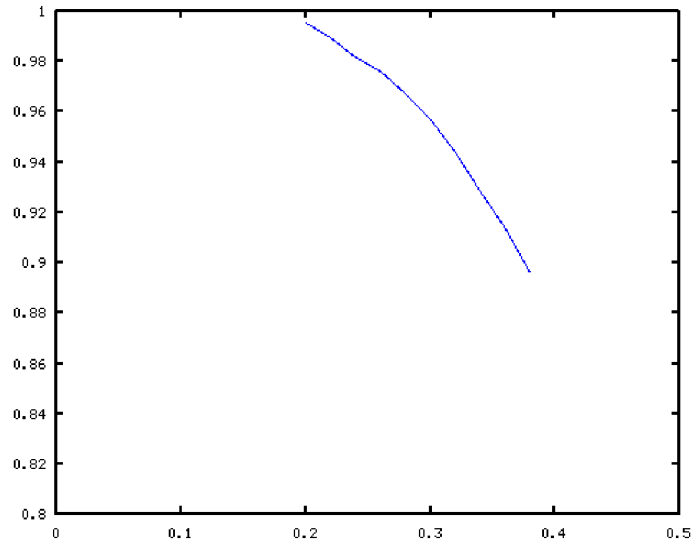

Fig. B.5. $\delta_2 \mapsto \mathbb{P}\left(V_{\delta_1,\delta_2} < S\right)$ pour $\delta_1 = 0.2$, $S = 2.2$, $N_1 = N_2 = 100$ et $n_1 = n_2 = 14$.

B.6 Perspectives

D'un point de vue industriel, la quantité d'intérêt représente l'énergie acoustique rayonnée depuis l'extérieur d'un avion à travers le fuselage vers les passagers. La section B.3 nous a permis d'établir que, dans le cas d'un objet en vibration, si la matrice du système est modélisée par une loi de Wishart, alors il est facile de trouver une dispersion associée à la loi de ces matrices telle que, pour toute dispersion inférieure, l'énergie acoustique rayonnée est inférieure à un seuil fixé. Cette étape peut même être conduite à différentes fréquences : la sensibilité du corps humain au bruit dépendant de la fréquence, nous pouvons imaginer un profil de dispersion δ à imposer à la loi des matrices, tel que l'énergie acoustique transmise par le fuselage soit inférieure à un seuil dépendant de la fréquence. Dans la section B.5, nous évoquons le cas de 2 plaques reliées à leur bord, pour la cabine de l'avion composée de plusieurs plaques boulonnées les unes aux autres. Le problème (B.49) signifie qu'à une dispersion donnée sur la loi de la première plaque, nous cherchons la dispersion autorisée sur la loi de la deuxième plaque telle que le coût global soit minimisé. Graphiquement, il semblerait que l'ensemble des états admissibles pour δ_2 soit un intervalle de $\mathbb{R}^+$.

L'idée d'allouer des quotas d'incertitudes à plusieurs constructeurs afin de garantir une faible probabilité de défaillance sur une quantité globale serait intéressante. Cependant, la modélisation des incertitudes par une loi de Wishart sur la matrice des sous-systèmes pose un problème d'ordre pratique. En effet, avec les moyens de construction actuels, il semble difficile de contractualiser sur la dispersion de la loi de Wishart. La pratique consiste plutôt à contractualiser sur les coefficients du modèle sous-jacent.

References

[1] http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm.

[2] P.R. Amestoy, I.S. Duff, and J.-Y. L'Excellent. Multifrontal parallel distributed symmetric and unsymmetric solvers. *Computer Methods in Applied Mechanics and Engineering*, 184(2–4):501–520, 2000.

[3] R. Amiet and W. R. Sears. The aerodynamic noise of small-perturbation subsonic flows. *Journal of Fluid Mechanics*, (44), 1928.

[4] A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *Journal of Non-Newtonian Fluid Mechanics*, 139(3):153 – 176, 2006.

[5] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, NY, third edition, 2003.

[6] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667 – 672, 2004.

[7] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):pp. 64–94, 2008.

[8] E. Bécache, A. S. BenDhia, and G. Legendre. Perfectly matched layers for the convected helmholtz equation. *SIAM Journal on Numerical Analysis*, 42(1):409–433, 2004.

[9] E. Bécache, A. S. Bonnet-Ben Dhia, and G. Legendre. Perfectly matched layers for the convected helmholtz equation. *SIAM Journal on Numerical Analysis*, 42(1):409–433, 2004.

[10] M. Beldi and A. Maghrebi. Some new results for the study of acoustic radiation within a uniform subsonic flow using boundary integral method. *Advanced Materials Research*, 488–489:383–395, 2012.

[11] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ., 1957.

[12] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ., 1961.

[13] P. Bettess. *Infinite Elements*. Penshaw Press: Cleadon, Sunderland, U.K., 1992.

[14] P. Bettess, D. W. Kelly, and O. C. Zienkiewicz. The coupling of the finite element method and boundary solution procedures. *Int. J. Numer. Meth. Engng.*, 11:355–375.

[15] P. Binev, A. Cohen, W. Dahmen, R. A. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Analysis*, 43(3):1457–1472, 2011.

[16] A. Björck and C. C. Paige. Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm. *SIAM J. Matrix Anal. Appl.*, 13:176–190, 1992.

[17] S. Boyaval. *Mathematical modelling and numerical simulation in materials science*. PhD thesis, Université Paris-Est, 2009.

[18] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N. C. Nguyen, and A. T. Patera. Reduced basis techniques for stochastic problems. *Archives of Computational Methods in Engineering*, 17(4), 2010.

[19] S. Boyaval, C. Le Bris, Y. Maday, N. C. Nguyen, and A. T. Patera. A reduced basis approach for variational problems with stochastic parameters: Application to heat conduction with variable Robin coefficient. *Computer Methods in Applied Mechanics and Engineering*, 198(41-44):3187 – 3206, 2009.

[20] H. Brakhage and P. Werner. Über das Dirichletsche Außenraum Problem für die Helmholtzsche Schwingungsgleichung. *Arch. der Math.*, 16:325–329, 1965.

[21] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Texts in Applied Mathematics. Springer, 2008.

[22] H. Brézis. *Analyse fonctionnelle*. Dunod, 1999.

[23] A. Buffa and R. Hiptmair. Regularized combined field integral equations. *Numer. Math.*, 100(1):1–19, 2005.

[24] A. Buffa, Y. Maday, A. T. Patera, C. Prud'homme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(3):595–603, 2012.

[25] E. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1-2):577–589, 2013.

[26] B. Carpentieri. *Sparse preconditioners for dense linear systems from electromagnetic applications*. PhD thesis, CERFACS, 2002.

[27] B. Carpentieri, I. Duff, L. Giraud, and G. Sylvand. Combining fast multipole techniques and an approximate inverse preconditioner for large electromagnetism calculations. *SIAM Journal on Scientific Computing*, 27(3):774–792, 2005.

[28] C. Carstensen, S.A. Funken, and E.P. Stephan. On the adaptive coupling of FEM and BEM in 2-D-elasticity. *Numerische Mathematik*, 77(2):187–221, 1997.

[29] Y. Chen, J. S. Hesthaven, Y. Maday, J. Rodriguez, and X. Zhu. Certified reduced basis method for electromagnetic scattering and radar cross section estimation. *Computer Methods in Applied Mechanics and Engineering*, 233–236(0):92 – 108, 2012.

[30] Y. Chen, J. S. Hesthaven, Y. Maday, and J. Rodríguez. Improved successive constraint method based a posteriori error estimate for reduced basis approximation of 2d Maxwell's problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43:1099–1116, 11 2009.

[31] M. Costabel. *Symmetric methods for the coupling of finite elements and boundary elements*. Preprint. Fachbereich Mathematik. Technische Hochschule Darmstadt. Fachber., TH, 1987.

[32] M. Costabel. Boundary integral operators on Lipschitz domains: Elementary results. *SIAM Journal on Mathematical Analysis*, 19(3):613–626, 1988.

[33] A. Delnevo and I. Terrasse. Code acti3s harmonique, justification mathématique, Partie I. Technical report, EADS, 2001.

[34] A. Delnevo and I. Terrasse. Code acti3s, justifications mathématiques, Partie II : presence d'un écoulement uniforme. Technical report, EADS, 2002.

[35] R. A. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989.

[36] R. A. DeVore, G. Petrova, and P. Wojtaszczyk. Greedy algorithms for reduced bases in Banach spaces. *Constructive Approximation*, 37(3):455–466, 2013.

[37] R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996.

[38] C. Domínguez, E.P. Stephan, and M. Maischak. FE/BE coupling for an acoustic fluid-structure interaction problem. residual a posteriori error estimates. *International Journal for Numerical Methods in Engineering*, 89(3):299–322, 2012.

[39] F. Dubois, E. Duceau, F. Maréchal, and I. Terrasse. Lorentz transform and staggered finite differences for advective acoustics. Technical report, EADS, 2002.

[40] S. Duprey. *Analyse Mathématique et Numérique du Rayonnement Acoustique des Turboréacteurs*. PhD thesis, EADS-CRC ; Institut Elie Cartan-Université Poincaré Nancy, 2005.

[41] V. Erhlacher. *Quelques modèles mathématiques en chimie quantique et propagation d'incertitudes*. PhD thesis, Université Paris-Est, 2012.

[42] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*. Number vol. 159 in Applied Mathematical Sciences. Springer, 2004.

[43] G. Fairweather, A. Karageorghis, and P. A. Martin. The method of fundamental solutions for scattering and radiation problems. *Engineering Analysis with Boundary Elements*, 27(7):759 – 769, 2003.

[44] M. Fares, J. S. Hesthaven, Y. Maday, and B. Stamm. The reduced basis method for the electric field integral equation. *Journal of Computational Physics*, 230(14):5532 – 5555, 2011.

[45] N. Garofalo and F.-H. Lin. Unique continuation for elliptic operators: A geometric-variational approach. *Communications on Pure and Applied Mathematics*, 40(3):347–366, 1987.

[46] L. Giraud and J. Langou. When modified Gram–Schmidt generates a well-conditioned set of vectors. *SIAM J. Matrix Anal. Appl.*, 22:521–528, 2002.

[47] H. Glauert. The effect of compressibility on the lift of an aerofoil. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 118(779):113–119, 1928.

[48] D. Goldberg. What every computer scientist should know about floating point arithmetic. *ACM Computing Surveys*, 23(1):5–48, 1991.

[49] M. E. Goldstein. *Aeroacoustics*. McGraw-Hill International Book Company, 1976.

[50] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996.

[51] M. F. Hamilton and D. T. Blackstock. *Nonlinear Acoustics: Theory and Applications*. Elsevier Science & Tech, 1998.

[52] F. Hecht, J. Morice, and O. Pironneau. freeFEM++, www.freefem.org/.

[53] R. Hiptmair. Coupling of finite elements and boundary elements in electromagnetic scattering. *SIAM J. Numer. Anal.*, 41(3):919–944.

[54] R. Hiptmair and P. Meury. *Stable FEM-BEM Coupling for Helmholtz Transmission Problems*. ETH, Seminar für Angewandte Mathematik, 2005.

[55] G. C. Hsiao and W. L. Wendland. *Boundary Element Methods: Foundation and Error Analysis*. John Wiley & Sons, Ltd, 2004.

[56] D. B. P. Huynh, A. T. Patera, G. Rozza, and S. Sen. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *Comptes Rendus Mathematique*, 345(8):473 – 478, 2007.

[57] J.M. Jin and V.V. Liepa. A note on hybrid finite element method for solving scattering problems. *IEEE Trans. Ant. Prop.*, 36(10):1486–1490, 1988.

[58] C. Johnson and J. C. Nédélec. On the coupling of boundary integral and finite element methods. *Mathematics of Computation*, 35(152):pp. 1063–1079, 1980.

[59] D. J. Knezevic, N. C. Nguyen, and A. T. Patera. Reduced Basis Approximation and A Posteriori Error Estimation for the Parametrized Unsteady Boussinesq Equations. *Mathematical Models & Methods in applied sciences*, 21(7):1415–1442, 2011.

[60] A. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541, 2001.

[61] P. Ladevèze. *Nonlinear computational structural mechanics: new approaches and non-incremental methods of calculation*. Mechanical engineering series. Springer, 1999.

[62] P. Langlois, S. Graillat, and N. Louvet. Compensated Horner Scheme. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2006.

[63] J. Langou. *Solving large linear systems with multiple right-hand sides*. PhD thesis, INSA, 2003.

[64] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics Vol. 28, No. 5, pp. 1302-1338*, october 2000.

[65] R. Leis. Zur Dirichletschen Randwertaufgabe des Außenraumes der Schwingungsgleichung. *Mathematische Zeitschrift*, 90:205–211, 1965.

[66] V. Levillain. *Couplage éléments finis-équations intégrales pour la résolution des équations de Maxwell en milieu hétérogène*. PhD thesis, Ecole Polytechnique, 1991.

[67] S. Lewy. L'aéroacoustique en aéronautique. In *Techniques de l'Ingénieur*.

[68] F. Leydecker, M. Maischak, E.P. Stephan, and M. Teltscher. Adaptive FE-BE coupling for an electromagnetic problem in R3- A residual error estimator. *Mathematical Methods in the Applied Sciences*, 33(18):2162–2186, 2010.

[69] M. J. Lighthill. On sound generated aerodynamically. I. General theory. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 211(1107):564–587, 1952.

[70] M. J. Lighthill. On sound generated aerodynamically. II. Turbulence as a source of sound. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 222(1148):1–32, 1954.

[71] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D. V. Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 331(2):153 – 158, 2000.

[72] L. Machiels, Y. Maday, A. T. Patera, C. Prud' homme, D. V. Rovas, G. Turinici, and K. Veroy. Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *CJ Fluids Engineering*, 124:70–80, 2002.

[73] Y. Maday, N. C. Nguyen, A. T. Patera, and S. Pau. A general multipurpose interpolation procedure: the magic points. *Communications On Pure And Applied Analysis*, 8(1):383–404, 2008.

[74] M. Maischak and E.P. Stephan. A FEM-BEM coupling method for a nonlinear transmission problem modelling coulomb friction contact. *Computer Methods in Applied Mechanics and Engineering*, 194(2-5):453–466, 2005.

[75] B. McDonald and A. Wexler. Finite-element solution of unbounded field problems. *IEEE Trans. Microwave Theory Tech.*, 20(12):841–847, 1972.

[76] W. C. H. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.

[77] R. V. Mises and H. Pollaczek-Geiringer. Praktische verfahren der gleichungsauflösung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 9(1):58–77, 1929.

[78] D. Mitsoudis, C. Makridakis, and M. Plexousakis. Helmholtz equation with artificial boundary conditions in a two-dimensional waveguide. *SIAM Journal on Mathematical Analysis*, 44(6):4320–4344, 2012.

[79] C. L. Morfey. Acoustic energy in non-uniform flows. *Journal of Sound and Vibration*, 14(2):159 – 170, 1971.

[80] J. C. Nédélec. *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*. Number vol. 144 in Applied Mathematical Sciences. Springer, 2001.

[81] A. Nouy. A priori tensor approximations for the numerical solution of high dimensional problems: alternative definitions. *The Seventh International Conference on Engineering Computational Technology (ECT2010), Valencia : Espagne*, 2010.

[82] D. P. O'Leary. The block conjugate gradient algorithm and related methods. *Linear Algebra and its Applications*, 29(0):293 – 322, 1980.

[83] O. Panich. On the question of the solvability of the exterior boundary value problems for the wave equation and maxwell's equations. *Usp. Mat. Nauk.*, 20A:221–226, 1965.

[84] A. T. Patera. private communication. 2012.

[85] A. T. Patera, C. Prud'homme, D. V. Rovas, and K. Veroy. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*, 2003.

[86] A. T. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2007.

[87] J. Periaux. Three dimensional analysis of compressible potential flows with the finite element method. *International Journal for Numerical Methods in Engineering*, 9(4):775–831, 1975.

[88] C. J. Powles and B. J. Tester. Scattering of sound from a monopole source by a steady cylindrical jet. 2007.

[89] C. J. Powles and B. J. Tester. Asymptotic and numerical solutions for shielding of noise sources by parallel coaxial jet flows. In *14th AIAA/CEAS Aeroacoustics Conference*, 2008.

[90] C. J. Powles and B. J. Tester. Asymptotic and numerical solutions for shielding of noise sources by parallel coaxial jet flows. In *14th AIAA/CEAS Aeroacoustics Conference*, 2008.

[91] A. Hirschberg S. W. Rienstra. *An Introduction to Acoustics*. Eindhoven University of Technology, 2004.

[92] Y. Saad and M. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[93] S. A. Sauter and C. Schwab. *Boundary Element Methods*. Springer Series in Computational Mathematics. Springer, 2010.

[94] S. Sen. Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems. *Numerical Heat Transfer, Part B: Fundamentals*, 54(5):369–389, 2008.

[95] S. Sen, K. Veroy, D.B.P. Huynh, S. Deparis, N.C. Nguyen, and A.T. Patera. "Natural norm" a posteriori error estimators for reduced basis approximations. *Journal of Computational Physics*, 217(1):37 – 62, 2006.

[96] C. Soize. A nonparametric model of random uncertainties for reduced matrix models in structural dynamics. *Probabilistic Engineering Mechanics*, 15(3):277–294, 2000.

[97] C. Soize and H. Chebli. Random uncertainties model in dynamic substructuring using a nonparametric probabilistic model. *Journal of Engineering Mechanics*, 129(4):449–457, 2003.

[98] M. L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Series in Statistics Series. Springer London, Limited, 1999.

[99] G Sylvand. *La méthode multipôle rapide en électromagnétisme : Performances, parallélisation, applications*. PhD thesis, Université de Nice-Sophia Antipolis, 2002.

[100] L. Tartar. *An Introduction to Sobolev Spaces and Interpolation Spaces*. Lecture notes of The Unione Matematica Italiana. Springer, 2007.

[101] V. Temlyakov. *Greedy Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2011.

[102] J. Utzmann, C.-D. Munz, M. Dumbser, E. Sonnendrücker, S. Salmon, S. Jund, and E. Frénod. *Numerical Simulation of Turbulent Flows and Noise Generation.* Springer, 2009.

[103] K. Veroy and A. T. Patera. Certified real-time solution of the parametrized steady incompressible navier-tokes equations: rigorous reduced-basis a posteriori error bounds. *International Journal for Numerical Methods in Fluids*, 47(8-9):773–788, 2005.

[104] K. Veroy, C. Prud'homme, and A. T. Patera. Reduced-basis approximation of the viscous Burgers equation: rigorous a posteriori error bounds. *Comptes Rendus Mathematique*, 337(9):619 – 624, 2003.

[105] O. von Estorff and M. Firuziaan. Coupled BEM/FEM approach for nonlinear soil/structure interaction. *Engineering Analysis with Boundary Elements*, 24(10):715–725, 2000.

[106] T. Von Petersdorff. Boundary integral equations for mixed Dirichlet, Neumann and transmission problems. *Mathematical Methods in the Applied Sciences*, 11(2):185–213, 1989.

[107] M. Yano. A space-time Petrov-Galerkin certified reduced basis method: Application to the boussinesq equations. *Submitted to SIAM Journal on Scientific Computing*, 2012.

[108] O. C. Zienkiewicz and P. Bettess. Dynamic fuid-structure interaction. numerical modelling of the coupled problem. *In Numerical Methods in Offshore Engineering*, pages 185–194, 1978.