Optimisation topologique de systèmes multiphysiques

Florian Feppon

université PARIS-SACLAY

NNT : 2019SACLX080

Thèse de doctorat

Shape and topology optimization of multiphysics systems

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'École polytechnique

Ecole doctorale n°574 Ecole doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à l'École polytechnique, le 16 décembre 2019, par

FLORIAN FEPPON

Composition du Jury :

Grégoire ALLAIRE Professeur, École polytechnique (CMAP)	Directeur de thèse
Samuel AMSTUTZ Professeur, École polytechnique (CMAP)	Examinateur
Julien CORTIAL Ingénieur de recherche, Safran	Invité
Charles DAPOGNY Chargé de recherche, Université Grenoble Alpes (LJK)	Co-encadrant de thèse
Sonia FLISS Enseignant chercheur, ENSTA (UMA)	Présidente du jury
Antoine HENROT Professeur, École des mines de Nancy	Rapporteur
Robert KOHN Professeur émerite, New York University (Courant Institute)	Rapporteur (absent)
Olivier PIRONNEAU Professeur émérite, Université Pierre et Marie Curie (LJLL)	Examinateur
Yannick PRIVAT Professeur, Université de Strasbourg	Examinateur
Nicole SPILLANE Chargé de recherche, École polytechnique (CMAP)	Examinatrice

NNT : 2019SACLX080

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École polytechnique

Laboratoire d'accueil : Centre de mathématiques appliquées de Polytechnique, UMR 7641 CNRS

Spécialité de doctorat : Mathématiques appliquées

Florian FEPPON

Shape and topology optimization of multiphysics systems

Date de soutenance : 16 décembre 2019

Après avis des rapporteurs : Antoine HENROT (Université de Lorraine)
Robert KOHN (Courant Institute, NYU)

Jury de soutenance :

Grégoire ALLAIRE	(CMAP) Directeur de thèse
Samuel AMSTUTZ	(CMAP) Examinateur
Julien CORTIAL	(Safran) Invité
Charles DAPOGNY	(LJK) Co-encadrant de thèse
Sonia FLISS	(ENSTA – UMA) Examinatrice
Antoine HENROT	(École des Mines de Nancy) Rapporteur
Olivier PIRONNEAU	(LJLL) Examinateur
Yannick PRIVAT	(Université de Strasbourg) Examinateur
Nicole SPILLANE	(CMAP) Examinatrice

ABSTRACT

This work is devoted to shape and topology optimization of multiphysics systems motivated by aeronautic industrial applications. Shape derivatives of arbitrary objective functionals are computed for a weakly coupled thermal fluid-structure model. A novel gradient flow type algorithm is then developed for solving generic constrained shape optimization problems without the need for tuning non-physical metaparameters. Motivated by the need for enforcing non-mixing constraints in the design of liquid-liquid heat exchangers, a variational method is developed in order to simplify the numerical evaluation of geometric constraints: it allows to compute line integrals on a mesh by solving a variational problem without requiring the explicit knowledge of these lines on the spatial discretization. All these ingredients allowed us to implement a variety of 2-d and 3-d multiphysics shape optimization test cases: from single, double or three physics problems in 2-d, to moderately large-scale 3-d test cases for structural design, thermal conduction, aerodynamic design and a fluid-structure interacting system. A final opening chapter derives high order homogenized equations for the Stokes system in a porous medium. These high order equations encompass the three classical homogenized regimes—namely Stokes, Brinkman and Darcy—associated with different obstacle's size scalings. They could allow, in future works, to develop new topology optimization methods for the design of fluid systems characterized by multi-scale patterns such as industrial heat exchangers.

Keywords: Topology optimization, remeshing, convective heat transfer, fluid-structure interaction, geometric constraints, high order homogenization.

RÉSUMÉ

Cette thèse est consacrée à l'optimisation de la topologie et de la forme de systèmes multiphysiques motivés par des applications de l'industrie aéronautique. Nous calculons les dérivées de forme de fonctions de coût arbitraires pour un modèle fluide, thermique et mécanique faiblement couplé. Nous développons ensuite un algorithme de type gradient adapté à la résolution de problèmes d'optimisation de formes sous contraintes qui ne requiert par de réglage de paramètres non physiques. Nous introduisons ensuite une méthode variationnelle qui permet de calculer des intégrales le long de rayons sur un maillage par la résolution d'un problème variationnel qui ne requiert pas la détermination explicite de ces lignes sur la discrétisation spatiale. Cette technique nous a ainsi permis d'imposer une contrainte de non-mélange de phases pour une application à l'optimisation d'échangeurs de chaleur bi-tubes. Tous ces ingrédients ont été employés pour traiter une variété de cas tests d'optimisation de formes pour des systèmes multi-physiques 2-d ou 3-d. Nous avons considéré des problèmes à une seule, deux ou bien trois physiques couplées en 2-d, et des problèmes de tailles relativement élevées en 3-d pour la mécanique, la conduction thermique, l'optimisation de profils aérodynamiques, et de la forme de systèmes en interaction fluide-structure. Un dernier chapitre d'ouverture est consacré à l'étude de modèles homogénéisées d'ordres élevés pour les équations de Stokes en milieu poreux. Ces équations d'ordres élevés englobent les trois régimes homogénéisés classiques—Stokes, Brinkman et Darcy–associés à divers rapports d'échelles pour la taille des obstacles. Elles pourraient permettre, lors de futurs travaux, de développer de nouvelles méthodes d'optimisation pour la conception de systèmes fluides caractérisés par des motifs multiéchelles, tels que les échangeurs thermiques industriels.

Mots clés : Optimisation topologique, remaillage, transfert thermique convectif, interaction fluide-structure, contraintes géométriques, modèles homogénéisés d'ordres élevés.

REMERCIEMENTS

ॐ भूर्भुवः स्वः	*oṃ bhūr bhuvaḥ suvaḥ*
तत्सवितुर्वरेण्यं	*tatsaviturvareṇyaṃ*
भर्गो देवस्य धीमहि	*bhargo devasyadhīmahi*
धियो यो नः प्रचोदयात् ॥	*dhiyo yo naḥ prachodayāt*

We choose the Supreme Light of the divine Sun; we aspire that it may impel our minds.
(Gayatri Mantra, Trad. Sri Aurobindo)

Mes remerciements vont en premier lieu à la providence qui a bien voulu créer les conditions propices à la réalisation de ce travail de par la réunion de bonnes volontés en un même endroit et en un même moment opportun. Certes, ces dispositions peuvent être comparées au fruit arrivé à maturation grâce à de nombreuses années de labeur; il suffit alors d'avoir la chance d'arriver au bon moment pour pouvoir le cueillir et en extraire la saveur. Cependant, malgré toutes ces conditions *nécessaires*, il s'en est fallu de peu pour que ce travail ne voie pas le jour sous sa forme actuelle. La recherche scientifique est une *écoute* alerte et vigilante permanente : il suffit d'une innocente autocensure de *rien du tout*, d'un *menu* attachement à une idée un peu trop personnelle, d'une *imperceptible* procrastination, d'une *anodine* discussion en moins avec un camarade doctorant éclairé pour que le fruit reste invisible au milieu du brouillard de l'opacité mentale. Je remercie donc cette grâce intérieure qui nourrit notre intuition (pour ne pas dire, nous souffle les preuves des théorèmes), nous guide vers des rencontres fructueuses, nous fait adorer la clarté, et nous réunit joyeusement à porter nos efforts vers une même direction.

Il convient ensuite de témoigner ma gratitude à l'ensemble des personnes qui m'ont accompagné et porté sur le chemin de cette thèse. Je remercie Grégoire pour avoir déroulé un long tapis rouge jonché de belles opportunités devant mes pieds depuis la fin 2014. Merci pour le dévouement, l'humanité, et la bienveillance avec laquelle tu encadres tes thésards. Un grand merci à Charles pour la sincérité de ton encadrement, pour l'abondante assistance que tu as pu m'apporter durant ces trois ans, pour ton accueil à Grenoble, pour l'effort de tes relectures et commentaires attentifs, et enfin plus encore pour les précieuses discussions et conseils à un thésard s'interrogeant sur sa vie. Je suis certain que la fin de cette thèse n'est qu'un début vers de nouvelles aventures communes, et si possible bien évidemment près des montagnes...

Je tiens à remercier sincèrement l'encadrement dont j'ai bénéficié à Safran Tech : Julien, pour son attachement caractéristique à la mise en place et au maintien de toutes les *conditions d'optimalité* nécessaires au bon déroulement de la thèse, puis Felipe et Christian pour leurs encouragements soutenus tout au long du travail[1]. Bien sûr, la vie à Safran Tech aurait été moins agréable sans la présence des permanents : je remercie entre autres Oana, Nissrine, Sophie, Noémie, Tonya, Wafa, Emma, Sébastien, Augustin, Fabien, Franck, Mohammed, Guy, Nicolas et Frédéric. L'environnement Safran Tech, à mi-chemin entre l'industrie et le monde académique, m'a paru fort enrichissant, aussi je recommanderais sans hésiter cette expérience CIFRE à de futurs doctorants.

Merci ensuite aux personnes du monde académique qui se sont intéressées de près à ce travail : notamment Yannick Privat pour son invitation à Strasbourg et ses encouragements, et Georgios pour avoir continué à dédier un peu de ton temps et de ton expérience pour nous. Je suis également reconnaissant envers Frédéric Hecht et Olivier Pironneau pour les échanges autour de mes travaux et les invitations au Laboratoire Jacques Louis Lions (LJLL). Je remercie chaleureusement Algiane, Pierre Jolivet et Vincent Moureau pour l'aide précieuse qu'ils m'ont apportée pour la réalisation des cas tests tridimensionnels.

Un grand merci à Antoine Henrot et Robert Kohn qui m'ont fait l'honneur de lire et rapporter mon travail. Je remercie de même tous les membres de mon jury de thèse qui ont accepté sans hésiter ce rôle et qui ont bien voulu offrir leur présence et leur attention le 16 décembre 2019.

Du côté du Centre de Mathématiques Appliquées (CMAP), je suis bien sûr reconnaissant envers Nasséra, Alex et Maud pour leur accueil, leur dévouement et leur patience infinies envers les doctorants trop peu soucieux des règles administratives (occupés qu'ils sont à faire des mathématiques, il leur faut environ trois ans pour assimiler toutes les règles...). Merci ensuite à toutes les autres personnes avec lesquelles j'ai eu l'opportunité d'échanger : entre autres Nicole Spillane, François Alouges, Samuel Amstutz, Lucas Chennel. Merci à Pierre et Sylvain pour leur disponibilité à toute épreuve en ce qui concerne les demandes informatiques. Merci à tous les doctorants du laboratoire avec lesquels j'ai eu

[1] Notons que la thèse, *in fine*, est aussi un travail soutenu.

le plaisir de partager ces trois années (en vous adressant mes plus sincères excuses si par hasard votre nom n'est pas cité). Merci tout particulièrement à Léa au moins pour une balade ensoleillée en vélo en Bretagne, Martin A. pour les moments de partages autour de nos passions artistiques, à Kevish pour son enthousiasme et sa simplicité d'*être* et à qui j'espère ne pas avoir (trop) abimé les genoux. Merci à Céline, Pamela, Juliette, Geneviève, Pierre C., Florian B., Fedor, Heythem, Nicolas, Corentin H. et Corentin C., Frédéric, toujours enthousiastes pour se retrouver, discuter, ou bien mettre de la vie dans les couloirs du laboratoire.

Merci à toutes les personnes du bureau 2016 qui ont subi ma présence : plus particulièrement Jean-Bernard, illustre maître du bureau, pour le partage de tes réflexions avisées sur l'actualité politique, Hadrien en tant qu'exemple vivant que la réussite peut naître de la désorganisation, Raphaël et Cédric. Aux personnes qui ont quitté ce bureau moins tôt : Mathieu toujours très réceptif à mon humour et qui nous a enseigné un moyen très astucieux pour maintenir en place une fenêtre de bureau récalcitrante. Merci à Julie, partante pour toutes les sorties sportives et qui parvient à nous convaincre (non sans usage d'une certaine insistance...) de venir l'accompagner par nuit noire courir sur des chemins boueux de forêt. Merci à ce cher Cheikh jamais de mauvaise humeur. Merci à Paul T. qui est sans aucun doute la personne m'ayant le mieux accompagné pendant cette thèse et ayant eu la plus grande indulgence devant mes erreurs... lorsqu'il jouait du violoncelle. Merci Paul J. pour m'avoir fait redécouvrir la beauté de la langue Française avec Cyrano de Bergerac. Aux autres doctorants qui bien assez vite deviendront à leur tour les plus anciens du bureau : Bowen, Louis, Alexis, Naoufal, Alann.

Je remercie les ami(e)s du groupe de Grégoire qui eux aussi dédient une partie de leur vie à l'optimisation de formes. Merci Mathilde, dont on envie le rayonnement permanent, pour tout ce débordement de bonne humeur que tu sais communiquer aux autres. Merci Perle pour les nombreux moments passés en école d'été, en conférence, (aux jardins botaniques), et d'avoir maintes fois œuvré pour nous rassembler entre doctorants de Safran ou bien du CMAP. Merci pour les figues (et les guêpes !), les pommes, et les analyses sur la psychologie de la personnalité féline. Merci Lalaina, Martin, Jeet, Alex, Beni, Alexis, Mathias pour les nombreuses expériences communes (notamment, de soupes, de fromages, mais par pitié sans crème fraiche ni gâteaux liquides) et d'échanges mathématiques enthousiastes.

Je suis bien également reconnaissant envers mes camarades doctorants de Safran. Je remercie Perle (pour m'avoir cité trois fois dans ses remerciements), Camille (pour la bienveillance), Adrien (qui reconnaîtra que l'Abondance est bien le meilleur des fromages), Maxence W. (pour son alignement et Baahubali), Yannis (pour son vocabulaire délicat), Clément B (pour avoir épargné mes genoux), Antony (pour une leçon d'anglais à ECCM), Thomas (pour son amour pour les ROM), Clément O. (pour ses recommandations de courses dans la vallée de la mère en thèse), Moubine (pour l'enthousiasme), et bien sûr Loïc J. (pour nous avoir transmis ton intérêt pour les aliments et boissons fermentées!). Un sincère remerciement pour l'ambiance empreinte de légèreté et d'humour qui a perduré durant ces trois années grâce à vos personnalités caractéristiques et colorées, et grâce aux nombreuses occasions qui nous ont réunis en dehors de Châteaufort.

J'adresse ensuite des remerciements sincères envers tous mes ami(e)s qui à leur manière ont contribué à maintenir la bonne santé mentale du thésard. Un *énorme*[2] merci à Maxence E. pour la fidélité et la sincérité de ton amitié qui me font te pardonner volontiers les cinq petites minutes d'attente systématiques. Merci Frédérique et Thibault (pour les déjeuners à l'INRIA après la piscine) Nicole et Benjamin (notamment pour une descente mémorable du pic rouge de Bassiès), Cécile, Sylvain (entre autres pour ce qui nous relie mutuellement à la France et l'Inde), Benjamin (merci pour ton invitation à une table ronde bien singulière), Christophe (au moins pour quelques bonnes rasades d'Orchata avec Anouk entre deux minisymposia, mais aussi pour la main gauche de Ravel), Amaury (avec qui il est toujours un plaisir de discuter), Chenzhang (pour ses conseils pianistiques et encouragements avisés), Guillaume (pour une visite à Zurich), Camille (pour m'avoir par exemple fait découvrir que les courgettes peuvent se croquer crues), et Marc (pour un beau week-end en Picardie).

Je remercie les précieux amis qui pensent toujours à moi depuis les U.S.A. (ou bien depuis Ferney-Voltaire International Airport) : *thank you* Viral, Dragosh, Wei Wei, Suhyoun, Parnika, Saviz, Salman, Ravi, Corbin, Rohit, Rameech, Chinmay, Arko ...

Enfin, je n'oublie pas l'éternelle *dream team* du lycée : merci Alex, Paul, David, Raphaël, Laurent, Cédric, Martin, Olivier de (presque) toujours parvenir à trouver le moyen d'être présent lors de mes venues parcimonieuses en contrées savoyardes.

Je termine par les remerciements à mes deux familles qui m'ont permis à leurs contacts de recharger et d'équilibrer régulièrement mes batteries d'énergie physique, mentale et émotionnelle. Ma famille

[2] énorme comment?

spirituelle tout d'abord, spécifiquement mes ami(e)s de l'École André Van Lysebeth auxquel(le)s le reliement me nourrit d'instant en instant. En particulier, merci Joumana pour la confiance que tu me témoignes depuis mon arrivée à Orsay. Et puis, je remercie bien affectueusement mes racines terrestres, toutes les personnes de ma famille de chair qui m'estiment, m'encouragent et colorent ma vie depuis son commencement. J'adresse en particulier une pensée sincère et profonde pour Kévin, qui me rappelle à l'humilité.

Et avant de commencer, je Te remercie Toi, ô cher(ère) Lecteur(trice), pour ce qui porte Tes yeux à parcourir ce travail!

CONTENTS

INTRODUCTION – RÉSUMÉ DE LA THÈSE

Cette thèse est consacrée à l'optimisation de la forme et de la topologie de systèmes multi-physiques motivés par des applications de l'industrie aéronautique. Nous employons principalement la méthode de Hadamard qui est particulièrement adaptée à la prise en compte de diverses contraintes industrielles détaillées ci-après. Cependant, nous effectuons dans un dernier chapitre indépendant des développements théoriques en ce qui concerne l'applicabilité future de la méthode d'optimisation topologique de systèmes fluides par homogénéisation dans un dernier chapitre indépendant.

L'optimisation des formes est l'art mathématique de générer des formes "optimales" qui satisfont au mieux un objectif proposé. Dans le contexte industriel, il s'agit de concevoir des systèmes physiques qui atteignent des performances optimales. Quelle est la forme de la structure la plus rigide utilisant une quantité donnée de matière ? Quel profil aérodynamique choisir pour générer une force de portance voulue ? Quelle distribution de fluides permet de réaliser le meilleur transfert de chaleur ? Comment concevoir la forme d'une coque de bateau ? Ces questions sont des problèmes industriels très classiques.

1 De l'optimisation de formes à l'optimisation topologique : la méthode de Hadamard, les méthodes par densité et par homogénéisation

La méthode de Hadamard tire ses origines du mémoire de 1908 [180]. Elle consiste à calculer la sensibilité du problème considéré à de petites déformations de la frontière (Figure 1). Ceci permet d'optimiser la forme via la détermination de déformations particulières qui donnent lieu à de meilleurs designs. De part

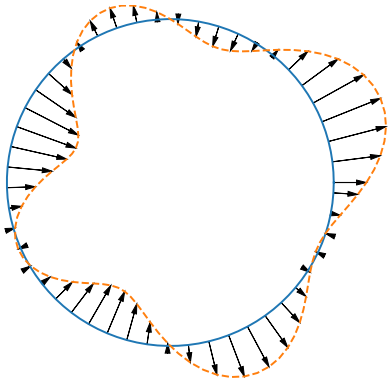
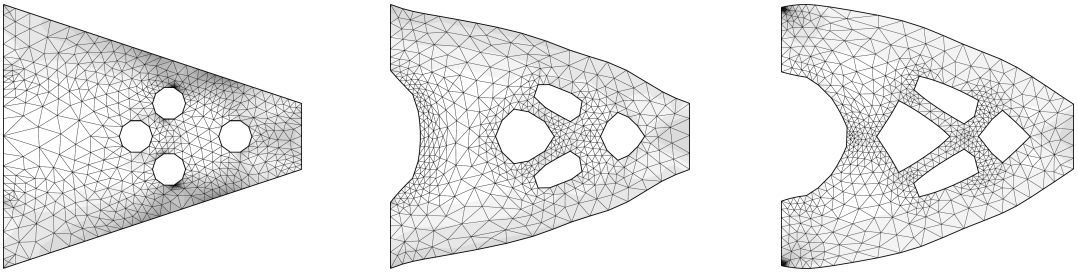


Figure 1: La méthode de variation de frontière de Hadamard : il s'agit d'évaluer la sensibilité du problème d'optimisation par rapport à de petites déformations de la forme.

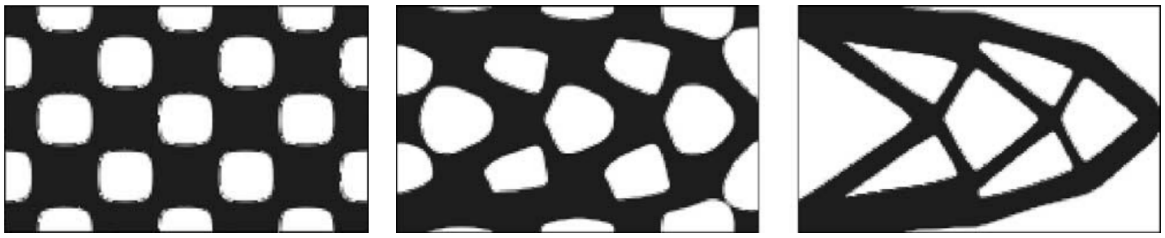
l'essor de l'informatique qui a rendu possible la simulation numérique de systèmes physiques toujours plus complexes, de nombreux développements autour de la méthode ont été effectués dans les années 1970 avec les travaux fondateurs de Céa [86, 87], Murat, Simon [241, 242], Dervieux et Palmerio [251, 252] et Pironneau [259, 260]. Ces auteurs ont été parmi les premiers à proposer des cadres théoriques et numériques pour la résolution de problèmes d'optimisation de formes contraints par des équations aux dérivées partielles (EDPs). Ces techniques ont très vite suscité un engouement dans l'industrie et ont été suivies de nombreux articles envisageant des applications physiques variées : pour la conception optimale des structures mécaniques [290, 182, 57], en théorie des coques [89], pour des problèmes de conduction thermique [121, 83, 91], et pour des applications en mécanique des fluides [90, 172, 191, 192, 226, 123].

Depuis lors, la discipline de l'optimisation de formes n'a cessé de se développer et de s'améliorer quant aux aspects théoriques [78, 79, 186] ou numériques [179, 229]. Les premiers algorithmes de calcul de formes optimales étaient basés sur la technique de déformation de maillages [260, 291, 34] : une forme candidate est discrétisée en un maillage, qui est déformé de manière itérative afin d'obtenir une géométrie améliorée. (Figure 2). Une avancée majeure a eu lieu au début des années 2000 avec l'introduction de la méthode des lignes de niveaux dans les algorithmes d'optimisation de formes [32, 311, 250]: la forme à optimiser n'est pas maillée explicitement mais plutôt capturée de manière implicite comme l'ensemble des valeurs négatives d'une fonction "*lignes de niveaux*", ce qui permet de capturer l'évolution de la forme sur un maillage fixe. Cette méthode permet l'apparition de changements topologiques complexes de la forme au cours des itérations, tels que la fusion de trous ou la fusion de frontières (Figure 3). Cette capacité de ces types d'algorithmes à gérer les changements topologiques permet de s'affranchir de tout *a priori* sur la topologie et sur la géométrie de la forme à optimiser, tel que le nombre ou la localisation des



(a) Forme initiale (b) Forme intermédiaire (c) Forme optimisée

Figure 2: La méthode de variation de frontière de Hadamard utilisant la déformation d'un maillage pour l'optimisation de la forme d'une console 2-d en flexion (figure extraite de [17]). Les changements topologiques sont difficiles à traiter numériquement : les formes initiales et finales ont le même nombre de trous.



(a) Forme initiale (b) Forme intermédiaire (c) Forme optimisée

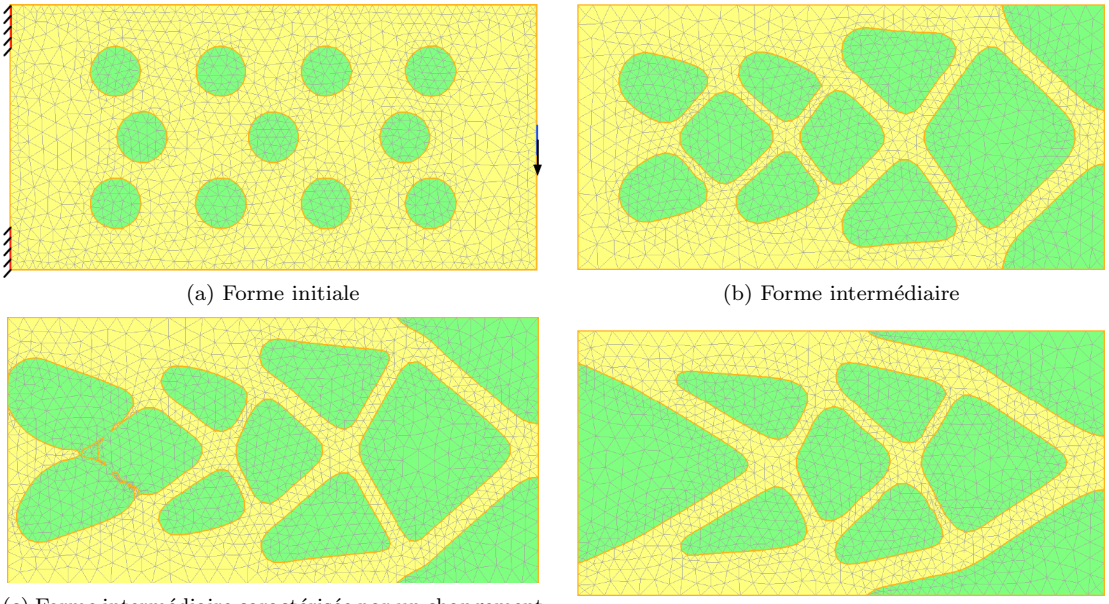
Figure 3: La méthode de variation de frontières de Hadamard utilisant la méthode des lignes de niveaux pour l'optimisation de la forme d'une console 2-d en flexion (figures extraites de [31]). Les changements topologiques sont possibles : certains trous ont fusionné entre l'initialisation et la forme finale.

trous (ou bien des arches en 3-d). Pour cette raison, les algorithmes basés sur la méthode des lignes de niveaux sont souvent considérés comme des techniques d'*optimisation topologique*, tandis que ceux basés sur la déformation d'un maillage de la forme est traditionnellement appelée *optimisation géométrique* [17]. Une tendance récente suggère d'intégrer des techniques de remaillage à la démarche d'optimisation topologique de façon à combiner les avantages des deux méthodes [24, 107]. Il est ainsi possible de garder une discrétisation explicite des formes au cours de l'optimisation tout en permettant à des changements topologiques d'avoir lieu (Figure 4).

L'optimisation topologique est une discipline très large qui bénéficie d'un large panel de techniques mathématiques et numériques. Nous pouvons mentionner, de manière non exhaustive, la méthode du gradient topologique [85, 145, 165, 289, 40], les méthodes par lignes de niveaux qui ne sont pas spécifiquement basées sur la méthode de Hadamard [133, 316], les méthodes de changement de phase [73, 299], les méthodes d'optimisations topologiques dites "évolutionnaires"' [193], et enfin les méthodes par densité [66] et par homogénéisation [18, 93] (nous renvoyons à [284] pour une revue plus complète du domaine). Parmi celles-ci, les méthodes par densité n'optimisent pas explicitement la frontière de la forme mais plutôt une fonction de densité ρ qui modélise la présence de matériau ($\rho = 1$) ou bien son absence ($\rho = 0$), ou bien des états "intermédiaires" ($0 < \rho < 1$). Des schémas numériques de pénalisation doivent être utilisés de façon à obtenir la convergence vers des formes "admissibles", c'est-à-dire des états purement "noirs et blancs" ($\rho = 0$ ou bien $\rho = 1$ comme sur la Figure 5). Les méthodes par homogénéisation remplacent quant à elles le problème d'optimisation de formes par la recherche d'une microstructure optimale pour un matériau composite effectif. Une première étape consiste à optimiser les paramètres de la microstructure tels que la densité de matière ou bien l'orientation des cellules. Dans un second temps, si aucune pénalisation n'a été utilisée, ces champs de paramètres sont interprétés afin de déterminer une forme qui approche convenablement la microstructure optimale (Figure 6) [254, 27, 166, 177].

2 Quelques enjeux actuels de l'optimisation topologique pour les applications de l'industrie aéronautique

Traditionnellement, les ingénieurs conçoivent les systèmes industriels grâce à l'aide de logiciels de Conception Assistée par Ordinateur (CAO). Les industries sont souvent dépendantes des formats CAO du fait de leur compatibilité avec toutes les étapes de conception, de la simulation numérique des pro-



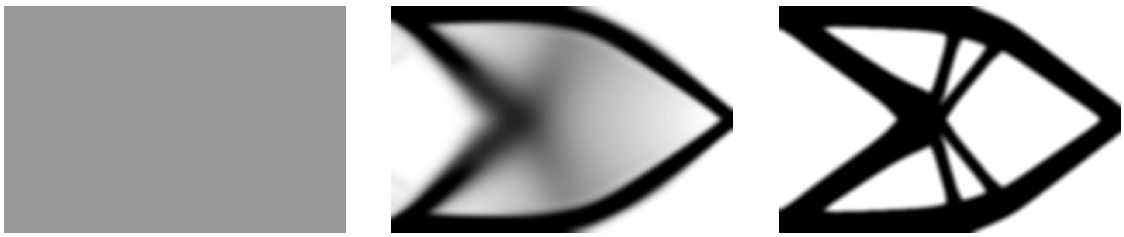
(a) Forme initiale

(b) Forme intermédiaire

(c) Forme intermédiaire caractérisée par un changement topologique

(d) Forme finale

Figure 4: La méthode de variation de frontière de Hadamard implémentée en utilisant la méthode des lignes de niveaux combinée à une technique de remaillage pour l'optimisation d'une console 2-d en flexion (figures extraites de [107]). Les changements topologiques sont permis tout en conservant une discrétisation explicite et conforme des formes.

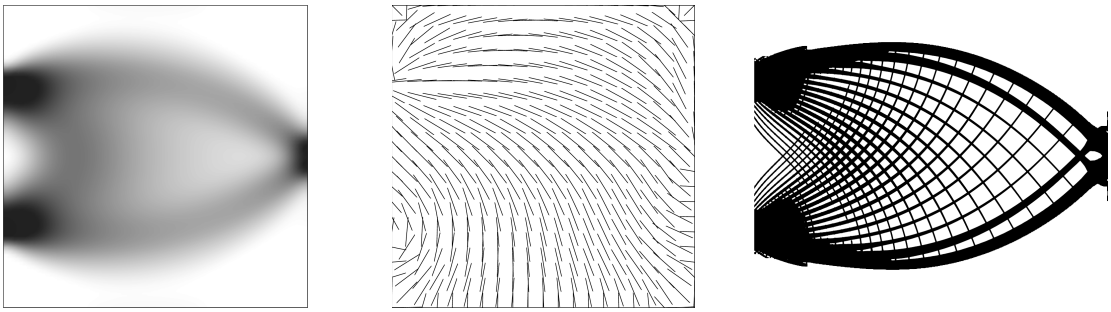


(a) Design initial

(b) Design intermédiaire

(c) Design optimisé

Figure 5: Optimisation topologique d'une console 2-d en flexion par une méthode de densité (ici la méthode SIMP—Solid Isotropic Material Penalization—). Figure obtenue à partir du code source décrit dans [43].



(a) Densité optimale

(b) Orientation optimale

(c) Forme interprétée

Figure 6: Optimisation topologique d'une console 2-d en flexion par une méthode d'homogénéisation. Figure extraite de [166].

cessus physiques par des codes industriels à la fabrication effective par des machines automatisées. Lors de l'optimisation d'une forme au moyen d'outils CAO, la topologie du système à concevoir est généralement proposée *a priori* par l'ingénieur. Elle est par exemple déterminée par un ensemble de points de contrôles de courbes de Bézier, par des paramètres de splines, ou bien par d'autres types de paramètres géométriques. Une optimisation paramétrique de ces paramètres de conception est alors effectuée pour améliorer le design initial [181]. Habituellement, la sensibilité du problème par rapport à la géométrie n'est pas connue de manière analytique. Celle-ci est parfois estimée par différentiation automatique (un outil très populaire en conception aérodynamique [94, 139, 130]) [235, 288].

Cependant, dans de nombreux cas, l'optimisation est effectuée par l'exploration d'un sous-ensemble suffisamment vaste de l'espace des paramètres appelé "plan d'expériences" (cela est fait par exemple avec le logiciel commercial Optimus [245]), voir par exemple [142, 286]. Puisque ces méthodes reposent sur

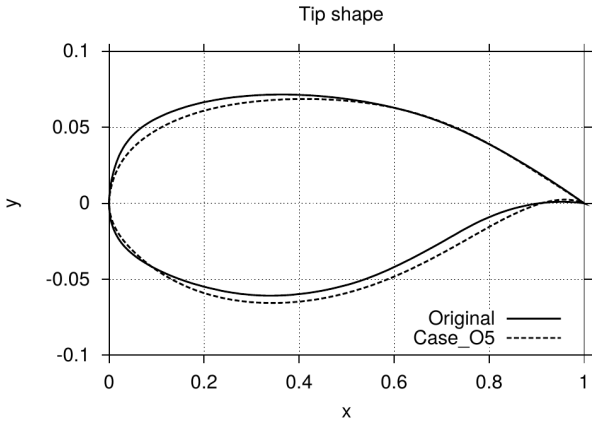


Figure 37. Tip wing section. Original vs optimized ones.

Figure 7: Optimisation de la CAO d'un profil aérodynamique pour le couple portance–traînée. Figure extraite de [142].

des choix importants de paramétrisation de la géométrie, elles ne conduisent généralement qu'à de très faibles modifications de la forme initialement proposée (Figure 7). En principe, l'optimisation topologique pourrait permettre d'obtenir des designs aux performances sensiblement meilleures, parmi un ensemble beaucoup moins restreint de formes admissibles.

Les développements de la fabrication additive depuis les années 1990 permettent aujourd'hui à l'industrie de fabriquer des systèmes à la géométrie toujours plus complexe, difficilement décrite par une paramétrisation CAO ; ceci suscite aujourd'hui un enthousiasme renouvelé pour l'optimisation topologique. Un grand nombre d'applications envisagées dans les travaux de recherche actuels sont issus de l'industrie aéronautique. Celles-ci constituent quelques-unes des motivations principales à l'origine de ce travail, par exemple lorsqu'il est question de la conception des composants de moteurs d'avion. Un des enjeux majeurs, en ce qui concerne l'application de l'optimisation topologique aux systèmes aéronautiques, est la nécessité de prendre en compte leur caractère intrinsèquement multi-physique : il est bien souvent souhaitable d'optimiser de tels systèmes au regard de critères impliquant des propriétés thermiques, mécaniques et hydrauliques couplées.

Un problème d'actualité est celui de l'optimisation topologique appliquée à la conception d'échangeurs de chaleur [255, 258, 189, 271, 275]. Les échangeurs de chaleurs sont des dispositifs utilisés dans les moteurs pour refroidir des fluides chauds en les transportant à proximité de certains gaz ou liquides réfrigérants. Les modèles industriels présentent généralement un assemblage de nombreux tubes et ailettes afin de maximiser la surface d'échange entre les phases chaudes et froides (Figure 8). Naturellement, il est aussi nécessaire de prendre en compte un ensemble de contraintes multi-physiques lors de la conception, telles que la perte de charge (la différence de pression entre l'entrée et la sortie du dispositif), ou bien la résistance des structures mécaniques en présence d'une charge thermique élevée.

De nombreux autres composants des moteurs aéronautiques pourraient certainement bénéficier de l'optimisation de formes, à commencer par le système de refroidissement par canaux internes des aubes de turbines [62, 162] (Figure 9).

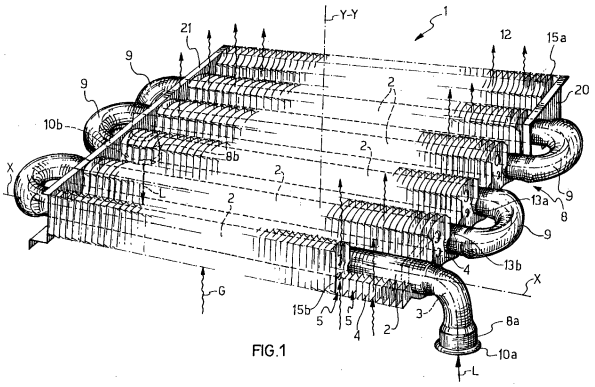


Figure 8: Échangeur de chaleur industriel liquide–gaz. Figure extraite de [169].

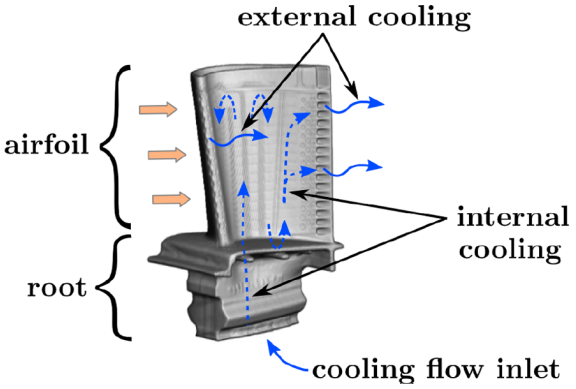


Figure 9: Système de refroidissement d'aubes de turbines avec présence de canaux internes réfrigérants. Figure extraite de [162].

3 Cadre de la thèse

À ce jour, la méthode de Hadamard utilisée à des fins d'optimisation topologique (dans laquelle on autorise des déformations libres de la forme optimisée) n'est pas encore suffisamment mature pour être intégrée à de véritables applications industrielles et multi-physiques. La plupart des cas tests traités dans la littérature se placent le plus souvent dans le cadre de l'élasticité linéaire dans [25, 108], ou dans celui de la conduction thermique [9, 325], et bien peu d'exemples numériques tridimensionnels, à résolution élevée, sont présentés [202]. Seulement récemment, la méthode a été appliquée à des physique plus complexes telles que la plasticité mécanique [230] ou bien l'électromagnétisme [160, 213]. Cette observation s'applique également à la communauté d'optimisation par méthodes de densités qui présente un nombre croissant de récents travaux portant sur des applications multidisciplinaires [116, 129, 128, 318, 319, 240, 288].

L'objectif de ce travail est de développer des outils théoriques et numériques pour l'optimisation de la forme et de la topologie de systèmes multi-physiques, guidé par la prise en compte d'un certain nombre d'exigences industrielles à long terme. À cet effet, nous utilisons principalement la méthode de variation de frontière de Hadamard ; un chapitre d'ouverture (le chapitre 7) est dédié à des développements mathématiques sur l'homogénéisation des milieux poreux qui pourraient permettre d'envisager des applications nouvelles pour l'optimisation de systèmes impliquant des fluides.

Par "exigences industrielles", nous entendons plusieurs besoins industriels qui sont autant d'axes majeurs pour notre travail et que nous décrivons dans les cinq prochains paragraphes. Les contributions de la thèse sont ensuite résumées chapitre par chapitre dans la section qui suit.

Systèmes multi-physiques

Comme en témoignent les applications industrielles évoquées dans la la section précédente, il existe une demande croissante en ce qui concerne l'optimisation topologique pour des applications mêlant plusieurs physiques en interaction. Notre étude se concentre sur des systèmes fluides présentant des propriétés hydrauliques, thermiques et mécaniques couplées. Ces derniers sont caractérisés par les variables physiques suivantes:

- $\boldsymbol{v}$ et p désignent les champs de vitesse et de pression associés à un ou plusieurs fluides évoluant à travers le système ;
- T pour le champs de température dans les phases solides et liquides ;
- $\boldsymbol{u}$ pour le champs de déplacement élastique des structures mécaniques solides mises en jeu dans le système.

Mathématiquement, ces variables sont déterminées par la résolution d'un système d'équations aux dérivées partielles, ces dernières provenant elles-mêmes du choix de la modélisation physique. Un cadre académique suffisamment représentatif pour nos applications, est décrit dans le chapitre 2.

En termes d'applications, une partie de notre travail est guidée par la conception d'échangeurs de chaleur. Une étude numérique dans ce contexte est proposée dans le chapitre 5, où nous optimisation la forme et la topologie de deux modèles d'échangeurs de chaleur bidimensionnels.

Optimisation sous contraintes

Lors de la conception d'un système industriel, il est d'usage de prendre en compte un cahier des charges dont les spécifications doivent être satisfaites dans les conditions d'utilisation réelles. Par exemple, les contraintes mécaniques ou thermiques d'une structure solide ne doivent pas dépasser un seuil donné afin d'éviter un endommagement prématuré. En d'autres termes, un problème de conception optimale consiste à déterminer la forme du système qui réalise la meilleure performance pour un ensemble donné de contraintes physiques à respecter.

De tels problèmes sont modélisés de manière générique par des programmes mathématiques de la forme :

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \\ s.t. \quad & \begin{cases} g_i(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = 0, & 1 \leq i \leq p, \\ h_j(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \leq 0, & 1 \leq j \leq q, \end{cases} \end{aligned} \tag{1}$$

où Γ est la frontière du système à optimiser, typiquement l'interface entre une structure solide et un fluide (ce qui laisse la possibilité de rendre compte de plusieurs phases liquides). J est une fonction de coût à minimiser qui quantifie la performance du système. Par exemple, la compliance d'une structure mécanique, la force de traînée générée par un obstacle, ou encore la chaleur emmagasinée par un système thermique sont des fonctions de coût classiques couramment rencontrées en optimisation de la forme. Les fonctions g_i et h_j modélisent des contraintes d'égalité et d'inégalité: elles renvoient à des valeurs cibles que certaines quantités physiques doivent atteindre (par exemple, un emplacement souhaité pour le centre de gravité de la structure) ou bien ne doivent pas dépasser (par exemple, une limite imposée sur la température du système, ou bien une taille caractéristique minimale). La fonction objectif J et les contraintes g_i, h_j dépendent à la fois de la forme Γ et des variables physiques $\boldsymbol{v}(\Gamma), p(\Gamma), \boldsymbol{u}(\Gamma), T(\Gamma)$, qui dépendent elles-mêmes de Γ via les équations d'états.

Dans le contexte des méthodes de densité, la variable optimisée ρ appartient à un espace vectoriel de type $\mathbb{R}^n$ après discrétisation, ce qui permet de résoudre des programmes mathématiques contraints par des méthodes très classiques d'optimisation du premier ordre [244, 298, 310]. Dans le cadre de la méthode de Hadamard, la variable d'optimisation Γ n'appartient pas à un espace vectoriel mais plutôt à une "variété" de dimension infinie ce qui rend la résolution de (1) beaucoup plus délicate. La plupart des travaux de la littérature s'appuyant sur la méthode de Hadamard convertissent (1) en un problème de minimisation sans contrainte via l'ajout de termes de pénalisation à la fonction objectif J; une technique couramment employée à cet égard est la méthode du Lagrangien augmenté [234, 107]. Des méthodes plus complexes, telles que l'optimisation linéaire successive (*Sequential Linear Programming, SLP*) ou la méthode des directions admissibles (*Method of Feasible Directions, MFD*) ont été utilisées avec succès dans un nombre restreint de travaux [135, 150]. Cependant, ces méthodes ne sont pas complètement satisfaisantes pour une utilisation industrielle car elles nécessitent un réglage fin de paramètres d'optimisation qui dépendent fortement de la situation considérée et qui peuvent être difficiles à ajuster, surtout en présence d'un grand nombre de contraintes.

Dans le chapitre 3, nous proposons un nouvel algorithme de type gradient pour la résolution de problèmes d'optimisation sous contraintes adapté au contexte de la méthode de Hadamard.

Algorithmes non-intrusifs

Dans un souci de compatibilité avec les outils industriels, le processus d'optimisation de formes doit être entièrement compatible avec la chaîne de conception du système. Par exemple, les outils d'optimisation topologique doivent idéalement pouvoir évaluer les variables d'état en faisant appel aux mêmes solveurs physiques que ceux utilisés lors de la phase de validation industrielle. De nombreuses méthodes d'optimisation topologique, notamment les méthodes de densité ou de lignes de niveaux, ne remplissent pas cette condition car elles reposent sur des modifications artificielles des équations d'état pour des raisons purement numériques.

Ce besoin constitue l'une des raisons de notre utilisation systématique pour tous nos cas tests numériques de l'algorithme d'évolution de maillages proposé par [25]. Cet algorithme illustré sur la Figure 4 et ses grands principes sont rappelés dans le chapitre 1. Cette technique, basée sur la méthode

des lignes de niveaux, permet de conserver une représentation maillée explicite de la forme optimisée durant toutes les itérations du processus d'optimisation, ce qui permet de résoudre les équations d'état sans aucune modification ou approximation de la physique considérée. Cette étape de résolution pourrait en principe, être effectuée par un solveur tiers. Pour nos applications, nous avons utilisé de manière intensive le logiciel open source `mmg` [108] pour les étapes de remaillage 2-d ou 3-d.

Une nécessité supplémentaire imposée par le souhait de non-intrusivité des algorithmes d'optimisation de formes est la minimisation de la quantité d'information demandée à l'utilisateur. En particulier, un utilisateur formulant un problème du type (1) devrait pouvoir renseigner des fonctions de coûts et des contraintes arbitraires sans avoir à connaître les détails mathématiques (plutôt techniques) du calcul des dérivées de forme au sens de Hadamard. Dans le chapitre 2, nous établissons des formules pour les dérivées de formes de fonctionnelles de coût *arbitraires*, qui ont l'avantage de pouvoir être assemblées numériquement de manière automatique à partir de de la seule connaissance des dérivées partielles (qui classiquement, sont très simples à calculer).

Contraintes géométriques

La prise en compte de contraintes géométriques quant aux formes optimisées est un besoin industriel classique en optimisation topologique du fait de la précision limitée des processus de fabrication. Elles peuvent par exemple se manifester par des contraintes sur l'épaisseur minimale d'une structure ou bien sur l'angle maximal de ses parties en porte-à-faux avec la direction verticale. Elles se présentent également lors de la conception des échangeurs de chaleur bi-tubes : dans le chapitre 5, nous modélisons une condition de non-mélange de deux phases fluides par une contrainte de distance géométrique.

De nombreux travaux ont analysé l'intégration de telles contraintes en optimisation topologique, voir par exemple [234], chapitre 3, pour une revue. Dans le contexte de l'optimisation de formes par la méthode de Hadamard, il est commode de modéliser les contraintes géométriques au moyen de la fonction de distance signée [30]. Cependant, l'implémentation pratique des dérivées de forme associées est très délicate et dépend sensiblement du type de discrétisation envisagée pour la forme optimisée. Dans le chapitre 4, nous proposons une nouvelle méthode variationnelle qui facilite l'évaluation numérique des dérivées de forme pour les contraintes géométriques de distance, et qui peut être implémentée de manière très commode par la méthode des éléments finis.

Calcul haute performance

Une dernière exigence industrielle importante concerne l'application de l'optimisation topologique à des problèmes physiques 3-d et discrétisé par des maillages ayant des résolutions élevées. Lors d'un processus d'optimisation, les équations d'état doivent être résolues de l'ordre d'une à plusieurs centaines de fois, ce qui est très coûteux en temps de calcul pour les problèmes dits de "grande échelle" (pour lesquels l'étape de résolution requiert la détermination d'un très grand nombre de variables). Ce coût est d'autant plus élevé lorsque les physiques mises en jeu comprennent des écoulements de fluides. Cette question a été assez largement traitée par la communauté utilisant les méthodes de densité [1, 10, 71, 232, 39], et elle commence à être également abordée dans le cadre des algorithmes d'optimisation topologiques dits "évolutionnaires" [227] ainsi que de ceux basés sur la méthode des lignes de niveaux [202]. Dans le chapitre 6, nous présentons et discutons notre implémentation en `FreeFEM` [183] de cas tests 3-d présentant des tailles de maillages relativement importantes (moyennant la résolution de systèmes linéaires jusqu'à 2 millions de degrés de liberté). À cet effet, nous exploitons le calcul parallèle et des techniques récentes de décomposition de domaines associées à des solveurs itératifs préconditionnés [238].

Dans un autre registre, les modèles industriels classiques de conceptions des échangeurs de chaleur (tel que celui illustré sur la Figure 8) suggèrent le besoin de générer des designs présentant des motifs multi-échelles. Dans le chapitre 7, nous proposons des modèles homogénéisés d'ordres supérieurs pour les écoulement de fluides en milieux poreux, qui pourraient permettre d'envisager de traiter ces applications par les méthodes d'optimisation topologique par homogénéisation dans de tels contextes.

4 Résumé par chapitre

Chapitre 1 : optimisation de formes par la méthode de variation de frontière de Hadamard

Ce chapitre préliminaire présente le contexte et le matériel technique de base concernant l'optimisation de formes par la méthode de Hadamard. Nous décrivons les ingrédients classiques requis pour la mise en œuvre d'un processus d'optimisation topologique via la méthode d'évolution de maillages de [25].

Plus précisément, nous commençons par un survol de la méthode de Hadamard et des techniques mathématiques classiques qui permettent de calculer les dérivées de forme de fonctions de coût dépendant de solutions d'équations aux dérivées partielles. Nous discutons ensuite du problème classique de l'adaptation de l'algorithme d'optimisation par la méthode du gradient au contexte de la dimension infinie. Une partie indépendante est dédiée à des rappels sur les définitions et les principales propriétés de la fonction de distance signée. Enfin nous résumons les différentes étapes de la méthode d'évolution de maillage de Allaire, Dapogny et Frey [25] pour l'optimisation de formes.

Chapitre 2 : Dérivées de formes au sens de Hadamard pour un problème hydraulique thermique mécanique faiblement couplé

Ce chapitre présente un modèle thermique fluide-structure simplifié qui fait office de référence pour toutes nos applications numériques: les variables $\boldsymbol{v}, p, T, \boldsymbol{u}$ sont définies comme les solutions de trois équations d'état qui sont faiblement couplées dans le sens où ces dernières peuvent être résolues successivement et indépendamment. Nous considérons les équations de Navier-Stokes incompressibles et stationnaires pour la vitesse et la pression du fluide, l'équation de convection-diffusion pour le champ de température au sein des phases fluides et solides, et le système de la thermo-élasticité linéaire pour le déplacement élastique de la structure solide. Ce modèle est une généralisation de plusieurs situations impliquant une seule physique qui sont couramment étudiées dans la littérature ; ils permettent en outre de traiter de nouveaux cas test couplés tels que l'interaction fluide-structure dans un régime de petites déformations, ou bien le transfert de chaleur par convection.

Une nouveauté importante proposée par cette partie est la détermination de formules de dérivées de forme pour des fonctions de coût *arbitraires*. Dans la littérature consacrée à la méthode de Hadamard, de telles formules reposent en effet sur des hypothèses sur la structure de la fonction de coût considérée. Ainsi, il s'avère nécessaire de refaire le calcul des dérivées de forme lors de chaque changement de fonction objectif ou contrainte, ce qui va à l'encontre d'une intégration systématique dans des logiciels industriels. Nos formules permettent au contraire d'automatiser le calcul des dérivées de forme : celles-ci sont assemblées après la résolution d'équations adjointes à partir de la seule connaissance des dérivées *partielles* de la fonction de coût, une information qui peut en principe être fournie facilement par un utilisateur externe.

Une autre contribution de ce chapitre concerne la prise en compte spécifique des problèmes d'interaction fluide-structure. Une propriété surprenante réside dans le couplage du système adjoint : l'égalité des contraintes normales sur l'interface fluide-structure (une condition de type "Neumann") se traduit par une égalité des valeurs des variables adjointes sur cette frontière (une condition de type "Dirichlet").

La validité de nos formules est enfin vérifiée numériquement sur plusieurs cas tests à deux dimensions impliquant simultanément une, deux ou trois des physiques mentionnées.

Le contenu de ce chapitre a été publié pour l'essentiel dans l'article [153]:

F. FEPPON, G. ALLAIRE, F. BORDEU, J. CORTIAL, AND C. DAPOGNY, *Shape optimization of a coupled thermal fluid–structure problem in a level set mesh evolution framework*, SeMA Journal, (2019), pp. 1–46.

Mentionnons cependant que nous avons ajouté de nombreux exemples numériques supplémentaires et amélioré certains cas tests de la publication [153].

Chapitre 3 : flots de gradient à noyaux pour l'optimisation de formes sous contrainte

Ce chapitre présente l'algorithme d'optimisation que nous avons spécifiquement développé dans la perspective de son application en optimisation de formes. Nous proposons un schéma d'optimisation qui peut être interprété comme la discrétisation d'un flot de gradient capable de "voir" les contraintes d'égalité et d'inégalité (il s'agit d'une généralisation de certains algorithmes d'optimisation par systèmes dynamiques [317, 300]). Ce flot de gradient est appelé *à noyau* car la direction d'optimisation s'écrit comme la somme d'un vecteur appartenant à l'espace tangent aux contraintes (*i.e.* au noyau de leur différentielle) qui vise à faire décroître les valeurs de la fonction objectif, et d'un vecteur orthogonal au premier et qui ramène le chemin d'optimisation à l'intérieur du domaine admissible. Le calcul de ces vecteurs dépend de la résolution d'un sous-problème quadratique qui indique quand décoller de la frontière des contraintes pour revenir à l'intérieur du domaine d'optimisation (Figure 10). Une caractéristique de l'algorithme est son aptitude à diminuer les valeurs de la fonction objectif tout en maintenant les contraintes satisfaites.

Nous établissons des propriétés de convergence de notre algorithme pour les trajectoires d'optimisation continues ; celles-ci garantissent l'amélioration de solutions optimisées jusqu'à ce qu'un minimum local

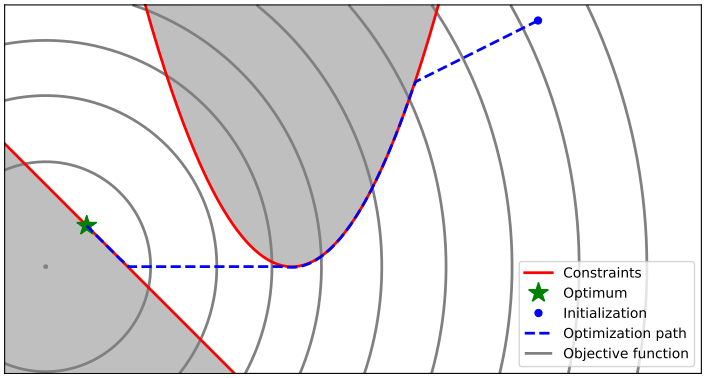


Figure 10: Illustration de notre flot de gradient à noyau pour la résolution des problèmes d'optimisation sous contraintes. Les trajectoires d'optimisation suivent toujours la meilleur direction de descente possible.

soit trouvé, à une discrétisation suffisamment fine près. Nous illustrons la robustesse et l'efficacité de notre méthode d'optimisation de formes sur un cas test de pont en chargement multiple présentant 10 contraintes d'optimisation.

La plupart du contenu de ce chapitre est à paraître dans la prépublication soumise [155]:

F. Feppon, G. Allaire, and C. Dapogny, *Null space gradient flows for constrained optimization with applications to shape optimization*, submitted, (2019).

Le chapitre contient en outre quelques comparaisons de notre méthode avec des algorithmes plus classiques sur des exemples académiques.

Chapitre 4 : une méthode variationnelle pour calculer les dérivées de formes de contraintes géométriques le long des rayons

Les applications envisagées de l'optimisation topologique aux échangeurs de chaleur bi-tubes nécessitent de prendre en compte une contrainte de non-mélange des phases fluides misent en jeu lors de la résolution du problème d'optimisation. Une approche très naturelle pour ce faire consiste à prescrire une distance minimale entre les deux phases non miscibles, ce qui peut être formulé mathématiquement grâce à la fonction de distance signée. Plus généralement, de nombreuses autres contraintes géométriques peuvent être formulées de manière analogue, telles que les contraintes d'épaisseurs minimales ou maximales.

Des travaux précédents [30, 234] ont proposé des expressions mathématiques pour calculer les dérivées de formes de fonctionnelles de coût faisant intervenir la fonction de distance signée, afin d'imposer des contraintes géométriques telles que l'épaisseur maximale ou minimale de la forme optimisée. Cependant, la mise en œuvre directe de ces formules est difficile car elle nécessite une intégration numérique le long des rayons normaux au bord de la forme (Figure 11), ainsi que des estimations précises du squelette et des courbures principales. L'implémentation de ces opérations est notoirement difficile et est très dépendante de la dimension (2-d ou 3-d) ainsi que du type de discrétisation envisagé pour capturer la forme optimisée (implicite ou explicite, sur maillage structuré ou non structuré).

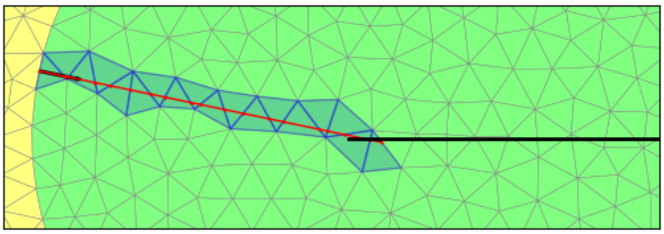


Figure 11: Les méthodes classiques pour calculer les dérivées de formes des contraintes géométriques (impliquant la fonction de distance signée) requièrent le calcul des rayons normaux à la forme (en *rouge*), ce qui nécessite de parcourir la discrétisation de maille en maille.

Dans ce chapitre, nous montrons que les dérivées de forme de contraintes géométriques peuvent être calculées au moyen de la résolution d'un problème variationnel, qui peut être aisément effectuée par la

méthode des éléments finis : la valeur des intégrales le long des rayons est obtenue à partir des valeurs de cette solution variationnelle aux sommets de la frontière de la forme. En toute généralité, notre méthode permet de calculer des intégrales le long des courbes caractéristiques d'un champ de vitesse β sans avoir à calculer explicitement ces courbes sur la discrétisation spatiale. L'implémentation s'en trouve considérablement simplifiée car elle ne nécessite pas le calculs de rayons, ni celui des courbures de la forme, mais uniquement la détermination d'un poids relativement arbitraire s'annulant approximativement sur le squelette.

Nous établissons le caractère bien posé de la formulation variationnelle proposée grâce à une analyse détaillée d'espaces à poids associés à l'opérateur d'advection $\beta \cdot \nabla$. Dans le contexte de l'optimisation de formes, β est le gradient de la fonction de distance signée au domaine sur lequel on souhaite imposer des contraintes géométriques, mais notre analyse permet de traiter également des champs de vecteurs plus généraux. Notre approche permet notamment de traiter des champs de vitesse ayant une divergence non bornée: les travaux classiques effectuant des analyses similaires [144] supposent généralement $\mathrm{div}(\beta) \in L^\infty(D)$ où D désigne le domaine de travail. Cette hypothèse est cependant systématiquement violée pour les applications considérées en optimisation de formes.

Nous montrons enfin l'efficacité de cette méthode variationnelle en revisitant l'implémentation des contraintes d'épaisseurs maximale et minimale en optimisation de formes. Nous retrouvons des résultats numériques analogues à ceux obtenus dans les travaux antérieurs [30, 234] s'appuyant sur le calcul explicite (difficile) des intégrales le long des rayons.

La majeure partie de ce chapitre est à paraître dans la publication [154]:

F. FEPPON, G. ALLAIRE, AND C. DAPOGNY, *A variational formulation for computing shape derivatives of geometric constraints along rays*, To appear in M2AN, (2019).

Chapitre 5 : optimisation topologique d'échangeurs thermiques 2-d

Les ingrédients développés dans les parties précédentes sont appliqués dans ce chapitre au problème de l'optimisation d'échangeurs de chaleur bidimensionnels.

Dans une première partie, nous mettons en œuvre nos méthodes pour l'optimisation topologique d'échangeurs de chaleur bi-tube dont le comportement physique est décrit par le modèle faiblement couplé du chapitre 2. Nous formulons une contrainte de non mélange des deux phases liquides au moyen d'une contrainte de géométrique impliquant la fonction de distance signée à l'une des phases, que nous traitons par la méthode variationnelle du précédent chapitre 4 (Figure 12a).

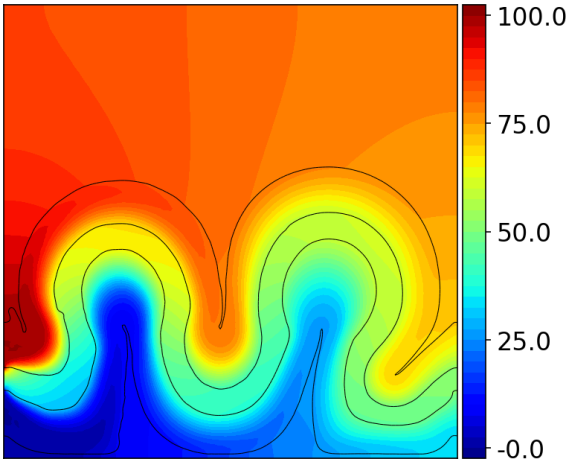
Dans une seconde partie, nous rendons compte de la réalisation d'un cas test d'optimisation de formes pour des échangeurs de chaleur 2-d issu d'une collaboration avec Safran Aero Boosters : l'objectif est de déterminer la forme optimale de la section de canaux d'huiles transverses et réfrigérés par un flux d'air froid en entrée. La physique de ce cas test est modélisée par les équations d'états introduites dans le chapitre 2, à un changement de condition aux limites près pour le champ de température. Nous avons également intégré une contrainte d'épaisseur minimale pour la phase huileuse, et une contrainte sur la perte de charge maximale pour la phase constituée d'air. Notre étude démontre les capacités de notre méthodologie numérique d'optimisation de formes pour générer des designs non classiques dans un contexte simplifié (Figure 12b).

Chapitre 6 : vers des applications 3-d et industrielles : stratégies d'implémentation pour une variété de cas tests numériques

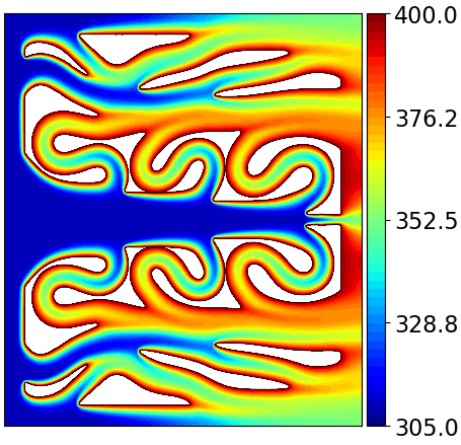
Ce chapitre traite de l'implémentation de cas tests numériques qui approchent des applications industrielles plus avancées.

Une première section décrit succinctement certains paradigmes d'implémentation de notre code d'optimisation topologique développé en `python` et `FreeFEM` [183] pour les applications de cette thèse. Nous discutons ensuite des détails d'implémentation spécifiques à la 3-d, avec en particulier l'utilisation inévitable de techniques de décomposition de domaine associées à des préconditionneurs spécifiques aux physiques considérées.

Dans une seconde partie, nous présentons plusieurs nouveaux résultats d'optimisation de formes tridimensionnelle pour différentes physiques. Nos cas tests considèrent l'optimisation de la forme de consoles mécaniques en traction ou torsion, d'une distribution de matériaux pour la conduction thermique (Figure 13), de profils aérodynamiques permettant de générer une portance maximale tout en

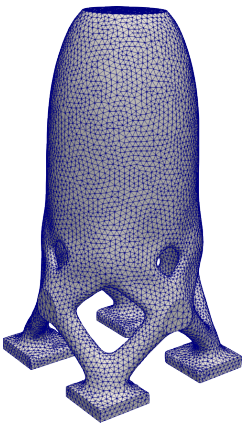


(a) Échangeur de chaleur 2-d avec contrainte de non-mélange.

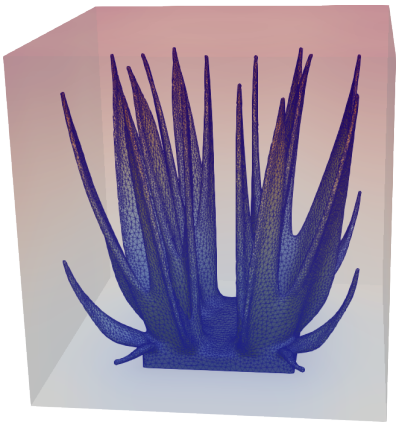


(b) Échangeur de chaleur 2-d issu d'un cas test proposé par SAFRAN Aero Boosters.

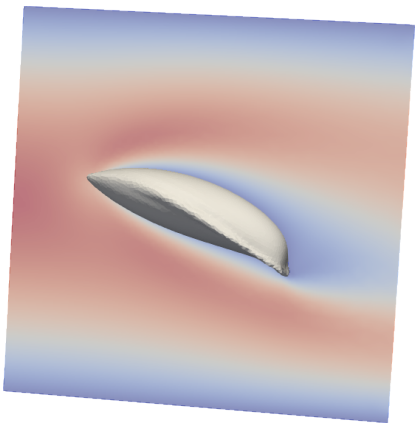
Figure 12: Deux cas tests d'optimisations considérés dans le chapitre 5.



(a) Structure mécanique



(b) Design optimisé pour la conduction thermique



(c) Design aérodynamique

Figure 13: Quelques formes optimisées 3-d obtenues dans le chapitre 6.

limitant les efforts de frictions. Nous traitons enfin un cas test plus complexe d'un système en interaction fluide-structure. Grâce aux travaux récents de Moulin et. al. [238], nous avons été en mesure de simuler la mécanique des fluides de systèmes discrétisés par des maillages présentant jusqu'à deux millions d'éléments.

Chapitre 7 : équations homogénéisées d'ordres élevés pour les problèmes perforés, vers l'optimisation topologique des fluides par la méthode d'homogénéisation

Ce dernier chapitre constitue une ouverture de la thèse vers l'utilisation de la méthode d'homogénéisation pour l'optimisation topologique de systèmes multiphysiques comportant des phases fluides.

Il est établi qu'en général, les solutions des problèmes d'optimisation de formes sont des matériaux composites. La théorie de l'homogénéisation permet de caractériser la physique effective des matériaux obtenus par des mélanges arbitrairement complexes de deux phases homogènes. Dans le contexte de la conception optimale des structures, celle-ci donne a donné lieu à des méthodes efficaces pour générer des formes approchant les microstructure optimales [27, 254] (telle que celle illustrée sur la Figure 6). Pour des applications en mécanique des fluides, la théorie de l'homogénéisation paraît *a priori* plus difficilement applicable : la littérature classique [103, 11, 15, 265] identifie trois régimes possibles de modèles homogénéisés, déterminés par le rapport d'échelles entre la taille et l'écartement des obstacles périodiques considérés. Aujourd'hui les méthodes d'optimisation topologique par densité utilisent quasi-exclusivement le modèle de Brinkman, qui se trouve être l'un des trois régimes homogénéisés possibles.

Dans cette partie, nous proposons des équations homogénéisées d'ordres élevés pour le système des équations de Stokes en milieu poreux qui permettent de capturer ces trois régimes simultanément. Ces modèles homogénéisés pourraient être utilisés pour l'optimisation topologique de systèmes fluides grâce à l'approche numérique proposée par [254, 27]. Notre motivation originale provient de la complexité des échangeurs thermiques industriels, dont la forme combine une silhouette macroscopique avec des motifs "microscopiques" en fines ailettes périodiques. L'optimisation de formes "classique" reposant par exemple sur la méthode de Hadamard permet de déterminer des formes optimisées *macroscopiques*, mais elle n'est pas très bien adaptée au cas où les géométries optimales sont caractérisées par de tels motifs multi-échelles.

L'un de nos résultats principaux montre que les trois régimes homogénéisés classiques peuvent être capturés par une unique équation homogénéisée d'ordre élevé : chacun des trois modèles est obtenu en passant à la limite de faible fraction volumique pour des régimes particuliers de taille d'obstacle. Plus généralement, des équations homogénéisées bien posées sont déterminées à tout ordre grâce à une méthode inspirée des travaux de Bakhvalov et Panasenko [53], Smyshlyaev et Cherednichenko [287] et Allaire, Lamacz et Rauch [33].

D'un point de vue pédagogique, nous calculons ces équations homogénéisées pour plusieurs problèmes elliptiques perforés successifs (dont les solutions s'annulent sur les petits obstacles distribués périodiquement) présentant un ordre de difficulté croissant. Nous considérons tout d'abord le cas du problème (scalaire) de Poisson perforé, puis nous étendons nos résultats au cadre vectoriel par la considération du problème analogue en élasticité. Nous étudions enfin le problème de Stokes pour lequel le terme de pression requiert un travail additionnel. Une caractéristique surprenante est l'apparition d'opérateurs *d'ordre impair* très étranges dans les équations homogénéisées d'ordres élevés : ce résultat n'est pas standard et ne semble pas avoir été observé dans la littérature classique proposant des termes correcteurs pour ces modèles.

Ce chapitre final ne contient pas de résultats numériques ; ceux-ci feront l'objet de futurs travaux.

INTRODUCTION – SUMMARY OF THE THESIS

This work is devoted to shape and topology optimization of multiphysics systems motivated by applications of the aeronautic industry. Due to several industrial constraints that are detailed further on, we primarily consider the method of Hadamard for such purposes, although topology optimization by the homogenization method is also investigated in a final independent chapter.

Shape optimization is the mathematical art of generating shapes that best fulfill a proposed objective. Industrial applications are generally concerned with the issue of determining physical systems that achieve optimal performance. What is the shape of the most rigid structure for a given prescribed amount of material? What is the most aerodynamic airfoil generating a prescribed lift force? Which distribution of fluid pipes does achieve the best heat transfer ? How to design the shape of a ship hull? These are very classical questions of industrial interest.

1 From shape to topology optimization: Hadamard's, density and homogenization methods

The method of Hadamard can be traced back to the 1908 seminal memoir [180]. It consists in computing the sensitivity of the problem to small deformations of the boundary (Figure 14), which allows to optimize shapes by finding particular deformations which gradually yield better designs. With the advent of

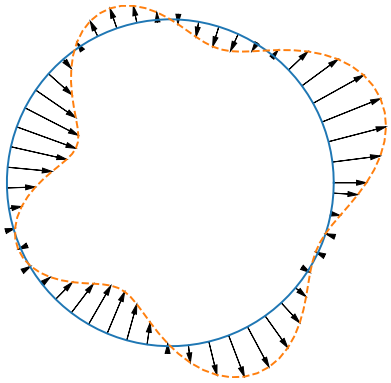
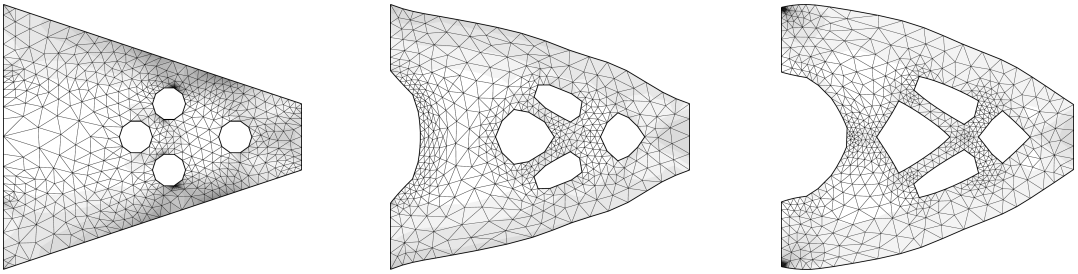


Figure 14: The boundary variation method of Hadamard: evaluating the sensitivity of the problem with respect to small shape deformations.

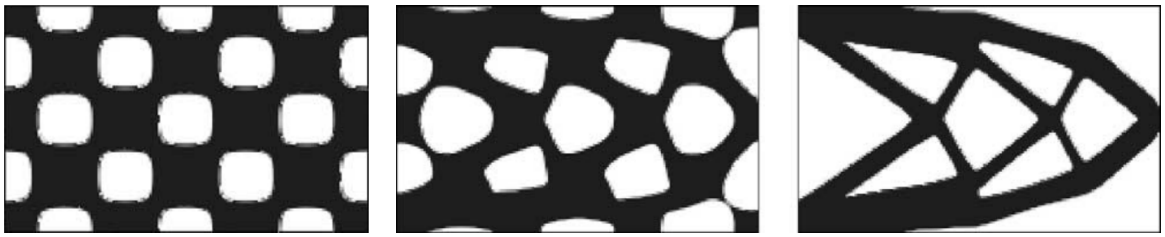
computers enabling to simulate numerically larger and larger physical systems, the method started to be the object of many developments in the 1970s with the seminal works of Céa [86, 87], Murat, Simon [241, 242], Dervieux and Palmerio [251, 252] and Pironneau [259, 260]. These authors were among the first to develop numerical and theoretical frameworks for solving shape optimization problems constrained by partial differential equations (PDEs). This very soon raised a growing interest in the industry with the publication of many works devoted to a variety of physical applications, dealing for instance with optimal design in mechanics [290, 182, 57], shell theory [89], heat conduction problems [121, 83, 91] and fluid mechanics [90, 172, 191, 192, 226, 123].

Since then, the field has kept improving both on theoretical [78, 79, 186] and computational aspects [179, 229]. The first numerical algorithms were based on mesh deformations [260, 291, 34]: a proposed initial guess is discretized by mean of a mesh which is deformed to a better shape (Figure 15). A major breakthrough arose in the beginning of the 2000s with the introduction of level set methods in shape optimization [32, 311, 250]: the shape to be optimized is not discretized explicitly by a mesh but rather described implicitly as the negative value set of a *level-set* function and evolved on a fixed mesh; this allows to handle complex topological changes of the optimized shape such as holes merging or boundaries collapsing (Figure 16). This ability to handle topological changes is very much desired in order to generate optimal designs without any *a priori* on the final shape, such as the number or the location of holes (of holes and arches in 3-d); for this reason the *level set method* used in conjunction with the method of Hadamard is often referred to as a *topology optimization* method, while the one based on mesh deformation is called geometric optimization [17]. A recent trend is to incorporate remeshing into shape and topology optimization in order to combine advantages of both methods [24, 107]: it allows to keep an explicit discretization of shapes throughout the optimization process while still allowing for topological changes (Figure 17).



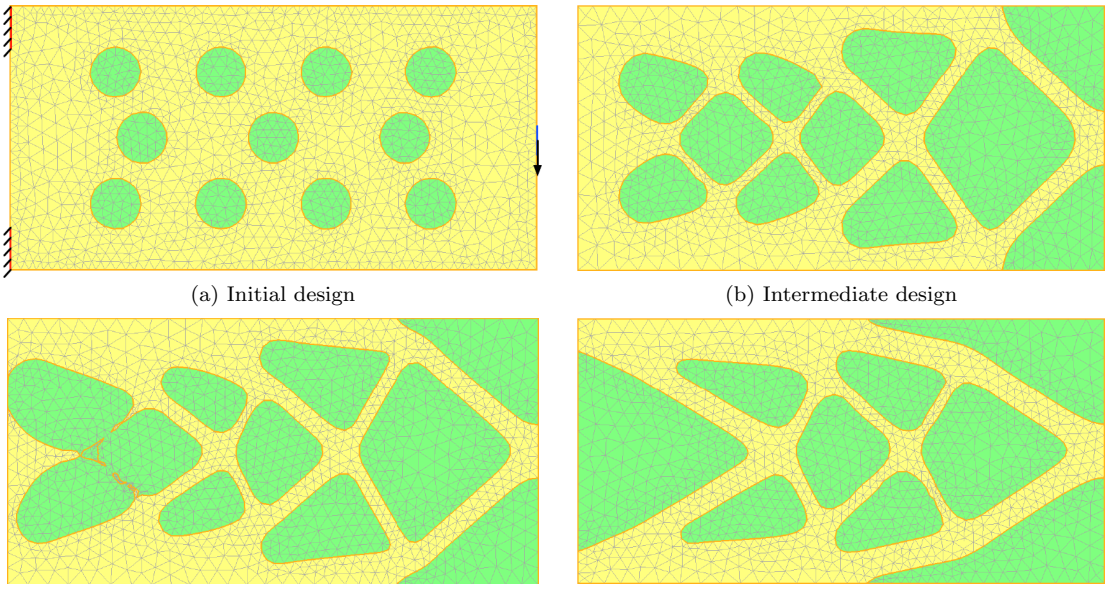
(a) Initial design (b) Intermediate design (c) Final design

Figure 15: Hadamard's method of boundary variation implemented for the optimization of the shape of a 2-d cantilever beam with the method of mesh deformation (figures from [17]). Topological changes are difficult to handle: the number of holes of the initial and final design is unchanged.



(a) Initial design (b) Intermediate design (c) Final design

Figure 16: Hadamard's method of boundary variation implemented for the optimization of the shape of a 2-d cantilever beam with the level set method (figures from [31]). Topological changes are handled: some holes have merged from the initial to the final design.

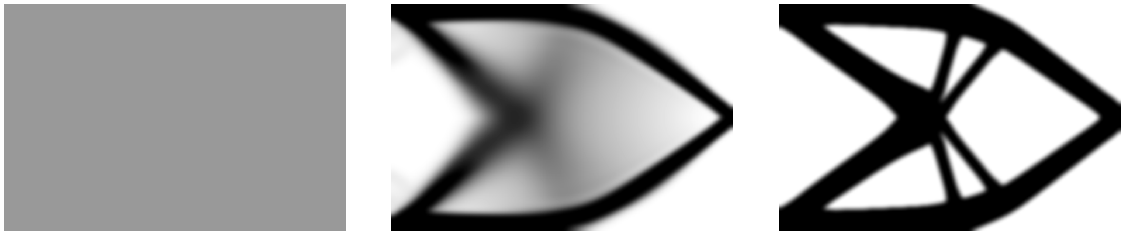


(a) Initial design (b) Intermediate design

(c) Intermediate design featuring a topological change (d) Final design

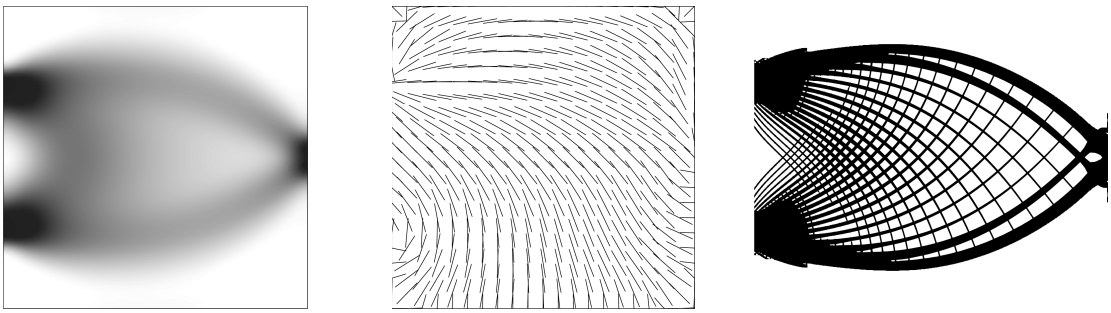
Figure 17: Hadamard's method of boundary variation implemented for the optimization of the shape of a 2-d cantilever beam with a level-set based mesh evolution method (figures from [107]). Topological changes are handled while keeping an explicit, conformal discretization of shapes.

Topology optimization is actually a rather large field which includes many other methods, such as the topological gradient method [85, 145, 165, 289, 40], level-set methods not based on the method of Hadamard [133, 316], phase field methods [73, 299], evolutionary topology optimization [193], density [66] and homogenization based methods [18, 93] (see [284] for a more complete review). Density based methods do not optimize the explicit location of the shape boundary but rather a density function ρ which represents the presence ($\rho = 1$) or the absence ($\rho = 0$), or intermediate states ($0 < \rho < 1$) of material; some penalization scheme is used in order to obtain convergence towards true shapes, i.e. black and white designs (Figure 18). Homogenization methods replace the issue of finding an optimal shape with the one of finding an optimal composite material characterized by a varying microstructure. A first step consists in optimizing the parameters of the microstructure such as material density or cell orientation fields. In a second step (if no penalization scheme is used), these fields are interpreted in order to determine a shape approximating in some sense the optimal microstructure (Figure 19) [254, 27, 166, 177].



(a) Initial design (b) Intermediate design (c) Final design

Figure 18: Topology optimization of a 2-d cantilever beam by a density method (here the so-called SIMP—Solid Isotropic Material Penalization—method). Figure obtained from the source code described in [43].



(a) Optimized density (b) Optimized orientation (c) Interpreted shape

Figure 19: Topology optimization of a 2-d cantilever beam by a homogenization method. Figure from [166].

2 Some current challenges in shape and topology optimization for the aeronautic industry

Today, engineers devise industrial systems most often with the assistance of Computer Aided Design (CAD) based geometry optimization software programs. Industries have been very much dependent on CAD formats because of their full compatibility with all stages of the design process, from physical numerical simulations on commercial software to actual manufacturing by automated machines. An *a priori* guess topology for the system to devise is first proposed by engineers. It is characterized for instance by a set of Bézier control points, spline surfaces parameters, or more general geometric parameters. Parametric optimization with respect to CAD parameters is then performed in order to improve the proposed design [181]. Usually, analytic shape sensitivities are not available; these are sometimes obtained by automatic differentiation (a very popular tool in aerodynamic design [94, 139, 130]) [235, 288].

However, the optimization is achieved in many cases by exploring a sufficiently large subset of the design space (e.g. with the commercial software Optimus [245]), see e.g. [142, 286]. Since these methods heavily rely on a parameterization choice of the shape geometry, they usually yield very small design updates of the initially proposed geometry (Figure 20). Such is fine for industrial applications to the extent that these small modifications may yield substantial gains of performance. However, it is easily

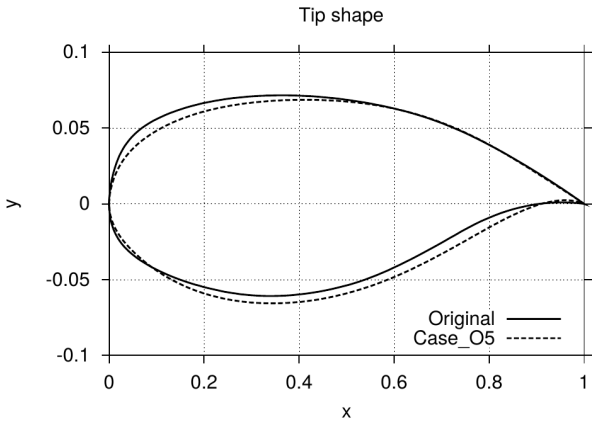


Figure 37. Tip wing section. Original vs optimized ones.

Figure 20: Lift–Drag optimization of a CAD–based airfoil design. Figure from [142].

imagined that even better performance could be obtained thanks to topology optimization since it could allow to seek very innovative new designs among much more unconstrained sets of shapes.

The rise of additive manufacturing since the 1990s has enabled the industry to be capable of fabricating more and more complex designs hardly described by CAD parameterizations, which feeds today a renewed enthusiasm for topology optimization. Many applications receiving currently a generous amount of effort are issued from the aeronautic industry and constitute some of the long term motivations at the origin of this work. One of the key challenges to overcome in order to make topology optimization applicable to aeronautic systems is the need for tackling inherent multiphysics aspects: coupled fluid, thermal, and mechanical constraints must very often be accounted for simultaneously when designing aircraft engines components.

A very representative issue drawing currently a substantial amount of attention in the topology optimization community lies in the design of heat exchangers [255, 258, 189, 271, 275]; these are devices used to cool down hot engine fluids by conveying them in the vicinity of some refrigerating gas or liquid. Industrial heat exchangers usually include many tubes and fins shaped in order to maximize the exchange surface area between hot and cold phases (Figure 21). Naturally, various additional multiphysics design constraints come into play, such as the need for controlling the loss of pressure induced by the system on the input fluid, or the mechanical resistance of the whole structure to the elevated thermal load.

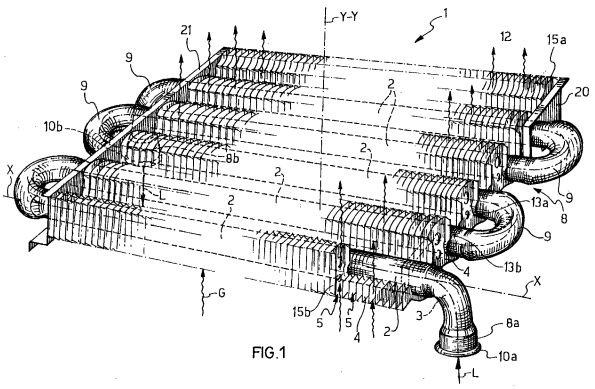


Figure 21: Industrial gas–liquid heat exchanger design. Figure from [169].

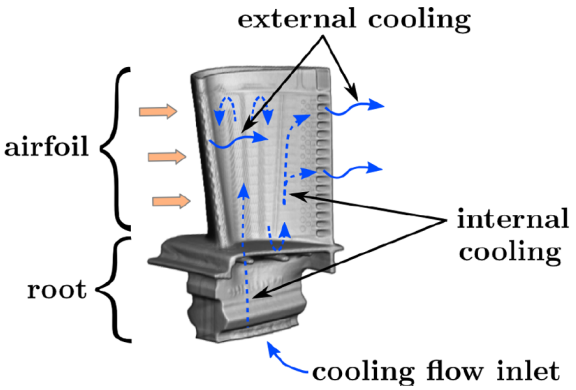


Figure 22: Turbine blade cooling system featuring internal cooling channels. Figure from [162].

It is suspected that many other components of aircraft engines could benefit as well from shape and topology optimization, such as the internal cooling channels system of turbine blades [62, 162] (Figure 22).

3 Scope of the thesis

To date, the method of Hadamard used as a topology optimization method (in which the geometry of the optimized design is allowed to deform freely) is not yet completely mature for true industrial applications. Most of the test cases featured in the literature are rather exclusively concerned with linear elasticity [25, 108] or heat conduction [9, 325], with very few large scale 3-d test cases [202]. It is only relatively recently that the method is being applied on more complicated physics such as, elastoplasticity [230] or electromagnetism [160, 213]. This trend is also visible in the the density based topology optimization community which also features an increasing number of works on multidisciplinary applications [116, 129, 128, 318, 319, 240, 288].

The objective of this work is to develop theory and methodologies for shape and topology optimization of multiphysics systems, keeping in view longer term industrial requirements. The content of this thesis is mainly concerned with the boundary variation method of Hadamard, but it also includes an opening chapter 7 investigating mathematical aspects of the homogenization method for future fluid applications.

By specific "industrial requirements", it is meant several identified major industrial needs that have guided our research and which are described in the next five paragraphs. Detailed contributions of the thesis are then summarized chapter by chapter in the next section.

Multiphysics systems

As motivated by the industrial applications of the previous section, there is an increasing demand for topology optimization of systems involving several interacting physics. Our study shall focus on systems featuring coupled fluid, thermal and mechanical properties. Mathematically, these are characterized by a set of physical variables denoted as follows in this entire manuscript:

- $\boldsymbol{v}$ and p for the velocity and pressure field associated to one or more fluid phases flowing through the system;
- T for the temperature field in solid and fluid phases;
- $\boldsymbol{u}$ for the elastic displacement of solid mechanical structures.

These variables are mathematically determined as the solutions of a set of Partial Differential Equations (PDEs), which are themselves derived from physical modelling choices. An academic setting for the characterization of $\boldsymbol{v}, p$, T and $\boldsymbol{u}$, sufficiently generic for our purposes, shall be described in chapter 2.

A significant motivation of the present work lies in the design of heat exchangers. A 2-d case study for the optimization of two different models of heat exchangers is proposed in chapter 5.

Constrained optimization

Most industrial systems feature a variety of load specifications that must be satisfied in realistic conditions of use. For instance, the overall mechanical stress or the temperature of a solid structure may be required to remain below a given bound in order to avoid premature fatigue. In other words, a design optimization problem reduces to determine the shape of a system that achieves the best performance subject to a given set of physical constraints.

Such problems can generically be modeled as mathematical programs of the form

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \\ s.t. \quad & \begin{cases} g_i(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = 0, & 1 \leq i \leq p, \\ h_j(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \leq 0, & 1 \leq j \leq q, \end{cases} \end{aligned} \tag{1}$$

where Γ denotes the boundary of the system to be optimized, typically the interface between a solid and a fluid structure (this may also include the case of a single phase). J refers to a given cost function which quantifies the system performance and which is to be minimized. Classical cost functions commonly encountered in shape optimization include the compliance of a mechanical structure, the drag force induced by an airfoil, or the heat stored into a thermal system. Equality and inequality constraints g_i and h_j model physical load specifications; they refer to target values some physical quantities need to reach (e.g. a desired location for the center of mass of the structure) or should not exceed (e.g. an upper bound limit for the overall temperature or a minimum feature size on the geometry). Both

objective function J and constraints g_i, h_j depend both on the shape Γ and on the physical variables $\boldsymbol{v}(\Gamma), p(\Gamma), \boldsymbol{u}(\Gamma), T(\Gamma)$, which depend themselves on Γ through physical state equations.

In density methods, the design variable ρ belongs to some euclidean space $\mathbb{R}^n$ after discretization, which makes mathematical program of the form of (1) solvable with classical first order optimization methods [244, 298, 310]. In the context of the method of Hadamard, the optimization variable Γ does not belong to any euclidean space but rather a kind of infinite dimensional manifold which makes the resolution of (1) more delicate. Most available works using the method of Hadamard convert (1) into an unconstrained minimization problem by adding penalization terms to the objective function J e.g. as in the Augmented Lagrangian Method [234, 107]. More complex methods, such as Sequential Linear Programming (SLP) or the Method of Feasible Directions (MFD) have been used with some success in few published works [135, 150]. However, these methods still present some unsatisfying qualities for industrial goals because they heavily rely on the selection of case dependent meta-parameters that can be difficult to tune, especially in the presence of a large number of constraints.

In chapter 3, a novel gradient flow type algorithm is described for the resolution of constrained optimization problems adapted to the context of the method of Hadamard.

Non-intrusiveness

For long term industrial applications, the shape optimization process should be fully compatible with the whole system conception chain. For instance, topology optimization tools should evaluate state variables by calling the same physical solvers used for industrial validation. Many topology optimization methods, including density based and level set methods, do not fulfill this condition, because they rely on artificial modifications of the state equations for pure numerical reasons.

Such need constituted one of the motivations for the systematic use of the level-set based mesh evolution algorithm of [25] (illustrated on Figure 17) in all our numerical topology optimization test cases. The main principles of the method are recalled in chapter 1. One of its main advantages is to preserve an explicit mesh discretization of the optimized shape; this allows to solve for state equations without any modification or approximation of the original physics, which could in principle be done by an external solver. For our remeshing purposes, we relied on the open source software `mmg` [108].

An additional non-intrusiveness constraint which retained our attention requires that only a strict minimal amount of information should be expected from an external user wishing to solve topology optimization test cases. In particular, the user should be allowed to specify arbitrary objective and constraints without being required the (rather technical) knowledge of the full derivation of Hadamard's shape sensitivities. In chapter 2, we provide mathematical expressions of shape sensitivities of *arbitrary* functionals which can be numerically automatically assembled from the sole knowledge of their partial derivatives.

Geometric distance constraints

Taking into account geometric distance constraints is a classical industrial need in topology optimization due to the finite precision of the manufacturing process. These can for instance manifest into minimum thickness or overhang constraints. They also occur in heat exchanger applications; in chapter 5, we shall see that a non-mixing condition between two different fluid phases can also be formulated as a geometric distance constraint.

Many works have been devoted to the integration of such constraints in topology optimization, see [234], Chapter 3, for a review. In the context of the method of Hadamard, geometric constraints can be taken into account thanks to formulations based on the signed distance function [30]. However the practical implementation of shape sensitivities is highly delicate and depends very much on the type of shape discretization used. In chapter 4, a new variational method is proposed for the numerical evaluation of shape sensitivities of geometric constraints, which can conveniently be implemented in any finite element setting.

Large scale problems

One last significant industrial requirement concerns the application of topology optimization to highly resolved 3-d physical systems. The physical state equations must be solved about a hundred times during a typical optimization process which is very computationally demanding for large scales problems, all the more when fluid mechanics is involved. This issue has been taken into account in various works relying on density based methods [1, 10, 71, 232, 39] and is starting to be addressed as well in evolutionary topology

optimization [227] and in level-set methods [202]. chapter 6 presents and discusses the implementation in `FreeFEM` [183] of moderately large 3-d numerical applications (involving up to 2 million degrees of freedom) exploiting parallel computing, preconditioned iterative solvers and recent high performance domain decomposition methods [238].

Finally, industrial heat exchanger designs such as the one illustrated on Figure 21 exemplify the need for generating optimized multiphysics designs featuring fine multiscale structures. In chapter 7, we derive higher order homogenized models for fluid systems which could allow for the future applicability of shape optimization by the homogenization method for such applications.

4 Summary of chapters

Chapter 1: Shape optimization based on Hadamard's boundary variation method

This preliminary chapter introduces background context and material on design optimization with the method of Hadamard. We outline well-established ingredients required for an implementation workflow of topology optimization using the level-set based mesh evolution method of [25]. We expose a summary of the method of Hadamard including classical mathematical techniques that allow to calculates the shape sensitivities of PDE dependent functionals. We mention the use of gradient based optimization adapted to the infinite dimensional context. We discuss the computation of the signed distance function and detail its mathematical properties. Finally we summarize the domain evolution method of [25] based on level set advection and a remeshing step.

Chapter 2: Hadamard's shape derivatives for a coupled thermal fluid structure problem

This chapter introduces a simplified thermal fluid-structure model which shall be of interest throughout this whole work: the variables $\boldsymbol{v}, p, T, \boldsymbol{u}$ are the solutions of three state equations which are weakly coupled in the sense that each of them can be solved successively and independently. This model is based on the steady state incompressible Navier-Stokes equations for the fluid velocity and pressure, convection diffusion for the temperature field, and linear thermo-elasticity for the elastic displacement. It includes previous literature studies where only one of the physics is active, while allowing for new coupled test cases such as fluid-structure interaction in a small deformation regime or convective heat transfer.

A significant novelty of this part is the derivation of Hadamard's shape derivatives formulas for *arbitrary* objective functionals, which we detail and compare to more classical techniques in details. In the literature concerned with the method of Hadamard, hypotheses on the form of the shape functional are usually assumed: this imposes to redo the full derivation of shape derivatives for each change of objective function or constraint, which is a burden for practical implementation in industrial software. Our formulas allows to automate the computation of shape derivatives: these are assembled after the resolution of adjoint equations from the sole knowledge of the partial derivatives of the objective functional, an information which can be in principle provided easily by an external user.

Another contribution is to include in our analysis a type of fluid-structure interaction problem. A surprising property lies in the coupling of the adjoint system involved in the computation of shape sensitivities: the equality of normal stress constraints at the fluid–structure interface turns into the matching of the respective adjoint variables values on this boundary.

The validity of our formulas is finally verified numerically on several 2-d test cases activating either one, two, or three physics simultaneously.

Most of the content of this chapter has been published in [153]:

F. Feppon, G. Allaire, F. Bordeu, J. Cortial, and C. Dapogny, *Shape optimization of a coupled thermal fluid–structure problem in a level set mesh evolution framework*, SeMA Journal, (2019), pp. 1–46.

Let us mention, however, that this chapter includes many more numerical test cases, and that some of the test cases of the published work [153] have been improved.

Chapter 3: Null space gradient flows for constrained shape optimization

The physical model and its shape optimization context being introduced, this chapter describes the optimization algorithm developed specifically for all our numerical test cases. We propose an optimization scheme that can be interpreted as the discretization of a gradient flow able to see equality and inequality constraints (a generalization of the dynamical system approaches of [317, 300]). The gradient flow is called

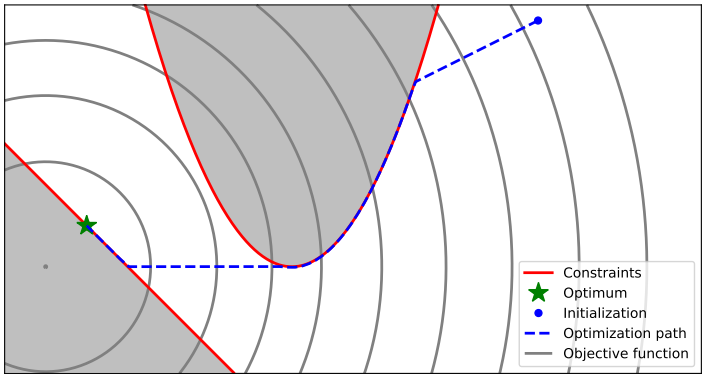


Figure 23: Illustration of our null space gradient flow for constrained optimization. Trajectories always follow the best feasible descent direction.

null space because the descent direction is decomposed into a *null space* step decreasing the objective function tangentially to the constraints, and a *range space* step orthogonal to the null space step and which gradually carries the optimization path towards the feasible domain. Optimization trajectories are guided by the resolution of dual quadratic subproblems which indicate when to come back into the interior of the optimization domain (Figure 23). One feature characterizing the algorithm is the ability to decrease objective function values while maintaining constraints satisfied.

Convergence properties of the continuous trajectories are established; they guarantee the improvement of optimized solutions until a local minimum is found up to the selection of a sufficiently small discretization step. The robustness and efficiency of our method for shape optimization is demonstrated numerically on multiple load bridge test cases featuring 10 constraints.

Most of the content of this chapter has been the object of the submitted preprint [155]:

F. FEPPON, G. ALLAIRE, AND C. DAPOGNY, *Null space gradient flows for constrained optimization with applications to shape optimization*, submitted, (2019).
The chapter contains in addition several numerical comparisons with classical algorithms on academic test cases.

Chapter 4: A variational method for computing shape derivatives of geometric constraints along rays

Heat exchangers applications featuring two distinct fluid phases motivate the need for taking into account non penetration constraints in shape optimization. A very natural approach consists in requiring each of the two phases to remain at a prescribed minimum distance from one another, which can be formulated mathematically conveniently by means of the signed distance function. More generally, many other geometric constraints can be formulated in this fashion, such as minimum or maximum thickness and minimum member's distance.

Previous works [30, 234] have proposed mathematical expressions for the derivatives of shape functionals depending on the signed distance function, and have demonstrated numerically their ability to enforce geometric constraints such as maximum or minimum thickness. However, their direct implementation is difficult because it requires numerical integration along the normal rays to the shape (Figure 24) and accurate estimates of shape skeleton or its principal curvatures. The implementation of these operations is notoriously difficult and very much depends on the dimension (2-d or 3-d) and on the type of shape discretization used (implicit or conforming, on a structured or unstructured mesh).

In this chapter, we demonstrate that shape derivatives of geometric constraints can in fact be obtained from a variational problem which can be solved conveniently by the finite element method: the value of integrals along the rays are retrieved from the variational solution values at boundary vertices. Our method amounts, in full generality, to compute integral quantities along the characteristic curves of a given velocity field β without requiring the explicit knowledge of these curves on the spatial discretization. The implementation is very much simplified because it requires neither the computations of rays nor of the shape curvatures but only that of a rather arbitrary weight vanishing approximately on the skeleton set.

The well-posedness of the proposed variational formulation is established thanks to a detailed analysis of the weighted graph space of the advection operator $\beta \cdot \nabla$. In the shape optimization context, β is

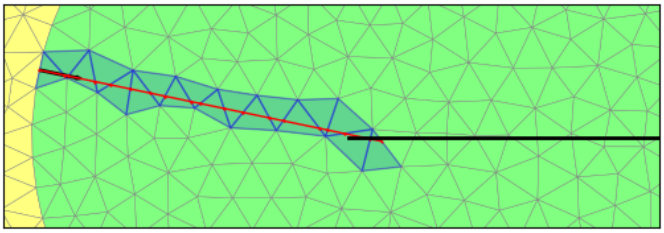


Figure 24: Classical methods for computing the shape derivatives of geometric distance constraints involve the computation of the normal rays of the shape (in *red*), which requires to travel along the mesh discretization.

given by the gradient of the signed distance function of the working domain, however our analysis also includes more general vector fields satisfying a suitable set of assumptions. One novelty of our approach is the ability to handle velocity fields with possibly unbounded divergence; classical works performing similar analysis [144] usually assume $\mathrm{div}(\beta) \in L^\infty(D)$ where D denotes the working domain, which in general does not hold for shape optimization applications.

The ease of implementation of our method is demonstrated for maximum and minimum thickness constraints in structural design: we have been able to retrieve numerical results analogous to those of previous works [30, 234] relying on explicit ray integrals.

Most of the content of this chapter is to appear in the publication [154]:

F. Feppon, G. Allaire, and C. Dapogny, *A variational formulation for computing shape derivatives of geometric constraints along rays*, To appear in M2AN, (2019).

Chapter 5: Topology optimization of 2-d heat exchangers

The material developed in the previous chapters is now applied to design optimization of 2-d heat exchangers.

In a first part, our methods are applied on the topology optimization of a liquid-liquid heat exchanger problem with the weakly coupled model introduced in chapter 2. A non-mixing constraint for two liquid phases exchanging their heat is formulated as a geometric distance constraint and treated with the variational method of the previous chapter 4 (Figure 25a).

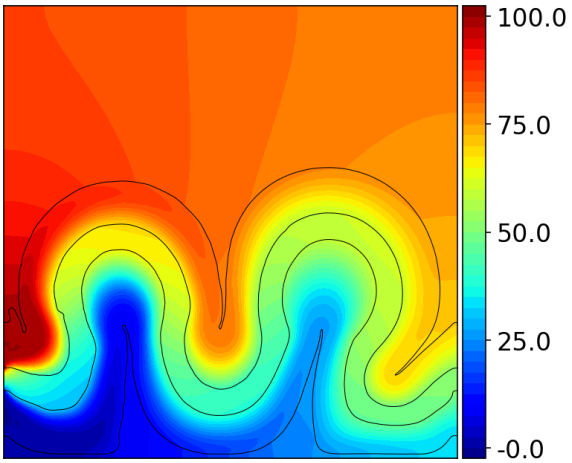
In a second part, we present a 2-d heat exchanger case study issued from a collaboration with Safran Aero Boosters: the objective is to determine the optimal shape of transverse oil pipe cross sections refrigerated by an input cold air flow. The test case fits well in the physical setting introduced in chapter 2 up to a small change of boundary conditions for the thermal profile. The optimization problem featured a minimum thickness constraint for the oil phase and a maximum pressure loss constraint for the air phase. Our study demonstrates the ability of our shape optimization method to generate a variety of non-classical designs in a simplified setting (Figure 25b).

Chapter 6: Towards 3-d and industrial applications: implementation recipes for a variety of numerical test cases

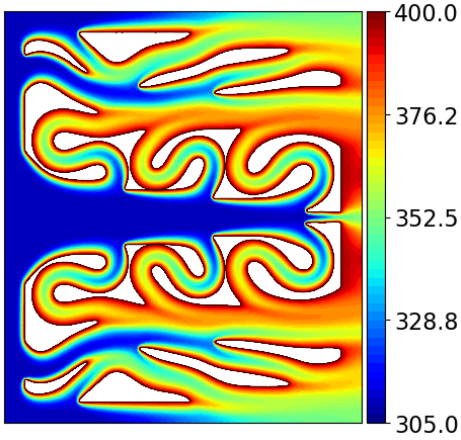
This chapter discusses the implementation of numerical test cases approaching more advanced industrial applications.

A first section describes succinctly programming paradigms of our topology optimization code developed in `python` and `FreeFEM` [183] in the context of this thesis. We then discuss delicate implementation details specific to the 3-d context, for instance the unavoidable use of domain decomposition techniques in conjunction with physics-dependent preconditioners.

The next section presents then a variety of new 3-d shape optimization results for different physics. Our case studies include the classical benchmark example of the cantilever beam subject to traction or torsion, optimal designs for heat conduction (Figure 13), optimal lift-drag profiles, up to more complex fluid-structure interaction test cases. Thanks to the work of Moulin et. al. [238], we have been able to solve problems featuring Navier-Stokes flows with mesh discretizations involving up to two millions of mesh elements.

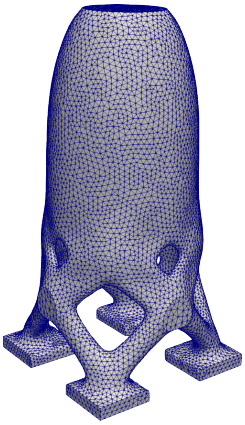


(a) 2-d heat exchanger design featuring a non-mixing constraint.

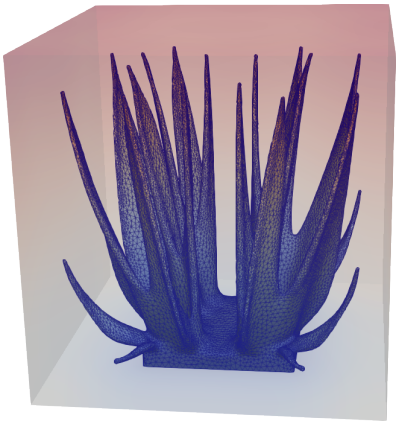


(b) 2-d heat exchanger design from the SAFRAN Aero Boosters case study.

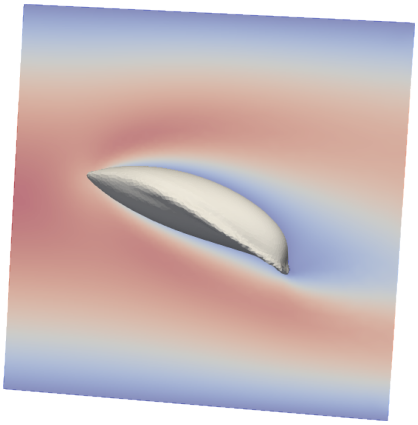
Figure 25: Two of the optimization test cases presented in chapter 5.



(a) Structural design



(b) Thermal design



(c) Aerodynamic design

Figure 26: A few 3-d test cases obtained in chapter 6.

Chapter 7: High order homogenized equations for perforated problems, towards fluid topology optimization by the homogenization method

This final chapter constitutes an opening of the thesis towards the use of the homogenization method for topology optimization of multiphysics applications involving fluids.

It is well known that in general, optimal designs in shape optimization are composite structures. The theory of homogenization allows to characterize the effective physics of the materials obtained by arbitrarily complex phase mixtures: this enables to develop practical shape and topology optimization algorithms [18]. In the context of optimal design for the mechanics of structures, the theory of homogenization is well known and efficient methods have been developed for generating shapes approaching optimal composite micro-structures [27, 254] (such as those depicted on Figure 19). For fluid applications, the theory of homogenization is less clear, because classical literature [103, 11, 15, 265] identifies three possible homogenized regimes depending on how periodic obstacles scale within their periodic cell. To date, density based methods rely exclusively on the Brinkman model for such purposes, which turns to be only one of these possible homogenized the three.

This part of the thesis investigates high order homogenized equations for the Stokes system in a porous medium which capture all three regimes. These homogenized models which could serve for topology optimization of fluid systems by applying the approach of [254, 27]. Our motivation originates from the observation of the complexity of industrial heat exchanger designs, which combine a macroscopic shape design together with highly resolved periodic blade structures. Classical shape optimization methods such as the one developed above are able to determine *macroscopic* geometries, but are not very well suited in the case where optimal geometries exhibit such multi-scale patterns.

One of our main results shows that the classical three homogenized regimes can be captured by a single higher order homogenized equation: each of them is retrieved, in low-volume fraction limits, for particular obstacle's shape scalings. More generally, well-posed homogenized equations are derived at any order thanks to a method inspired from the work of Bakhvalov and Panasenko [53], Smyshlyaev and Cherednichenko [287] and Allaire et. al. [33].

From a pedagogical perspective, we provide derivations and a mathematical analysis of high order homogenized equations for several elliptic perforated problems (whose solutions vanish on small periodic holes) featuring an increasing order of complexity. We first consider the case of the perforated scalar Poisson problem, the extension to the vectorial case is then treated by studying the analogous perforated problem in elasticity, before turning eventually to the Stokes problem where the pressure term requires some additional work. A very striking feature is the occurrence of strange differential operators of *odd order* in these homogenized equations: this fact is not standard and does not seem to have been noticed in classical literature seeking higher order model corrections.

This final chapter does not include numerical results, which are the object of future works.

CHAPTER 1

SHAPE AND TOPOLOGY OPTIMIZATION BASED ON HADAMARD'S BOUNDARY VARIATION METHOD

Contents

This chapter is a review of classical background material on shape and topology optimization based on the method of Hadamard. The first section 1.1 is a reference section for the notation that is used throughout this whole thesis. Then section 1.2 summarizes common definitions and properties of shape derivatives. For later comparison with the method proposed in chapter 2, classical techniques are recalled for the derivation of shape derivatives in the context of PDE constrained problem. Section 1.3 is devoted to the signed distance function, a central object for taking into account geometric constraints in shape optimization. Important properties that shall be used intensively in chapter 4 are referenced, including the sensitivity of the signed distance function with respect to the domain. Finally, section 1.4 focuses on the practical integration of the previous ingredients into numerical algorithms. Classical implementation steps of the Hadamard's method for topology optimization are summarized. The gradient method and its adaptation to the (infinite dimensional) shape optimization context is reviewed as a preliminary to the dedicated chapter 3. Finally, we discuss the important issue of numerically representing and evolving shapes. We mention mesh deformation and level set methods, and we summarize the level-set based mesh evolution algorithm of [24] which we used for all our numerical shape optimization test cases.

1.1 Notation

In what follows, for d an integer and a given open set $D \subset \mathbb{R}^d$, we denote by

$$L^p(D) := \left\{ v \text{ measurable } \mid \int_D |v|^p \mathrm{d}x < +\infty \right\}$$

the space of p-integrable functions and by $L^\infty(D)$ the space of bounded functions almost everywhere on D. The Sobolev space $W^{m,p}(D)$ for $m \in \mathbb{N}$ and $p \in (1, +\infty)$ (see [76]) is the space of functions with derivatives up to the order k in $L^p(D)$.

$$W^{m,p}(D) := \{ v \text{ measurable } \mid \forall 0 \leq k \leq m, \forall 1 \leq i_1 \leq \ldots \leq i_k \leq d, \partial^k_{i_1 \ldots i_k} v \in L^p(D) \}.$$

The Hilbert space $W^{m,2}(D)$ shall also be denoted by $H^m(D)$.

The norm naturally associated to a Banach space V is denoted by $||\cdot||_V$, for instance

$$\forall v \in W^{m,p}(D),\ ||v||_{W^{m,p}(D)} = \left(\sum_{0\le k\le m} \sum_{1\le i_1\dots i_k\le d} |\partial^k_{i_1\dots i_k} v|^p \right)^{1/p}.$$

When V is a Hilbert space, its scalar product is denoted by $\langle\cdot,\cdot\rangle_V$. Its natural norm is then $||\cdot||_V := \langle\cdot,\cdot\rangle_V^{1/2}$. Space of vector valued functions whose components belong to $L^p(D)$, $W^{m,p}(D)$ or $H^m(D)$ are respectively denoted by $L^p(D,\mathbb{R}^d)$, $W^{m,p}(D,\mathbb{R}^d)$ and $H^m(D,\mathbb{R}^d)$. The dual space of a Banach space V is denoted by V^*, and the identity mapping by $I : V \to V$ (the dependence with respect to V being implicit if clear from the context).

The usual euclidean norm on $\mathbb{R}^d$ is denoted by $||\cdot||$ (without index notation) and the scalar product of two vectors $\boldsymbol{a},\boldsymbol{b} \in \mathbb{R}^d$ by $\boldsymbol{a}\cdot\boldsymbol{b}$. We shall also use the notation

$$\boldsymbol{a}\cdot M\cdot\boldsymbol{b} := \boldsymbol{a}^T M\boldsymbol{b}$$

for scalar products in $\mathbb{R}^d$ featuring a product by a $d\times d$ matrix $M\in\mathbb{R}^{d\times d}$. In general, elements $\boldsymbol{f}\in\mathbb{R}^d$ that must be thought of as vectors are written in boldface, while a non bold notation $x\in\mathbb{R}^d$ is preferred those that must be rather considered as points.

Finally, the gradient of a differentiable function f is the *column* vector $\nabla f := (\partial_i f)_{1\le i\le d}$. The Jacobian matrix of a differentiable vector field $\boldsymbol{f}$ is the $d\times d$ matrix $\nabla\boldsymbol{f} := (\partial_j f_i)_{1\le i,j\le d}$.

1.2 The boundary variation method of Hadamard for shape sensitivity of PDE constrained problems

This section reviews the classical theory of shape differentiation by the method of Hadamard for PDE constrained functionals. Definitions of shape derivatives and the Hadamard's structure theorem are recalled in section 1.2.1. Reference formulas for the shape differentiation of volume and boundary integrals are summarized in section 1.2.2. Finally, the classical steps involved in the calculation of shape derivatives for PDE constrained functionals are reviewed in section 1.2.3. Sketch of proofs are often included for the reader's convenience, who will find more complete material in the classical textbooks [17, 184, 291].

1.2.1 Shape derivatives in the sense of Hadamard

Shape optimization is concerned with the problem of finding the shape of a bounded Lipschitz domain $\Omega\subset\mathbb{R}^d$ ($d=2$ or 3 in usual applications) which minimizes a given cost function J:

$$\min_{\Omega}\ J(\Omega). \tag{1.2.1}$$

For the applications considered in this thesis, we shall also be interested in the slightly more general case where the shape to optimize is a Lipschitz codimension one manifold Γ, possibly with non empty boundary. For instance, in chapters 2 and 6, Γ represents the interface between two subdomains occupied by fluid and solid phases.

The purpose of the method of Hadamard [184, 17, 291] is to introduce a notion of differentiation with respect to the position of the shape Ω (or Γ) which ultimately allows to solve (1.2.1) (or more complex problems such as (1) in the introduction) with gradient based optimization algorithms. The principle of the method is to consider shape deformations of the form

$$\Omega_{\boldsymbol{\theta}} = (I+\boldsymbol{\theta})\Omega,\ \text{where } \boldsymbol{\theta}\in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d),\ ||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}<1, \tag{1.2.2}$$

$$\Gamma_{\boldsymbol{\theta}} = (I+\boldsymbol{\theta})\Gamma,\ \text{where } \boldsymbol{\theta}\in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d),\ ||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}<1, \tag{1.2.3}$$

in either case where the shape to optimize is a bounded domain Ω or an interface Γ.

The variable $\boldsymbol{\theta}$ denotes a small vector field which moves all points of Ω from x to the deformed location $x+\boldsymbol{\theta}(x)$, as is illustrated on Figure 1.1 below. In order to estimate the sensitivity of the shape with respect to such deformations, it is sufficient to consider "small" displacements $\boldsymbol{\theta}$; this smallness is generally measured with the norm $||\cdot||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}$ of the set $W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$, which is the space of Lipschitz bounded vector fields [76]:

$$W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d) := \{\boldsymbol{\theta}\in L^\infty(\mathbb{R}^d,\mathbb{R}^d)\,|\,\nabla\boldsymbol{\theta}\in L^\infty(\mathbb{R}^d,\mathbb{R}^{d\times d})\}.$$

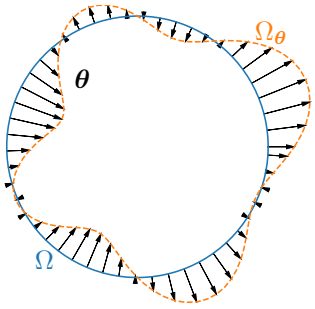


Figure 1.1: Deformation of a domain Ω with the method of Hadamard. A small vector field $\boldsymbol{\theta}$ is used to deform Ω into $\Omega_{\boldsymbol{\theta}} = (I + \boldsymbol{\theta})\Omega$.

A classical result states that $I + \boldsymbol{\theta}$ is a Lipschitz diffeomorphism whenever $\boldsymbol{\theta}$ has a norm smaller than one:

Lemma 1.1 (see [17], Lemma 6.13 p.129)**.** *For any $\boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ such that $||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} < 1$, the map $I + \boldsymbol{\theta}$ is a bijection satisfying $(I + \boldsymbol{\theta})^{-1} - I \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$.*

Sketch of proof. Formally, the inverse map is given by

$$(I + \boldsymbol{\theta})^{-1} = \sum_{k=0}^{+\infty} (-1)^k \overbrace{\boldsymbol{\theta} \circ \cdots \circ \boldsymbol{\theta}}^{k \text{ times}},$$

where the above series is convergent in the norm of $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$. For a complete proof, see [17]. □

The shape derivative of a given functional $J(\Omega)$ is then defined by differentiation of $J(\Omega_{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$:

Definition 1.1. A shape functional $J(\Omega)$ is said shape differentiable if the mapping

$$\begin{array}{rcl} W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d) & \longrightarrow & \mathbb{R} \\ \boldsymbol{\theta} & \longmapsto & J(\Omega_{\boldsymbol{\theta}}) \end{array} \tag{1.2.4}$$

is Fréchet differentiable at $\boldsymbol{\theta} = 0$, *i.e.* if there exists a continuous linear form

$$\mathrm{D}J(\Omega) \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)^* \tag{1.2.5}$$

such that the following asymptotics holds true:

$$J(\Omega_{\boldsymbol{\theta}}) = J(\Omega) + \mathrm{D}J(\Omega)(\boldsymbol{\theta}) + o(\boldsymbol{\theta}), \qquad \text{where} \quad \frac{|o(\boldsymbol{\theta})|}{||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}} \xrightarrow{\boldsymbol{\theta} \to 0} 0. \tag{1.2.6}$$

Remark 1.1. In case where the shape to optimize is an interface Γ, a functional $J(\Gamma)$ is said shape differentiable if $\boldsymbol{\theta} \mapsto J(\Gamma_{\boldsymbol{\theta}})$ is differentiable and the shape derivative $\mathrm{D}J(\Gamma)(\boldsymbol{\theta})$ is defined analogously to (1.2.6).

Remark 1.2. It will be convenient to write (1.2.6) with a $\mathrm{d}/\mathrm{d}\boldsymbol{\theta}$ differential notation:

$$\left.\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=0} [J(\Omega_{\boldsymbol{\theta}})](\boldsymbol{\theta}) := \mathrm{D}J(\Omega)(\boldsymbol{\theta}),$$

where with a little abuse of notations, we have also denoted by $\boldsymbol{\theta}$ the direction in which $\boldsymbol{\theta} \mapsto J(\Omega_{\boldsymbol{\theta}})$ is differentiated.

Remark 1.3. $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)^*$ is the dual space of $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$. The notation (1.2.5) requires consequently the existence of some constant $C(\Omega)$ independent of $\boldsymbol{\theta}$ such that

$$\forall \boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d),\ |\mathrm{D}J(\Omega)(\boldsymbol{\theta})| \leq C(\Omega)||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}.$$

An important theoretical result known as Hadamard's structure theorem states that under suitable smoothness assumptions, the shape derivative $\mathrm{D}J(\Omega)(\boldsymbol{\theta})$ of a functional $J(\Omega)$ depends only on the normal trace component $\boldsymbol{\theta}\cdot\boldsymbol{n}$ of the vector field $\boldsymbol{\theta}$ deforming the bounded open set Ω.

Proposition 1.1 (Hadamard's structure theorem [17, 184])**.** *Let Ω a smooth bounded open set of $\mathbb{R}^d$ and $J(\Omega)$ a shape differentiable functional. If $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ are such that $\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\in\mathcal{C}^1(\mathbb{R}^d,\mathbb{R}^d)$ and $\boldsymbol{\theta}_1\cdot\boldsymbol{n}=\boldsymbol{\theta}_2\cdot\boldsymbol{n}$ on $\partial\Omega$, then it holds*

$$\mathrm{D}J(\Omega)(\boldsymbol{\theta}_1)=\mathrm{D}J(\Omega)(\boldsymbol{\theta}_2).$$

1.2.2 Shape derivatives of volume and surface functionals

The purpose of this part is to recall classical theorems involved in shape derivative calculus of volume and surface integrals. The main tool underlying all subsequent properties is the classical change of variable formula [184, 147, 212].

Proposition 1.2. *If Φ is a Lipschitz diffeomorphism of $\mathbb{R}^d$ and $\Omega\subset\mathbb{R}^d$ an open set, then for any $f\in L^1(\Phi(\Omega))$, $f\circ\Phi$ belongs to $L^1(\Omega)$ and it holds*

$$\int_{\Phi(\Omega)} f\mathrm{d}x=\int_{\Omega} f\circ\Phi\,||\nabla\Phi||\mathrm{d}x. \tag{1.2.7}$$

We shall use as well its variant for the change of variables on co-dimension one surfaces [17, 184]:

Proposition 1.3. *Let Γ a $\mathcal{C}^1$ codimension one surface and Φ a $\mathcal{C}^1$ diffeomorphism of $\mathbb{R}^d$. Then for any function $f\in L^1(\Phi(\Gamma))$, it holds $f\circ\Phi\in L^1(\Gamma)$ and*

$$\int_{\Phi(\Gamma)} f\mathrm{d}s=\int_{\Gamma} f\circ\Phi\,|\det(\nabla\Phi)|\,|(\nabla\Phi)^{-T}\boldsymbol{n}|\mathrm{d}s, \tag{1.2.8}$$

where $\boldsymbol{n}$ is any normal vector field to Γ.

Sketch of proof. Formula (1.2.8) can be deduced from the generalization of (1.2.7) to arbitrary manifolds [212]. In that case, the change of variable from Γ to $\Phi(\Gamma)$ reads:

$$\int_{\Phi(\Gamma)} f\mathrm{d}s=\int_{\Gamma} f\circ\Phi\,|\det(\nabla\Phi_{|\mathcal{T}(s)})|\mathrm{d}s$$

where $|\det(\nabla\Phi_{|\mathcal{T}(s)})|$ is the determinant of the restriction of $\nabla\Phi$ to the tangent space $\mathcal{T}(s)$ of Γ at s and onto the tangent space $\mathcal{T}(\Phi(s))$ of $\Phi(\Gamma)$ at $\Phi(s)$. It is clear that $\mathcal{T}(\Phi(s))$ is given by

$$\mathcal{T}(\Phi(s))=\{\nabla\Phi(s)\boldsymbol{\xi}\,|\,\boldsymbol{\xi}\in\mathcal{T}(s)\}.$$

Denote $\boldsymbol{n}'(s)$ a normal vector to $\mathcal{T}(\Phi(s))$. The Jacobian matrix $\nabla\Phi(s)$ can be written in the form of a block matrix with respect to the decomposition $\mathbb{R}^d=\mathbb{R}\boldsymbol{n}(s)\oplus\mathcal{T}(s)$ onto $\mathbb{R}^d=\mathbb{R}\boldsymbol{n}'(s)\oplus\mathcal{T}(\Phi(s))$:

$$\nabla\Phi(s)=\begin{array}{c} \overbrace{\phantom{\nabla\Phi(s)\boldsymbol{n}(s)}}^{\mathbb{R}\boldsymbol{n}(s)}\ \overbrace{}^{\mathcal{T}(s)} \\ \left[\begin{array}{c:c} \nabla\Phi(s)\boldsymbol{n}(s) & 0 \\ \hdashline & \nabla\Phi(s)_{|\mathcal{T}(s)} \end{array}\right] \end{array}\begin{array}{l}\}\,\mathbb{R}\boldsymbol{n}'(s)\\ \}\,\mathcal{T}(\Phi(s))\end{array}$$

This implies that the determinant of $\nabla\Phi(s)$ is given by $\det(\nabla\Phi(s))=\boldsymbol{n}'(s)\cdot\nabla\Phi(s)\cdot\boldsymbol{n}(s)\det(\nabla\Phi_{|\mathcal{T}(s)})$. Finally, the relation $\boldsymbol{n}'(s)\cdot\nabla\Phi(s)\cdot\boldsymbol{\xi}=\boldsymbol{\xi}\cdot\nabla\Phi(s)^T\cdot\boldsymbol{n}'(s)=0$ which holds true for any $\boldsymbol{\xi}\in\mathcal{T}(\Phi(s))$ implies that $\boldsymbol{n}(s)$ is proportional to $\nabla\Phi(s)^T\boldsymbol{n}'(s)$. This yields the following expression for $\boldsymbol{n}'(s)$ (up to a sign change):

$$\boldsymbol{n}'(s)=\frac{\nabla\Phi^{-T}\boldsymbol{n}(s)}{||\nabla\Phi^{-T}(s)\boldsymbol{n}(s)||}, \tag{1.2.9}$$

from where we infer $<\nabla\Phi(s)\boldsymbol{n}(s),\boldsymbol{n}'(s)>=1/||\nabla\Phi^{-T}(s)\boldsymbol{n}(s)||$ and the result. □

Proposition 1.4. *Let Ω be a bounded open set of $\mathbb{R}^d$. For any $f \in W^{1,1}(\mathbb{R}^d)$, the functional $J(\Omega)$ defined by*

$$J(\Omega) := \int_\Omega f(x)\mathrm{d}x$$

is shape differentiable and it holds

$$\mathrm{D}J(\Omega)(\boldsymbol{\theta}) = \int_\Omega \mathrm{div}(f\boldsymbol{\theta})\mathrm{d}x = \int_\Omega (\nabla f \cdot \boldsymbol{\theta} + f\mathrm{div}(\boldsymbol{\theta}))\mathrm{d}x, \qquad \boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d).$$

If in addition Ω is smooth then the above formula can be rewritten as

$$\mathrm{D}J(\Omega)(\boldsymbol{\theta}) = \int_{\partial\Omega} f\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s, \qquad \boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d).$$

where $\boldsymbol{n}$ denotes the outward normal to Ω.

Sketch of proof. The application of the change of variable formula yields

$$J(\Omega_{\boldsymbol{\theta}}) = \int_{(I+\boldsymbol{\theta})\Omega} f\mathrm{d}x = \int_\Omega f\circ(I+\boldsymbol{\theta})\ \det(I+\nabla\boldsymbol{\theta})\mathrm{d}x = \int_\Omega (\nabla f\cdot\boldsymbol{\theta} + f\mathrm{div}(\boldsymbol{\theta}))\mathrm{d}x + o(\boldsymbol{\theta}).$$

The boundary integral formula follows by integration by part. □

The derivation of the shape derivatives of surface integrals is slightly more involved and requires some notions of differential geometry and tangential calculus [293]. For a given smooth codimension one surface Γ with a prescribed smooth normal vector field $\boldsymbol{n}$, we denote by ∇_Γ the tangential gradient:

$$\forall g \in H^1(\mathbb{R}^d),\ \nabla_\Gamma g := \nabla g - (\boldsymbol{n}\cdot\nabla g)\boldsymbol{n}.$$

The tangential divergence of a vector field $\boldsymbol{f} \in H^1(\mathbb{R}^d, \mathbb{R}^d)$ is defined by the trace of the linear operator $\nabla\boldsymbol{f}$ restricted to the tangent spaces of Γ; it reads

$$\mathrm{div}_\Gamma(\boldsymbol{f}) := \mathrm{div}(\boldsymbol{f}) - \boldsymbol{n}\cdot\nabla\boldsymbol{f}\cdot\boldsymbol{n}.$$

Let us now recall the classical definition of principal curvatures for the codimension one manifold $\partial\Omega$ (see [293] for proofs and further material on differential geometry).

Proposition 1.5 (Principal curvatures)**.** *Let Γ be a $\mathcal{C}^2$ manifold and let $\boldsymbol{n}$ be any differentiable unit vector field normal to Γ. The gradient of the normal $\nabla\boldsymbol{n}$ satisfies:*

1. $\forall y \in \Gamma,\ \nabla\boldsymbol{n}(y)\cdot\boldsymbol{n}(y) = 0$,
2. $\forall y \in \Gamma,\ \nabla\boldsymbol{n}^T = \nabla\boldsymbol{n}$.

In other words, for any $y \in \Gamma$, $\nabla\boldsymbol{n}(y)$ is a symmetric matrix which leaves the tangent space of Γ at y invariant. Consequently, it can be diagonalized with $d-1$ eigenvectors $(\boldsymbol{\tau}_i(y))_{1\le i\le d-1} \in \mathbb{R}^{d\times d}$ belonging to the tangent space at y and associated to $d-1$ eigenvalues $(\kappa_i(y))_{1\le i\le d-1} \in \mathbb{R}^d$:

$$\forall y \in \Gamma,\ \nabla\boldsymbol{n}(y) = \sum_{i=1}^{d-1} \kappa_i(y)\boldsymbol{\tau}_i(y)\boldsymbol{\tau}_i(y)^T.$$

The real numbers $(\kappa_i(y))_{1\le i\le d-1}$ and their related basis of eigenvectors $(\boldsymbol{\tau}_i(y))_{1\le i\le d-1}$ are respectively called principal curvatures and principal directions of Γ at y.
The mean curvature of Γ is the real number $\kappa(y)$ defined (up to a sign change) by

$$\kappa(y) := \sum_{i=1}^{d-1} \kappa_i(y) = \mathrm{div}(\boldsymbol{n}(y)).$$

For the derivation of shape derivatives of boundary integrals, we shall also need the following identity which is a variant of the Stokes formula on manifolds:

Lemma 1.2. *Let Γ be a smooth codimension one surface of $\mathbb{R}^d$ with boundary $\partial\Gamma$, $\boldsymbol{f} \in H^1(\Gamma, \mathbb{R}^d)$ and $g \in H^1(\Gamma)$. Then it holds*

$$\int_\Gamma g \mathrm{div}_\Gamma(\boldsymbol{f}) \mathrm{d}s = \int_\Gamma (-\nabla_\Gamma g \cdot \boldsymbol{f} + \kappa g(\boldsymbol{f}\cdot\boldsymbol{n}))\mathrm{d}s + \int_{\partial\Gamma} g\boldsymbol{f}\cdot\boldsymbol{\tau}\mathrm{d}s,$$

where $\boldsymbol{\tau}$ is the outward normal to $\partial\Gamma$ tangent to Γ.

Sketch of proof. The projection of $\boldsymbol{f}$ tangent to Γ is $\boldsymbol{f}_\Gamma := \boldsymbol{f} - (\boldsymbol{f}\cdot\boldsymbol{n})\boldsymbol{n}$. Stokes formula on manifolds then reads [212]

$$\int_\Gamma g \mathrm{div}_\Gamma(\boldsymbol{f}_\Gamma)\mathrm{d}s = -\int_\Gamma \nabla_\Gamma g\cdot\boldsymbol{f}_\Gamma \mathrm{d}s + \int_{\partial\Gamma} g\boldsymbol{f}_\Gamma\cdot\boldsymbol{\tau}\mathrm{d}l.$$

The results follows from the following identities

$$\mathrm{div}_\Gamma(\boldsymbol{f}) = \mathrm{div}_\Gamma(\boldsymbol{f}_\Gamma) + \kappa\boldsymbol{f}\cdot\boldsymbol{n}, \tag{1.2.10}$$

$$\nabla_\Gamma g\cdot\boldsymbol{f}_\Gamma = \nabla_\Gamma g\cdot\boldsymbol{f}, \tag{1.2.11}$$

$$\boldsymbol{f}_\Gamma\cdot\boldsymbol{\tau} = \boldsymbol{f}\cdot\boldsymbol{\tau}. \tag{1.2.12}$$

□

The previous lemma 1.2 allows to compute shape derivatives of surface integrals.

Proposition 1.6. *Let Γ a smooth codimension one surface of $\mathbb{R}^d$ with boundary $\partial\Gamma$. For any $f \in W^{2,1}(\mathbb{R}^d)$, the functional $J(\Gamma)$ defined by*

$$J(\Gamma) := \int_\Gamma f\mathrm{d}s$$

is shape differentiable and the shape derivative reads

$$\begin{aligned}\mathrm{D}J(\Gamma)(\boldsymbol{\theta}) &= \int_\Gamma (\mathrm{div}(f\boldsymbol{\theta}) - \boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{n}f)\mathrm{d}s \\ &= \int_\Gamma \left(\frac{\partial f}{\partial \boldsymbol{n}} + \kappa f\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s + \int_{\partial\Gamma} f\boldsymbol{\theta}\cdot\boldsymbol{\tau}\mathrm{d}l,\end{aligned}$$

where $\boldsymbol{\tau}$ denotes the outward normal to $\partial\Gamma$ tangent to Γ.

Proof. The first equality follows easily from the application of (1.2.8) and differentiation with respect to $\boldsymbol{\theta}$ of

$$\int_{\Gamma_{\boldsymbol{\theta}}} f\mathrm{d}s = \int_\Gamma f\circ(I+\boldsymbol{\theta})\,|\det(I+\nabla\boldsymbol{\theta})|\,||(I+\nabla\boldsymbol{\theta})^{-T}\boldsymbol{n}||\mathrm{d}s. \tag{1.2.13}$$

The use of the previous lemma 1.2 yields the second equality because

$$\begin{aligned}\int_\Gamma \mathrm{div}(f\boldsymbol{\theta})\mathrm{d}s &= \int_\Gamma (\mathrm{div}_\Gamma(f\boldsymbol{\theta}) + \boldsymbol{n}\cdot\nabla(f\boldsymbol{\theta})\cdot\boldsymbol{n})\mathrm{d}s \\ &= \int_\Gamma \left(\kappa f(\boldsymbol{\theta}\cdot\boldsymbol{n}) + \frac{\partial f}{\partial\boldsymbol{n}}(\boldsymbol{\theta}\cdot\boldsymbol{n}) + \boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{n}f\right)\mathrm{d}s + \int_{\partial\Gamma} f(\boldsymbol{\theta}\cdot\boldsymbol{\tau})\mathrm{d}l.\end{aligned} \tag{1.2.14}$$

□

The last result of this part provides formulas for the shape derivatives of surface integrals involving the normal (which also depends on the shape). It is slightly less classical but is obtained thanks to very similar techniques.

Proposition 1.7. *Let Γ a smooth codimension one surface of $\mathbb{R}^d$ with boundary $\partial\Gamma$ and differentiable normal vector field $\boldsymbol{n}$. For any $\boldsymbol{f} \in W^{2,1}(\mathbb{R}^d, \mathbb{R}^d)$, the functional $J(\Gamma)$ defined by*

$$J(\Gamma) := \int_\Gamma \boldsymbol{f}\cdot\boldsymbol{n}\mathrm{d}s$$

is shape differentiable and the shape derivative reads

$$\begin{aligned}\mathrm{D}J(\Gamma)(\boldsymbol{\theta}) = &= \int_\Gamma (\boldsymbol{n}\cdot\nabla\boldsymbol{f}\cdot\boldsymbol{\theta} - \boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{f} + (\boldsymbol{f}\cdot\boldsymbol{n})\mathrm{div}(\boldsymbol{\theta}))\mathrm{d}s \\ &= \int_\Gamma \mathrm{div}(\boldsymbol{f})(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s + \int_{\partial\Gamma}[(\boldsymbol{f}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{\tau}) - (\boldsymbol{f}\cdot\boldsymbol{\tau})(\boldsymbol{\theta}\cdot\boldsymbol{n})]\mathrm{d}s,\end{aligned} \tag{1.2.15}$$

where $\boldsymbol{\tau}$ denotes the outward normal to $\partial\Gamma$ tangent to Γ.

Proof. The proposition implicitly understands that the transported unit normal $\boldsymbol{n_\theta}$ to $\Gamma_{\boldsymbol{\theta}} = (I + \boldsymbol{\theta})\Gamma$ is given by (see formula (1.2.9))

$$\boldsymbol{n_\theta} \circ (I + \boldsymbol{\theta}) := \frac{(I + \nabla\boldsymbol{\theta})^{-T}\boldsymbol{n}}{||(I + \nabla\boldsymbol{\theta})^{-T}\boldsymbol{n}||}.$$

Under this circumstance, $J(\Gamma_{\boldsymbol{\theta}})$ reads

$$\begin{aligned} J(\Gamma_{\boldsymbol{\theta}}) &= \int_{\Gamma_{\boldsymbol{\theta}}} \boldsymbol{f} \cdot \boldsymbol{n_\theta} \mathrm{d}s = \int_\Gamma \boldsymbol{f} \circ (I + \boldsymbol{\theta}) \cdot \boldsymbol{n_\theta} \circ (I + \boldsymbol{\theta}) \, |\det(I + \nabla\boldsymbol{\theta})| \, ||(I + \nabla\boldsymbol{\theta})^{-T}\boldsymbol{n}|| \mathrm{d}s \\ &= \int_\Gamma \boldsymbol{f} \circ (I + \boldsymbol{\theta}) \cdot (I + \nabla\boldsymbol{\theta})^{-T} \cdot \boldsymbol{n} \, |\det(I + \nabla\boldsymbol{\theta})| \mathrm{d}s. \end{aligned}$$

Differentiating this expression with respect to $\boldsymbol{\theta}$ yields then the first equality. We then use the fact that $\boldsymbol{\theta} \cdot \nabla\boldsymbol{n} \cdot \boldsymbol{f} = \boldsymbol{f} \cdot \nabla\boldsymbol{n} \cdot \boldsymbol{\theta}$ and $\nabla\boldsymbol{n} \cdot \boldsymbol{n} = 0$ in order to write

$$\int_\Gamma (\boldsymbol{n}\cdot\nabla\boldsymbol{f}\cdot\boldsymbol{\theta} - \boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{f})\mathrm{d}s = \int_\Gamma (\boldsymbol{n}\cdot\nabla\boldsymbol{f}\cdot\boldsymbol{n}(\boldsymbol{\theta}\cdot\boldsymbol{n}) - \boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{n}(\boldsymbol{f}\cdot\boldsymbol{n}) + \boldsymbol{\theta}\cdot\nabla_\Gamma(\boldsymbol{f}\cdot\boldsymbol{n}) - \boldsymbol{f}\cdot\nabla_\Gamma(\boldsymbol{\theta}\cdot\boldsymbol{n}))\mathrm{d}s. \quad (1.2.16)$$

Applying now lemma 1.2 to treat the term depending on $\mathrm{div}(\boldsymbol{\theta})$ in the first line of (1.2.15), we obtain:

$$\begin{aligned} \int_\Gamma (\boldsymbol{f}\cdot\boldsymbol{n})\mathrm{div}(\boldsymbol{\theta})\mathrm{d}s &= \int_\Gamma [(\boldsymbol{f}\cdot\boldsymbol{n})\mathrm{div}_\Gamma(\boldsymbol{\theta}) + (\boldsymbol{f}\cdot\boldsymbol{n})\boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{n}]\mathrm{d}s \\ &= \int_\Gamma (-\nabla_\Gamma(\boldsymbol{f}\cdot\boldsymbol{n})\cdot\boldsymbol{\theta} + \kappa(\boldsymbol{f}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{n}) + (\boldsymbol{f}\cdot\boldsymbol{n})\boldsymbol{n}\cdot\nabla\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s + \int_{\partial\Gamma} (\boldsymbol{f}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{\tau})\mathrm{d}s. \quad (1.2.17) \end{aligned}$$

It remains to sum (1.2.16) and (1.2.17) to obtain:

$$\mathrm{D}J(\Gamma)(\boldsymbol{\theta}) = \int_\Gamma (-\nabla_\Gamma(\boldsymbol{\theta}\cdot\boldsymbol{n})\cdot\boldsymbol{f} + \kappa(\boldsymbol{f}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{n}) + \boldsymbol{n}\cdot\nabla\boldsymbol{f}\cdot\boldsymbol{n}(\boldsymbol{\theta}\cdot\boldsymbol{n}))\mathrm{d}s + \int_{\partial\Gamma} (\boldsymbol{f}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{\tau})\mathrm{d}s \quad (1.2.18)$$

which is equivalent to the formula (1.2.15) after a second application of lemma 1.2. □

Remark 1.4. In the case where $\Gamma = \partial\Omega$ is the boundary of a bounded smooth open set Ω, the result is trivial because Stokes formula allows to write

$$\int_\Gamma \boldsymbol{f}\cdot\boldsymbol{n}\mathrm{d}s = \int_{\partial\Omega} \boldsymbol{f}\cdot\boldsymbol{n}\mathrm{d}s = \int_\Omega \mathrm{div}(\boldsymbol{f})\mathrm{d}x$$

from where the shape derivative can be obtained directly by applying proposition 1.4.

1.2.3 Shape derivatives of PDE constrained functionals

In this last subsection we review classical methods for the calculation of shape derivatives of functionals depending on the solutions to Partial Differential Equations (PDEs). For simplicity, we consider the model problem of the Laplace equation

$$\begin{cases} -\Delta u = f \text{ in } \Omega \\ \quad u = 0 \text{ on } \Gamma_D \\ \dfrac{\partial u}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_N \end{cases} \quad (1.2.19)$$

set on a Lipschitz domain Ω, where the boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ is divided into respective Dirichlet and Neumann parts Γ_D and Γ_N (the setting is represented on Figure 1.2). Let $V = \{v \in H^1(\Omega) \,|\, v = 0 \text{ on } \Gamma_D\}$. Recall that $u(\Omega) \in V$ is the unique function solving the variational problem

$$\text{Find } u \in V \text{ such that } \forall v \in V, \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x = \int_\Omega f v \mathrm{d}x. \quad (1.2.20)$$

In order to make the integral term of the right hand side shape differentiable, the source term f is assumed to belong to $H^1(\mathbb{R}^d)$, although the problem makes sense in general for $f \in H^{-1}(\Omega)$.

We are concerned with the derivation of the shape derivative of a functional $J(\Omega, u(\Omega))$ depending both on the shape Ω and on the solution $u(\Omega)$ to (1.2.19) (the dependence of u with respect to Ω shall be omitted when the context is clear). Both Neumann and Dirichlet boundaries Γ_N and Γ_D are allowed

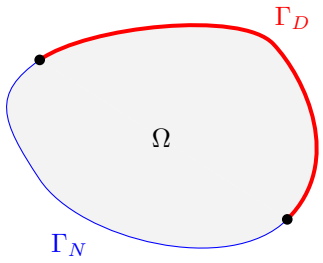


Figure 1.2: Setting for the Poisson problem of (1.2.19).

to deform. As is customary in the classical shape optimization literature, we assume (in this chapter only) $J(\Omega, u(\Omega))$ can be written in the form of a volume integral:

$$J(\Omega, u(\Omega)) := \int_\Omega j(x, u(x))\mathrm{d}x \tag{1.2.21}$$

where $j : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a $\mathcal{C}^1$ function satisfying $|\partial_u j(x,u)| \leq c|u|$ for some constant $c > 0$.

As a preliminary to the derivations of chapter 2 involving more complex PDEs, the full, rigorous, calculation of shape derivative based on Lagrangian derivatives is reviewed and compared to the so-called "fast" but only formal derivation method of Céa in the next paragraph. We shall see later in chapter 2 how to adapt (and even shorten, see the proof of chapter 2, proposition 2.2) these derivations to obtain formulas for *arbitrary* functionals $J(\Omega, u(\Omega))$.

Rigorous derivation based on Lagrangian derivatives

In this part, we follow the derivation of [184] which allows to obtain the shape sensitivity of $J(\Omega, u(\Omega))$ in a "safe" mathematically rigorous manner. The first step is to perform the usual change of variable $x = (I + \boldsymbol{\theta})(y)$:

$$J(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}})) = \int_{\Omega_{\boldsymbol{\theta}}} j(x, u(\Omega_{\boldsymbol{\theta}})(x))\mathrm{d}x = \int_\Omega j((I+\boldsymbol{\theta})(y), u(\Omega_{\boldsymbol{\theta}}) \circ (I+\boldsymbol{\theta})(y))\, \det(I + \nabla\boldsymbol{\theta})\mathrm{d}y. \tag{1.2.22}$$

This change of variable brings into play the transported solution $u_{\boldsymbol{\theta}} \in V$ defined by

$$u_{\boldsymbol{\theta}} := u(\Omega_{\boldsymbol{\theta}}) \circ (I + \boldsymbol{\theta}).$$

Importantly, the function $u_{\boldsymbol{\theta}}$ has the nice property to be defined on the *fixed space* V, while $u(\Omega_{\boldsymbol{\theta}})$ is defined on the space $H^1(\Omega_{\boldsymbol{\theta}})$ which depends on $\boldsymbol{\theta}$. The same change of variable in (1.2.20) yields a variational formulation satisfied by $u_{\boldsymbol{\theta}}$:

Find $u_{\boldsymbol{\theta}} \in V$ such that $\forall v \in V$,

$$\int_\Omega (I + \nabla\boldsymbol{\theta})^{-T}\nabla u_{\boldsymbol{\theta}} \cdot (I + \nabla\boldsymbol{\theta})^{-T}\nabla v \det(I + \nabla\boldsymbol{\theta})\mathrm{d}x = \int_\Omega f \circ (I + \boldsymbol{\theta}) \det(I + \nabla\boldsymbol{\theta})\, v\mathrm{d}x, \tag{1.2.23}$$

We have used the elementary but very useful lemma 1.3 below when performing change of variables involving gradients:

Lemma 1.3. *Let $f \in H^1(\mathbb{R}^d)$ and $\boldsymbol{f} \in H^1(\mathbb{R}^d, \mathbb{R}^d)$ be respectively scalar and vectorial functions, and $\boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ with $||\boldsymbol{\theta}||_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} < 1$. It holds*

$$(\nabla f) \circ (I + \boldsymbol{\theta}) = (I + \nabla\boldsymbol{\theta})^{-T}\nabla(f \circ (I + \boldsymbol{\theta})) \tag{1.2.24}$$

$$(\nabla \boldsymbol{f}) \circ (I + \boldsymbol{\theta}) = \nabla(\boldsymbol{f} \circ (I + \boldsymbol{\theta}))(I + \nabla\boldsymbol{\theta}). \tag{1.2.25}$$

Remark 1.5. The apparent asymmetry between (1.2.24) and (1.2.25) is only related to an unfortunate convention for the gradient notation ∇. For a vector field $\boldsymbol{f}$, the notation $\nabla\boldsymbol{f} := (\partial_j f_i)_{1\leq i,j\leq d}$ for the Jacobian matrix is not consistent with the one for the gradient $\nabla f := (\partial_i f)_{1\leq i\leq d}$ considered as a column vector (the Jacobian matrix of f is a linear form, hence it should be a row vector). We shall however maintain this convention commonly assumed, introducing a different notation in chapter 3 in contexts where a clear distinction between gradients and differential is needed.

Remark 1.6. From a practical point of view, (1.2.24) and (1.2.25) imply that when differentiating a variational formulation with respect to the shape:

1. the differentiation of a scalar gradient term ∇f yields a term $-\nabla\boldsymbol{\theta}^T\nabla f$;
2. the differentiation of a vectorial gradient term $\nabla \boldsymbol{f}$ yields a term $-\nabla \boldsymbol{f}\nabla\boldsymbol{\theta}$;
3. the differentiation of $\mathrm{d}x$ yields a term $\mathrm{div}(\boldsymbol{\theta})\mathrm{d}x$;
4. test functions v are not differentiated (because $v\circ(I+\boldsymbol{\theta})^{-1}$ can be chosen in (1.2.20)) (see [17])

These properties could be used formally e.g. for obtaining (1.2.28) below directly from the differentiation of (1.2.20) without writing actually (1.2.23).

The variational formulation (1.2.23) can be rewritten into an equation of the form $F(\boldsymbol{\theta},u_{\boldsymbol{\theta}})=0$ where F is the functional

$$\begin{array}{rcccl} F & : & W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)\times V & \to & V^* \\ & & (\boldsymbol{\theta},u) & \mapsto & v\mapsto \int_\Omega (I+\nabla\boldsymbol{\theta})^{-T}\nabla u\cdot(I+\nabla\boldsymbol{\theta})^{-T}\nabla v\det(I+\nabla\boldsymbol{\theta})\mathrm{d}x \\ & & & & \qquad -\int_\Omega f\circ(I+\boldsymbol{\theta})\det(I+\nabla\boldsymbol{\theta})\,v\mathrm{d}x. \end{array} \tag{1.2.26}$$

It can be verified that the partial derivative $\partial F/\partial u$ at $(0,u(\Omega))$ is the linear operator associated to the bilinear form $(u,v)\mapsto\int_\Omega\nabla u\cdot\nabla v\mathrm{d}x$, which is invertible from standard Lax-Milgram theory (see [184, 185]). This allows to apply the implicit function theorem ([212], Chap. 1, Theorem 5.9), which states that $\boldsymbol{\theta}\mapsto u_{\boldsymbol{\theta}}$ is differentiable at $\boldsymbol{\theta}=0$ (see also chapter 2, lemma 2.1 for more details). Its Fréchet derivative in a direction $\boldsymbol{\theta}$ is denoted

$$\dot{u}(\boldsymbol{\theta}) := \left.\frac{\mathrm{d}u_{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=0}(\boldsymbol{\theta}) \tag{1.2.27}$$

(the dependence with respect to Ω is omitted) and it is called the *Lagrangian derivative* of $u(\Omega)$. A variational formulation for $\dot{u}(\boldsymbol{\theta})$ is obtained by differentiating (1.2.23) with respect to $\boldsymbol{\theta}$:

$$\forall v\in V,\ \int_\Omega\nabla\dot{u}(\boldsymbol{\theta})\cdot\nabla v\mathrm{d}x=\int_\Omega(\nabla\boldsymbol{\theta}+\nabla\boldsymbol{\theta}^T-\mathrm{div}(\boldsymbol{\theta})I)\nabla u\cdot\nabla v\mathrm{d}x+\int_\Omega\mathrm{div}(f\boldsymbol{\theta})v\mathrm{d}x. \tag{1.2.28}$$

Remark 1.7. The Lagrangian derivative $\dot{u}(\boldsymbol{\theta})$ is the Fréchet derivative of $\boldsymbol{\theta}\mapsto u(\Omega_{\boldsymbol{\theta}})\circ(I+\boldsymbol{\theta})$. Another possible way to differentiate the domain dependent function $u(\Omega)$ is to consider the (Gâteaux) differential of the map $\boldsymbol{\theta}\mapsto u(\Omega_{\boldsymbol{\theta}})(x)$ for a fixed x, which is denoted

$$\forall x\in\Omega,\ u'(\boldsymbol{\theta})(x):=\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}u(\Omega_{\boldsymbol{\theta}})(x)$$

and is called the *Eulerian derivative* of $u(\Omega)$. Although this definition could seem more natural at first glance, the Eulerian derivative does not systematically exists in the same solution space than $u(\Omega)$. Indeed, a formal computation shows that Eulerian and Lagrangian derivatives are related by the formula

$$u'(\boldsymbol{\theta})(x)=\dot{u}(\boldsymbol{\theta})(x)-\nabla u(x)\cdot\boldsymbol{\theta}(x).$$

In our case, the Lagrangian derivative $\dot{u}(\boldsymbol{\theta})$ is an element of $H^1(\Omega)$, however $\nabla u(x)$ only belongs to $L^2(\Omega)$, which implies that $u'(\boldsymbol{\theta})$ has less regularity than $\dot{u}(\boldsymbol{\theta})$.

Coming back to $J(\Omega_{\boldsymbol{\theta}},u(\Omega_{\boldsymbol{\theta}}))$ in (1.2.22), the use of the chain rules yields then

$$J(\Omega_{\boldsymbol{\theta}},u(\Omega_{\boldsymbol{\theta}}))=J(\Omega)+\int_\Omega\left(\nabla_x j(x,u(x))\cdot\boldsymbol{\theta}+j(x,u(x))\mathrm{div}(\boldsymbol{\theta})+\frac{\partial j}{\partial u}(x,u(x))\dot{u}(\boldsymbol{\theta})\right)\mathrm{d}x+o(\boldsymbol{\theta}). \tag{1.2.29}$$

The formula is not yet satisfactory because (i) it involves the computation of $\dot{u}(\boldsymbol{\theta})$ which requires the resolution of (1.2.28) for any vector field $\boldsymbol{\theta}$, and (ii) it does not satisfy the Hadamard's structure theorem of proposition 1.1. The classical trick is to introduce an adjoint variable $p\in V$ solution to the variational problem

$$\text{Find } p\in V \text{ such that } \forall v\in V,\ \int_\Omega\nabla p\cdot\nabla v\mathrm{d}x=\int_\Omega\frac{\partial j}{\partial u}(x,u(x))v\mathrm{d}x. \tag{1.2.30}$$

The above equation means that p is the solution of the Poisson problem

$$\begin{cases} -\Delta p = \dfrac{\partial j}{\partial u}(x,u(x)) \text{ in } \Omega \\ \quad p = 0 \text{ on } \Gamma_D \\ \dfrac{\partial p}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_N. \end{cases} \tag{1.2.31}$$

Inserting $v = \dot{u}(\boldsymbol{\theta})$ in (1.2.30) and using (1.2.28) with $v = p$ yields indeed

$$\int_\Omega \frac{\partial j}{\partial u}(x,u(x))\dot{u}(\boldsymbol{\theta})\mathrm{d}x = \int_\Omega \nabla p \cdot \nabla \dot{u}(\boldsymbol{\theta})\mathrm{d}x = \int_\Omega (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nabla u \cdot \nabla p \mathrm{d}x + \int_\Omega \mathrm{div}(f\boldsymbol{\theta})p\mathrm{d}x.$$

Replacing the above formula into (1.2.29) yields the well known volumetric expression of the shape derivative of $J(\Omega, u(\Omega))$.

Proposition 1.8. *Assume $\Omega \subset D$ is a Lipschitz bounded open set and $f \in H^1(\mathbb{R}^d)$. The functional $J(\Omega, u(\Omega))$ defined by (1.2.21) is shape differentiable and the shape derivative reads*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}[J(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}}))](\boldsymbol{\theta}) = \int_\Omega \left(\frac{\partial j}{\partial x}(x,u(x)) \cdot \boldsymbol{\theta} + j(x,u(x))\mathrm{div}(\boldsymbol{\theta})\right)\mathrm{d}x \\ + \int_\Omega [(\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nabla u \cdot \nabla p + p\mathrm{div}(f\boldsymbol{\theta})]\mathrm{d}x. \end{aligned} \tag{1.2.32}$$

Remark 1.8. Classically and importantly, no further regularity of the state solutions u and p is required for the volume expression (1.2.32) to make sense. However, further regularity is needed for either rewriting it in the form of a boundary integral (see below).

The expression (1.2.32) is better than (1.2.29) in the sense that it does not depend on $\dot{u}(\boldsymbol{\theta})$, however it can still be simplified in order to obtain a boundary expression satisfying Hadamard's structure theorem. Such is obtained through an integration by part eliminating the derivatives of $\boldsymbol{\theta}$. Higher order derivatives of u and p are going to appear in the process: specifically, we need H^2 regularity of u and p, which generally holds if Ω is smooth except on a neighborhood of the interface $\Gamma_D \cap \Gamma_N$ between the Neumann and Dirichlet boundary (see [176] and [114] for a rigorous treatment of such issue in shape optimization).

Therefore, we assume for now that Ω is a smooth domain and that a neighborhood ω of $\Gamma_D \cap \Gamma_N \subset \omega$ is fixed: $\boldsymbol{\theta} = 0$ on ω. Therefore, it holds $u, p \in H^2(\Omega\backslash\omega)$ which allows us to perform the following integration by part:

$$\begin{aligned} \int_\Omega [(\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nabla u \cdot \nabla p \mathrm{d}x = -\int_\Omega \boldsymbol{\theta} \cdot (\mathrm{div}(\nabla p \otimes \nabla u + \nabla u \otimes \nabla p) - \nabla(\nabla u \cdot \nabla p))\mathrm{d}x \\ + \int_{\partial\Omega} [(\boldsymbol{\theta} \cdot \nabla p)(\nabla u \cdot \boldsymbol{n}) + (\boldsymbol{\theta} \cdot \nabla u)(\nabla p \cdot \boldsymbol{n}) - (\nabla u \cdot \nabla p)(\boldsymbol{\theta} \cdot \boldsymbol{n})]\mathrm{d}s. \end{aligned} \tag{1.2.33}$$

A small calculation allows to simplify the volume integrand:

$$\begin{aligned} -[\mathrm{div}(\nabla p \otimes \nabla u + \nabla u \otimes \nabla p) - \nabla(\nabla u \cdot \nabla p))] &= -\Delta u \nabla p - \Delta p \nabla u \\ &= f\nabla p + \frac{\partial j}{\partial u}(\cdot, u)\nabla u. \end{aligned} \tag{1.2.34}$$

Then the volume integrals involving j in (1.2.32) and (1.2.34) can be rewritten as follows:

$$\begin{aligned} \int_\Omega \left(\frac{\partial j}{\partial x}(x,u(x)) \cdot \boldsymbol{\theta} + j(x,u(x))\mathrm{div}(\boldsymbol{\theta}) + \nabla u \cdot \boldsymbol{\theta}\frac{\partial j}{\partial u}(x,u(x))\right)\mathrm{d}x \\ = \int_\Omega \mathrm{div}(j(x,u(x))\boldsymbol{\theta})\mathrm{d}x = \int_{\partial\Omega} j(x,u(x))\,\boldsymbol{\theta} \cdot \boldsymbol{n}\mathrm{d}s. \end{aligned} \tag{1.2.35}$$

The volume terms involving p can also be treated in the same manner:

$$\int_\Omega (p\mathrm{div}(f\boldsymbol{\theta}) + \nabla p \cdot \boldsymbol{\theta} f)\mathrm{d}x = \int_\Omega \mathrm{div}(fp\boldsymbol{\theta})\mathrm{d}x = \int_{\partial\Omega} fp\,\boldsymbol{\theta} \cdot \boldsymbol{n}\mathrm{d}s.$$

Finally, using the boundary conditions satisfied by u and p, the surface integral term of (1.2.33) can be rewritten

$$\int_{\partial\Omega}[(\boldsymbol{\theta}\cdot\nabla p)(\nabla u\cdot\boldsymbol{n})+(\boldsymbol{\theta}\cdot\nabla u)(\nabla p\cdot\boldsymbol{n})-(\nabla u\cdot\nabla p)(\boldsymbol{\theta}\cdot\boldsymbol{n})]\mathrm{d}s \\ =\int_{\Gamma_D}\frac{\partial u}{\partial\boldsymbol{n}}\frac{\partial p}{\partial\boldsymbol{n}}\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s-\int_{\Gamma_N}\nabla u\cdot\nabla p\,(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s.$$

All in all, we have obtained

Proposition 1.9. *Assume Ω is smooth and $f\in H^1(\mathbb{R}^d)$. If $\boldsymbol{\theta}=0$ on a neighborhood of $\Gamma_D\cap\Gamma_N$, then the shape derivative of $J(\Omega,u(\Omega))$ given by (1.2.32) rewrites as a boundary integral involving only the normal trace component $\boldsymbol{\theta}\cdot\boldsymbol{n}$ of $\boldsymbol{\theta}$:*

$$\left.\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=0}[J(\Omega_{\boldsymbol{\theta}},u(\Omega_{\boldsymbol{\theta}}))](\boldsymbol{\theta})=\int_{\partial\Omega}(j(x,u(x))+fp)\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s+\int_{\Gamma_D}\frac{\partial u}{\partial\boldsymbol{n}}\frac{\partial p}{\partial\boldsymbol{n}}\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s-\int_{\Gamma_N}\nabla u\cdot\nabla p\,(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s. \tag{1.2.36}$$

Remark 1.9. As emphasized in e.g. [168, 312], the volume expression (1.2.32) requires less regularity on the data (only the one obtained from the variational setting) than (1.2.36). For example, it is well-known that u_f,u_s may fail to be H^2 functions when the domains Ω_s or Ω_f involve corners [76, 176]; in such case (1.2.32) remains valid while the surface expression (1.2.36) may become invalid. Furthermore, some authors have found evidence that the Fréchet derivatives of shape functional are better approximated when discretizing the volume form [188]. Let us note, however, that regularity assumptions are still needed (i.e. $\mathrm{D}J(\Omega)$ continuous over $H^1(D,\mathbb{R}^d)$ and not only $W^{1,\infty}(D,\mathbb{R}^d)$) if one wants to identify $\mathrm{D}J(\Omega)$ with a H^1 scalar product.

Céa's "fast" derivation method

The previous result (1.2.36) shall now be retrieved with Céa's classical method, which is faster in the sense that it circumvents the need for computing the Lagrangian derivative $\dot{u}$ (eqn. (1.2.27)). However, the computation is only formal and is prone to errors. This part follows [17], section 6.4.3, in a slightly more complicated context. The principle of the method is to introduce a Lagrangian function

$$\mathcal{L}(\Omega,\widehat{u},\widehat{p},\widehat{\lambda}):=\int_\Omega j(x,\widehat{u}(x))\mathrm{d}x+\int_\Omega(f+\Delta\widehat{u})\widehat{p}\mathrm{d}x+\int_{\Gamma_D}\widehat{\lambda}\widehat{u}\mathrm{d}s+\int_{\Gamma_N}\widehat{\lambda}\frac{\partial\widehat{u}}{\partial\boldsymbol{n}}\mathrm{d}s \tag{1.2.37}$$

defined for $\widehat{u},\widehat{p},\widehat{\lambda}\in H^1(\mathbb{R}^d)$ (they are defined on $\mathbb{R}^d$ and not on Ω to make them independent from the shape Ω). The variable $\widehat{\lambda}$ is a Lagrange multiplier for boundary constraints, and $\widehat{p}$ is a Lagrange multiplier for the PDE constraint (1.2.19) satisfied by the solution $u(\Omega)$. In other words, the partial derivatives of the Lagrangian $\mathcal{L}$ with respect to $\widehat{p}$ and $\widehat{\lambda}$ cancel at $\widehat{u}=u$ where u is the solution to the Poisson problem (1.2.19):

$$\forall\widehat{p},\widehat{\lambda}\in H^1(\mathbb{R}^d),\ \frac{\partial\mathcal{L}}{\partial\widehat{p}}(\Omega,u,\widehat{p},\widehat{\lambda})=0\text{ and }\frac{\partial\mathcal{L}}{\partial\widehat{\lambda}}(\Omega,u,\widehat{p},\widehat{\lambda})=0.$$

The values of $\widehat{p}$ and $\widehat{\lambda}$ are sought in order to cancel the partial derivative $\partial\mathcal{L}/\partial\widehat{u}$ for $\widehat{u}=u$. Performing a double integration by part, it holds:

$$\forall v\in H^1(\mathbb{R}^d),\ \frac{\partial\mathcal{L}}{\partial\widehat{u}}(\Omega,u,\widehat{p},\widehat{\lambda})(v)=\int_\Omega\frac{\partial j}{\partial u}(x,\widehat{u}(x))v\mathrm{d}x+\int_\Omega\widehat{p}\Delta v\mathrm{d}x+\int_{\Gamma_D}\widehat{\lambda}v\mathrm{d}s+\int_{\Gamma_N}\widehat{\lambda}\frac{\partial v}{\partial\boldsymbol{n}}\mathrm{d}s \\ =\int_\Omega\frac{\partial j}{\partial u}(x,\widehat{u}(x))v\mathrm{d}x+\int_\Omega v\Delta\widehat{p}\mathrm{d}x+\int_{\partial\Omega}\left(\widehat{p}\frac{\partial v}{\partial\boldsymbol{n}}-v\frac{\partial\widehat{p}}{\partial\boldsymbol{n}}\right)\mathrm{d}s+\int_{\Gamma_D}\widehat{\lambda}v\mathrm{d}s+\int_{\Gamma_N}\widehat{\lambda}\frac{\partial v}{\partial\boldsymbol{n}}\mathrm{d}s. \tag{1.2.38}$$

Therefore this partial derivative vanishes by setting $\widehat{p}=p$ where p is the adjoint variable defined by (1.2.30), and

$$\widehat{\lambda}=\lambda:=\begin{cases}\dfrac{\partial p}{\partial\boldsymbol{n}}\text{ on }\Gamma_D\\ -p\text{ on }\Gamma_N.\end{cases}$$

The boundary expression for the shape derivative of $J(\Omega, u(\Omega)$ follows then by remarking that

$$J(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}})) = \mathcal{L}(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}}), \widehat{p}, \widehat{\lambda})$$

for any $\widehat{p}, \widehat{\lambda} \in H^1(\mathbb{R}^d)$, which allows to write

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}[J(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}}))](\boldsymbol{\theta}) &= \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}[\mathcal{L}(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}}), \widehat{p}, \widehat{\lambda})] \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathcal{L}(\Omega_{\boldsymbol{\theta}}, u(\Omega), \widehat{p}, \widehat{\lambda}))(\boldsymbol{\theta}) + \frac{\partial\mathcal{L}}{\partial\widehat{u}}(\Omega, u(\Omega), \widehat{p}, \widehat{\lambda})(u'(\boldsymbol{\theta})). \end{aligned} \tag{1.2.39}$$

where

$$u'(\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} u(\Omega_{\boldsymbol{\theta}}) \tag{1.2.40}$$

is the *Eulerian* derivative of $u(\Omega)$ (remark 1.7). Setting $\widehat{p} = p$ and $\widehat{\lambda} = \lambda$ in (1.2.39) allows to cancel the term involving $u'(\boldsymbol{\theta})$, which yields then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=0} [J(\Omega_{\boldsymbol{\theta}}, u(\Omega_{\boldsymbol{\theta}}))](\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}[\mathcal{L}(\Omega_{\boldsymbol{\theta}}, u(\Omega), p, \lambda)] \\ &= \int_{\partial\Omega} j(x, u(x))\, \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s + \int_{\Gamma_D} \left(\kappa + \frac{\partial}{\partial\boldsymbol{n}}\right)\left(\frac{\partial p}{\partial\boldsymbol{n}} u\right) \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s - \int_{\Gamma_N} \mathrm{div}(p\nabla u)\, \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s \\ &= \int_{\partial\Omega} j(x, u(x))\, \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s + \int_{\Gamma_D} \frac{\partial u}{\partial\boldsymbol{n}}\frac{\partial p}{\partial\boldsymbol{n}}\, \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s + \int_{\Gamma_N} (fp - \nabla u\cdot\nabla p)\, \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s. \end{aligned}$$

Note the use of proposition 1.7 with the assumption that $\boldsymbol{\theta} = 0$ on $\partial\Gamma_N$ in the shape differentiation of the term

$$\int_{\Gamma_N} \lambda\frac{\partial u}{\partial\boldsymbol{n}}\mathrm{d}s = \int_{\Gamma_N} (\lambda\nabla u)\cdot\boldsymbol{n}\mathrm{d}s.$$

This expression coincides with (1.2.36) remembering the boundary conditions satisfied by u and p. Albeit this method allows to avoid to explicit Eulerian or Lagrangian derivatives $u'(\boldsymbol{\theta})$ and $\dot{u}(\boldsymbol{\theta})$, it remains rather technical and does not work without hypothesis on the objective functional. Furthermore, this method remains formal (and may even yield wrong formulas, see e.g. [253]) because (i) the Eulerian derivative $u'(\boldsymbol{\theta})$ (eqn. (1.2.40)) has less regularity than $u(\Omega)$, which could imply

$$\frac{\partial\mathcal{L}}{\partial\widehat{u}}(\Omega, u(\Omega), p, \lambda)(u'(\boldsymbol{\theta})) \neq 0,$$

and (ii) the Lagrangian $\mathcal{L}$ defined in (1.2.37) implicitly assume very high regularity for u and p.

1.3 On the signed distance function and its main properties

The signed distance function to a domain (as illustrated on Figure 1.3) is a mathematical object commonly used in a variety of applications of the level-set method [96, 248, 280, 247] (reviewed in section 1.4.2 below). In the field of shape optimization based on the method of Hadamard, it also allows to formulate geometric distance constraints such as minimum or maximum thickness [234, 30]. Its definition and regularity properties are summarized in section 1.3.1. Section 1.3.2 recalls then differentiability results with respect to the domain and their use for obtaining the shape derivatives of distance constraints. Finally, a few words are given in section 1.3.3 regarding the numerical computation of the signed distance function on discretization meshes. The reader is referred to [64, 119, 122, 23] for proofs and much more exhaustive material.

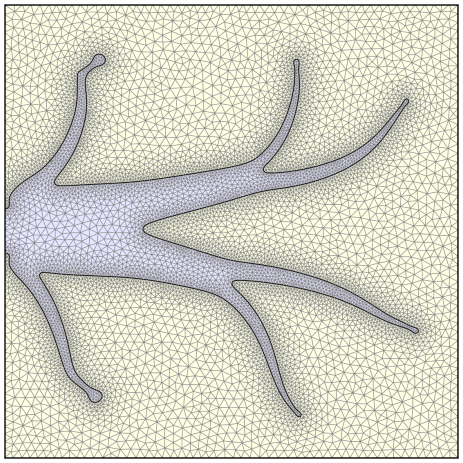
1.3.1 Definition of the signed distance function and its first two derivatives

Definition 1.2 (Signed distance function)**.** The signed distance function to a bounded open domain $\Omega \subset \mathbb{R}^d$ is the function

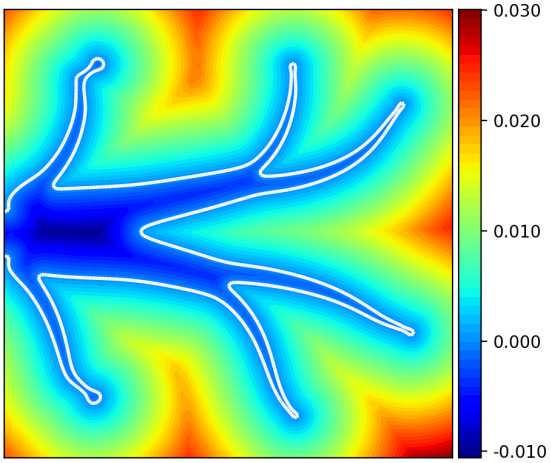
$$d_\Omega \,:\, \mathbb{R}^d \to \mathbb{R}$$

defined for any $x \in \mathbb{R}^d$ by

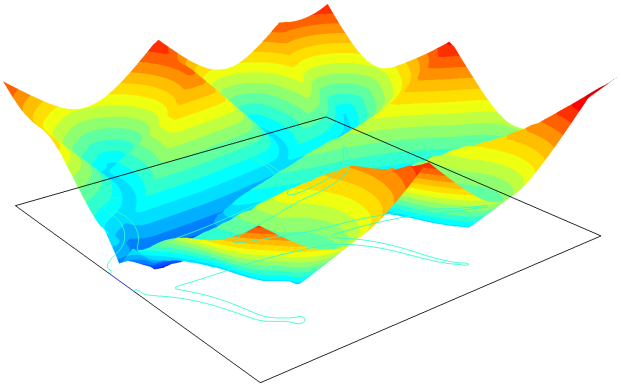
$$d_\Omega(x) = \begin{cases} -\inf\limits_{y\in\partial\Omega} ||x-y|| & \text{if } x \in \Omega, \\ \inf\limits_{y\in\partial\Omega} ||x-y|| & \text{if } x \notin \Omega. \end{cases}$$



(a) Meshed subdomain $\Omega \subset D$ (in blue) of a computational domain D.



(b) Isocontours of the signed distance function d_Ω.



(c) 3-d plot of d_Ω.

Figure 1.3: Example of signed distance function d_Ω numerically computed on a meshed domain.

It is obvious that (i) $d_\Omega(x) = 0$ for any $x \in \partial\Omega$, and (ii) that the infimum involved in the definition of Ω is attained in possibly several points of $\partial\Omega$ that are called *projections* onto $\partial\Omega$. It is useful to distinguish points that admit a unique projection from the others.

Definition 1.3 (Skeleton set and projection)**.**

1. The set of points $x \in \mathbb{R}^d$ for which the minimization problem

$$\min_{y\in\partial\Omega} ||x - y|| \tag{1.3.1}$$

admits several minimizers is called the *skeleton* of Ω and is denoted by Σ.

2. For any $x \in \mathbb{R}^d\backslash\Sigma$, the unique minimizer of (1.3.1) is denoted $p_{\partial\Omega}(x)$ and is called the (orthogonal) *projection* of x onto $\partial\Omega$, in that case it holds

$$\forall x \in \mathbb{R}^d\backslash\Sigma,\ d_\Omega(x) = \begin{cases} -||x - p_{\partial\Omega}(x)|| \text{ if } x \in \Omega, \\ \ \ ||x - p_{\partial\Omega}(x)|| \text{ if } x \notin \Omega. \end{cases}$$

The skeleton Σ of a domain Ω and the projection $p_{\partial\Omega}$ are illustrated on Figure 1.4.

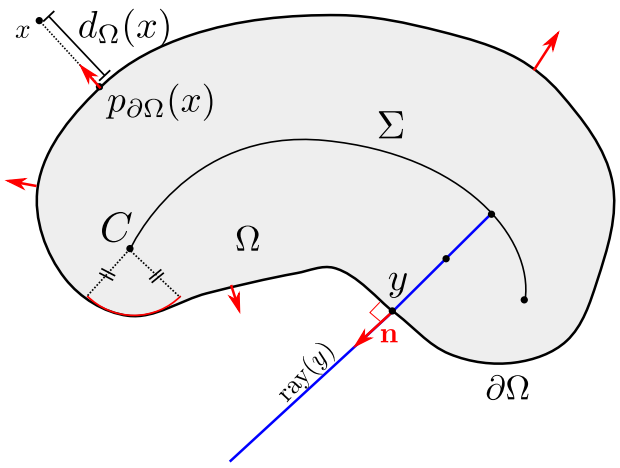


Figure 1.4: Shape $\Omega \subset \mathbb{R}^d$, skeleton Σ with normal rays $(\text{ray}(y))_{y\in\partial\Omega}$, outward normal vector $\boldsymbol{n}$, center of curvature C and orthogonal projection of a point $x \in D$ onto $\partial\Omega$.

It is elementary to show that without any further assumption, d_Ω is a Lipschitz map with Lipschitz constant smaller than one. Therefore, from Rademacher's theorem [146, 147], d_Ω is differentiable almost everywhere, and it can be proved that d_Ω is actually differentiable on $\mathbb{R}^d \setminus (\Sigma \cup \partial\Omega)$ which implies that Σ has zero Lebesgue measure (see [122, 146]). It is possible to be more explicit in the case where Ω is a $\mathcal{C}^1$ domain:

Proposition 1.10 (Differentiability of d_Ω)**.** *Assume Ω is a $\mathcal{C}^1$ domain with outward normal $\boldsymbol{n}$. The signed distance function d_Ω is differentiable at any point $x \in \mathbb{R}^d\backslash\Sigma$, and it is not differentiable on Σ. The gradient ∇d_Ω is an extension of the unit normal vector $\boldsymbol{n}$ to $\partial\Omega$ pointing outward Ω:*

$$\forall x \in \mathbb{R}^d\backslash\Sigma,\ \nabla d_\Omega(x) = \boldsymbol{n}(p_{\partial\Omega}(x)). \tag{1.3.2}$$

In particular, d_Ω solves the so-called "Eikonal" equation:

$$\begin{cases} ||\nabla d_\Omega|| = 1 \text{ in } \mathbb{R}^d \setminus \Sigma, \\ \quad d_\Omega = 0 \text{ on } \partial\Omega. \end{cases} \tag{1.3.3}$$

Formulas for the differential of the projection map $p_{\partial\Omega}$ and the Hessian $\nabla^2 d_\Omega$ are available under the additional assumption that Ω is a domain of class $\mathcal{C}^2$, referring to [2, 122] for less regular contexts. In that case, the use of the basis $(\boldsymbol{\tau}_i)_{1\le i\le d-1}$ associated with the principal curvatures $(\kappa_i)_{1\le i\le d-1}$ of $\partial\Omega$ (proposition 1.5) yields (see [36, 64, 82, 157] for proofs):

Proposition 1.11 (Differentiability of $p_{\partial\Omega}$ and Hessian of d_Ω)**.** *If Ω is a $\mathcal{C}^2$ bounded domain, then the following properties hold:*

1. *for any point $x \in \mathbb{R}^d \backslash \overline{\Sigma}$, one has:*

$$1 + \kappa_i(p_{\partial\Omega}(x))d_\Omega(x) > 0, \;\; i = 1, ..., d-1;$$

2. *the projection $p_{\partial\Omega}$ is differentiable on $\mathbb{R}^d \backslash \overline{\Sigma}$ and*

$$\forall x \in \mathbb{R}^d \backslash \overline{\Sigma}, \;\; \nabla p_{\partial\Omega}(x) = \sum_{i=1}^{d-1} \frac{1}{1 + \kappa_i(p_{\partial\Omega}(x))d_\Omega(x)} \boldsymbol{\tau}_i(p_{\partial\Omega}(x))\boldsymbol{\tau}_i(p_{\partial\Omega}(x))^T; \tag{1.3.4}$$

3. *the signed distance function d_Ω is twice differentiable on $\mathbb{R}^d \backslash \overline{\Sigma}$ and the Hessian $\nabla^2 d_\Omega$ is given by*

$$\forall x \in \mathbb{R}^d \backslash \overline{\Sigma}, \; \nabla^2 d_\Omega(x) = \sum_{i=1}^{d-1} \frac{\kappa_i(p_{\partial\Omega}(x))}{1 + \kappa_i(p_{\partial\Omega}(x))d_\Omega(x)} \boldsymbol{\tau}_i(p_{\partial\Omega}(x))\boldsymbol{\tau}_i(p_{\partial\Omega}(x))^T.$$

If in addition, Ω is a $\mathcal{C}^3$ domain, then $\mathbb{R}^d \setminus \overline{\Sigma}$ has zero Lebesgue measure (see [223]).

Remark 1.10. The projection $p_{\partial\Omega}$ is not differentiable on the boundary $\partial\Sigma$ of the skeleton Σ. Indeed (see Figure 1.4 and [82, 157]), a point $C \in \mathbb{R}^d$ belongs to the closure $\overline{\Sigma}$ either because it lies on Σ, or because it is a *center of curvature* of $\partial\Omega$, i.e. there exists $y \in \Pi_{\partial\Omega}(C)$ and $i = 1, ..., d-1$ such that

$$1 + d_\Omega(C)\kappa_i(y) = 0\,.$$

Remark 1.11. Formula (1.3.4) can be generalized to arbitrary embedded smooth manifolds, see e.g. [157, 156] where it is used in the context of matrix manifolds for the differentiation of matrix decompositions.

1.3.2 Differentiability of the signed distance function with respect to the domain and shape derivatives of distance functionals

Delfour and Zolésio [120] followed by [23, 107] demonstrated that the function d_Ω is differentiable with respect to the shape Ω in the following sense:

Proposition 1.12 ([23], Proposition 3.5)**.** *For any $x \notin \Sigma$, the map $\boldsymbol{\theta} \mapsto d_{(I+\boldsymbol{\theta})\Omega}(x)$ is Gâteaux-differentiable at $\boldsymbol{\theta}$ as an application from $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ into R^d and its derivative reads*

$$d'_\Omega(\boldsymbol{\theta})(x) = -\boldsymbol{\theta}(p_{\partial\Omega}(x)) \cdot \boldsymbol{n}(p_{\partial\Omega}(x)).$$

Sketch of proof. A formal proof of this result is obtained as follows. Recall the signed distance function d_Ω satisfies

$$\begin{cases} ||\nabla d_\Omega(x)|| = 1 \text{ in } \mathbb{R}^d \backslash \Sigma \\ \quad d_\Omega(x) = 0 \text{ on } \partial\Omega. \end{cases}$$

Differentiating the first equality with respect to Ω yields formally $\nabla d_\Omega \cdot \nabla(d'_\Omega(\boldsymbol{\theta})) = 0$, which means that $d'_\Omega(\boldsymbol{\theta})$ is constant along the rays normal to the shape. One then rewrites the second equality in a weak sense on the deformed domain $\Omega_{\boldsymbol{\theta}} = (I + \boldsymbol{\theta})\Omega$:

$$\forall \phi \in H^1(\mathbb{R}^d), \; \int_{\partial\Omega_{\boldsymbol{\theta}}} \phi \circ (I + \boldsymbol{\theta})^{-1} d_{\Omega_{\boldsymbol{\theta}}}(y)\mathrm{d}y = 0.$$

Differentiating this equation with respect to $\boldsymbol{\theta}$ then yields

$$\forall \phi \in H^1(\mathbb{R}^d), \; \int_{\partial\Omega} \phi \left(\kappa d_\Omega + \nabla d_\Omega \cdot \boldsymbol{n}\right) \boldsymbol{\theta} \cdot \boldsymbol{n}\mathrm{d}y + \int_{\partial\Omega} \phi d'_\Omega(\boldsymbol{\theta})\mathrm{d}y = 0,$$

from where $d'_\Omega(\boldsymbol{\theta}) = -\boldsymbol{\theta} \cdot \boldsymbol{n}$ on $\partial\Omega$ follows. □

In practice, this result is used for taking into account geometric constraints expressed as an integral criterion involving the signed distance function. Let $D \subset \mathbb{R}^d$ be a "hold-all" domain containing $\Omega \subset D$ as a subdomain. A geometric constraint functional $P(\Omega)$ is typically written in the form

$$P(\Omega) := \int_D j(d_\Omega(x))\mathrm{d}x \tag{1.3.5}$$

where $j : \mathbb{R} \to \mathbb{R}$ is a given $\mathcal{C}^1$ function. Maximum thickness, minimum thickness, minimum member's distance [30, 234], or the stress energy of a multimaterial medium involving a smooth interface [23] can be formulated in this form. The application of proposition 1.12 yields that the shape derivative of $P(\Omega)$ is (in a Gâteaux differentiability sense):

$$P'(\Omega)(\boldsymbol{\theta}) = \int_D j'(d_\Omega(x))d'_\Omega(\boldsymbol{\theta})(x)\mathrm{d}x. \tag{1.3.6}$$

This formula is, at first glance, not satisfying because it does not satisfy Hadamard's structure theorem (proposition 1.1). A more explicit expression is obtained thanks to a suitable change of variable which involves integrals along the normal rays to the shape. More precisely, for $y \in \partial\Omega$, the ray emerging from y is defined to be the one-dimensional segment

$$\mathrm{ray}(y) := \left\{x \in D\backslash\overline{\Sigma},\ \ p_{\partial\Omega}(x) = y\right\}.$$

The application of the coarea formula (decomposing D onto the level sets of $p_{\partial\Omega}$, see [23]), or a suitable change of variable (detailed later in chapter 4) yields:

$$P'(\Omega)(\boldsymbol{\theta}) = \int_{\partial\Omega}\left(\int_{z\in\mathrm{ray}(y)} j'(d_\Omega(z))\prod_{i=1}^{d-1}(1+\kappa_i(y)d_\Omega(z))\mathrm{d}z\right)\boldsymbol{\theta}(y)\cdot\boldsymbol{n}(y)\mathrm{d}y$$

which verifies Hadamard's structure theorem. This formula remains however very delicate to implement in practice; we shall introduce a new characterization for $P'(\Omega)(\boldsymbol{\theta})$ in chapter 4 which allows to evaluate the above integral without the need for explicitly calculating $\mathrm{ray}(y)$ nor the shape curvatures $\kappa_i(y)$ from the discretization of $\partial\Omega$.

1.3.3 Numerical methods for the computation of the signed distance function

In practical PDE constrained shape optimization algorithms, the domain $\Omega \subset D$ to optimize is known only in a discrete form. Most common types of discretization can be broadly divided into two categories:

1. explicit discretization: the shape Ω is explicitly discretized into a finite element mesh, such as in the example of Figure 1.3a.

2. implicit discretization: the shape Ω is implicitly represented by nodal values of a *level set* function ϕ discretized on a *fixed* meshed domain D:

   $$\Omega = \{x \in D \,|\, \phi(x) < 0\}.$$

   In that case, the shape Ω is retrieved from the 0 isocontour of the level set ϕ, as visible on Figure 1.3b.

Various numerical algorithms exist for computing the signed distance function d_Ω in these various discrete settings. These generally rely on the fact that d_Ω is a viscosity solution of the Eikonal equation (1.3.3), which can be computed as a steady-state solution of the so called "redistanciation" equation:

$$\partial_t\phi + \mathrm{sign}(\phi)(||\nabla\phi|| - 1) = 0. \tag{1.3.7}$$

It can be shown that the viscosity solution $\phi(t,\cdot)$ of (1.3.7) (see [61]) converges indeed to d_Ω where $\Omega = \{x \in D \,|\, \phi(0,x) < 0\}$. This can be used:

1. to devise numerical schemes on *fixed grid* transforming an input level set function $\phi(0,\cdot)$ into a signed distance function d_Ω, see e.g. [247], chapter 7. This is of particular interest in implicit methods because the signed distance function is a particular level set function characterized by good numerical properties [247];

2. to devise numerical algorithms on triangular or tetrahedral mesh to build the signed distance function of a subdomain discretized as a submesh, provided the distance has been computed first on vertex adjacent to the subdomain boundary. This approach is e.g. followed by [111] and implemented in the software `mshdist`, which we used in our own implementations of shape optimization test cases.

Several other methods exist for the computation of the signed distance function such as the Fast Marching Method (see [280], or [204] for a version on simplicial meshes), or the Fast Sweeping method [323]; the reader is referred to [107], section 1.3.1. for a review.

1.4 A CLASSICAL SHAPE OPTIMIZATION NUMERICAL WORKFLOW USING A LEVEL-SET BASED MESH EVOLUTION METHOD

This last section describes the classical main steps involved in the numerical implementation of the method of Hadamard in order to solve shape optimization problems of the form

$$\min_{\Omega} \; J(\Omega), \tag{1.4.1}$$

where J may depend on the solutions to some partial differential equations as in (1.2.21).

In practice, additional equality or inequality constraints come into play in most applications. For now, it is sufficient to assume that these can be treated by penalization techniques which reduce constrained problems to unconstrained one of the same form of (1.4.1), This approach has been actually commonly used in literature dealing with problems featuring only a few constraints [190, 203, 29, 234]. For instance, if there is only one constraint $P(\Omega) = 0$, the Augmented Lagrangian Method [244] considers

$$\min_{\Omega} \; J(\Omega) + \lambda P(\Omega) + \frac{1}{2}\mu P(\Omega)^2, \tag{1.4.2}$$

where λ and μ are tuning parameters which may be updated in the course of the optimization process. This method is unsatisfying for a number of reasons; for constrained optimization with an arbitrary number of constraints, all our numerical examples rely in fact on a different algorithm which is described in details in the dedicated chapter 3.

The main steps of a typical topology optimization algorithm for the minimization of PDE constrained functionals are summarized in algorithm 1.1 below. These algorithms involve two classical cornerstones: the computation of a descent direction (a vector field $\boldsymbol{\theta} \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ yielding improved design shapes), and the representation and practical updates of shapes in the numerical setting. These two steps are reviewed in details in the next two sections. section 1.4.1 reviews the well known necessity to regularize and extend the shape derivative and its consistency with respect to the gradient method. Then, the problem of numerically evolving shapes is discussed in sections 1.4.1 and 1.4.2, which includes a summary of the level set mesh evolution method which we used in our numerical test cases in the latter section.

Algorithm 1.1 Shape optimization with the method of Hadamard and a level set based mesh evolution algorithm.

Generate an initial mesh $\mathcal{T}_0$ for a computational domain D adapted to a first guess subdomain $\Omega_0 \subset D$.
for $n = 0, 1, 2 \dots$ **do**
1. Solve the physical equations posed on Ω_n (e.g. (1.2.19)).
2. Assemble the shape derivative $\mathrm{D}J(\Omega_n)$ (e.g. (1.2.36) and section 1.2.2) after solving adjoint equations (such as (1.2.30)).
3. Identify the shape derivative $\mathrm{D}J(\Omega_n)$ to a gradient $\nabla J(\Omega_n)$ by solving an extension regularization problem of the form of (1.4.4) (see section 1.4.1).
4. Update the shape according to the descent direction $\boldsymbol{\theta}_n = -\Delta t \nabla J(\Omega_n)$ where Δt is a small discretization step. This can be done by using mesh deformations (with (1.4.3)) or with the level set based mesh evolution algorithm of algorithm 1.2 (with (1.4.12)), see section 1.4.2.

end for

1.4.1 Gradient based optimization in the context of the method of Hadamard: identification, regularization, and extensions of shape derivatives

Numerical algorithms for solving shape optimization problems of the form

$$\min_{\Omega} \; J(\Omega).$$

compute a minimizing sequence of domains $(\Omega_n)_{n\in\mathbb{N}}$ which gradually decrease the objective function $J(\Omega)$. In the context of the method of Hadamard, the domain Ω_{n+1} is obtained by deformation of Ω_n along a vector field $\boldsymbol{\theta}_n \in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$:

$$\Omega_{n+1} = (I+\boldsymbol{\theta}_n)\Omega_n, \quad \forall n \geq 0. \tag{1.4.3}$$

The principle of the gradient method is to use the knowledge of the shape derivative in order to build a descent direction $\boldsymbol{\theta}_n$ that decreases the objective function $J(\Omega)$ in some optimal sense.

In this infinite dimensional context, the shape derivative $\mathrm{D}J(\Omega)$ of the objective function $J(\Omega)$ at a given domain Ω is not strictly speaking a descent direction: indeed, $\mathrm{D}J(\Omega)$ belongs to the dual space $W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)^*$ and not to $W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ itself. This issue is numerically related to the extension and regularization step of the shape derivative, which is very classical issue acknowledged in a number of works since the early developments of shape optimization [87, 86, 49, 48, 81, 34, 117, 236]. Practically, a descent direction can be obtained by identifying the linear form $\mathrm{D}J(\Omega) \in W^{1,\infty}(D,\mathbb{R}^d)^*$ to an actual gradient $\nabla J(\Omega) \in W^{1,\infty}(D,\mathbb{R}^d)$. To do so, it is sufficient to consider a Hilbert space V sufficiently "small" for the inclusion $V \subset W^{1,\infty}(\Omega,\mathbb{R}^d)$ to hold. The converse inclusion for the dual spaces, $W^{1,\infty}(\Omega,\mathbb{R}^d)^* \subset V^*$, implies the continuity of $\mathrm{D}J(\Omega)$ on V. Therefore, denoting by $\langle\cdot,\cdot\rangle_V$ the scalar product of V, the standard Riesz representation theorem [76, 49, 81, 117] yields the existence of a unique element $\nabla J(\Omega) \in V$ such that

$$\forall \boldsymbol{\theta} \in V,\ \langle \nabla J(\Omega), \boldsymbol{\theta}\rangle_V = \mathrm{D}J(\Omega)(\boldsymbol{\theta}). \tag{1.4.4}$$

The opposite of the vector field $\nabla J(\Omega) \in V \subset W^{1,\infty}(D,\mathbb{R}^d)$ can then serve as a descent direction to obtain a better shape. Indeed, it holds for any sufficiently small step size $\Delta t > 0$

$$\begin{aligned} J((I-\Delta t\nabla J(\Omega))\Omega) &= J(\Omega) - \Delta t \mathrm{D}J(\Omega)(\nabla J(\Omega)) + o(\Delta t) \\ &= J(\Omega) - \Delta t \langle \nabla J(\Omega), \nabla J(\Omega)\rangle_V + o(\Delta t), \end{aligned} \tag{1.4.5}$$

with $-\langle \nabla J(\Omega), \nabla J(\Omega)\rangle_V \leq 0$, which shows that updating Ω by deformation with the vector field $\boldsymbol{\theta} = -\Delta t \nabla J(\Omega)$ leads to a non-increase of the objective function. Furthermore, $-\nabla J(\Omega)$ is a "best" descent direction in the sense that it is better than any other element of V having the same norm: in chapter 3, we recall that if $\mathrm{D}J(\Omega) \neq 0$, then

$$-\frac{\nabla J(\Omega)}{||\nabla J(\Omega)||_V} = \min_{\substack{\boldsymbol{\theta}\in V,\\ ||\boldsymbol{\theta}||_V \leq 1}} \mathrm{D}J(\Omega)(\boldsymbol{\theta}).$$

As for the choice of the Hilbert space $V \subset W^{1,\infty}(D,\mathbb{R}^d)$ used in the identification (1.4.4), one can set $V = H^m(D,\mathbb{R}^d)$ with $m > 1 + d/2$, equipped with its standard inner product: indeed, the inclusion $H^m(D,\mathbb{R}^d) \subset W^{1,\infty}(D,\mathbb{R}^d)$ holds as a consequence of the Sobolev embedding theorem [76]. In this case, the identification problem (1.4.4) reduces to a linear elliptic problem of order $2m$. What is more, when $\mathrm{D}J(\Omega)(\boldsymbol{\theta})$ can be written in the form of a boundary integral, namely when there exists $v_J(\Omega) \in L^1(\partial\Omega)$ such that

$$\forall \boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d),\ \mathrm{D}J(\Omega)\boldsymbol{\theta} = \int_{\partial\Omega} v_J(\Omega)\ \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s, \tag{1.4.6}$$

then $\nabla J(\Omega) \in H^m(D,\mathbb{R}^d)$ is a regularized extension of the field $v_J(\Omega)\boldsymbol{n}$ to the whole domain D, which turns to be very convenient for practical numerical algorithms.

A very common strategy in the literature (see for instance [34, 49, 81, 153, 117, 236]) , though, consists in taking simply $V = H^1(D,\mathbb{R}^d)$ with the inner product

$$\forall \boldsymbol{\theta},\boldsymbol{\theta}' \in V,\ \langle \boldsymbol{\theta},\boldsymbol{\theta}'\rangle_V := \int_D (\gamma^2 \nabla\boldsymbol{\theta} : \nabla\boldsymbol{\theta}' + \boldsymbol{\theta}\cdot\boldsymbol{\theta}')\mathrm{d}x, \tag{1.4.7}$$

where $\gamma > 0$ is a user-defined parameter which can physically be interpreted as a length-scale for the regularity of the deformations $\boldsymbol{\theta}$; typically, γ is of the order of the minimum mesh element size of the

discretization. Variants can be considered for tuning more finely the smoothness of such extensions, or to prescribe non optimizable boundaries by setting zero Dirichlet boundary conditions for $\nabla J(\Omega)$ in (1.4.6). This choice of V is *a priori* only formal because V is not rigorously a subspace of $W^{1,\infty}(D,\mathbb{R}^d)$: it works only if $\mathrm{D}J(\Omega)$ is continuous over $H^1(D,\mathbb{R}^d)$. This condition is actually not a limitation because it turns to be often satisfied for PDE constrained problems under suitable regularity assumptions on Ω and the data. Furthermore, this choice is very convenient numerically because this space is easily discretized with $\mathbb{P}_1$ finite elements. Since this leads to very good results in practice, we rely on this strategy in our own numerical implementation.

Last, it is possible to even reduce the cost of the computation of $\nabla J(\Omega)$ to the one of the resolution of a simple scalar elliptic problem if one considers $V = \{v\nabla d_\Omega \,|\, v \in H^1(\Omega)\}$ equipped with the scalar product $\langle\cdot,\cdot\rangle_V$ defined by

$$\forall v, w \in H^1(\Omega),\ \langle v\nabla d_\Omega, w\nabla d_\Omega\rangle_V := \int_\Omega (\gamma\nabla v\cdot\nabla w + vw)\mathrm{d}x. \tag{1.4.8}$$

Remembering that $\nabla d_\Omega(x) = \boldsymbol{n}(p_{\partial\Omega}(x))$ is a continuous extension of the outward vector field normal to $\partial\Omega$ (proposition 1.10), (1.4.4) reduces to the identification of a single normal component v for the updating vector field $\boldsymbol{\theta} = -v\nabla d_\Omega$. In practice, one resorts to (1.4.8) when using boundary integral expressions of shape derivatives such as (1.2.36), and to (1.4.7) when relying rather on volumetric expressions of the form of (1.2.32).

1.4.2 Numerical representations and updates of shapes: mesh deformations, level set methods, and level set based mesh evolution method

One of the key issues of shape optimization algorithms lies in the numerical implementation of the shape update step (1.4.3). This step involves (i) a numerical representation of the shape Ω_n and (ii) a process that allows to obtain a new shape Ω_{n+1} in this representation from a displacement vector field $\boldsymbol{\theta}_n$. Furthermore, in a context involving partial differential equations, the numerical representation of Ω should be compatible with physical solvers relying e.g. on the Finite Element method.

In what follows, two complementary numerical representations of shapes together with their related method for updating them are briefly reviewed: explicit methods which evolve a meshed representation, and *level set* methods. Then, we briefly describe the main steps of the level set mesh evolution algorithm of Allaire, Dapogny and Frey [25], which combines advantages of both representations.

Explicit mesh representations and nodal displacements

The first implementations of the method of Hadamard for shape optimization were considering explicit discretizations of shapes [259, 86, 34]. An open domain Ω is typically represented by a simplicial mesh (a set of triangles or tetrahedra) which is globally *conforming* in the sense that the intersection between any two triangles or tetrahedra is either empty, or reduced to a vertex, or an edge, or a triangle of the mesh (Figure 1.5).

This representation is convenient for computing solutions to partial differential equations on Ω with the finite element method domain such as (1.2.19), as well as volume or surface integrals involved in the evaluation of shape derivatives such as (1.2.32) and (1.2.36). Furthermore, it is easily amenable to *small* deformation updates by translations of mesh vertices. If Ω is discretized into a mesh involving N vertices $x_1,\dots,x_N$, then a deformed shape $(I+\boldsymbol{\theta})\Omega$ with $\boldsymbol{\theta}\in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ is obtained by applying the transformation

$$x_i \leftarrow x_i + \boldsymbol{\theta}(x_i),\quad 1\le i\le N.$$

The process is not computationally expensive and can be iterated in order to implement (1.4.3) for obtaining a minimizing sequence Ω_n from descent updates $\boldsymbol{\theta}_n$ until a final design. However, it is well-known [259] that in very few iterations, the quality of the deformed mesh may quickly decrease due to the occurrence of nearly flat or inverted mesh elements, calling for the use of remeshing, or for the use of other numerical shape representations.

Level set methods: advection and Hamilton-Jacobi equations

The level set method was initially introduced by S. Osher and J. Sethian for solving free boundary problems in fluid mechanics [248], before being used for shape optimization in [31, 32, 250, 281, 311].

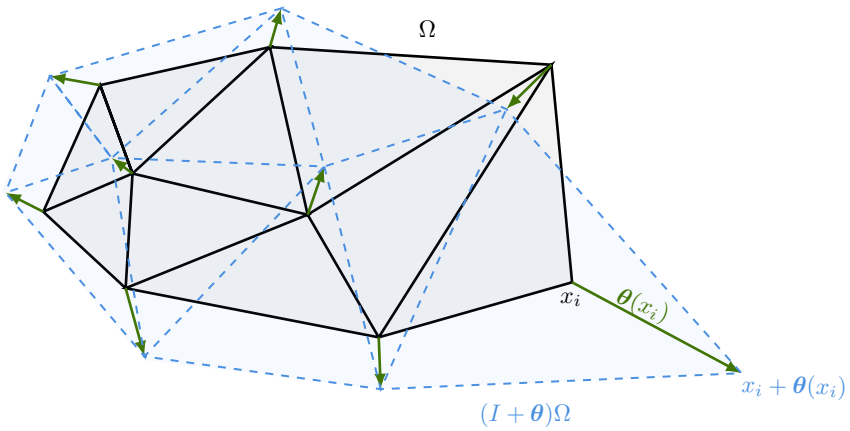


Figure 1.5: Discretization of a 2-d domain Ω in a simplicial mesh and its deformation by application of a displacement vector field $\boldsymbol{\theta}$.

Given a fixed computational domain D, it amounts to representing any subdomain $\Omega \subset D$ as the negative subdomain of a scalar "level set" function $\phi : D \to \mathbb{R}$:

$$\begin{cases} \phi(x) < 0 & \text{if } x \in \Omega, \\ \phi(x) = 0 & \text{if } x \in \partial\Omega, \\ \phi(x) > 0 & \text{if } x \in D \setminus \Omega. \end{cases} \tag{1.4.9}$$

The motion of a domain $\Omega(t)$ in D according to a vector velocity field $\boldsymbol{\theta}(x)$ translates then in terms of an associated level set function $\phi(t, x)$ as the following advection equation:

$$\begin{cases} \dfrac{\partial \phi}{\partial t}(t, x) + \boldsymbol{\theta}(t, x) \cdot \nabla \phi(t, x) = 0, & x \in D, \\ \phi(0, x) = \phi_0(x), & x \in D, \end{cases} \tag{1.4.10}$$

where ϕ_0 is any level set function for Ω. This representation is very convenient for describing the evolution of domains at the discrete level on *fixed* meshes, because topological changes are handled naturally and automatically (Figure 1.6).

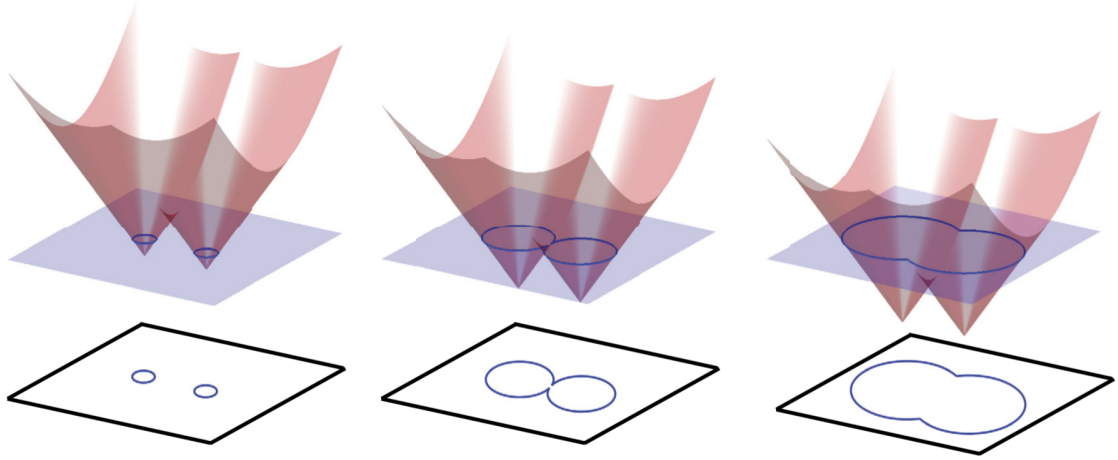


Figure 1.6: Motion of an implicit domain featuring a topological change and its representation by level set functions. Figure from [170].

When the time dependent domain $\Omega(t)$ evolves according to a motion $\boldsymbol{\theta}(t, x)$ normal to the interface $\partial\Omega(t)$, *i.e.* $\boldsymbol{\theta}(t, x) = v(t, x)\boldsymbol{n}$ on $\partial\Omega$ for some scalar function $v(t, x)$, (1.4.10) rewrites as the following Hamilton-Jacobi equation:

$$\frac{\partial \phi}{\partial t}(t, x) + v(t, x)||\nabla \phi(t, x)|| = 0, \;\; t > 0, \;\; x \in \Omega. \tag{1.4.11}$$

Generally, the level set function is initialized to the signed distance function of the initial domain and reinitialized from times to times, which allows to reduce numerical diffusion and round off errors [96, 247]. In applications where the computational domain D is discretized into a finite difference or finite volume cartesian grid, one usually relies on the Hamilton-Jacobi equation (1.4.11) for the evolution of implicit domains $\Omega(t)$ represented by level set functions $\phi(t,\cdot)$. Indeed, efficient and accurate numerical schemes are available for the resolution of (1.4.11) on structured grids [249, 247]. In contexts involving finite element unstructured meshes with level set functions ϕ discretized as a set of nodal values, it can be more convenient to solve the evolution equation (1.4.10), e.g. with the method of characteristics. This is the method we favored for all our numerical examples; we relied on the implementation available in the software `advect` issued from [80] and which shares common features with the semi-Lagrangian method of J. Strain [295].

In shape optimization algorithms (such as in algorithm 1.1) relying on the level set method, either of the evolution equations (1.4.10) or (1.4.11) is solved with respectively $\boldsymbol{\theta}(t,x) \equiv \boldsymbol{\theta}_n(x)$ or $v(t,x) \equiv v_n(x)\boldsymbol{n}(t,x)$, where the (time-independent) descent directions $\boldsymbol{\theta}_n(x)$ and $v_n(x)$ are computed to evolve the current shape Ω_n into the next iterate Ω_{n+1}.

Let us note that the level set method for moving domain boundaries (either with (1.4.10) or (1.4.11)) actually involves domain updates different of those (1.4.3) considered by the method of Hadamard : if the current domain Ω_n is to be deformed according to a descent direction $\boldsymbol{\theta}_n \in W^{1,\infty}(D,\mathbb{R}^d)$, then the updated domain Ω_{n+1} is obtained by

$$\Omega_{n+1} = \widetilde{\rho}_{\Omega_n}(\boldsymbol{\theta}_n),$$

where for a given subdomain $\Omega \subset D$, $\widetilde{\rho}_\Omega$ is the mapping defined by

$$\widetilde{\rho}_\Omega(\boldsymbol{\theta}) := \{x \in D \,|\, \phi(1,x) < 0\}, \qquad \boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d) \tag{1.4.12}$$

with $\phi(t,x)$ the solution to either (1.4.10) of (1.4.11). The mapping $\widetilde{\rho}_\Omega$ is different from the one $\rho_\Omega(\boldsymbol{\theta}) := (I+\boldsymbol{\theta})\Omega$ considered in the Hadamard's method. However, it can be shown that there is no loss of consistency when using either $\widetilde{\rho}_\Omega(\boldsymbol{\theta})$ or $\rho_\Omega(\boldsymbol{\theta})$ in the implementation of first order optimization methods. In fact, when $\phi(t,\cdot)$ is obtained with the advection equation (1.4.10), then $\widetilde{\rho}_\Omega$ is the update map considered in the *speed* method of Zolésio for boundary variations [291], which is equivalent at first order to the method of Hadamard. When $\phi(t,\cdot)$ is obtained by (1.4.11), $\widetilde{\rho}_\Omega$ corresponds to moving the points of $\partial\Omega$ along *bicharacteristics* associated to the Hamilton-Jacobi equation, which is also a second order perturbation of the method of Hadamard (see [21] where this is discussed in details). An equivalent way to formulate this statement is that the domain update maps ρ_Ω and $\widetilde{\rho}_\Omega$ can be interpreted as two admissible *retractions* converting descent directions into new optimization points, this point is further explained in chapter 3.

A common difficulty of the implicit representation on a fixed computational mesh is that it does not make readily possible to solve for original physical equations such as (1.2.19) posed on the implicit domain Ω: indeed, the location of Ω is known only through the values of the level set function ϕ. Classical level set based shape optimization methods use the so-called "*ersatz*" or fictitious material approach [32] (other interpolation methods are possible in such level set descriptions, such as XFEM or cutFEM methods [140, 309]). For instance, in the context of the Laplace problem considered in section 1.2.3, if only the Neumann boundary Γ_N is optimizable and $\Gamma_D \subset \partial D$, this amounts to replace (1.2.19) by an approximate problem set on the whole domain D:

$$\begin{cases} -\mathrm{div}(A_\varepsilon(\Omega)\nabla u_\varepsilon) = f \text{ in } D \\ \qquad\qquad\qquad u_\varepsilon = 0 \text{ on } \Gamma_D. \end{cases} \tag{1.4.13}$$

The parameter A_ε is a fictitious conductivity set equal to 1 in the domain Ω and to a (small) "ersatz" value ε in the complementary $D\backslash\Omega$. The solution u_ε can be computed on the mesh adapted to D, and it satisfies $u_\varepsilon \to u$ in the L^2 norm on Ω as $\varepsilon \to 0$. Using regularized Heaviside and Dirac functions [247], it becomes possible to assemble finite element matrices associated with (1.4.13) and to estimate volume integrals on Ω or surface integrals on $\partial\Omega$ for the implementation of shape optimization algorithms [31, 311].

Finally, let us remark that if the level set method allows *large* dmomain deformations up to topological changes to occur, these are not natively accounted for by the method of Hadamard. Indeed, shape derivatives (definition 1.1) quantify the sensitivity of objective functionals with respect to *small* deformations only, without topology changes. Therefore, topological changes occurring in the course of

optimization updates with the level set method are the consequence of the finite resolution of both the spatial and time discretization. For instance, very thin parts of an evolved domain $\Omega(t)$ typically disappear because either of some numerical diffusion smearing out small details, or because the advection step chosen for the numerical resolution of (1.4.10) is too large with respect to the smallest mesh element size. There is no guarantee that an objective function $J(\Omega)$ keeps decreasing when such an event suddenly occurs, however it can be expected that the optimization carries on smoothly afterwards until another topological change happens or until convergence.

The level set based mesh evolution algorithm of Allaire, Dapogny and Frey [24]

In all numerical examples of the present work, we use the level set based mesh evolution method introduced by Allaire, Dapogny and Frey [24]. The main idea is to combine both numerical representations of the domain $\Omega \subset D$ seen as a subset of a computational domain D (see Figure 1.3):

- on the one hand, a computational mesh $\mathcal{T}$ of D is available, in which Ω is explicitly discretized as a meshed subdomain;
- on the other hand, Ω is *implicitly* described, using the level set method: e.g. it can be seen as the negative subdomain of a function $\phi : D \to \mathbb{R}$.

It becomes then possible to consistently alternate between both descriptions of the domain Ω depending on the nature of the ongoing operation: finite element resolutions are carried out using the meshed description, while the motion of the phases is tracked using the level set method.

The idea of remeshing shapes in between shape optimization iterations was first proposed by Persson and implemented on a test case in his thesis [256]. For our applications, we rely on the subsequent work of [108] for implicit domain meshing. The main features of this method are now outlined in the case of two space dimensions; referring to [25] for further details. Our numerical shape optimization test cases relied on the library `mmg` in which the remeshing step of a subdomain described by a level set is implemented both in 2-d and in 3-d.

In what follows, we assume a vector field $\boldsymbol{\theta}$ is given at the nodes of the computational mesh $\mathcal{T}$ discretizing D and $\Omega \subset D$ (Figure 1.7a). The objective of the algorithm is to compute a new mesh $\mathcal{T}'$ adapted to the discretization of the evolved domain $\Omega' := \widetilde{\rho}_\Omega(\boldsymbol{\theta})$, where $\widetilde{\rho}_\Omega$ (eqn. (1.4.12)) encloses the advection step as dictated by the transport equation (1.4.10). The main steps of the algorithm are summarized in algorithm 1.2 and illustrated on Figure 1.8.

Algorithm 1.2 Domain update by mean of a level set based mesh evolution method.

1. A level set function ϕ associated with the domain $\Omega \subset D$ is computed at the nodes of the mesh $\mathcal{T}$ (Figure 1.7b). As is classical in level set methods, we choose $\phi = d_\Omega$ to be given by the signed distance function of Ω, although any other level set function smooth in a band around $\partial\Omega$ would be sufficient.
2. ϕ is advected on the fixed mesh D by solving the evolution equation (1.4.10) (Figure 1.8a).
3. The new domain Ω' (Figure 1.8b) is discretized by splitting mesh elements of $\mathcal{T}$ according to the zero level set of the updated level set. This part is purely combinatorial and yields a conforming but poor quality mesh.
4. Remeshing operations (explained in details [108, 25, 107, 153]) are performed iteratively to improve the mesh quality and the approximation of the discrete surface $\partial\Omega'$. Such is possible thanks to the computation of a metric based on the Haussdorff distance to the surface $\partial\Omega'$. This yields a new high quality mesh $\mathcal{T}'$ adapted to the new subdomain Ω' (Figure 1.8c).

Finally, let us , let us mention the existence of the *Deformable Simplicial Complex* (DSC) method for shape and topology optimization [99, 98], which is an alternative method for evolving explicit meshed representations of shapes. The essential principle of the method is to translate vertices according to the deformation field, and to perform elementary mesh operations to repair invalid or poor quality elements. The method is substantially different, in that no adaptive refinement of the meshed interface Γ is performed according to a surface metric. Furthermore, topological changes are not dictated by the level set advection but are the consequences of the remeshing operation. Let us notice that as an essential user constraint, the remeshing software `mmg` implementing the method of [25] forbids itself to alter the topology of the zero isosurface of an input level set function to mesh.

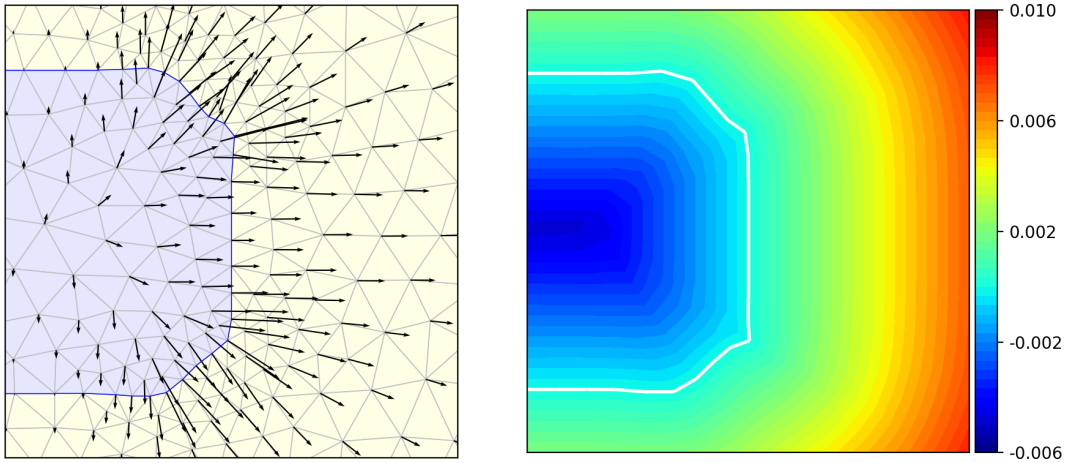


Figure 1.7: Initial setting of the level set based mesh evolution algorithm of [24]. (a) A vector field $\boldsymbol{\theta}$ is defined at the vertices of the computational mesh $\mathcal{T}$ for the background domain D in which $\Omega \subset D$ is explicitly; (b) the signed distance function d_Ω is computed on the mesh $\mathcal{T}$ as a level set $\phi = d_\Omega$ representing Ω.

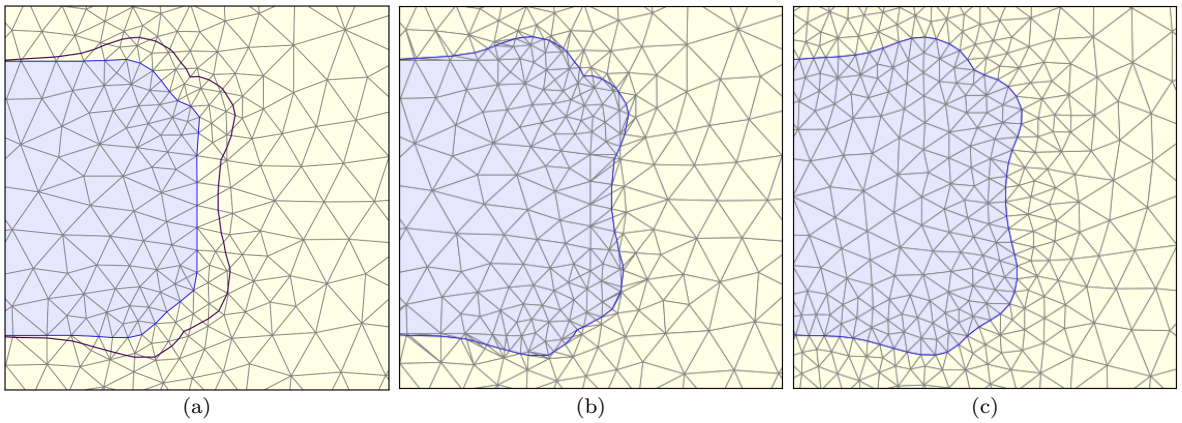


Figure 1.8: Mesh evolution algorithm: (b) the level set function $\phi = d_\Omega$ associated with the subdomain Ω is advected on $\mathcal{T}$; (c) A poor quality mesh $\tilde{\mathcal{T}}$ of the evolved domain Ω' is obtained by splitting $\mathcal{T}$ according to the updated level set function ; (d) $\tilde{\mathcal{T}}$ is iteratively remeshed into a new mesh $\mathcal{T}'$ of sufficient quality for finite element analysis.

CHAPTER 2

HADAMARD'S SHAPE DERIVATIVES FOR A COUPLED THERMAL FLUID STRUCTURE PROBLEM

Contents

Note : most of the content of this chapter has been published in [153]:

F. Feppon, G. Allaire, F. Bordeu, J. Cortial, and C. Dapogny, *Shape optimization of a coupled thermal fluid–structure problem in a level set mesh evolution framework*, SeMA Journal, (2019), pp. 1–46.

However, the introduction has been rewritten in order to fit the outline of the thesis, and the numerical section 2.5 has been substantially improved and enlarged with additional test cases.

2.1 Introduction

One of the ultimate motivations guiding this thesis is the optimization of mechanical structures subjected to thermal loads and cooled down by fluids. In this chapter, we investigate shape and topology optimization based on the method of Hadamard for a weakly coupled model of heat propagation, fluid flow and structure deformation. This model is detailed in section 2.2 and serves as a basis for all the numerical shape optimization test cases considered in this thesis; it is based on the incompressible Stokes or Navier Stokes equations in the fluid domain, on the convection-diffusion equation for heat propagation in both fluid and solid domains, and on the linearized thermoelasticity system for the mechanical displacement of the solid domain. Our main result is the calculation of shape derivatives of *arbitrary* objective functionals in volume and surface form; these are given in propositions 2.3 and 2.4 respectively. These formulas are implemented and verified numerically on 2-d test cases in section 2.5 with the level set based mesh evolution method of [25] (summarized in chapter 1, section 1.4.2). Our 2-d numerical test cases involve either one, two or three of the aforementionned physics simultaneously; more complex numerical test cases including 3-d problems being specifically detailed in chapter 4.

The main originality of the present work is the consideration of coupled heat propagation, fluid flow and structure deformations for topology optimization with the method of Hadamard. There are few previous contributions [318] using such complete models for shape and topology optimization, and they all rely on very different methods for parametrizing shapes and topologies, namely SIMP [65, 66] or variable density methods [72]. Some simpler variants of this full three-physics model have been more extensively studied. For example, there are quite many works dealing with convective heat transfer problems (involving coupled fluid and thermal equations, without coupled elasticity) with density based methods [224, 116, 134, 324, 275, 129, 118, 258, 263] or variants of level set methods [9, 104, 315] (not explicitly based on the method of Hadamard). Fluid-structure interacting systems (without taking into account thermal effects) have also been considered and in slightly fewer works [319, 240, 42, 194, 221, 208] featuring only 2-d results, as well as thermoelastic structures [197, 285, 151, 314, 101, 28].

Likewise, there is a relatively small amount of work considering level set methods for topology optimization of fluid systems [315, 194, 133] (which do not rely explicitly on the method of Hadamard), although its use in the context of geometry optimization (in which the shape is explicitly meshed but its topology is fixed) is a very classical problem, see the monographs [191, 260, 236, 261] and the more recent work [112].

From a numerical point of view, one specificity of our contribution is the implementation of the Hadamard method with the level set based remeshing algorithm of Allaire, Dapogny and Frey [24] (reviewed in chapter 1, section 1.4.2). This allows—in sharp contrast with density and more classical level set methods—to optimize such coupled systems without introducing a relaxed formulation for the description of a mixture of the fluid and solid domains. In the context of density methods (using namely SIMP for linear elasticity or the Brinkman penalization approach for the Navier-Stokes equations [72]), the entire domain D is assumed to be filled with a porous material containing a volume fraction $\eta_f(x)$ of fluid and $1-\eta_f(x)$ of solid at every point x. Then the state equations posed in the fluid domain Ω_f or in the solid domain Ω_s are replaced with an approximate version (inspired from the homogenization theory) to be solved on the whole domain $D = \Omega_s \cup \Omega_f$ with coefficients depending on the volume fraction η_f (if not on a local microstructure, see [18]). The nature of the approximation may depend on the physical model and the type of boundary condition considered at the interface $\Gamma = \partial\Omega_s \cap \partial\Omega_f$ (we discuss this point further in chapter 6, section 6.1.3); for instance the classical approximation of the standard linear elasticity system with constant Hooke's tensor A,

$$\begin{cases} -\mathrm{div}(A\nabla \boldsymbol{u}) = \boldsymbol{f}_s \text{ in } \Omega_s \\ A\nabla \boldsymbol{u}\cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega_s \end{cases}$$

is an elasticity system with variable Hooke's tensor $\widetilde{A}(\eta_f)$,

$$-\mathrm{div}(\widetilde{A}(\eta_f)\nabla \boldsymbol{u}) = \boldsymbol{f}_s \text{ in } D.$$

Likewise, the Stokes problem

$$\begin{cases} -\Delta \boldsymbol{v} + \nabla p = \boldsymbol{f}_f \text{ in } \Omega_f \\ \mathrm{div}(\boldsymbol{v}) = 0 \text{ in } D \\ \boldsymbol{v} = 0 \text{ on } \partial\Omega_f \end{cases}$$

is commonly relaxed as a Brinkman model with variable inverse permeability $\widetilde{\alpha}(\eta_f)$ [72]

$$\begin{cases} -\Delta \boldsymbol{v} + \nabla p + \widetilde{\alpha}(\eta_f)\boldsymbol{v} = \boldsymbol{f} \text{ in } D, \\ \mathrm{div}(\boldsymbol{v}) = 0 \text{ in } D. \end{cases}$$

The relaxed Hooke's tensor $\widetilde{A}$ and the inverse permeability $\widetilde{\alpha}$ must vary between respectively $\widetilde{A}(0) = A$ and $\widetilde{\alpha}(0) \simeq +\infty$ in the pure solid part corresponding to $\eta_f = 0$, and $\widetilde{A}(1) \simeq 0$ and $\widetilde{\alpha(1)} = 0$ in the pure fluid part corresponding to $\eta_f = 1$. A common difficulty characterizing these density approaches is the need for determining adequate interpolation laws in order to penalize intermediate densities and to obtain in practice convergence towards "black and white" designs (featuring only $\eta_f = 0$ or $\eta_f = 1$ in the computational domain D). One law must be proposed per homogenized coefficient, which makes it delicate to set when considering complex multiphysics problems involving several or many of these coefficients.

On the contrary, our level set approach keeps distinct fluid and solid domains throughout the optimization process, Ω_f and Ω_s, and makes the fluid-solid interface Γ the main optimization variable. The

method of Hadamard allows to estimate the sensitivity of the original physics with respect to local variations of the sharp interface Γ. Keeping track of a clear and neat interface with the use of remeshing avoids various problems such as mass conservation, load transmission or boundary layers description [281, 276]. In our context, the level set based mesh evolution algorithm of [24, 25] alleviates these difficulties while enabling shapes to change their topology in the course of optimization iterations. In numerical practice, the solid-fluid interface Γ is evolved with the level set method and remeshed at every iteration for the finite element analysis. This is particularly well suited for fluid-solid applications, as no alteration of the physical equations is required in contrast with density methods or fixed mesh level set methods, requiring ersatz materials. Another advantage is that this allows, in principle, to solve for each physics non intrusively, for instance with industrial solvers external to the shape optimization implementation.

The chapter is organized as follows. Section 2.2 details the considered weakly coupled model. The coupling is weak in the sense that it allows to solve for each physical equations successively and independently from the knowledge of already computed state variables. The main simplifying hypothesis is that mechanical deformations and displacements are small. In particular, the fluid domain is fixed at first order; it is independent of the deformation of the structure. This allows to solve, first, the incompressible Stokes or Navier-Stokes equations in the fluid domain. Second, we solve a convection-diffusion equation for the temperature in the fluid and solid domains with the previously computed velocity. Third, the linearized thermoelasticity system is solved for the mechanical displacement of the solid domain: the forces at play are a combination of applied external loads, of the fluid stress prescribed on the fluid-solid interface, and of the thermal dilation induced by the temperature. This weak coupling is of course a major simplification and it dramatically reduces the computational cost since no monolithic coupled system has to be solved. Let us note that despite these simplifications, this model is sufficiently general to include previous classical studies when considering only a subset of these physics separately.

The next two sections 2.3 and 2.4 are concerned with the derivations of Hadamard shape sensitivities formulas which are at the basis of our gradient-based topology optimization algorithm used in our numerical simulations.The first section 2.3 introduces a simplified model made of two scalar Laplace equations, mimicking the weak coupling between the solid and the fluid subdomains. This section is purely pedagogical: it deals with the same mathematical difficulties which could be somewhat hidden behind the relative complexity of the full multiphysics system treated in section 2.4. Two major contributions are especially highlighted.

First, we explain how to adapt the classical shape differentiation methods in order to consider *arbitrary* shape functionals. Indeed, as reviewed in chapter 1, section 1.2.3 the most commonly used method is the so-called Lagrangian method of Céa [84, 17]. It brings simplifications with respect to the "rigorous calculation" (see chapter 1, section 1.2.1) but it relies on the specific knowledge of a formula for the objective function. In other words, if the objective function changes (say from a volume integral to a surface integral), the calculation of the shape derivative has to be performed again from the beginning. Here, following the lead of Murat and Simon [242], we rather rely on a Lagrangian approach which allows us to treat very general objective functions, without precise formulas, under a mild assumption on the existence of some partial derivatives. Obviously, it is in the more complex framework of the full three-physic problem that our proposed Lagrangian approach is really more efficient than the classical Céa's method.

Second, we detail the treatment of the coupling induced by the fluid on the structure, which features a quite surprising mathematical phenomenon. Classically, the order of the coupling is reversed for the adjoint system: for the direct problem, one first solves the fluid equation, and in a second step the solid equation; for the adjoint problem, it is the opposite, namely elasticity is solved first, followed by fluid mechanics. While the fluid and solid equations are coupled by a one-sided interface transmission condition of Neumann type in the direct problem (which accounts for the force exerted by the fluid on the solid), the adjoint equations are, on the contrary, coupled by a Dirichlet interface condition. This 'transformation' of the Neumann interface condition for the direct problem into a Dirichlet interface condition for the adjoint seems new to the best of our knowledge.

These derivations are explained in details in section 2.3 on the simplified model, before being extended to the full multiphysics setting in section 2.4 for our practical applications.

Eventually, 2-d applicative examples are considered in section 2.5 for the numerical validation of our derived analytical shape derivatives. Several optimizations are performed on test cases featuring either one, two, or three physics enabled simultaneously. Constrained optimization problems are solved in these situations with the help of a constrained optimization algorithm detailed in the dedicated chapter 3. Our preliminary test cases are drawn from classical literature with only one physics involved:

we consider the classical cantilever beam in structural mechanics, an optimal heat conduction test case, and classical optimal drag profile designs in fluid mechanics. We then address problems featuring two physics simultaneously. Our first example is a fluid-structure interaction problem, without taking into accounts thermal effects. It is inspired from a similar example in [319, 221]. Our second test case is a convective heat transfer problem where the elastic deformation of the solid phase is neglected. It was already studied in [225] with a different approach (based on a variable density relaxation of the problem). Our third test case is concerned with optimal design in thermoelasticity and is based on the previous work [314]. Finally, we present a last test case which comprises the three physics and is new, to the best of our knowledge. All three examples work nicely in the sense that the objective function is indeed minimized while respecting optimization constraints constraints, and that the obtained optimal shapes are significantly different from the initial ones sharing improved performances. These results are preliminary to the more complex heat transfer and 3-d applications considered in chapter 6.

2.2 SETTING OF THE THREE-PHYSIC PROBLEM

Let $D = \Omega_s \cup \Omega_f$ be a fixed, open bounded domain in $\mathbb{R}^d$ ($d = 2$ or 3 in applications), arising as the disjoint reunion of a 'fluid' phase Ω_f and a 'solid' phase $\Omega_s = D \backslash \overline{\Omega_f}$ (see Figure 2.1), separated by an interface $\Gamma := \partial\Omega_f \cap \partial\Omega_s$ which is to be optimized. Throughout this chapter, the normal vector $\boldsymbol{n}$ to Γ is pointing outward the fluid domain Ω_f.

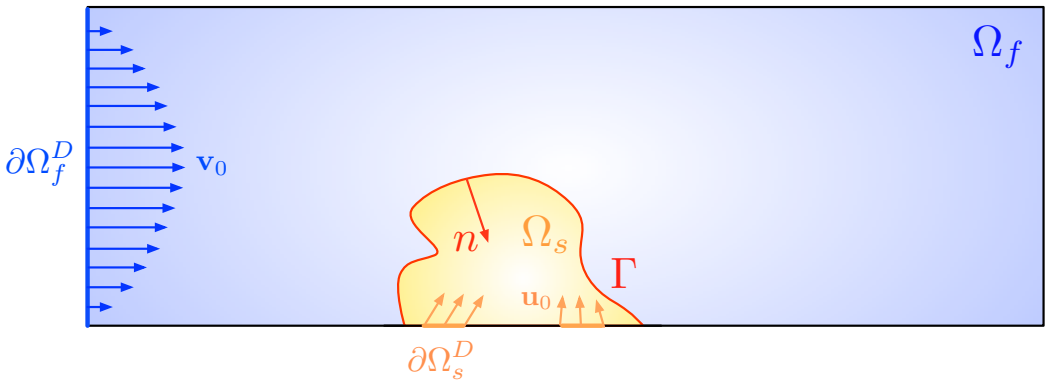


Figure 2.1: Setting of the considered three-physic problem (not all the regions of ∂D featured in (2.2.1) (2.2.2) and (2.2.3) are represented, see the numerical section 2.5 for more complete settings).

The domain D is described by three physical variables which are governed by three coupled models:

- the motion of the fluid inside Ω_f described by the velocity and pressure fields $\boldsymbol{v}$ and p;
- the diffusion of heat inside the whole domain D, and its transport by convection in the fluid domain, resulting in a temperature field T;
- the deformation of the solid region Ω_s as a result of the stress exerted by the fluid part and of the dilation induced by thermoelastic effects, characterized by a mechanical displacement $\boldsymbol{u}$.

The physical equations chosen for the modeling of $(\boldsymbol{v}, p)$, T and $\boldsymbol{u}$ with their relevant set of boundary conditions are described in strong form in sections 2.2.1 to 2.2.3. The shape optimization setting considered with the method of Hadamard is then introduced in section 2.2.4. Notation for the setting of boundary conditions as well as for all the physical parameters involved in the state equations are summarized respectively in Tables 2.1 and 2.2 below.

$D = \Omega_s \cup \Omega_f$	Whole domain featuring fluid and solid phases
Ω_f	Fluid phase
Ω_s	Solid phase
$\Gamma = \partial\Omega_s \cup \partial\Omega_f$	Fluid-structure interface
$\partial D = \partial\Omega_T^N \cup \partial\Omega_T^D$	Boundary of the global domain D
$\partial\Omega_T^D$	Dirichlet (isothermal) boundary condition for the temperature variable ($T = T_0$ on $\partial\Omega_T^D$)
$\partial\Omega_T^N$	Neumann (adiabatic or isoflux) boundary condition for the temperature variable ($-k_f \partial T_f/\partial \boldsymbol{n} = h$ on $\partial\Omega_T^N \cap \partial\Omega_f$ and $-k_s \partial T_s/\partial \boldsymbol{n} = h$ on $\partial\Omega_T^N \cap \partial\Omega_s$)
$\partial\Omega_s = \partial\Omega_s^D \cup \partial\Omega_f^D \cup \Gamma$	Boundary of the solid domain Ω_s
$\partial\Omega_s^D$	Dirichlet boundary for the solid variable $\boldsymbol{u}$
$\partial\Omega_s^N$	Neumann boundary for the solid variable $\boldsymbol{u}$
$\partial\Omega_f = \partial\Omega_f^D \cup \partial\Omega_f^D \cup \Gamma$	Boundary of the fluid domain Ω_f
$\partial\Omega_f^D$	Dirichlet (inlet) boundary for the fluid variable ($\boldsymbol{v} = \boldsymbol{v}_0$ on $\partial\Omega_f^D$)
$\partial\Omega_f^N$	Neumann (outlet) boundary for the fluid variable ($\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = 0$ on $\partial\Omega_f^N$)

Table 2.1: Domains D, Ω_s and Ω_f with their associated boundary conditions

2.2.1 Steady state incompressible Navier-Stokes equation for the fluid variable

The fluid domain Ω_f is filled with a fluid characterized by its viscosity ν and density ρ; its velocity $\boldsymbol{v}$ and pressure p satisfy the incompressible steady-state Navier-Stokes equations:

$$\begin{cases} -\mathrm{div}(\sigma_f(\boldsymbol{v}, p)) + \rho \nabla \boldsymbol{v}\, \boldsymbol{v} = \boldsymbol{f}_f & \text{in } \Omega_f \\ \mathrm{div}(\boldsymbol{v}) = 0 & \text{in } \Omega_f \\ \boldsymbol{v} = \boldsymbol{v}_0 & \text{on } \partial\Omega_f^D \\ \sigma_f(\boldsymbol{v}, p)\boldsymbol{n} = 0 & \text{on } \partial\Omega_f^N \\ \boldsymbol{v} = 0 & \text{on } \Gamma, \end{cases} \tag{2.2.1}$$

where $\nabla \boldsymbol{v}$ is the Jacobian matrix $(\nabla \boldsymbol{v})_{ij} = \partial_j \boldsymbol{v}_i$, $i, j = 1, ..., d$. The fluid stress tensor $\sigma_f(\boldsymbol{v}, p)$ is given by the Newton law

$$\sigma_f(\boldsymbol{v}, p) = 2\nu e(\boldsymbol{v}) - pI \quad \text{with} \quad e(\boldsymbol{v}) = (\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T)/2,$$

where I is the $d \times d$ identity matrix.

Remark 2.1. The nonlinear term $\nabla \boldsymbol{v}\, \boldsymbol{v}$ is often written $(\boldsymbol{v} \cdot \nabla)\boldsymbol{v}$ in the literature, however this convention is not consistent with our notation $\boldsymbol{v} \cdot \nabla \boldsymbol{v} := \boldsymbol{v}^T \nabla \boldsymbol{v}$ (section 1.1). We shall therefore keep the notation $\nabla \boldsymbol{v}\, \boldsymbol{v}$ considered to be the matrix product of the Jacobian matrix $\nabla \boldsymbol{v}$ with the vector $\boldsymbol{v}$.

In (2.2.1), $\boldsymbol{f}_f$ is an applied body force (e.g. gravity); the boundary of the fluid phase is the disjoint reunion

$$\partial\Omega_f = \partial\Omega_f^D \cup \partial\Omega_f^N \cup \Gamma$$

of a Dirichlet (or inlet) part $\partial\Omega_f^D$ where the flow enters with a given velocity $\boldsymbol{v} = \boldsymbol{v}_0$, a Neumann (or outlet) part $\partial\Omega_f^N$ where zero normal stress is observed, and the interface Γ with the solid domain Ω_s. At this stage it is assumed that the deformation of the solid domain is sufficiently small so that no slip boundary conditions hold on Γ: $\boldsymbol{v} = 0$; see remark 2.2 below about this point. Therefore, the variables $(\boldsymbol{v}, p)$ depend solely on the geometry of the fluid domain Ω_f.

$\boldsymbol{v} \in H^1(\Omega_f, \mathbb{R}^d)$	Fluid velocity
$p \in L^2(\Omega_f)$	Fluid pressure
$\nu \in \mathbb{R}_+^*$	Fluid viscosity
$\rho \in \mathbb{R}_+^*$	Fluid density
$\boldsymbol{f}_f \in H^1(\Omega_f, \mathbb{R}^d)$	Volume force in the fluid phase
$\boldsymbol{v}_0 \in H^{1/2}(\partial\Omega_f, \mathbb{R}^d)$	Input inlet or outlet velocity
$T \in H^1(D)$	Temperature field
$T_s \in H^1(\Omega_s)$	Restriction of T to the solid domain
$T_f \in H^1(\Omega_f)$	Restriction of T to the fluid domain
$c_p \in \mathbb{R}_+^*$	Heat capacity of the fluid
$k_f \in \mathbb{R}_+^*$	Conductivity in fluid
$k_s \in \mathbb{R}_+^*$	Conductivity in solid
$Q_f \in L^2(\Omega_f)$	External input heat flux in the fluid domain
$Q_s \in L^2(\Omega_s)$	External input heat flux in he solid domain
$T_0 \in H^{1/2}(\partial D)$	Input temperature on the total boundary
$h \in L^2(\partial\Omega_T^N)$	Input entering heat flux on the total boundaries
$\boldsymbol{u} \in H^1(\Omega_s, \mathbb{R}^d)$	Elastic displacement
$\mu \in \mathbb{R}_+^*$	Lamé coefficient
$\lambda \in \mathbb{R}_+^*$	Lamé coefficient
$T_{\text{ref}} \in \mathbb{R}_+^*$	Reference temperature in the solid
$\alpha \in \mathbb{R}_+^*$	Thermal expansion coefficient
$\boldsymbol{f}_s \in H^1(\Omega_s, \mathbb{R}^d)$	Volume force in the solid phase
$\boldsymbol{u}_0 \in H^{1/2}(\partial\Omega_s^D, \mathbb{R}^d)$	Prescribed displacement
$\boldsymbol{g} \in L^2(\partial\Omega_s^N)$	Input traction force

Table 2.2: Physical parameters considered in the weakly coupled model

2.2.2 Convection diffusion for the temperature variable

The fluid velocity $\boldsymbol{v}$ determines the physical behavior of the temperature T in the whole domain D, as a result of convection and diffusion effects inside the fluid domain Ω_f, and of pure diffusion inside the solid domain Ω_s. Denoting by k_f and k_s the thermal conductivity inside Ω_f and Ω_s respectively, and by c_p the thermal capacity of the fluid, the temperature field T is determined by the steady-state

convection-diffusion equations:

$$\begin{cases} -\mathrm{div}(k_f \nabla T_f) + \rho c_p \boldsymbol{v} \cdot \nabla T_f = Q_f & \text{in } \Omega_f \\ -\mathrm{div}(k_s \nabla T_s) = Q_s & \text{in } \Omega_s \\ T = T_0 & \text{on } \partial\Omega_T^D \\ -k_f \dfrac{\partial T_f}{\partial \boldsymbol{n}} = h & \text{on } \partial\Omega_T^N \cap \partial\Omega_f \\ -k_s \dfrac{\partial T_s}{\partial \boldsymbol{n}} = h & \text{on } \partial\Omega_T^N \cap \partial\Omega_s \\ T_f = T_s & \text{on } \Gamma \\ -k_f \dfrac{\partial T_f}{\partial \boldsymbol{n}} = -k_s \dfrac{\partial T_s}{\partial \boldsymbol{n}} & \text{on } \Gamma, \end{cases} \tag{2.2.2}$$

where we use the subscripts $_f$ and $_s$ for the restrictions T_f and T_s of T to Ω_f and Ω_s respectively. In (2.2.2), Q_f and Q_s are volumic sources inside Ω_f and Ω_s; the boundary $\partial D = \partial\Omega_T^N \cup \partial\Omega_T^D$ is split into a Dirichlet part, where a temperature T_0 is imposed and a Neumann part where a given incoming heat flux h is applied on $\partial\Omega_T^N$. The temperature T as well as the heat flux are continuous across the interface Γ between Ω_f and Ω_s.

2.2.3 Thermoelasticity with fluid structure interaction for the elastic variable

Finally, the fluid variables $(\boldsymbol{v}, p)$ and the temperature T together determine the displacement $\boldsymbol{u}$ of the solid domain Ω_s, which is assumed to be an isotropic thermoelastic material with Lamé coefficients λ, μ, thermal expansion parameter α and temperature at rest T_{ref}. This variable $\boldsymbol{u}$ is characterized by the equations of linear thermoelasticity:

$$\begin{cases} -\mathrm{div}(\sigma_s(\boldsymbol{u}, T_s)) = \boldsymbol{f}_s & \text{in } \Omega_s \\ \boldsymbol{u} = \boldsymbol{u}_0 & \text{on } \partial\Omega_s^D \\ \sigma_s(\boldsymbol{u}, T_s)\boldsymbol{n} = \boldsymbol{g} & \text{on } \partial\Omega_s^N \\ \sigma_s(\boldsymbol{u}, T_s)\boldsymbol{n} = \sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} & \text{on } \Gamma. \end{cases} \tag{2.2.3}$$

where the solid stress tensor is given by

$$\sigma_s(\boldsymbol{u}, T_s) = Ae(\boldsymbol{u}) - \alpha(T_s - T_{\mathrm{ref}})I \text{ with } Ae(\boldsymbol{u}) = 2\mu e(\boldsymbol{u}) + \lambda \mathrm{Tr}(e(\boldsymbol{u}))I, \tag{2.2.4}$$

and $\boldsymbol{f}_s$ is an applied body force. In (2.2.3), the boundary $\partial\Omega_s$ is split into respectively a Dirichlet part $\partial\Omega_s^D$ where a displacement $\boldsymbol{u} = \boldsymbol{u}_0$ is prescribed, a Neumann part $\partial\Omega_s^N$ where a stress $\boldsymbol{g}$ is imposed, and the interface Γ with the fluid domain. On this latter boundary, the solid is submitted to the pressure force imposed by the fluid, which translates into the equality $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = \sigma_s(\boldsymbol{u}, T_s) \cdot \boldsymbol{n}$ between the normal fluid and solid stresses.

Remark 2.2. The above model is a simplified version of a genuine fluid-solid-thermic coupling between the phases Ω_f and Ω_s. A more accurate description of fluid-structure interaction would feature a vanishing fluid velocity $\boldsymbol{v}$ on the *deformed* interface $(I + \boldsymbol{u})(\Gamma)$, namely:

$$\boldsymbol{v}(x + \boldsymbol{u}(x)) = 0, \quad x \in \Gamma, \tag{2.2.5}$$

see e.g. [267], or [319] in an optimization context. Likewise, the equality between normal stresses $\sigma_s(\boldsymbol{u}, T_s)\boldsymbol{n} = \sigma_f(\boldsymbol{v}, p)\boldsymbol{n}$ should hold on the deformed interface. In the present work, the displacement $\boldsymbol{u}$ of the solid phase is assumed to be small enough so that the influence of the interface deformation on the physical behavior of the fluid can be neglected. Thanks to this simplification, the system (2.2.1) to (2.2.3) is only weakly coupled: its resolution is achieved by solving first the fluid system (2.2.1), then using the resulting fluid velocity $\boldsymbol{v}$ in the heat transfer equation (2.2.2), and finally using the fluid stress $\sigma_f(\boldsymbol{v}, p)$ and the temperature T_s to solve (2.2.3).

2.2.4 Shape optimization setting

The final goal of the thesis is the resolution of general constrained optimization problems of the form

$$\min_{\Gamma} \quad J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma))$$
$$s.t. \begin{cases} g_i(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = 0, & 1 \le i \le p, \\ h_j(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \le 0, & 1 \le j \le q, \end{cases} \tag{2.2.6}$$

where J, g_i and h_j are *arbitrary* shape objective and constraint functionals set by a user. For its resolution, we shall rely on the null space gradient flow algorithm detailed in chapter 3, which (like any other first order optimization method) requires the knowledge of the shape derivatives of the above functionals. In the context of the method of Hadamard (reviewed in chapter 1, section 1.2), this means computing the Fréchet derivative of the mapping

$$\boldsymbol{\theta} \mapsto J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma, \boldsymbol{\theta}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}})),$$

where we recall that $\boldsymbol{\theta}$ is a deformation field belonging in general to $W^{1,\infty}(D, \mathbb{R}^d)$ and $\Gamma_{\boldsymbol{\theta}} = (I + \boldsymbol{\theta})\Gamma$ (and so on for the constraints g_i and h_j). The setting in the biphasic domain $D = \Omega_f \cup \Omega_s$ is illustrated on Figure 2.2.

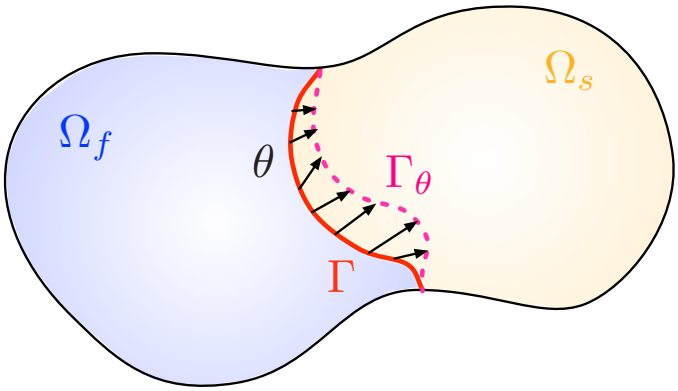


Figure 2.2: Deformation of a partition $D = \Omega_f \cup \Omega_s$ using Hadamard's method.

For simplicity, we assume in our context that only the interface Γ is subject to optimization, and not the boundary ∂D of the total domain, therefore the admissible space for deformations $\boldsymbol{\theta}$ is not $W^{1,\infty}(D, \mathbb{R}^d)$ but rather the subspace:

$$W_0^{1,\infty}(D, \mathbb{R}^d) = \{\boldsymbol{\theta} \in L^\infty(D, \mathbb{R}^d) |\, \nabla\boldsymbol{\theta} \in L^\infty(D, \mathbb{R}^d \times \mathbb{R}^d) \text{ and } \boldsymbol{\theta} = 0 \text{ on } \partial D\}. \tag{2.2.7}$$

The condition $\boldsymbol{\theta} = 0$ on ∂D implies that junction points corresponding to the intersection $\partial D \cap \Gamma$ are fixed. The analysis to follow can of course be extended to more general shape variations (*e.g.* only $\boldsymbol{\theta} \cdot \boldsymbol{n} = 0$ on ∂D or even $\boldsymbol{\theta} \in W^{1,\infty}(D, \mathbb{R}^d)$), at the expense of including extra terms in shape derivatives accounting for the variations of boundary conditions or for the tangential displacements of the junction points [114].

2.3 SHAPE DERIVATIVES FOR A SIMPLIFIED SCALAR FLUID STRUCTURE INTERACTION PROBLEM

One of the salient features of the system (2.2.1) to (2.2.3) is the equality of normal stresses $\sigma_s(\boldsymbol{u}, T_s)\boldsymbol{n} = \sigma_f(\boldsymbol{v}, p)\boldsymbol{n}$ imposed on the optimized interface Γ. The calculation of shape derivatives in this context is not a completely standard issue to the best of our knowledge, and the result has an interesting structure: the interface conditions for the adjoint systems are different from those appearing in the state equations (see the expressions (2.4.8) and (2.4.10)). To illustrate this fact, we first consider a simplified scalar problem which gathers with lighter notations all the essential mathematical points.

This preliminary study contains as a particular case the classical Poisson problem considered in chapter 1, (1.2.19), which will also allow us to highlight how the derivation of shape derivatives can be adapted, in quite general contexts, for arbitrary objective functionals.

2.3.1 Presentation of the simplified problem and of its variational formulation

We consider the following setting: the fluid variable $\boldsymbol{v}$ is replaced by a scalar variable u_f solving the Poisson equation (2.3.1) in Ω_f with homogeneous Dirichlet boundary conditions on the interface Γ. The elastic variable $\boldsymbol{u}$ is replaced by the solution u_s to another Poisson equation (2.3.2) posed in the complement Ω_s. Body sources $f_f \in H^1(\Omega_f)$, $f_s \in H^1(\Omega_s)$, with the dimension of a force, are applied in

both cases. Note that such H^1 regularity is needed when computing shape derivatives in the sequel (see the discussion of chapter 1, section 1.2.3), although weaker regularity is enough for state equations to be well-posed. For the sake of physical units (u_f and u_s have respectively the dimension of a velocity and of a displacement), the Laplace operators involved in (2.3.1) and (2.3.2) are scaled with constants ν and μ that assume respectively the role of a viscosity in the fluid domain, and of a Lamé coefficient in the solid domain. The two systems for u_f and u_s are weakly coupled: the equation (2.3.1) for u_f does not depend on u_s and in the system (2.3.2) for u_s, equality of normal fluxes is imposed at the interface Γ, as the exact counterpart in the scalar setting to the continuity of normal stresses $\sigma_s(\boldsymbol{u}, T_s)\boldsymbol{n} = \sigma_f(\boldsymbol{v}, p)\boldsymbol{n}$. For simplicity, homogeneous Dirichlet boundary conditions are imposed on the boundary ∂D of the total domain $D = \Omega_f \cup \Omega_s$ in both systems for u_f and u_s:

$$\begin{cases} -\nu\Delta u_f = f_f & \text{in } \Omega_f \\ u_f = 0 & \text{on } \Gamma \\ u_f = 0 & \text{on } \partial\Omega_f \backslash \Gamma, \end{cases} \tag{2.3.1}$$

$$\begin{cases} -\mu\Delta u_s = f_s & \text{in } \Omega_s \\ \mu \dfrac{\partial u_s}{\partial \boldsymbol{n}} = \nu \dfrac{\partial u_f}{\partial \boldsymbol{n}} & \text{on } \Gamma \\ u_s = 0 & \text{on } \partial\Omega_s \backslash \Gamma. \end{cases} \tag{2.3.2}$$

We recall our convention that the normal vector $\boldsymbol{n}$ in (2.3.2) is pointing outward Ω_f (note that the transmission boundary condition in (2.3.2) reads the same if $\boldsymbol{n}$ is pointing outward Ω_s). We consider the minimization problem

$$\min_{\Gamma} J(\Gamma, u_f(\Gamma), u_s(\Gamma)), \tag{2.3.3}$$

where J is a given cost function upon which we shall impose adequate regularity conditions in due time.

Remark 2.3. This simplified setting reduces to the single Poisson problem considered in chapter 1, section 1.2.3 with $\Omega = \Omega_f$, $\Gamma_N = \emptyset$ in the particular case where the objective function $J(\Gamma, u_f(\Gamma))$ does not depend on the solid variable (in that case (2.3.2) can be ignored). Although in the present context, no Neumann boundary condition is considered for u_f for simplicity, we shall still retrieve the results of chapter 1, propositions 1.8 and 1.9 because our way to obtain shape derivatives is rather insensitive to the type of boundary conditions considered.

In the sequel, the dependence of the state variables u_f and u_s with respect to Γ is made explicit—using the notations $u_s(\Gamma)$ and $u_f(\Gamma)$ as in (2.3.3)—when it is needed (e.g. as in. (2.3.12) below). In order to discuss the precise mathematical setting for (2.3.1) (2.3.2), the following spaces of functions on the subdomains Ω_s and Ω_f are introduced:

$$V_s(\Gamma) = \{v_s \in H^1(\Omega_s) \,|\, v_s = 0 \text{ on } \partial\Omega_s \backslash \Gamma\}, \tag{2.3.4}$$

$$V_f(\Gamma) = \{v_f \in H^1(\Omega_f) \,|\, v_f = 0 \text{ on } \partial\Omega_f \backslash \Gamma\}, \tag{2.3.5}$$

$$V_{f,0}(\Gamma) = \{v_f \in H^1(\Omega_f) \,|\, v_f = 0 \text{ on } \partial\Omega_f\}. \tag{2.3.6}$$

We also consider the subspace $H^{1/2}_{00}(\Gamma)$ of $H^{1/2}(\Gamma)$ composed of restrictions to Γ of functions in $V_f(\Gamma)$,

$$H^{1/2}_{00}(\Gamma) = \{v|_\Gamma \mid v \in V_f(\Gamma)\}, \tag{2.3.7}$$

and its dual space $H^{-1/2}_{00}(\Gamma)$. Roughly speaking, any element $v \in H^{1/2}_{00}(\Gamma)$ has an extension v_f to Ω_f vanishing on $\partial\Omega_f \setminus \Gamma$ [218].

In this framework, the state variables u_s and u_f in (2.3.1) and (2.3.2) are the unique solutions to the following variational problems:

$$\text{Find } u_f \in V_{f,0}(\Gamma) \text{ such that } \forall v_f \in V_{f,0}(\Gamma), \int_{\Omega_f} \nu \nabla u_f \cdot \nabla v_f \mathrm{d}x = \int_{\Omega_f} f_f v_f \mathrm{d}x, \tag{2.3.8}$$

$$\text{Find } u_s \in V_s(\Gamma) \text{ such that } \forall v_s \in V_s(\Gamma), \int_{\Omega_s} \mu \nabla u_s \cdot \nabla v_s \mathrm{d}x = \int_{\Omega_s} f_s v_s \mathrm{d}x - \int_{\Gamma} \nu \frac{\partial u_f}{\partial \boldsymbol{n}} v_s \mathrm{d}s, \tag{2.3.9}$$

where the minus sign in the last term of the right-hand side of (2.3.9) is due to our convention whereby $\boldsymbol{n}$ is pointing outward Ω_f.

A comment is in order about the meaning of (2.3.9). In general, the normal derivative $\partial v/\partial \boldsymbol{n}$ on the interface Γ of an arbitrary function $v \in V_{f,0}(\Gamma)$ is not defined, because there is no trace theorem for the gradient of functions in $H^1(\Omega_f)$. However, u_f is not a mere function of $H^1(\Omega_f)$: (2.3.8) implies that $\Delta u_f = f_f \in L^2(\Omega_f)$. Therefore, the last term of (2.3.9) is defined, for any $v \in H_{00}^{1/2}(\Gamma)$, by

$$-\int_\Gamma \nu \frac{\partial u_f}{\partial \boldsymbol{n}} v \mathrm{d}s := -\int_{\Omega_f} (\nu \Delta u_f\, \tilde{v} + \nu \nabla u_f \cdot \nabla \tilde{v}) \mathrm{d}x = \int_{\Omega_f} (f_f \tilde{v} - \nu \nabla u_f \cdot \nabla \tilde{v}) \mathrm{d}x, \tag{2.3.10}$$

where $\tilde{v} \in V_f(\Gamma)$ is any extension of v to Ω_f satisfying $\tilde{v} = v$ on Γ. Note that, for smooth u_f, (2.3.10) is just Green's formula. Since from (2.3.8), the right-hand side of (2.3.10) does not depend on the choice of such extension $\tilde{v}$, this identity actually *defines* $\partial u_f/\partial \boldsymbol{n}$ as an element of the dual $H_{00}^{-1/2}(\Gamma)$; see e.g. [218, 176].

The variational formulation (2.3.9) associated to (2.3.2) thus rewrites:

Find $u_s \in V_s(\Gamma)$ such that $\forall v_s \in V_s(\Gamma)$,

$$\int_{\Omega_s} \mu \nabla u_s \cdot \nabla v_s \mathrm{d}x = \int_{\Omega_s} f_s v_s \mathrm{d}x + \int_{\Omega_f} f_f \tilde{v}_s \mathrm{d}x - \int_{\Omega_f} \nu \nabla u_f \cdot \nabla \tilde{v}_s \mathrm{d}x \tag{2.3.11}$$

where $\tilde{v}_s \in V_f(\Gamma)$ is any extension of v_s to Ω_f such that $\tilde{v}_s = v_s$ on Γ.

Remark 2.4. When $\Gamma \cap \partial D = \emptyset$, which happens if for instance Ω_s is strictly included in D ($\Omega_s \subset\subset D$), then $H_{00}^{1/2}(\Gamma)$ and $H_{00}^{-1/2}(\Gamma)$ coincide with the more usual fractional Sobolev spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ respectively; see [218, 302] about these technicalities.

2.3.2 A fully Lagrangian setting for computing shape derivatives of arbitrary objective functionals

Although very common and widely used in the literature (see e.g. [32, 253]), an issue with Céa's method as exposed in chapter 1, section 1.2.3 is that the computation of the shape derivatives depend very much on the assumptions made on the nature of the considered objective functional J. Different type of functionals may lead to different strong forms for the adjoint equations (see section 2.6.2 where this fact is exemplified), which imposes to redo the analytical derivation whenever the objective function is modified, and to update the numerical implementation accordingly.

In this section, we use a fully Lagrangian setting to derive rigorously the shape derivative of very general objective functionals in the simplified setting of section 2.3.1, in the spirit of the seminal work of Murat and Simon [242]. The shape sensitivities of the state variables $u_f(\Gamma), u_s(\Gamma)$ are calculated *first* in order to obtain the shape derivative of an arbitrary objective functional in volume form. Then, under sufficient regularity assumptions, the well-known Hadamard structure theorem together with suitable integration by parts yield general shape derivative formulas in the classical form of a boundary integral.

A modified objective functional and Lagrangian derivative of the state variables

The starting remark is that the functional J, although appearing naturally in the formulation of the optimization problem (2.3.3) is not so convenient for the mathematical analysis. Indeed, the domain of definition of $J(\Gamma, \cdot, \cdot)$ is $V_{f,0}(\Gamma) \times V_s(\Gamma)$, a functional space which depends on the first argument Γ. In order to address this issue, the classical idea is to work within a Lagrangian framework rather than a Eulerian one. Therefore, we consider a fixed reference interface Γ, and we introduce a modified functional $\mathfrak{J}$ obtained by "transporting" J on a fixed space: for any $(\boldsymbol{\theta}, \widehat{u}_f, \widehat{u}_s) \in W_0^{1,\infty}(D, \mathbb{R}^d) \times V_{f,0}(\Gamma) \times V_s(\Gamma)$, we define

$$\mathfrak{J}(\boldsymbol{\theta}, \widehat{u}_f, \widehat{u}_s) := J(\Gamma_{\boldsymbol{\theta}}, \widehat{u}_f \circ (I + \boldsymbol{\theta})^{-1}, \widehat{u}_s \circ (I + \boldsymbol{\theta})^{-1}). \tag{2.3.12}$$

Conversely, this allows to rewrite the objective functional J as:

$$\begin{aligned} J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}})) &= \mathfrak{J}(\boldsymbol{\theta}, u_f(\Gamma_{\boldsymbol{\theta}}) \circ (I + \boldsymbol{\theta}), u_s(\Gamma_{\boldsymbol{\theta}}) \circ (I + \boldsymbol{\theta})) \\ &= \mathfrak{J}(\boldsymbol{\theta}, u_{f,\boldsymbol{\theta}}, u_{s,\boldsymbol{\theta}}), \end{aligned} \tag{2.3.13}$$

where (following the notation of chapter 1, section 1.2) we have denoted $\Gamma_{\boldsymbol{\theta}} := (I + \boldsymbol{\theta})(\Gamma)$, $u_{s,\boldsymbol{\theta}} := u_s(\Gamma_{\boldsymbol{\theta}}) \circ (I + \boldsymbol{\theta})$ and $u_{f,\boldsymbol{\theta}} := u_f(\Gamma_{\boldsymbol{\theta}}) \circ (I + \boldsymbol{\theta})$ the transported functions on their reference domains Ω_s and Ω_f. The identity (2.3.13) is the key motivation for introducing $\mathfrak{J}$: indeed, as it is classical in shape optimization, the transported functions $u_{f,\boldsymbol{\theta}}, u_{s,\boldsymbol{\theta}}$ are differentiable with respect to $\boldsymbol{\theta}$ without additional regularity assumptions [184, 242]. More precisely, the following lemma holds:

Lemma 2.1. *The mappings $\boldsymbol{\theta} \mapsto u_{f,\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \mapsto u_{s,\boldsymbol{\theta}}$, from $W_0^{1,\infty}(D,\mathbb{R}^d)$ into $V_{f,0}(\Gamma)$ and $V_s(\Gamma)$ are Fréchet differentiable at $\boldsymbol{\theta}=0$ and their Fréchet derivatives $\dot{u}_f(\boldsymbol{\theta})$ and $\dot{u}_s(\boldsymbol{\theta})$ in the direction $\boldsymbol{\theta}$ are the unique solutions to the following variational problems:*

Find $\dot{u}_f(\boldsymbol{\theta}) \in V_{f,0}(\Gamma)$ such that $\forall v_f \in V_{f,0}(\Gamma)$,

$$\int_{\Omega_f} \nu \nabla \dot{u}_f(\boldsymbol{\theta}) \cdot \nabla v_f \mathrm{d}x = \int_{\Omega_f} (\mathrm{div}(f_f\boldsymbol{\theta})v_f + \nu(\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nabla u_f \cdot \nabla v_f)\mathrm{d}x, \quad (2.3.14)$$

Find $\dot{u}_s(\boldsymbol{\theta}) \in V_s(\Gamma)$ such that $\forall v_s \in V_s(\Gamma)$,

$$\int_{\Omega_s} \mu \nabla \dot{u}_s(\boldsymbol{\theta}) \cdot \nabla v_s \mathrm{d}x = \int_{\Omega_s} (\mathrm{div}(f_s\boldsymbol{\theta})v_s + (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\mu\nabla u_s \cdot \nabla v_s)\mathrm{d}x - \int_\Gamma \nu \frac{\partial \dot{u}_f(\boldsymbol{\theta})}{\partial \boldsymbol{n}} v_s \mathrm{d}s, \quad (2.3.15)$$

where $-\frac{\partial \dot{u}_f(\boldsymbol{\theta})}{\partial \boldsymbol{n}} \in H_{00}^{-1/2}(\Gamma)$ is defined for any $v \in H_{00}^{1/2}(\Gamma)$ by:

$$-\int_\Gamma \nu \frac{\partial \dot{u}_f(\boldsymbol{\theta})}{\partial \boldsymbol{n}} v \mathrm{d}s = \int_{\Omega_f} (\mathrm{div}(f_f\boldsymbol{\theta})\tilde{v} + (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nu\nabla u_f \cdot \nabla\tilde{v})\mathrm{d}x - \int_{\Omega_f} \nu\nabla\dot{u}_f(\boldsymbol{\theta})\cdot\nabla\tilde{v}\mathrm{d}x. \quad (2.3.16)$$

for any extension $\tilde{v} \in V_f(\Gamma)$ of v such that $v = \tilde{v}$ on Γ.

Proof. The proof is classical, so we content ourselves with a very brief description of the main ideas, which are very similar to those outlined in chapter 1, section 1.2.3. We first perform a change of variables in the variational formulations (2.3.8) and (2.3.11) so that the integrals involved there are written on fixed domains. Taking $v_f \circ (I+\boldsymbol{\theta})^{-1} \in V_{f,0}(\Gamma_{\boldsymbol{\theta}})$ as a test function in (2.3.8) for arbitrary $v_f \in V_{f,0}(\Gamma)$, this yields a variational formulation for $u_{f,\boldsymbol{\theta}}$:

$$\forall v_f \in V_{f,0}(\Gamma), \int_{\Omega_f} \nu A(\boldsymbol{\theta})\nabla u_{f,\boldsymbol{\theta}} \cdot \nabla v_f \mathrm{d}x = \int_{\Omega_f} f_f \circ (I+\boldsymbol{\theta}) v_f |\det(I+\nabla\boldsymbol{\theta})| \mathrm{d}x, \quad (2.3.17)$$

where $A(\boldsymbol{\theta})$ is the $d \times d$ matrix $A(\boldsymbol{\theta}) = |\det(I+\nabla\boldsymbol{\theta})|(I+\nabla\boldsymbol{\theta})^{-1}(I+\nabla\boldsymbol{\theta})^{-T}$. Note that Γ corresponds to the reference configuration in (2.3.17) while it is the deformed configuration in (2.3.8).

Now, for a given $v_s \in V_s(\Gamma)$ and any extension $\tilde{v}_s \in V_f(\Gamma)$ such that $\tilde{v}_s = v_s$ on Γ, the function $\tilde{v}_s \circ (I+\boldsymbol{\theta})^{-1} \in V_f(\Gamma_{\boldsymbol{\theta}})$ is an extension of $v_s \circ (I+\boldsymbol{\theta})^{-1} \in V_s(\Gamma_{\boldsymbol{\theta}})$ satisfying $\tilde{v}_s \circ (I+\boldsymbol{\theta})^{-1} = v_s \circ (I+\boldsymbol{\theta})^{-1}$ on $\Gamma_{\boldsymbol{\theta}}$. Therefore taking $v_s \circ (I+\boldsymbol{\theta})^{-1}$ as a test function in (2.3.11) and performing a change of variables yields a variational formulation for $u_{s,\boldsymbol{\theta}}$:

$$\forall v_s \in V_s(\Gamma), \int_{\Omega_s} \mu A(\boldsymbol{\theta})\nabla u_{s,\boldsymbol{\theta}} \cdot \nabla v_s \mathrm{d}x = \int_{\Omega_s} f_s \circ (I+\boldsymbol{\theta}) |\det(I+\nabla\boldsymbol{\theta})| \mathrm{d}x + \int_{\Omega_f} f_f \circ (I+\boldsymbol{\theta}) \tilde{v}_s |\det(I+\nabla\boldsymbol{\theta})| \mathrm{d}x - \int_{\Omega_f} \nu A(\boldsymbol{\theta})\nabla u_{f,\boldsymbol{\theta}} \cdot \nabla \tilde{v}_s \mathrm{d}x \quad (2.3.18)$$

for any extension $\tilde{v}_s \in V_f(\Gamma)$ satisfying $\tilde{v}_s = v_s$ on Γ.

Eventually, a classical use of the implicit function theorem, as in [184], reveals that the mappings $\boldsymbol{\theta} \mapsto u_{f,\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \mapsto u_{s,\boldsymbol{\theta}}$, from $W_0^{1,\infty}(D,\mathbb{R}^d)$ into $V_{f,0}(\Gamma)$ and $V_s(\Gamma)$ respectively, are Fréchet differentiable in the neighborhood of $\boldsymbol{\theta}=0$. Eqns. (2.3.14) and (2.3.15) are then simply obtained by differentiating (2.3.17) and (2.3.18) with respect to $\boldsymbol{\theta}$. □

Remark 2.5. The functions $\dot{u}_f(\boldsymbol{\theta})$ and $\dot{u}_s(\boldsymbol{\theta})$, defined by lemma 2.1, are the *Lagrangian* derivatives of u_f and u_s with respect to variations of Γ, a notion of derivative which is directly compatible with the variational setting of the PDEs (2.3.1) and (2.3.2). Recall that $\dot{u}_s(\boldsymbol{\theta})$ and $\dot{u}_f(\boldsymbol{\theta})$ do not coincide with the perhaps more physical 'Eulerian' derivatives $u_f'(\boldsymbol{\theta}), u_s'(\boldsymbol{\theta})$, that are the differentials of the mappings $\boldsymbol{\theta} \mapsto u_f(\Gamma_{\boldsymbol{\theta}})$ and $\boldsymbol{\theta} \mapsto u_s(\Gamma_{\boldsymbol{\theta}})$ (without composition by $(I+\boldsymbol{\theta})$); see e.g. [242, 184, 17] and chapter 1, section 1.2.

Adjoint system and volume expression of the shape derivative

Assuming that the transported objective function $\mathfrak{J}$ has continuous partial derivatives at

$$(\boldsymbol{\theta}, \widehat{u}_s, \widehat{u}_f) = (0, u_s(\Gamma), u_f(\Gamma)),$$

equation (2.3.13) and the chain rule imply that the mapping $\boldsymbol{\theta} \mapsto J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))$ is differentiable with respect to $\boldsymbol{\theta}$ and that its derivative at $\boldsymbol{\theta} = 0$ reads:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[\mathfrak{J}(\boldsymbol{\theta}, u_{f,\boldsymbol{\theta}}, u_{s,\boldsymbol{\theta}})\Big](\boldsymbol{\theta}) = \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) + \frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(\dot{u}_f(\boldsymbol{\theta})) + \frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(\dot{u}_s(\boldsymbol{\theta})). \quad (2.3.19)$$

Remark that, to keep notations as light as possible, we omit the point $(0, u_f(\Gamma), u_s(\Gamma))$ where the partial derivatives of $\mathfrak{J}$ are evaluated in (2.3.19) and below. The occurrence of the Lagrangian derivatives $\dot{u}_s(\boldsymbol{\theta})$ and $\dot{u}_f(\boldsymbol{\theta})$ in (2.3.19) are now classically eliminated by introducing adequate adjoint states $p_f \in V_f(\Gamma)$ and $p_s \in V_s(\Gamma)$, defined in weak form by the following variational problems:

$$\text{Find } p_s \in V_s(\Gamma) \text{ such that } \forall v_s \in V_s(\Gamma), \int_{\Omega_s} \mu \nabla p_s \cdot \nabla v_s \mathrm{d}x = \frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(v_s), \quad (2.3.20)$$

$$\text{Find } p_f \in V_f(\Gamma) \text{ such that } p_f = p_s \text{ on } \Gamma \text{ and } \forall v_f \in V_{f,0}(\Gamma), \int_{\Omega_f} \nu \nabla p_f \cdot \nabla v_f \mathrm{d}x = \frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(v_f). \quad (2.3.21)$$

Interestingly, the equality between normal derivatives $\nu \frac{\partial u_f}{\partial \boldsymbol{n}} = \mu \frac{\partial u_s}{\partial \boldsymbol{n}}$ featured in the system (2.3.2) for the state variables translates into the rather surprising boundary condition $p_f = p_s$ on Γ in (2.3.24) for the adjoint variables p_f. This boundary condition can be obtained in at least two ways. In the appendix, we provide in section 2.6.2 a derivation of this adjoint boundary condition based on a Lagrangian with Céa's method. In section 2.3.2 below, we also provide an equivalent mixed formulation for the state equations (2.3.1) and (2.3.2) which features different spaces for the 'primal' state variables and the 'adjoint' test functions. This difference of functional spaces is at the root of this change of boundary conditions at the interface.

Remark 2.6. The variational problems (2.3.20) and (2.3.21) make sense for general objective functions. Let us remark that they may lead to different adjoint equations when written in strong forms, which is the way they are obtained with Céa's method (this is one of the reasons why the derivation of shape derivatives must be repeated from the beginninthe beginning when changing the type of the objective functional). For instance, if we consider an objective functional depending on $u_s(\Gamma)$ and $u_f(\Gamma)$ through a volume integral (see section 2.6.2 below for the complete derivation of shape sensitivities with the method of Céa):

$$J(\Gamma, u_s(\Gamma), u_f(\Gamma)) := \int_{\Omega_f} j_f(u_f)\mathrm{d}x + \int_{\Omega_s} j_s(u_s)\mathrm{d}x \quad (2.3.22)$$

for two $\mathcal{C}^2$ functions $j_s, j_f : \mathbb{R} \to \mathbb{R}$ with bounded second order derivatives, then

$$\frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(v_s) = \int_{\Omega_s} j_s'(u_s) v_s \mathrm{d}x, \quad \frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(v_f) = \int_{\Omega_f} j_f'(u_f) v_f \mathrm{d}x$$

which yields the following strong form for the adjoint system (2.3.20) and (2.3.21):

$$\begin{cases} -\mu \Delta p_s = j_s'(u_s) \text{ in } \Omega_s \\ \mu \dfrac{\partial p_s}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma \\ p_s = 0 \text{ on } \partial\Omega_s \setminus \Gamma, \end{cases} \quad (2.3.23)$$

$$\begin{cases} -\nu \Delta p_f = j_f'(u_f) \text{ in } \Omega_f \\ p_f = p_s \text{ on } \Gamma \\ p_f = 0 \text{ on } \partial\Omega_f \setminus \Gamma. \end{cases} \quad (2.3.24)$$

However, if $J(\Gamma, u_f(\Gamma), u_s(\Gamma))$ now depends on the gradient of one of the variables, for instance

$$J(\Gamma, u_f(\Gamma), u_s(\Gamma)) := \frac{1}{2}\int_{\Omega_s} \mu |\nabla u_s|^2 \mathrm{d}x,$$

then

$$\frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(v_s) = \int_{\Omega_s} \mu \nabla u_s \cdot \nabla v_s \mathrm{d}x = -\int_{\Omega_s} \mu \Delta u_s v_s \mathrm{d}x + \int_{\partial\Omega_s} \mu \frac{\partial u_s}{\partial \boldsymbol{n}} v_s \mathrm{d}s, \quad \frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(v_f) = 0,$$

which yields a different strong form for the adjoint equation associated with p_s:

$$\begin{cases} -\mu\Delta p_s = -\mu\Delta u_s \text{ in } \Omega_s \\ \mu\dfrac{\partial p_s}{\partial \boldsymbol{n}} = \mu\dfrac{\partial u_s}{\partial \boldsymbol{n}} \text{ on } \Gamma \\ p_s = 0 \text{ on } \partial\Omega_s \setminus \Gamma, \end{cases} \tag{2.3.25}$$

$$\begin{cases} -\nu\Delta p_f = 0 \text{ in } \Omega_f \\ p_f = p_s \text{ on } \Gamma \\ p_f = 0 \text{ on } \partial\Omega_f \setminus \Gamma. \end{cases} \tag{2.3.26}$$

The adjoint variables p_s and p_f allow to obtain an expression independent of $\dot{u}_s(\boldsymbol{\theta})$ and $\dot{u}_f(\boldsymbol{\theta})$ for the shape derivative of J as in chapter 1, proposition 1.4:

Proposition 2.1. *Assume that the transported objective function $\mathfrak{J}$ given by (2.3.12) has continuous partial derivatives at $(\boldsymbol{\theta}, u_s, u_f) = (0, u_s(\Gamma), u_f(\Gamma))$. Then, the mapping $\boldsymbol{\theta} \mapsto J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))$, from $W_0^{1,\infty}(D, \mathbb{R}^d)$ into $\mathbb{R}$, is Fréchet differentiable at $\boldsymbol{\theta} = 0$ and its derivative reads:*

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) &= \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) \\ &+ \int_{\Omega_f} \Big[\mathrm{div}(f_f\boldsymbol{\theta})p_f + (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nu\nabla u_f \cdot \nabla p_f\Big]\mathrm{d}x \\ &+ \int_{\Omega_s} \Big[\mathrm{div}(f_s\boldsymbol{\theta})p_s + (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\mu\nabla u_s \cdot \nabla p_s\Big]\mathrm{d}x, \end{aligned} \tag{2.3.27}$$

where p_f and p_s are the adjoint states defined by (2.3.20) and (2.3.21).

Remark 2.7. Since formula (2.3.27) for the shape derivative of J involves domain integrals, it is called a volume expression of the shape derivative. In the next subsection, it is shown that it can equivalently be written in terms of surface integrals (which are more obviously satisfying the so-called Hadamard structure theorem stated in chapter 1, proposition 1.1).

Proof. We first insert $v_s = \dot{u}_s(\boldsymbol{\theta})$ and $v_f = \dot{u}_f(\boldsymbol{\theta})$ in the adjoint equations (2.3.20) and (2.3.21) to obtain:

$$\frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(\dot{u}_s(\boldsymbol{\theta})) = \int_{\Omega_s} \mu\nabla p_s \cdot \nabla\dot{u}_s(\boldsymbol{\theta})\mathrm{d}x, \tag{2.3.28}$$

$$\frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(\dot{u}_f(\boldsymbol{\theta})) = \int_{\Omega_f} \nu\nabla p_f \cdot \nabla\dot{u}_f(\boldsymbol{\theta})\mathrm{d}x. \tag{2.3.29}$$

Then taking $v_s = p_s$ in the variational formulation (2.3.15) for $\dot{u}_s(\boldsymbol{\theta})$ yields:

$$\int_{\Omega_s} \mu\nabla p_s\cdot\nabla\dot{u}_s(\boldsymbol{\theta})\mathrm{d}x = \int_{\Omega_s} (\mathrm{div}(f_s\boldsymbol{\theta})p_s + (\nabla\boldsymbol{\theta}+\nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\mu\nabla u_s\cdot\nabla p_s)\mathrm{d}x - \int_\Gamma \nu\frac{\partial\dot{u}_f(\boldsymbol{\theta})}{\partial\boldsymbol{n}}p_s\mathrm{d}s. \tag{2.3.30}$$

Now remarking that $\tilde{v} = p_f \in V_f(\Gamma)$ is an extension of $v = p_s = p_f \in H_{00}^{1/2}(\Gamma)$, taking $\tilde{v} = p_f$ in (2.3.16) implies:

$$\int_{\Omega_f} \nu\nabla p_f \cdot \nabla\dot{u}_f(\boldsymbol{\theta})\mathrm{d}x = \int_{\Omega_f} (\mathrm{div}(f_f\boldsymbol{\theta})p_f + (\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta})I)\nu\nabla u_f \cdot \nabla p_f)\mathrm{d}x + \int_\Gamma \nu\frac{\partial\dot{u}_f(\boldsymbol{\theta})}{\partial\boldsymbol{n}}p_s\mathrm{d}s. \tag{2.3.31}$$

The desired formula (2.3.27) follows then by summation of (2.3.30) and (2.3.31). □

Obtaining the surface expression of the shape derivative

The Hadamard's structure theorem (chapter 1, proposition 1.1) states that under additional regularity on the optimized interface Γ and on the considered vector fields $\boldsymbol{\theta}$, the shape derivative of a sufficiently smooth objective function J depends only on the normal component $\boldsymbol{\theta}\cdot\boldsymbol{n}$ on Γ. In this section, we highlight how this remark allows to obtain a surface expression for the shape derivative of J from the volume expression (2.3.27) in a simple way. To achieve this, we classically rely on three regularity assumptions, which we assume to be satisfied throughout this section:

1. The considered deformations $\boldsymbol{\theta}$ are smooth, e.g. of class $\mathcal{C}^1$;
2. The state and adjoint variables u_s, u_f, p_s, p_f enjoy H^2 regularity in their domain of definition; this is for instance the case when D, Ω_s, Ω_f and Γ are smooth enough; see e.g. [76].
3. The partial derivative $\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}$ is sufficiently "regular", in the sense that there exist $\boldsymbol{f}_{\mathfrak{J}} \in L^1(D, \mathbb{R}^d)$ and $\boldsymbol{g}_{\mathfrak{J}} \in L^1(\Gamma, \mathbb{R}^d)$ such that

$$\forall \boldsymbol{\theta} \in W_0^{1,\infty}(D, \mathbb{R}^d),\ \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_D \boldsymbol{f}_{\mathfrak{J}} \cdot \boldsymbol{\theta} \mathrm{d}x + \int_\Gamma \boldsymbol{g}_{\mathfrak{J}} \cdot \boldsymbol{\theta} \mathrm{d}s. \tag{2.3.32}$$

   The uniqueness of the decomposition (2.3.32), when it exists, is straightforward. The existence of such a structure is typically obtained by performing integration by parts on $\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}$ using, in turn, the H^2 regularity of the state and adjoint variables u_s, u_f, p_s, p_f.

In the following and under these assumptions, we denote by

$$\forall \boldsymbol{\theta} \in W_0^{1,\infty}(D, \mathbb{R}^d), \quad \overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) := \int_\Gamma (\boldsymbol{g}_{\mathfrak{J}} \cdot \boldsymbol{n})(\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}s, \tag{2.3.33}$$

the part of $\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}$ that depends only on $\boldsymbol{\theta} \cdot \boldsymbol{n}$. The main result of this section is the following proposition:

Proposition 2.2. *Under the above assumptions, the shape derivative (2.3.27) rewrites as an integral over the boundary Γ involving only the normal component $\boldsymbol{\theta} \cdot \boldsymbol{n}$:*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) &= \overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) \\ &+ \int_\Gamma \left[f_f p_f - f_s p_s - \nu \nabla u_f \cdot \nabla p_f + \mu \nabla u_s \cdot \nabla p_s + 2\nu \frac{\partial u_f}{\partial \boldsymbol{n}} \frac{\partial p_f}{\partial \boldsymbol{n}} - 2\mu \frac{\partial u_s}{\partial \boldsymbol{n}} \frac{\partial p_s}{\partial \boldsymbol{n}} \right] (\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}s. \end{aligned} \tag{2.3.34}$$

Proof. The regularity assumptions allow to perform integration by parts in the volume expression (2.3.27), which yields:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, u_f(\Gamma_{\boldsymbol{\theta}}), u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) \\ &= \int_\Gamma \boldsymbol{g}_{\mathfrak{J}} \cdot \boldsymbol{\theta} \mathrm{d}s + \int_\Gamma [f_f p_f (\boldsymbol{\theta} \cdot \boldsymbol{n}) + \nu (\boldsymbol{\theta} \cdot \nabla u_f)(\nabla p_f \cdot \boldsymbol{n}) + \nu (\boldsymbol{n} \cdot \nabla u_f)(\nabla p_f \cdot \boldsymbol{\theta}) - \nu (\nabla u_f \cdot \nabla p_f)(\boldsymbol{\theta} \cdot \boldsymbol{n})] \mathrm{d}s \\ &\quad - \int_\Gamma [f_s p_s (\boldsymbol{\theta} \cdot \boldsymbol{n}) + \mu (\boldsymbol{\theta} \cdot \nabla u_s)(\nabla p_s \cdot \boldsymbol{n}) + \mu (\boldsymbol{n} \cdot \nabla u_s)(\nabla p_s \cdot \boldsymbol{\theta}) - \mu (\nabla u_s \cdot \nabla p_s)(\boldsymbol{\theta} \cdot \boldsymbol{n})] \mathrm{d}s \\ &\quad + \int_D \boldsymbol{\Lambda} \cdot \boldsymbol{\theta} \mathrm{d}x, \end{aligned} \tag{2.3.35}$$

for some function $\boldsymbol{\Lambda} \in L^1(D, \mathbb{R}^d)$ which does not need to be written explicitly. Hadamard's structure theorem implies that (2.3.35) must vanish for vector fields $\boldsymbol{\theta}$ compactly supported in Ω_s or in Ω_f, or for vector fields $\boldsymbol{\theta}$ which are tangential to Γ. Therefore, $\boldsymbol{\Lambda} = 0$, and one obtains (2.3.34) by computing (2.3.35) with $\boldsymbol{\theta}$ normal to Γ. □

Formula (2.3.34) is called a surface expression of the shape derivative of J. It is retrieved in section 2.6.2 for a particular case of an objective functional J with Céa's method. An asset of the above Lagrangian method is that it depends neither on the nature of the objective function J (e.g. if it depends on u_f, u_s or their gradients), nor on the type of boundary conditions satisfied by the state variables u_f and u_s. Both expressions (2.3.27) and (2.3.34) are convenient to implement because they require minimal inputs from the user: namely, the expression of the partial derivatives of $\mathfrak{J}$ with respect to $\boldsymbol{\theta}$ (for (2.3.27) and (2.3.34)) and to the state variables u_s, u_f (for solving the adjoint system (2.3.14) and (2.3.15)).

Remark 2.8. Propositions 2.1 and 2.2 respectively reduce to propositions 1.8 and 1.9 in the context of the Poisson problem (1.2.19) considered in chapter 1, by replacing u and p by u_f and p_f and setting $u_s = 0, p_s = 0$. Notice that in the derivation of (2.3.34), we never used the zero Dirichlet boundary condition of u_f, p_f and u_s, p_s on respectively $\partial\Omega_f \backslash \Gamma$ and $\partial\Omega_s \backslash \Gamma$. Therefore, it can be seen that (2.3.34) is also valid in the context of proposition 1.9 (featuring Neumann boundary conditions, $\Gamma_N \neq \emptyset$).

Remark 2.9. In practice, one can obtain very quickly the surface expression (2.3.34) of the shape derivative from the volume expression (2.3.27) by

1. replacing volume integrals by surface integrals, adding a minus sign when these involve Ω_s in order to account the orientation of the normal $\boldsymbol{n}$ outward to Ω_f;
2. replacing $\nabla\boldsymbol{\theta}$ by $\boldsymbol{n}\boldsymbol{n}^T\,(\boldsymbol{\theta}\cdot\boldsymbol{n})$.

Let us now illustrate these results with the calculation of the shape derivative of the 'solid compliance', which is not of the form (2.3.22) considered previously:

$$J(\Gamma, u_f(\Gamma), u_s(\Gamma)) = \int_{\Omega_s} \mu|\nabla u_s|^2 \mathrm{d}x. \tag{2.3.36}$$

The associated transported objective function $\mathfrak{J}$ via (2.3.12) reads:

$$\mathfrak{J}(\boldsymbol{\theta}, \widehat{u}_f, \widehat{u}_s) = \int_{(I+\boldsymbol{\theta})\Omega_s} \mu|\nabla(\widehat{u}_s \circ (I+\boldsymbol{\theta})^{-1})|^2 \mathrm{d}x = \int_{\Omega_s} \mu|(I+\nabla\boldsymbol{\theta})^{-T}\nabla\widehat{u}_s|^2|\det(I+\nabla\boldsymbol{\theta})|\mathrm{d}x. \tag{2.3.37}$$

Lemma 2.2. *The functional $\mathfrak{J}$ defined in (2.3.37) has continuous partial derivatives at*

$$(\boldsymbol{\theta}, \widehat{u}_s, \widehat{u}_f) = (0, u_s(\Gamma), u_f(\Gamma)),$$

which are given by:

$$\frac{\partial\mathfrak{J}}{\partial\widehat{u}_f}(v_f) = 0, \qquad \frac{\partial\mathfrak{J}}{\partial\widehat{u}_s}(v_s) = \int_{\Omega_s} 2\mu\nabla u_s\cdot\nabla v_s \mathrm{d}x, \tag{2.3.38}$$

$$\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\Omega_s} (\mu|\nabla u_s|^2\mathrm{div}(\boldsymbol{\theta}) - 2\mu\nabla u_s\cdot\nabla\boldsymbol{\theta}\cdot\nabla u_s)\mathrm{d}x, \tag{2.3.39}$$

$$\overline{\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \int_\Gamma \left[-\mu|\nabla u_s|^2 + 2\mu\left|\frac{\partial u_s}{\partial\boldsymbol{n}}\right|^2\right](\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s. \tag{2.3.40}$$

Therefore the solid compliance (2.3.36) $\boldsymbol{\theta}\mapsto J(\Gamma_{\boldsymbol{\theta}}, u_s(\Gamma_{\boldsymbol{\theta}}), u_f(\Gamma_{\boldsymbol{\theta}}))$ is differentiable with respect to $\boldsymbol{\theta}\in W_0^{1,\infty}(D,\mathbb{R}^d)$, and the shape derivative is given by proposition 2.1 in volume form or proposition 2.2 in surface form.

Remark 2.10. Note that in this particular case the adjoint state p_s defined by (2.3.20) satisfies $p_s = 2u_s$, but no such property holds for p_f (the problem is not self-adjoint). Unlike the usual case of the standard Poisson equation, there is no obvious way to write the solid compliance (2.3.36) in a form (2.3.22) involving only volumic integrals in u_s, u_f without occurrence of their gradient as in section 2.6.2. For example, the formulation as the work of external forces,

$$J(\Gamma, u_s(\Gamma), u_f(\Gamma)) = \int_{\Omega_s} f_s u_s \mathrm{d}x - \int_\Gamma \nu\frac{\partial u_f}{\partial\boldsymbol{n}} u_s \mathrm{d}s, \tag{2.3.41}$$

involves the gradient of u_f; it is therefore less convenient to handle than (2.3.36), since the partial derivative

$$\frac{\partial\mathfrak{J}}{\partial u_f} : v \mapsto -\int_\Gamma \nu\frac{\partial v}{\partial\boldsymbol{n}} u_s \mathrm{d}s$$

is not defined on $V_{f,0}$, but merely in a subspace of smoother functions (a treatment could be considered however, by considering different solution spaces, see e.g. [294, 216]). We shall also discuss this point when calculating the shape derivative of the lift functional in chapter 6, section 6.2.3.

A mixed variational formulation for the state and adjoint problem

We conclude this section with a remark which may shed some light on the a priori surprising boundary condition $p_s = p_f$ on Γ for the adjoint systems (2.3.20) and (2.3.21). The key observation is that the systems (2.3.1) and (2.3.2) may be formally described by means of a variational formulation for the couple (u_s, u_f) which features different functional spaces for the solution and test functions (a so-called

Petrov-Galerkin variational formulation, see e.g. [144]). The latter is simply obtained by summing (2.3.8) and (2.3.11):

Find $(u_f, u_s) \in H^1_0(\Omega_f) \times H^1(\Omega_s)$ such that $\forall v \in H^1(D)$,

$$\int_{\Omega_s} \mu \nabla u_s \cdot \nabla v \mathrm{d}x + \int_{\Omega_f} \nu \nabla u_f \cdot \nabla v \mathrm{d}x = \int_{\Omega_s} f_s v \mathrm{d}x + \int_{\Omega_f} f_f v \mathrm{d}x. \quad (2.3.42)$$

Problem (2.3.42) implicitly encloses the transmission condition $\mu \frac{\partial u_s}{\partial \boldsymbol{n}} = \nu \frac{\partial u_f}{\partial \boldsymbol{n}}$ on Γ, and the need to resort to extensions of test functions defined on Ω_s in the variational formulation (2.3.11) for u_s is reflected here in that the test function v belongs to $H^1(D)$.

As is customary (see for instance section 2.6.2), the adjoint system is obtained by formally taking the linear transpose of the mixed variational formulation (2.3.42) (with a different right-hand side), which exchanges the roles of the functional spaces for the sought solution and the test functions:

Find $p \in H^1(D)$ such that $\forall (v_f, v_s) \in H^1_0(\Omega_f) \times H^1(\Omega_s)$,

$$\int_{\Omega_s} \nu \nabla p \cdot \nabla v_s \mathrm{d}x + \int_{\Omega_f} \nu \nabla p \cdot \nabla v_f \mathrm{d}x = \frac{\partial \mathfrak{J}}{\partial \widehat{u}_s}(v_s) + \frac{\partial \mathfrak{J}}{\partial \widehat{u}_f}(v_f). \quad (2.3.43)$$

The above equation is in turn equivalent to the triangular system (2.3.20) and (2.3.21) for the restrictions p_s and p_f of p on Ω_s and Ω_f, where the transmission condition $p_s = p_f$ on Γ is implicitly contained in the requirement that p be an element of $H^1(D)$.

Remark 2.11. The above argument is only formal because it is not obvious that the variational problem (2.3.42) be well-posed. It can however be made rigorous by changing the functional spaces featured in there, more precisely, by searching for (u_f, u_s) in

$$\left\{(u_f, u_s) \in H^1_0(\Omega_f) \times V_s(\Gamma),\ \ \mathbb{1}_{\Omega_f} \nu \nabla u_f + \mathbb{1}_{\Omega_s} \nu \nabla u_s \in H(\mathrm{div}, D)\right\},$$

where $\mathbb{1}_{\Omega_f}$ (resp. $\mathbb{1}_{\Omega_s}$) is the characteristic function of Ω_f (resp. Ω_s), and by searching for v in $L^2(D)$. It is then possible to prove that the *inf-sup* condition of the Banach-Necas-Babuska theorem holds (see [144]) in the case of this new version of (2.3.42), which guarantees its well-posedness.

2.4 SHAPE DERIVATIVES FOR THE THREE-PHYSIC PROBLEM

We now briefly describe how the methodology presented in section 2.3 applies to the weakly coupled, multiphysics system (2.2.1) to (2.2.3). Let us introduce the functional spaces which are required, respectively, for the Navier-Stokes equations

$$V_{\boldsymbol{v},p}(\Gamma) = \{(\boldsymbol{w}, q) \in H^1(\Omega_f, \mathbb{R}^d) \times L^2(\Omega_f)/\mathbb{R} \,|\, \boldsymbol{w} = 0 \text{ on } \partial\Omega_f\},$$

for the thermal equation

$$V_T(\Gamma) = \{S \in H^1(D) \,|\, S = 0 \text{ on } \partial\Omega_T^D\},$$

for the thermo-mechanical equations

$$V_{\boldsymbol{u}}(\Gamma) = \{\boldsymbol{r} \in H^1(\Omega_s, \mathbb{R}^d) \,|\, \boldsymbol{r} = 0 \text{ on } \partial\Omega_s^D\}.$$

Note that, as is customary in the theory of the Navier-Stokes equations, the quotient space $L^2(\Omega_f)/\mathbb{R}$, associated to the pressure field, gathers square integrable functions defined up to an additive constant. We consider as well the affine spaces associated to the non-homogeneous Dirichlet boundary data $\boldsymbol{v}_0 \in H^{1/2}(\partial\Omega_f^D, \mathbb{R}^d)$, $\boldsymbol{u}_0 \in H^{1/2}(\partial\Omega_s^D, \mathbb{R}^d)$ and $T_0 \in H^{1/2}(\partial\Omega_T^D)$ featured in (2.2.1) to (2.2.3):

$$\begin{aligned}
\boldsymbol{v}_0 + V_{\boldsymbol{v},p}(\Gamma) &= \{(\boldsymbol{w}, q) \in H^1(\Omega_f, \mathbb{R}^d) \times L^2(\Omega_f)/\mathbb{R} \,|\, \boldsymbol{w} = \boldsymbol{v}_0 \text{ on } \partial\Omega_f^D \text{ and } \boldsymbol{w} = 0 \text{ on } \Gamma\},\\
T_0 + V_T(\Gamma) &= \{S \in H^1(D) \,|\, S = T_0 \text{ on } \partial\Omega_T^D\},\\
\boldsymbol{u}_0 + V_{\boldsymbol{u}}(\Gamma) &= \{\boldsymbol{r} \in H^1(\Omega_s, \mathbb{R}^d) \,|\, \boldsymbol{r} = \boldsymbol{u}_0 \text{ on } \partial\Omega_s^D\}.
\end{aligned}$$

Finally, we shall make use of the trace space

$$H^{1/2}_{00}(\Gamma, \mathbb{R}^d) = \{\boldsymbol{r}|_\Gamma \mid \boldsymbol{r} \in H^1(\Omega_f, \mathbb{R}^d) \text{ and } \boldsymbol{r} = 0 \text{ on } \partial\Omega_f \backslash \Gamma\}, \quad (2.4.1)$$

and its dual $H_{00}^{-1/2}(\Gamma,\mathbb{R}^d)$. The state variables $\boldsymbol{v},p,T,\boldsymbol{u}$ are the solutions to the following variational problems: for the Navier-Stokes equations (2.2.1), find $(\boldsymbol{v},p)\in\boldsymbol{v}_0+V_{\boldsymbol{v},p}(\Gamma)$ such that

$$\forall(\boldsymbol{w},q)\in V_{\boldsymbol{v},p}(\Gamma)\quad \int_{\Omega_f}[\sigma_f(\boldsymbol{v},p):\nabla\boldsymbol{w}+\rho\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v}-q\mathrm{div}(\boldsymbol{v})]\,\mathrm{d}x=\int_{\Omega_f}\boldsymbol{f}_f\cdot\boldsymbol{w}\mathrm{d}x; \tag{2.4.2}$$

for the thermal equation (2.2.2), find $T\in T_0+V_T(\Gamma)$ such that, for any $S\in V_T(\Gamma)$,

$$\int_{\Omega_s}k_s\nabla T\cdot\nabla S\mathrm{d}x+\int_{\Omega_f}(k_f\nabla T\cdot\nabla S+\rho c_pS\boldsymbol{v}\cdot\nabla T)\mathrm{d}x=\int_{\Omega_s}Q_sS\mathrm{d}x+\int_{\Omega_f}Q_fS\mathrm{d}x+\int_{\partial\Omega_T^N}hS\mathrm{d}s; \tag{2.4.3}$$

for the thermo-mechanical equations (2.2.3), find $\boldsymbol{u}\in\boldsymbol{u}_0+V_{\boldsymbol{u}}(\Gamma)$ such that

$$\forall\boldsymbol{r}\in V_{\boldsymbol{u}}(\Gamma),\quad \int_{\Omega_s}\sigma_s(\boldsymbol{u},T_s):\nabla\boldsymbol{r}\mathrm{d}x=\int_{\Omega_s}\boldsymbol{f}_s\cdot\boldsymbol{r}\mathrm{d}x+\int_{\partial\Omega_s^N}\boldsymbol{g}\cdot\boldsymbol{r}\mathrm{d}s-\int_\Gamma\boldsymbol{r}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{n}\mathrm{d}s. \tag{2.4.4}$$

Let us comment on the well-posedness of the coupled system of variational problems (2.4.2), (2.4.3) and (2.4.4). As in section 2.3, the volumic source terms are assumed to enjoy H^1 regularity in their domain: $\boldsymbol{f}_f\in H^1(\Omega_f,\mathbb{R}^d)$, $\boldsymbol{f}_s\in H^1(\Omega_s,\mathbb{R}^d)$, $Q_f\in H^1(\Omega_f)$, $Q_s\in H^1(\Omega_s)$. The surface fluxes h, $\boldsymbol{g}$ are assumed to belong to L^2 spaces. At first, the classical theory for the Navier-Stokes equation states that (2.4.2) is well-posed as soon as the Reynolds numbers $\mathrm{Re}:=||\boldsymbol{v}_0||_{H^{1/2}(\partial\Omega_f^D,\mathbb{R}^d)}\rho/\nu$ is sufficiently small; see [304].

The variational formulation (2.4.3) of the thermal problem is not well-posed in utter generality because of the lack of coercivity induced by the advection term $\int_{\Omega_f}\rho c_pS\boldsymbol{v}\cdot\nabla T\mathrm{d}x$ and of the presence of inhomogeneous Dirichlet boundary conditions. However, in usual applications [255, 225], it is customary to impose a Dirichlet boundary condition $T=T_{0,f}$ at the inlet of the computational domain (i.e. where $\boldsymbol{v}\cdot\boldsymbol{n}<0$) and a Neumann boundary condition $-k_f\nabla T\cdot\boldsymbol{n}=0$ at the outlet ($\boldsymbol{v}\cdot\boldsymbol{n}>0$). This together with the incompressibility condition $\mathrm{div}(\boldsymbol{v})=0$ is easily shown to imply the coercivity of the bilinear form featured in (2.4.3); see e.g. [75].

Eventually, the well-posedness of the linear elasticity problem (2.4.4) results from the Lax-Milgram theorem, the only subtle point is that, as in section 2.3, $\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{n}$ is an element of the dual space of $H_{00}^{1/2}(\Gamma,\mathbb{R}^d)$: if $\boldsymbol{v},p$ were regular enough, the following integration by parts would hold true:

$$\forall\boldsymbol{r}\in V_{\boldsymbol{u}}(\Gamma),\ -\int_\Gamma\boldsymbol{r}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{n}\mathrm{d}s=\int_{\Omega_f}(-\mathrm{div}(\sigma_f(\boldsymbol{v},p))\cdot\boldsymbol{r}-\sigma_f(\boldsymbol{v},p):\nabla\boldsymbol{r})\mathrm{d}x. \tag{2.4.5}$$

Hence the normal stress $\sigma_f(\boldsymbol{v},p)\boldsymbol{n}$ can be understood mathematically as an element of $H_{00}^{-1/2}(\Gamma,\mathbb{R}^d)$, defined by

$$\forall\boldsymbol{r}\in H_{00}^{1/2}(\Gamma,\mathbb{R}^d),\ -\int_\Gamma\boldsymbol{r}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{n}\mathrm{d}s=\int_{\Omega_f}(\boldsymbol{f}_f\cdot\tilde{\boldsymbol{r}}-\rho\tilde{\boldsymbol{r}}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v}-\sigma_f(\boldsymbol{v},p):\nabla\tilde{\boldsymbol{r}}+\tilde{q}\mathrm{div}(\boldsymbol{v}))\mathrm{d}x \tag{2.4.6}$$

for any extension $(\tilde{\boldsymbol{r}},\tilde{q})\in H^1(\Omega_f,\mathbb{R}^d)\times(L^2(\Omega_f)/\mathbb{R})$ satisfying $\tilde{\boldsymbol{r}}=\boldsymbol{r}$ on Γ. Note that although it is not fully necessary, we consider also an extension $\tilde{q}$ of the pressure field to maintain a complete analogy with (2.3.10). This turns to be convenient in the calculation of the shape derivative performed in section 2.6.1.

Throughout this section, we assume that the above conditions for the well-posedness of the coupled system of variational problems (2.4.2), (2.4.3) and (2.4.4) are fulfilled.

In the above context, we aim at solving the minimization problem (2.2.6) where the velocity $\boldsymbol{v}(\Gamma)$, pressure $p(\Gamma)$, temperature $T(\Gamma)$ and elastic displacement $\boldsymbol{u}(\Gamma)$ are the solutions to (2.2.1) to (2.2.3).

In order to compute shape derivatives with respect to variations of a given interface Γ, we introduce as in section 2.3.2 a transported functional $\mathfrak{J}$ defined on the fixed functional space

$$W_0^{1,\infty}(D,\mathbb{R}^d)\times H^1(\Omega_f,\mathbb{R}^d)\times(L^2(\Omega_f)/\mathbb{R})\times H^1(D)\times H^1(\Omega_s,\mathbb{R}^d)$$

by:

$$\begin{aligned}\forall\boldsymbol{\theta}\in W_0^{1,\infty}(D,\mathbb{R}^d),(\widehat{\boldsymbol{v}},\widehat{p},\widehat{T},\widehat{\boldsymbol{u}})\in H^1(\Omega_f,\mathbb{R}^d)\times(L^2(\Omega_f)/\mathbb{R})\times H^1(D)\times H^1(\Omega_s,\mathbb{R}^d),\\ \mathfrak{J}(\boldsymbol{\theta},\widehat{\boldsymbol{v}},\widehat{p},\widehat{T},\widehat{\boldsymbol{u}})=J(\Gamma_{\boldsymbol{\theta}},\widehat{\boldsymbol{v}}\circ(I+\boldsymbol{\theta})^{-1},\widehat{p}\circ(I+\boldsymbol{\theta})^{-1},\widehat{T}\circ(I+\boldsymbol{\theta})^{-1},\widehat{\boldsymbol{u}}\circ(I+\boldsymbol{\theta})^{-1}).\end{aligned} \tag{2.4.7}$$

The only requirement made on J is that the associated functional $\mathfrak{J}$ has continuous partial derivatives at $(\boldsymbol{\theta},\widehat{\boldsymbol{v}},\widehat{p},\widehat{T},\widehat{\boldsymbol{u}})=(0,\boldsymbol{v}(\Gamma),p(\Gamma),T(\Gamma),\boldsymbol{u}(\Gamma))$. Under this assumption, arguing as in section 2.3.2 and

section 2.6.2, we define the adjoint variables $\boldsymbol{w}, q, S, \boldsymbol{r}$ as follows. The elasticity adjoint variable $\boldsymbol{r} \in V_{\boldsymbol{u}}(\Gamma)$ is the solution of

$$\int_{\Omega_s} Ae(\boldsymbol{r}) : \nabla \boldsymbol{r}' \mathrm{d}x = \frac{\partial \mathfrak{J}}{\partial \widehat{\boldsymbol{u}}}(\boldsymbol{r}') \quad \forall \boldsymbol{r}' \in V_{\boldsymbol{u}}(\Gamma)\,. \tag{2.4.8}$$

The thermal adjoint variable $S \in V_T(\Gamma)$ is the solution of

$$\int_{\Omega_s} k_s \nabla S \cdot \nabla S' \mathrm{d}x + \int_{\Omega_f} (k_f \nabla S \cdot \nabla S' + \rho c_p S \boldsymbol{v} \cdot \nabla S') \mathrm{d}x = \int_{\Omega_s} \alpha \mathrm{div}(\boldsymbol{r}) S' \mathrm{d}x + \frac{\partial \mathfrak{J}}{\partial \widehat{T}}(S) \quad \forall S' \in V_T(\Gamma)\,. \tag{2.4.9}$$

The fluid adjoint variables $(\boldsymbol{w}, q) \in H^1(\Omega_f, \mathbb{R}^d) \times (L^2(\Omega_f)/\mathbb{R})$ are the solution of

$$\begin{aligned} &\boldsymbol{w} = \boldsymbol{r} \text{ on } \Gamma \text{ and } \forall (\boldsymbol{w}', q') \in V_{\boldsymbol{v},p}(\Gamma) \\ &\int_{\Omega_f} \Big(\sigma_f(\boldsymbol{w}, q) : \nabla \boldsymbol{w}' + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{w}' \cdot \boldsymbol{v} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{w}' - q' \mathrm{div}(\boldsymbol{w}) \Big) \mathrm{d}x = \\ &\qquad \int_{\Omega_f} -\rho c_p S \nabla T \cdot \boldsymbol{w}' \mathrm{d}x + \frac{\partial \mathfrak{J}}{\partial (\boldsymbol{v}', p')}(\boldsymbol{w}', q'), \end{aligned} \tag{2.4.10}$$

and we recall our convention whereby the point $(0, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma))$ where the partial derivatives of J are evaluated is omitted.

As expected, the cascade dependency $(\boldsymbol{v}, q) \to T \to \boldsymbol{u}$ in the state variables (the variable on the right of the arrow depends on that on the left) is reversed into $\boldsymbol{r} \to S \to (\boldsymbol{w}, q)$ for the adjoint variables, which reflects the fact that the adjoint system is formally the linearized transpose of the state problem.

Remark 2.12. The existence and uniqueness of a solution $(\boldsymbol{w}, q, S, \boldsymbol{r})$ in $H^1(\Omega_f, \mathbb{R}^d) \times (L^2(\Omega_f)/\mathbb{R}) \times H^1(D) \times H^1(\Omega_s, \mathbb{R}^d)$ to the adjoint system (2.4.8) to (2.4.10) follows from the same considerations as in the case of the state system (2.4.2) to (2.4.4), except when it comes to the *linearized* Navier-Stokes equation (2.4.10). The latter is well-posed provided the Reynolds number is sufficiently small; see [171], Chap. IV about this point.

Remark 2.13. Let us consider the case of a particular objective functional of the form

$$J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = \int_{\Omega_f} j_f(\boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma)) \mathrm{d}x + \int_{\Omega_s} j_s(\boldsymbol{u}(\Gamma), T(\Gamma)) \mathrm{d}x, \tag{2.4.11}$$

where $j_f : \mathbb{R}^d_{\boldsymbol{v}} \times \mathbb{R}_p \times \mathbb{R}_T \to \mathbb{R}$ and $j_s : \mathbb{R}^d_{\boldsymbol{u}} \times \mathbb{R}_T \to \mathbb{R}$ are smooth and satisfy adequate growth conditions. The adjoint equations (2.4.8) to (2.4.10) rewrite respectively in strong form, for the elasticity system,

$$\begin{cases} -\mathrm{div}(Ae(\boldsymbol{r})) = \dfrac{\partial j_s}{\partial \boldsymbol{u}} & \text{in } \Omega_s \\ \boldsymbol{r} = 0 & \text{on } \partial\Omega_s^D \\ Ae(\boldsymbol{r})\boldsymbol{n} = 0 & \text{on } \partial\Omega_s^N \cup \Gamma, \end{cases}$$

for the thermal equation,

$$\begin{cases} -\mathrm{div}(k_s \nabla S_s) = \alpha \mathrm{div}(\boldsymbol{r}) + \dfrac{\partial j_s}{\partial T_s} & \text{in } \Omega_s \\ -\mathrm{div}(k_f \nabla S_f) - \rho c_p \boldsymbol{v} \cdot \nabla S_f = \dfrac{\partial j_f}{\partial T_f} & \text{in } \Omega_f \\ S = 0 & \text{on } \partial\Omega_T^D \\ k_s \dfrac{\partial S_s}{\partial \boldsymbol{n}} = 0 & \text{on } \partial\Omega_T^N \cap \partial\Omega_s \\ k_f \dfrac{\partial S_f}{\partial \boldsymbol{n}} + \rho c_p (\boldsymbol{v} \cdot \boldsymbol{n}) S_f = 0 & \text{on } \partial\Omega_T^N \cap \partial\Omega_f \\ S_s = S_f & \text{on } \Gamma \\ -k_s \dfrac{\partial S_s}{\partial \boldsymbol{n}} = -k_f \dfrac{\partial S_f}{\partial \boldsymbol{n}} & \text{on } \Gamma, \end{cases}$$

for the Navier-Stokes equations,

$$\begin{cases} -\mathrm{div}(\sigma_f(\boldsymbol{w},q)) + \rho(\nabla \boldsymbol{v}^T\,\boldsymbol{w} - \nabla \boldsymbol{w}\,\boldsymbol{v}) = -\rho c_p S \nabla T_f + \dfrac{\partial j_f}{\partial \boldsymbol{v}} & \text{in } \Omega_f \\ -\mathrm{div}(\boldsymbol{w}) = \dfrac{\partial j_f}{\partial p} & \text{in } \Omega_f \\ \boldsymbol{w} = 0 & \text{on } \partial\Omega_f^D \\ \sigma_f(\boldsymbol{w},q)\boldsymbol{n} + \rho(\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{w} = 0 & \text{on } \partial\Omega_f^N \\ \boldsymbol{w} = \boldsymbol{r} & \text{on } \Gamma. \end{cases} \tag{2.4.12}$$

Note the "surprising" fact in the linearized adjoint system (2.4.12) for the Navier-Stokes equations that the interface condition for the velocity variable $\boldsymbol{w}$ is of Dirichlet type on Γ, while it was of Neumann type for the direct elasticity problem (2.2.3).

A very similar analysis to that of section 2.3 yields the shape derivative of J in the present physical context; proofs are postponed to section 2.6.1.

Proposition 2.3. *Assume that the transported objective function $\mathfrak{J}$, defined by (2.4.7), has continuous partial derivatives at $(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}, \widehat{p}, \widehat{T}, \widehat{\boldsymbol{u}}) = (0, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))$. Then the objective function J, considered in (2.2.6), is differentiable with respect to $\boldsymbol{\theta} \in W_0^{1,\infty}(D,\mathbb{R}^d)$ and the derivative reads*

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) \\ &\quad = \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) + \int_{\Omega_f} [\boldsymbol{w}\cdot\mathrm{div}(\boldsymbol{f}_f\otimes\boldsymbol{\theta}) - (\sigma_f(\boldsymbol{v},p) : \nabla\boldsymbol{w} + \rho\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v})\mathrm{div}(\boldsymbol{\theta})]\,\mathrm{d}x \\ &\qquad + \int_{\Omega_f} [\sigma_f(\boldsymbol{v},p) : (\nabla\boldsymbol{w}\nabla\boldsymbol{\theta}) + \sigma_f(\boldsymbol{w},q) : (\nabla\boldsymbol{v}\nabla\boldsymbol{\theta}) + \rho\boldsymbol{w}\cdot(\nabla\boldsymbol{v}\nabla\boldsymbol{\theta})\cdot\boldsymbol{v}]\mathrm{d}x \\ &\qquad - \int_{\Omega_s} \mathrm{div}(\boldsymbol{\theta})(k_s\nabla T\cdot\nabla S)\mathrm{d}x - \int_{\Omega_f} \mathrm{div}(\boldsymbol{\theta})(k_f\nabla T\cdot\nabla S + \rho c_p(\boldsymbol{v}\cdot\nabla T)S)\mathrm{d}x \\ &\quad + \int_{\Omega_s} k_s(\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T)\nabla T\cdot\nabla S\mathrm{d}x + \int_{\Omega_f} \left[k_f(\nabla\boldsymbol{\theta} + \nabla\boldsymbol{\theta}^T)\nabla T\cdot\nabla S + \rho c_p\boldsymbol{v}\cdot(\nabla\boldsymbol{\theta}^T\nabla T)S\right]\mathrm{d}x \\ &\qquad + \int_{\Omega_s} \mathrm{div}(Q_s\boldsymbol{\theta})S\mathrm{d}x + \int_{\Omega_f} \mathrm{div}(Q_f\boldsymbol{\theta})S\mathrm{d}x \\ &\quad + \int_{\Omega_s} [-\mathrm{div}(\boldsymbol{\theta})\sigma_s(\boldsymbol{u},T) : \nabla\boldsymbol{r} + \sigma_s(\boldsymbol{u},T) : (\nabla\boldsymbol{r}\nabla\boldsymbol{\theta}) + Ae(\boldsymbol{r}) : (\nabla\boldsymbol{u}\nabla\boldsymbol{\theta}) + \boldsymbol{r}\cdot\mathrm{div}(\boldsymbol{f}_s\otimes\boldsymbol{\theta})]\,\mathrm{d}x, \end{aligned} \tag{2.4.13}$$

where $\boldsymbol{r}, S, \boldsymbol{w}, q$ are the adjoint states defined by (2.4.8) to (2.4.10).

Proposition 2.4. *If in addition the state and adjoint variables $\boldsymbol{v}, T_s, T_f, \boldsymbol{u}, \boldsymbol{w}, S, \boldsymbol{r}$ (resp. p, q) have H^2 (resp H^1) regularity in their domain of definition, and if the partial derivative $\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}$ has a decomposition of the form (2.3.32), then (2.4.13) rewrites as an integral over the interface Γ depending only on the normal component $\boldsymbol{\theta}\cdot\boldsymbol{n}$ of $\boldsymbol{\theta}$:*

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) \\ &\quad = \overline{\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}}(\boldsymbol{\theta}) + \int_\Gamma (\boldsymbol{f}_f\cdot\boldsymbol{w} - \sigma_f(\boldsymbol{v},p) : \nabla\boldsymbol{w} + \boldsymbol{n}\cdot\sigma_f(\boldsymbol{w},q)\nabla\boldsymbol{v}\cdot\boldsymbol{n} + \boldsymbol{n}\cdot\sigma_f(\boldsymbol{v},p)\nabla\boldsymbol{w}\cdot\boldsymbol{n})(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s \\ &\quad + \int_\Gamma \left(k_s\nabla T_s\cdot\nabla S_s - k_f\nabla T_f\cdot\nabla S_f + Q_fS - Q_sS_s - 2k_s\frac{\partial T_s}{\partial\boldsymbol{n}}\frac{\partial S_s}{\partial\boldsymbol{n}} + 2k_f\frac{\partial T_f}{\partial\boldsymbol{n}}\frac{\partial S_f}{\partial\boldsymbol{n}}\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s \\ &\qquad + \int_\Gamma (\sigma_s(\boldsymbol{u},T_s) : \nabla\boldsymbol{r} - \boldsymbol{f}_s\cdot\boldsymbol{r} - \boldsymbol{n}\cdot Ae(\boldsymbol{r})\nabla\boldsymbol{u}\cdot\boldsymbol{n} - \boldsymbol{n}\cdot\sigma_s(\boldsymbol{u},T_s)\nabla\boldsymbol{r}\cdot\boldsymbol{n})\,(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s, \end{aligned} \tag{2.4.14}$$

where $\overline{\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}}$ denotes the part of $\partial\mathfrak{J}/\partial\boldsymbol{\theta}$ that depends only on the normal trace $\boldsymbol{\theta}\cdot\boldsymbol{n}$ (see (2.3.33)).

Remark 2.14. Formula (2.4.13) is a volume expression of the shape derivative, while formula (2.4.14) is a surface expression of the same derivative. Formula (2.4.14) can be simplified a little by using the

following identities on Γ, which arise as consequences of the boundary conditions featured in (2.2.1) to (2.2.3):

$$\boldsymbol{n} \cdot \sigma_f(\boldsymbol{w}, q)\nabla \boldsymbol{v} \cdot \boldsymbol{n} = \sigma_f(\boldsymbol{w}, q) : \nabla \boldsymbol{v}, \tag{2.4.15}$$

$$\boldsymbol{n} \cdot \sigma_f(\boldsymbol{v}, p)\nabla \boldsymbol{w} \cdot \boldsymbol{n} - \boldsymbol{n} \cdot \sigma_s(\boldsymbol{u}, T_s)\nabla \boldsymbol{r} \cdot \boldsymbol{n} = \sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{w} - \sigma_s(\boldsymbol{v}, p) : \nabla \boldsymbol{r}. \tag{2.4.16}$$

The above equations (2.4.13) and (2.4.14) generalize more classical shape derivatives expressions for each of the physics considered individually: the elastic, thermic, and fluid parts coincide with expressions stated in e.g. [32, 168] for the elasticity, [314] for pure thermoelasticity, [253] for the thermal conductivity terms, and [259, 100, 112] for the Navier Stokes equations. However some terms of formula (2.4.14) vanishing for particular instances of the objective function J (for example, $Ae(\boldsymbol{r}) \cdot \boldsymbol{n} = 0$ for objective functions written as a volume integral depending only on $\boldsymbol{u}$ and not on its gradient), may be missing in previous works.

The shape derivative formulas (2.4.13) and (2.4.14) may seem rather complex at first glance, however they can be implemented once and for all. In a practical implementation, they allow then to automate very conveniently the numerical assembly of shape derivatives, since only analytical expressions of the partial derivatives of the objective functionals are required. These are usually obtained very easily in weak form, which turns to be also convenient for the resolution of the adjoint equations (2.4.8) to (2.4.10).

Remark 2.15. In the case where the Stokes system is considered instead of the Navier-Stokes equations in (2.2.1), the above formulas hold by setting $\rho = 0$ in the terms involving the fluid state and adjoint variables $\boldsymbol{v}$ and $\boldsymbol{w}$.

2.5 Numerical test cases

This section is devoted to the presentation of several 2-d test cases which in particular allow to verify numerically the shape derivatives formulas of propositions 2.3 and 2.4. This verification is based on the assumption that their correct implementation should make objective functions decrease and constraints become gradually satisfied, in accordance with the expected behavior of our null space algorithm for constrained optimization (detailed in chapter 3).

Here, we demonstrate on various multiphysics examples how the previous ideas can be effectively implemented in order to address a wide range of topology optimization problems.

This part deviates from the published work [153] in several aspects. First, supplementary test cases are treated in sections 2.5.2 to 2.5.5. Second, the test case of sections 2.5.6 and 2.5.7 that were originally treated in [153] have been respectively revised (in order to illustrate differences between the use of volume and surface expressions of the shape derivative) or improved (constraints are treated explicitly with our null space gradient flow optimization algorithm).

2.5.1 A few details about the numerical implementation

Our numerical implementation follows algorithm 1.1 outlined in chapter 1 with the level set based mesh evolution method of [25] summarized in algorithm 1.2. We rely on the open-source `FreeFem++` environment for the resolution of Finite Element problems [183] (see [34] for its use in the context of structural optimization and [112] about its use in the context of fluid flow optimization). Since much more details shall be provided in chapter 6, we content ourselves to provide here only a brief overview of our implementation.

The Navier-Stokes system (2.4.2) is solved by using a mixed formulation, where the space $V_{\boldsymbol{v},p}$ is discretized with $\mathbb{P}_1$-bubble $\times$ $\mathbb{P}_1$ elements. The equations (2.2.2) and (2.2.3) for the temperature T and the elastic displacement $\boldsymbol{u}$ are solved with $\mathbb{P}_1$ finite elements. In all the considered examples, the Reynolds number and the velocity $\boldsymbol{v}$ of the fluid are sufficiently small so that the convergence and stability of our numerical schemes are guaranteed without the need for more complex numerical strategies, e.g. upwinding methods. Note that, when solving the fluid-structure interaction problem (2.2.3), a numerical estimate of the normal stress $\sigma_f(\boldsymbol{v}, p)\boldsymbol{n}$ is used as a boundary load in the discretization of the variational formulation (2.4.4). Naturally, a mixed formulation analogous to (2.3.42) could be implemented for the triplet $(\boldsymbol{v}, p, \boldsymbol{u})$, which would avoid computing normal derivatives at the boundary.

Both surface and volume expressions of shape derivatives are considered for the computation of a descent direction as described in chapter 1, section 1.4.1. Once the shape derivative $\mathrm{D}J(\Gamma)$ of the considered objective function J (or the constraints) is assembled, the identification problem (1.4.4) (also called extension and regularization of the interface velocity) is solved using `FreeFEM`. We rely on the inner

product (1.2.24) when using the surface expression (2.4.14) for the shape derivative, and on (1.2.25) when using the volume expression (2.4.13). In several cases, we observed without explanation that the use of the volume expression was beneficial, either for obtaining better shapes or smoother convergence curves for the objective function and constraints. A few comparisons illustrating these facts shall be provided hereafter. Note that this matter has been discussed in a number of works, see e.g. [188, 168].

The open-source library `mmg` [108] is used when it comes to the isosurface discretization and quality-oriented remeshing operations outlined in chapter 1, section 1.4.2. Let us mention that the meshes obtained are in general non symmetric even if the discretized shape is symmetric. Since this can affect slightly the quality of the optimized shape, the level set function associated to the shape of the next iteration is symmetrized before remeshing in test cases where symmetry is to be expected (see chapter 6 for the details). This allows to avoid large symmetry loss induced by the accumulation of small numerical errors. The computation of the signed distance function to a meshed domain (see section 1.3.1) is performed by using the open-source algorithm `mshdist`; see [111].

In all our examples, the considered shape and topology optimization problems feature equality or inequality constraints, for instance on the volume of one of the two phases Ω_f, Ω_s. These constraints are gradually enforced and maintained thanks to the null space gradient flow algorithm detailed in chapter 3. For now, let us content ourselves with saying that optimization trajectories are calculated by mean of a gradient flow (a dynamical system based on the gradient of the objective and constraint functions) that is able to detect and to handle equality and inequality constraints.

In the following, we treat a variety of test cases which are all subcases of the full three physics model. The first three examples are benchmark test cases of the literature featuring only one physics. The next three problems involve two physics: pure fluid-structure interaction, thermoelasticity, and heat convection. The final example involves all physics simultaneously.

2.5.2 Cantilever beam in linearized elasticity

Our first example is the very classical cantilever beam test case which has been commonly used to illustrate many topology optimization algorithms, e.g. with the Hadamard method [32, 34] or the SIMP method [43]. We consider the classical compliance minimization problem of finding a structure $\Omega_s \subset D$ enclosed in a domain $D = [0,2] \times [0,1]$ fixed at top and bottom left boundaries and subject to a traction load $\boldsymbol{g}$ on the middle of the right-hand boundary (the setting is illustrated on Figure 2.3). The objective

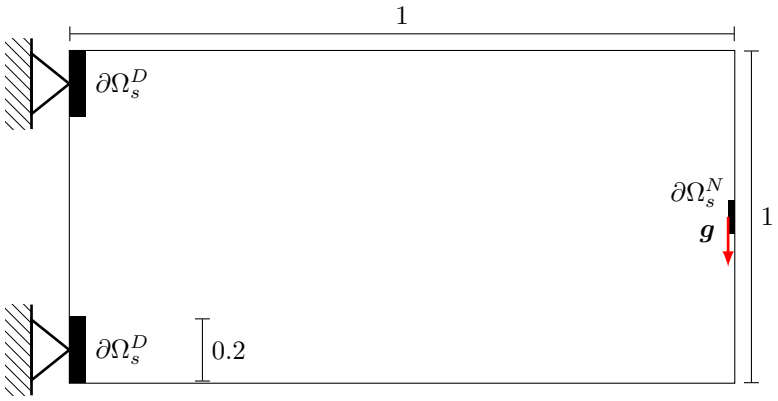


Figure 2.3: Setting of the cantilever optimization problem of section 2.5.2.

is to minimize the compliance of the structure subject to a volume constraint:

$$\begin{aligned} \min_{\Omega_s \subset D} \quad & J(\Omega_s, \boldsymbol{u}(\Omega_s)) := \int_{\Omega_s} Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_s) := \int_{\Omega_s} \mathrm{d}x = V_{target}. \end{aligned} \tag{2.5.1}$$

where V_{target} is a target volume set to 0.6 in our implementation.

In this setting, the formulas (2.3.38) and (2.3.39) needed for the shape derivative of J and the definition of the adjoint systems read:

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\Omega_s} (Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{div}(\boldsymbol{\theta}) - 2Ae(\boldsymbol{u}) : (\nabla \boldsymbol{u} \nabla \boldsymbol{\theta})) \mathrm{d}x, \tag{2.5.2}$$

$$\overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = \int_\Gamma (-Ae(\boldsymbol{u}) : e(\boldsymbol{u}) + 2\boldsymbol{n} \cdot (Ae(\boldsymbol{u}))\nabla \boldsymbol{u} \cdot \boldsymbol{n})(\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}s, \tag{2.5.3}$$

$$\frac{\partial \mathfrak{J}}{\partial \widehat{\boldsymbol{u}}}(\boldsymbol{r}') = \int_{\Omega_s} 2Ae(\boldsymbol{u}) : e(\boldsymbol{r}') \mathrm{d}x. \tag{2.5.4}$$

The shape derivative of the volume functional $\mathrm{Vol}(\Omega_s)$ reads simply

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \mathrm{Vol}(\Omega_{s,\boldsymbol{\theta}}) = \int_{\Omega_s} \mathrm{div}(\boldsymbol{\theta}) \mathrm{d}x = -\int_\Gamma \boldsymbol{\theta} \cdot \boldsymbol{n} \mathrm{d}s, \tag{2.5.5}$$

where the minus sign comes from our convention that $\boldsymbol{n}$ is pointing inside Ω_s. For this example, the values of the Lamé parameters are set to $\lambda = 12.96$ and $\mu = 5.56$.

Intermediate shapes obtained with our implementation by using the *surface* expression for the shape derivative are displayed on Figure 2.4. An instance of a mesh associated to one of the intermediate shapes where a topological change occurs is visible on Figure 2.5, which further illustrates a key feature of our numerical method: an explicitly meshed, black and white description of the shape is available during the whole optimization process. The history curves for the objective and constraint function values are displayed on Figure 2.6. These illustrate the ability of our optimization algorithm to decrease the objective function while maintaining the constraint satisfied. Observe that very good shapes are obtained within approximately a thirty iterations, and that the oscillations visible for the objective function J at subsequent iterations are the result of numerical noise around the optimum.

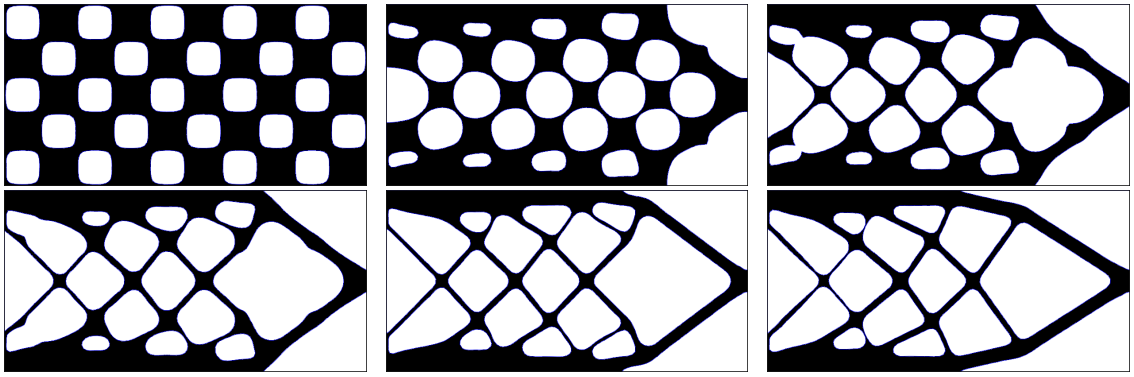


Figure 2.4: Intermediate iterations 0, 8, 15, 19, 30 and 110 for the cantilever test case of section 2.5.2.

2.5.3 Optimal shapes for pure heat conduction

We now address a different problem where only thermal effects are considered, for which the only state variable active is the temperature T. This test case has been treated by Marck and Privat [225] with a density based topology optimization method. We consider a domain $D = [0.1, 0.1]$ with two phases Ω_s and Ω_f representing two conductive materials with respective conductivity constants $k_s = 1$ and $k_f = 100$. The whole domain is heated with a uniform volume heat source $Q_s = Q_f = 10^4$. The most conductive phase Ω_f is connected to a "cold" Dirichlet boundary $\partial\Omega_T^D$ featuring $T = T_0$ (with $T_0 = 0$ in our implementation). All other parts of the boundary are adiabatic (no heat flux escaping the domain). The setting is represented on Figure 2.7.

The objective is to find the distribution of material Ω_f which minimizes the average temperature of the whole domain using a limited amount of volume V_{target} (set to $V_{target} = 0.15$ in our case):

$$\begin{aligned} \min_\Gamma \quad & J(\Gamma, T(\Gamma)) = \int_D T \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_f) \leq V_{target}. \end{aligned} \tag{2.5.6}$$

For this problem, the partial derivatives of the objective function J read simply

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = 0 \tag{2.5.7}$$

$$\frac{\partial \mathfrak{J}}{\partial \widehat{T}}(S') = \int_D S' \mathrm{d}x. \tag{2.5.8}$$

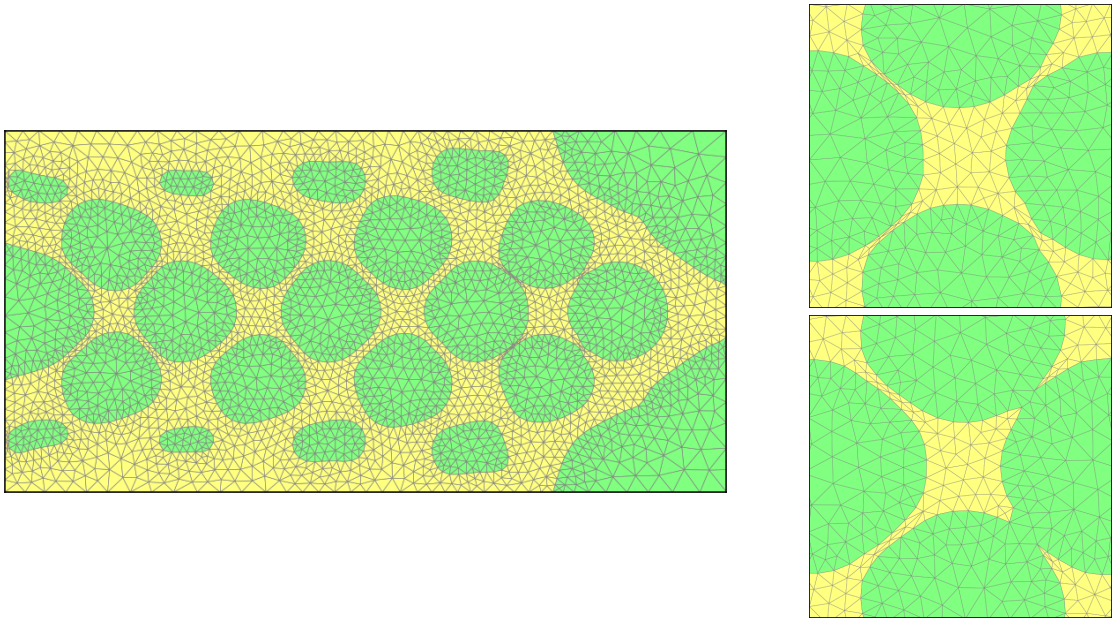
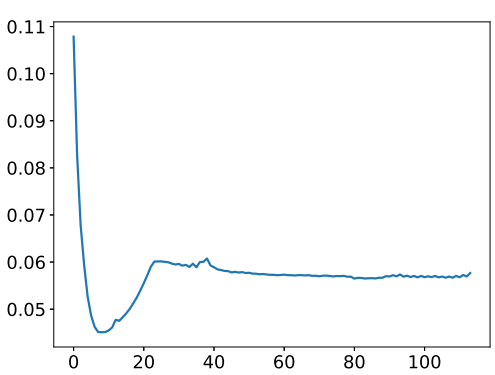
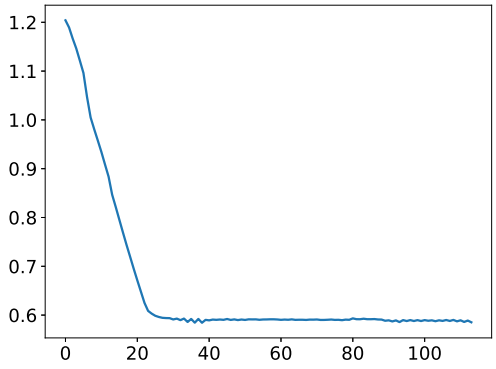


Figure 2.5: Mesh of an intermediate optimization iteration (on the *left*) and zoom on a region featuring a change of topology (on the *right*) for the cantilever test case of section 2.5.2.



(a) Objective function J. Final value: J =0.058.



(b) Volume constraint $\mathrm{Vol}(\Omega_s)$. Final value: $\mathrm{Vol}(\Omega_s)$ =0.59.

Figure 2.6: Convergence history for the cantilever test case of section 2.5.2.

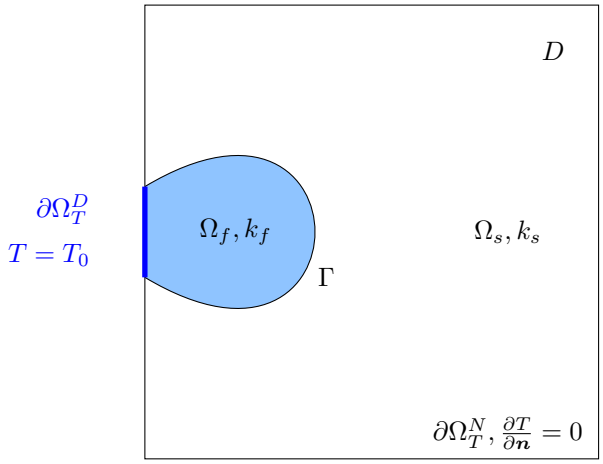


Figure 2.7: Setting of the optimal heat conduction problem of section 2.5.3.

Initial and final shapes obtained by using the surface expression (2.4.14) of the shape derivative are displayed on Figure 2.8. A few intermediate iterations are visible on Figure 2.10. It is interesting to compare them to those obtained with the *volume* expression (2.4.13) of the shape derivative. Both final shapes are similar, although the optimization path is a little bit different with smaller details appearing more quickly with the volume expression of the shape derivative, and a final shape slightly better for a similar number of iterations. The convergence curves for the objective and constraint functionals are plotted on Figure 2.13. It is interesting to observe that although the volume constraint in (2.5.6) is an *inequality* constraint. The process converges without oscillations characterizing many other optimization algorithms using active set strategies [297]. The details of the method allowing to obtain this smoothness of convergence are explained in chapter 3, section 3.4.1.

A mesh of the final shape is shown on Figure 2.12. Let us mention that local parameters were prescribed in the remeshing tool `mmg` in order to obtain a (small) constant mesh edge size for the discretization of the optimized boundary Γ.

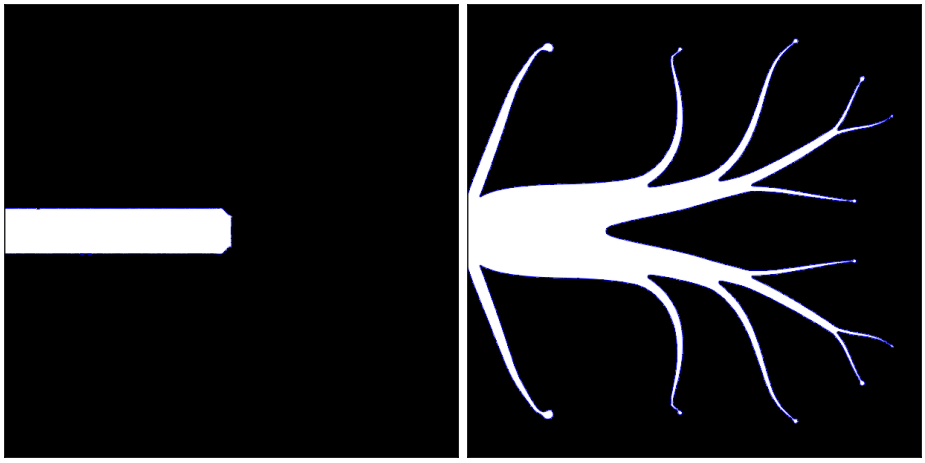


Figure 2.8: Initial and final configurations for the pure heat conduction test case of section 2.5.3. The fluid material is represented in white.

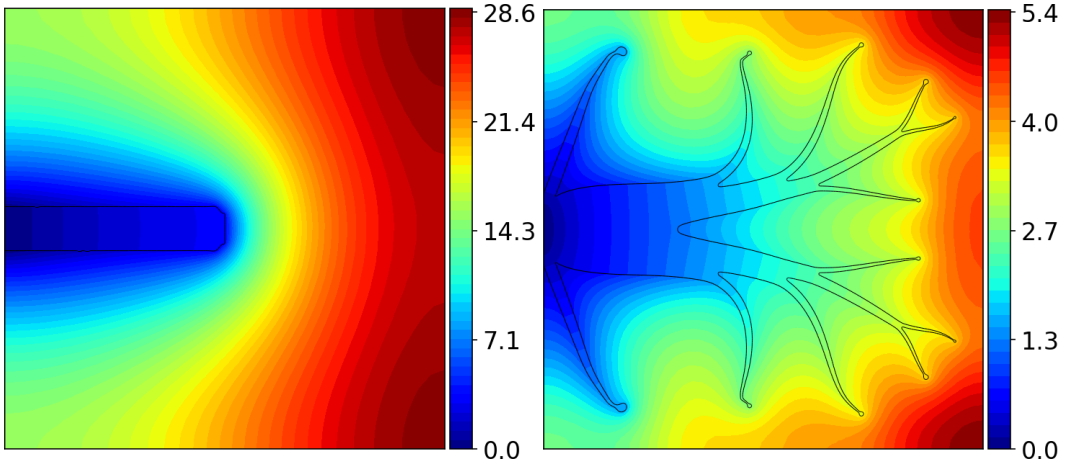


Figure 2.9: Initial and final temperature fields for the pure heat conduction test case of section 2.5.3.

2.5.4 Optimal drag profiles for Stokes and Navier-Stokes flows

The final example featuring only one physics is concerned with the classical problem of finding the shape of an obstacle minimizing the drag force induced by a surrounding flow [259, 172]. Here we follow the setting proposed in [113]. The global domain is a rectangle $D = [0, 1] \times [0, 1]$. A flow $\boldsymbol{v}$ is entering on the left side with a velocity $\boldsymbol{v} = \boldsymbol{e}_x$ where $\boldsymbol{e}_x$ is the unit horizontal direction. The viscosity and density parameters of the flow are given by $\nu = 5e - 3$ and $\rho = 1$, which corresponds to a Reynolds number $\mathrm{Re} = 200$. In order to limit the effects induced by the bottom and top walls, a slip boundary condition

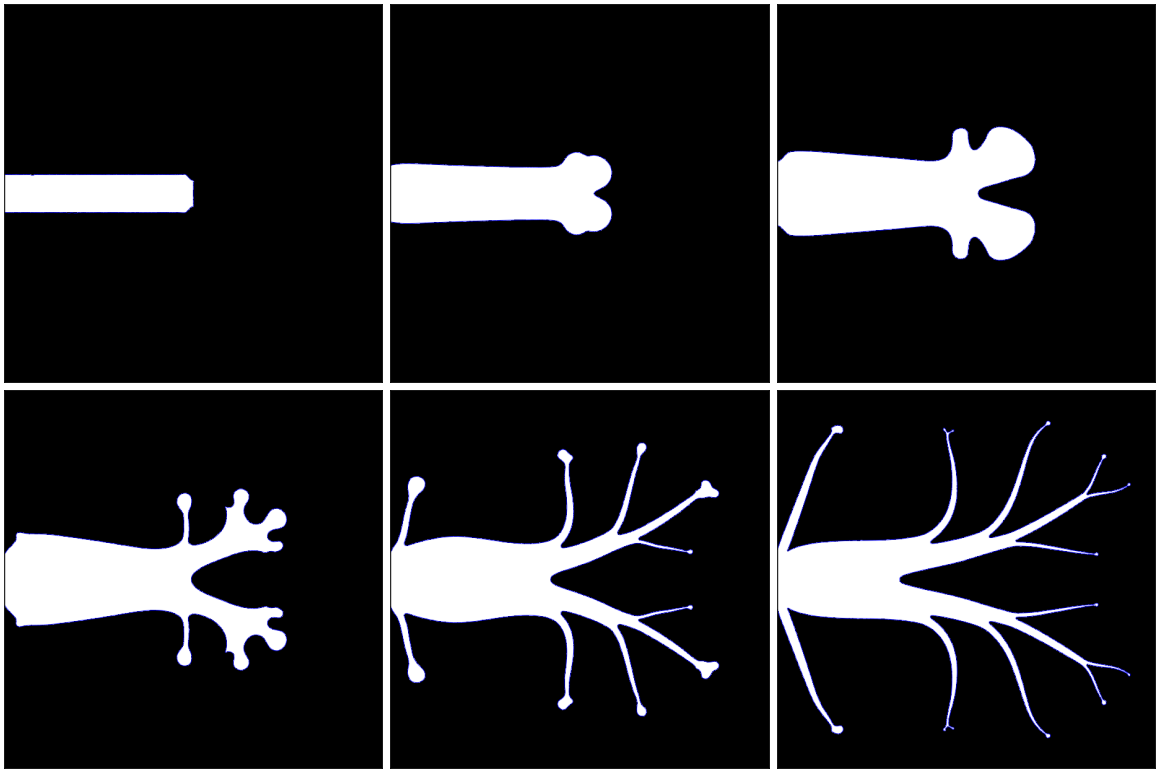


Figure 2.10: Intermediate iterations 0, 10, 25, 40, 90 and 200 for the pure heat conduction test case of section 2.5.3 using the *surface* expression (2.4.14) of the shape derivative. Final objective function value: $J = 2.8$.

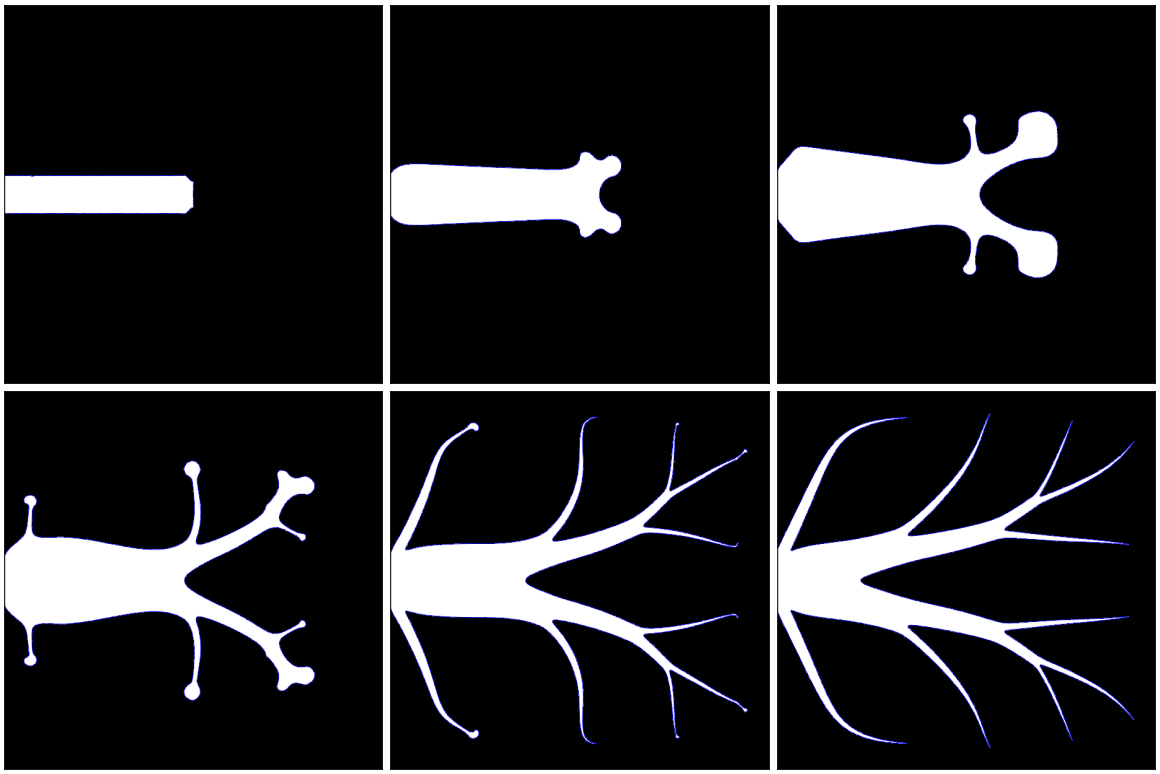


Figure 2.11: Intermediate iterations 0, 10, 25, 40, 90 and 190 for the pure heat conduction test case of section 2.5.3 using the *volume* expression of the shape derivative. Final objective function value: $J = 2.6$.

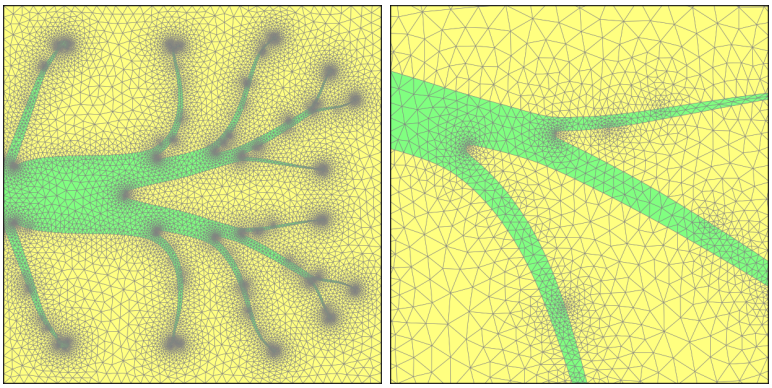
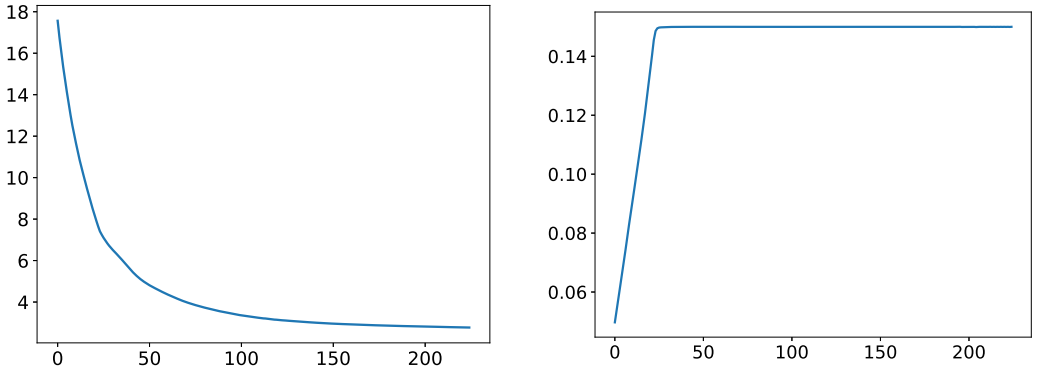


Figure 2.12: Final mesh and zoom on a part of it for the heat conduction test case of section 2.5.3.



(a) Objective function J. Final value: $J = 2.8$.

(b) Constraint $\mathrm{Vol}(\Omega_f)$. Final value: $\mathrm{Vol}(\Omega_f) = 0.15$.

Figure 2.13: Convergence history for the heat conduction test case of section 2.5.3.

$\boldsymbol{v} \cdot \boldsymbol{n} = 0$ is assumed on these boundaries[1]. The flow leaves the domain on the right boundary with a zero normal stress boundary condition: $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = 0$. The setting is illustrated on Figure 2.14).

The objective is to minimize the energy dissipation of the fluid generated by the solid structure, subject to a constraint on its volume, as well as on its center of mass fixed to the center $\boldsymbol{x}_0$ of the domain:

$$\min_{\Gamma} \quad J(\Gamma, \boldsymbol{v}(\Gamma)) = \int_{\Omega_f} 2\nu e(\boldsymbol{v}) : e(\boldsymbol{v}) \mathrm{d}x$$
$$s.t. \quad \begin{cases} \mathrm{Vol}(\Omega_f) = V_{target} \\ \boldsymbol{X}(\Omega_s) := \dfrac{1}{|\Omega_s|} \displaystyle\int_{\Omega_s} \boldsymbol{x} \mathrm{d}x = \boldsymbol{x}_0. \end{cases} \tag{2.5.9}$$

We have set $V_{target} = 0.03$ and $\boldsymbol{x}_0 = (0.5, 0.5)$ in our implementation. The partial derivatives of the objective function allowing to obtain its shape derivative are given by

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\Omega_f} [-2\nu e(\boldsymbol{v}) : \nabla \boldsymbol{v} \nabla \boldsymbol{\theta} + 2\nu e(\boldsymbol{v}) : e(\boldsymbol{v}) \mathrm{div}(\boldsymbol{\theta})] \mathrm{d}x, \tag{2.5.10}$$

$$\overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = -\int_{\Gamma} 2\nu e(\boldsymbol{v}) : e(\boldsymbol{v}) \, (\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}s, \tag{2.5.11}$$

$$\frac{\partial \mathfrak{J}}{\partial (\boldsymbol{v}, p)}(\boldsymbol{w}', q') = \int_{\Omega_f} 4\nu e(\boldsymbol{v}) : e(\boldsymbol{w}') \mathrm{d}x. \tag{2.5.12}$$

The reader may verify that these equations plugged in the formulas (2.4.13) yield shape derivative formulas identical to those derived for the same problem e.g. in [112]. The derivative of the center of mass $\boldsymbol{X}(\Omega_s)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \boldsymbol{X}(\Omega_{s,\boldsymbol{\theta}})(\boldsymbol{\theta}) = \frac{1}{|\Omega_s|} \int_{\Omega_s} (\boldsymbol{X}(\Omega_s) - \boldsymbol{x}) \mathrm{div}(\boldsymbol{\theta}) \mathrm{d}x = \frac{1}{|\Omega_s|} \int_{\Gamma} (\boldsymbol{X}(\Omega_s) - \boldsymbol{s}) \boldsymbol{\theta} \cdot \boldsymbol{n} \mathrm{d}s.$$

[1] This boundary condition is not exactly the one considered in (2.2.1) but the computation of shape derivatives remains identical up to the account of the same boundary condition $\boldsymbol{w} \cdot \boldsymbol{n} = 0$ for the adjoint variable on this boundary

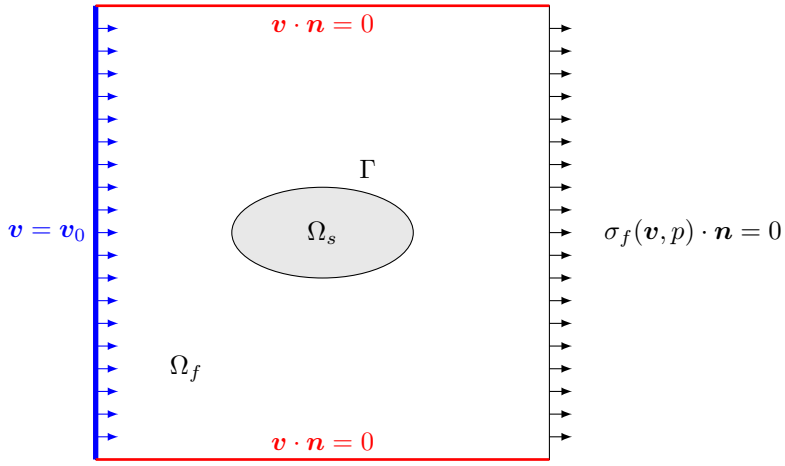
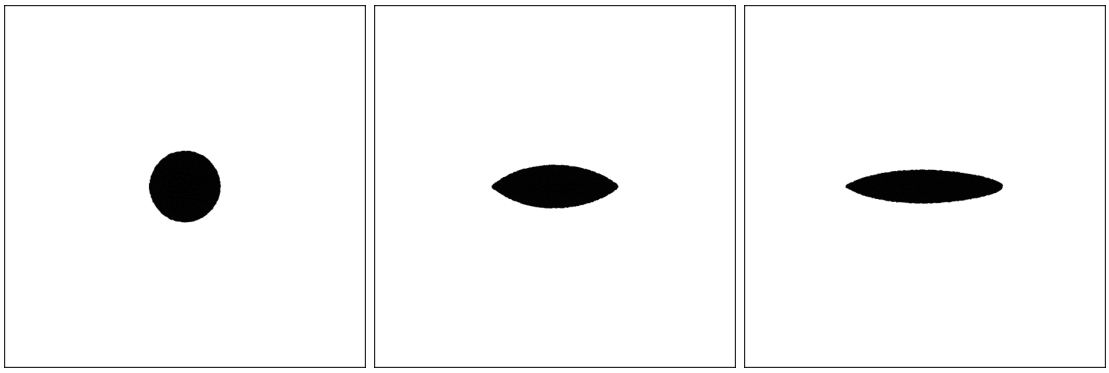


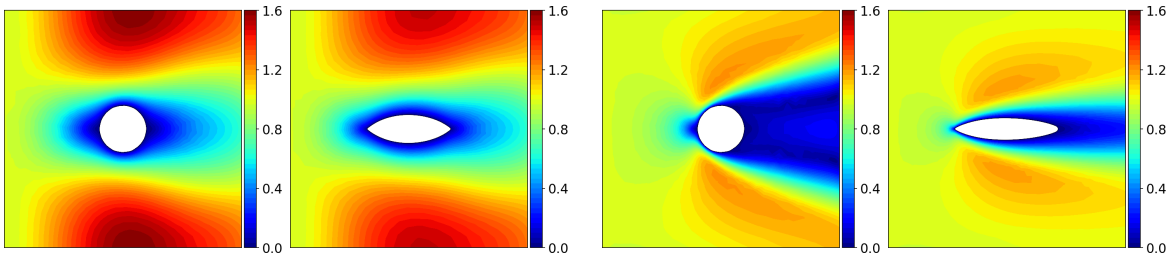
Figure 2.14: Setting of the minimum drag flow problem of section 2.5.4.

Final optimized shapes are shown on Figure 2.15 below in both situations where the fluid is characterized by the Stokes or Navier-Stokes equations. The celebrated rugby ball shape can be recognized for the Stokes problem [259]. Velocity fields are plotted on Figure 2.16. Convergence curves for objective and constraint functions are visible on Figure 2.19. We note that our constrained optimization algorithm was able to compute these shapes without the oscillations of the constraints obtained in [110] with the Augmented Lagrangian Method. A few intermediate iterations and the mesh of the final shape for the Navier Stokes case are respectively shown on Figs. 2.17 and 2.18.



(a) Initial (b) Stokes (c) Navier-Stokes

Figure 2.15: Initial and final configurations for the minimum drag test case of section 2.5.4.

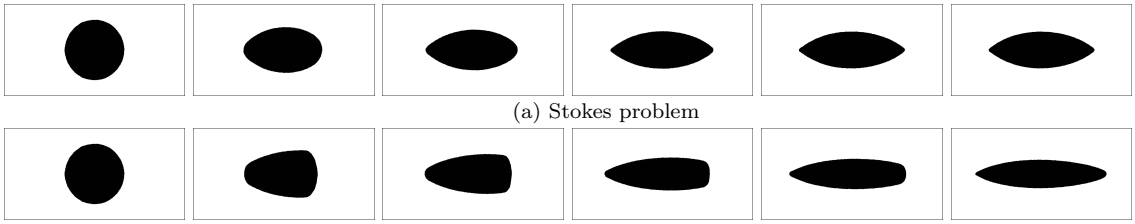


(a) Initial (Stokes) (b) Final (Stokes) (c) Initial (Navier-Stokes) (d) Navier-Stokes

Figure 2.16: Initial and final norm fields for the fluid velocity.

2.5.5 Minimum compliance problem in thermoelasticity

In this paragraph, we reproduce the test case of Xia and Wang [314] for compliance minimization in thermoelasticity. A structure $\Omega_s \subset D$ to be found in the rectangular domain $D = [0, 2] \times [0, 1]$ is



(a) Stokes problem

(b) Navier Stokes problem

Figure 2.17: Intermediate iterations 0, 5, 10, 20, 30 and 70 for the minimum drag test case of section 2.5.4.

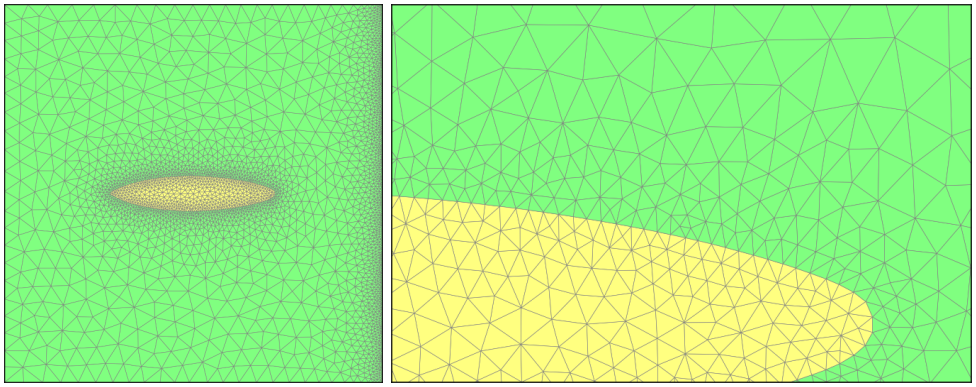
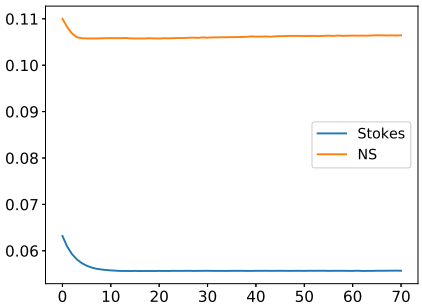
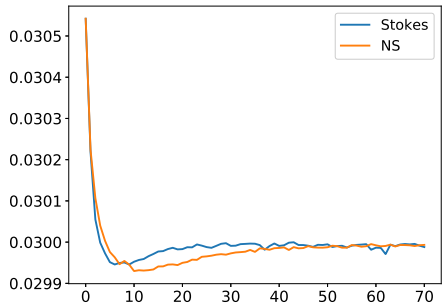


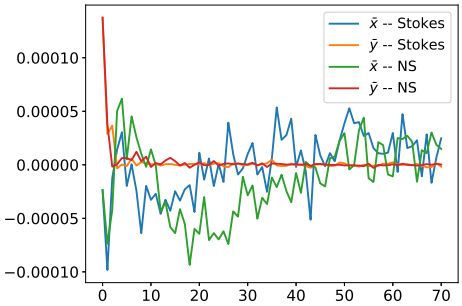
Figure 2.18: Mesh of the *Navier-Stokes* minimum drag profile and zoom on a part of the mesh. The mesh was refined near the boundary of the obstacle and the outlet boundary to capture the flow variability.



(a) Objective function J. Final values: $J = 0.056$ (Stokes) and $J = 0.11$ (Navier-Stokes)



(b) Volume constraint $\mathrm{Vol}(\Omega_s)$. Final values (both cases): $\mathrm{Vol}(\Omega_s) = 0.03$.



(c) Center of mass constraint $\boldsymbol{X}(\Omega_s) = (\overline{x}, \overline{y})$. Final values (both cases): $\overline{x} = 2.5e-05$ and $\overline{y} = -1.8e-06$.

Figure 2.19: Convergence history for the test case of section 2.5.3.

clamped on the left and right sides of the domain, and subjected to a traction load on the middle of the bottom boundary. It is made of an elastic material characterized by Lamé parameters $\lambda = 11510$, $\mu = 7673$, thermoelastic coefficient $\alpha = 0.77$ and reference temperature $T_{ref} = 0$.

A constant temperature field $T = T_{ref} + \Delta T$ is applied on the whole structure, which induces thermal expansion. The setting is reproduced on Figure 2.20. We aim to solve the minimum compliance minimization problem subject to a volume inequality constraint:

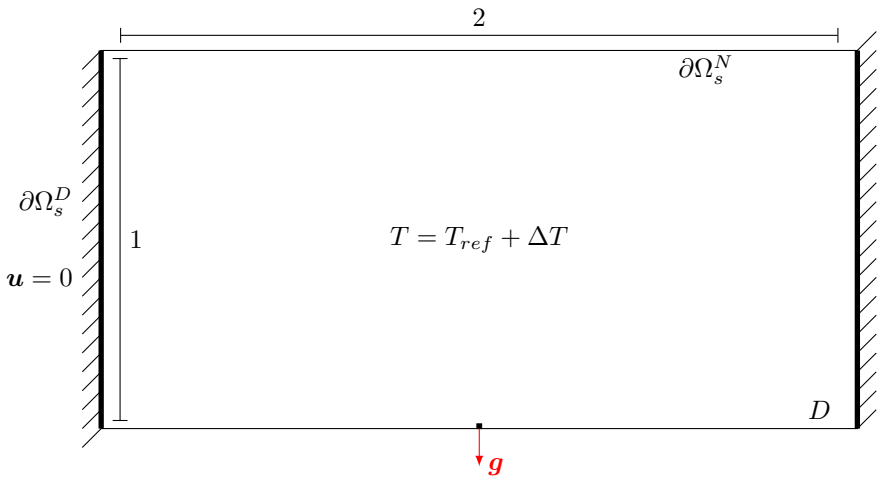


Figure 2.20: Setting for the thermoelastic compliance minimization problem of section 2.5.5, issued from [314].

$$\begin{aligned} \min_{\Omega_s \subset D} \quad & J(\Omega_s, \boldsymbol{u}(\Omega_s)) := \int_{\Omega_s} Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_s) := \int_{\Omega_s} \mathrm{d}x \leq V_{target}. \end{aligned}$$

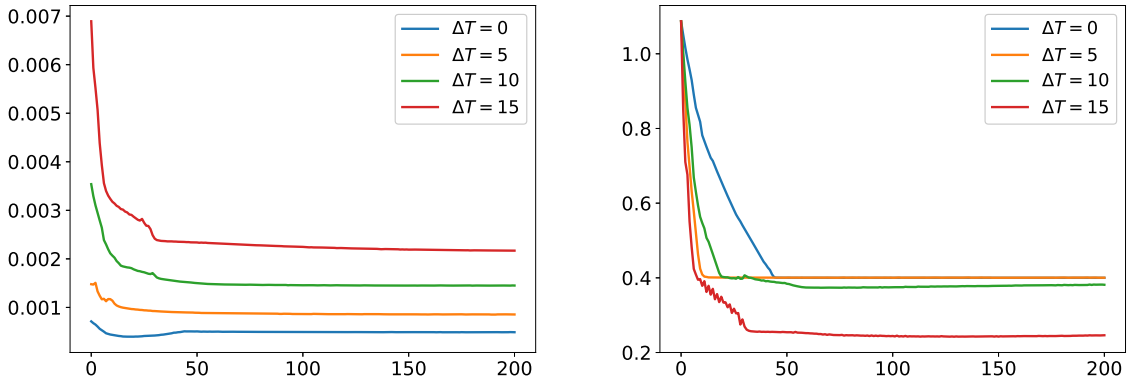
Strictly speaking, the problem still involves only one physics (the heat conduction problem (2.2.2) does not need to be solved since the temperature field is prescribed). We could have made it multiphysics by applying e.g. heat sources Q_f and Q_s, however we keep the setting of Xia and Wang [314] for the sake of comparison. The force applied in [314] has a value $F = 1$, which is set in our implementation by prescribing a traction force density $\boldsymbol{g} = -1/\varepsilon$ on a small portion of size $\varepsilon = 0.0125$ on the boundary. The upper bound volume is set to $V_{target} = 0.4$.

Following [314], we solve the optimization problem for four values of ΔT ($\Delta T = 0$, 5, 10 or 20). On Figure 2.21, we plot the convergence history curves for the objective function and the volume constraint for each test case. Very interestingly, and as observed in [314], we retrieve the fact that the volume constraint $\mathrm{Vol}(\Omega_s) \leq V_{target}$ is saturated for the first two test cases $\Delta T = 0$ and $\Delta T = 5$, and is not saturated otherwise. Notice as well in our case the smooth convergence of these curves obtained with our null space gradient flow, in contrast to those of the original paper which relied on a variant of the Augmented Lagrangian Method. In particular, our algorithm is able to quickly detect if the volume constraint needs to remain saturated or not.

Optimized shapes including intermediate iterations are shown on Figure 2.22, and the corresponding final compliance and volume values are shown in Table 2.3. Note that our numerical values do not coincide exactly with those of Wang because their original physical parameters were multiplied by nondimensionalization constants more compatible with our setting. However we clearly retrieve very similar optimized shapes.

2.5.6 A steady-state fluid-structure interaction problem

We now address a true multiphysics problem involving two physics. A fluid is flowing through a pipe, where it is pushing on a vertical beam of solid clamped at its bottom; see Figure 2.23 for a schematic of the test case. Thermal effects are neglected (namely, (2.2.2) is ignored), so that (2.2.3) reduces to a standard linear elasticity system with the pressure load induced by the fluid. The objective is to minimize



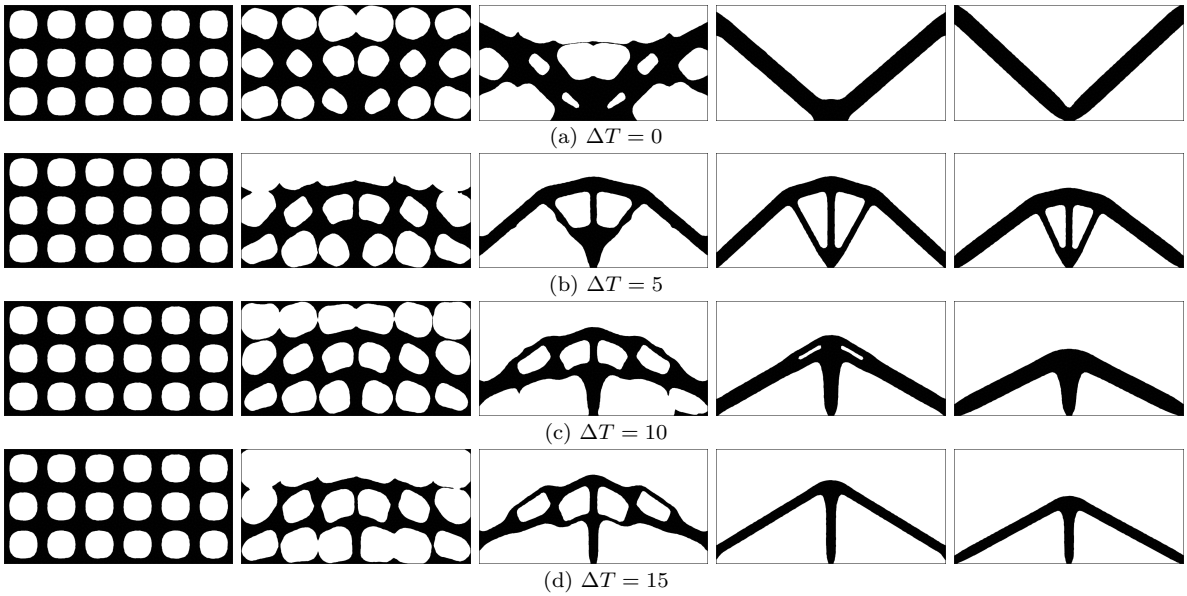
(a) Objective function J.

(b) Volume constraint $\text{Vol}(\Omega_s) \leq 0.4$.

Figure 2.21: Convergence history for the thermoelasticity test case of section 2.5.5.

ΔT	Final J	Final $\text{Vol}(\Omega_s)$
0	0.00049	0.4
5	0.00085	0.4
10	0.0015	0.38
15	0.0022	0.25

Table 2.3: Optimized compliance and volume values for the thermoelasticity test case of section 2.5.5. The results are analogous to those of [311].



(a) $\Delta T = 0$

(b) $\Delta T = 5$

(c) $\Delta T = 10$

(d) $\Delta T = 15$

Figure 2.22: Intermediate iterations 0, 4, 13, 50 and 200 for each case of the thermoelasticity problem of section 2.5.5.

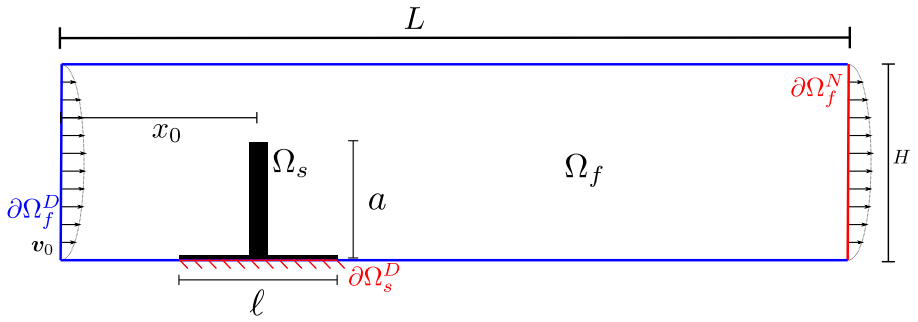


Figure 2.23: Physical setting of the fluid-structure optimization problem of section 2.5.6. During optimization, the black domain cannot be reduced but only enlarged by adding reinforcements.

L	H	ℓ	a	ν	ρ	λ	μ
2	0.5	0.4	0.3	0.005	1	0.00529	0.0476

Table 2.4: Numerical values of the physical parameters for the fluid-structure problem of section 2.5.6

the compliance of the solid phase Ω_s subject to a volume constraint, that is:

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{u}(\Gamma)) = \int_{\Omega_s} Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_s) = V_{target}. \end{aligned} \tag{2.5.13}$$

This example was previously considered by Yoon [319] with a different density-based (SIMP) method. In our case, we set $V_{target} = 0.025|D|$ where $|D| = 1$ is the volume of the total domain.

The numerical values of the considered physical parameters are given in Table 2.4. The velocity profile $\boldsymbol{v}_0$ imposed at the entrance of the pipe is parabolic, with maximum amplitude $v_{\max} = 1$, and value 0 at the upper and lower walls. The Reynolds number is equal to $\mathrm{Re} = \frac{\rho h v_{\max}}{\nu} = 60$. The elastic displacement is set to 0 on a horizontal segment of length ℓ supporting the beam.

In this part, we slightly ellaborate from our study previously published in [153] in that we highlight some numerical differences between the use of the surface (2.4.14) and volume expression (2.4.13) when computing the shape derivative. The optimization problem (2.5.13) is solved for both cases.

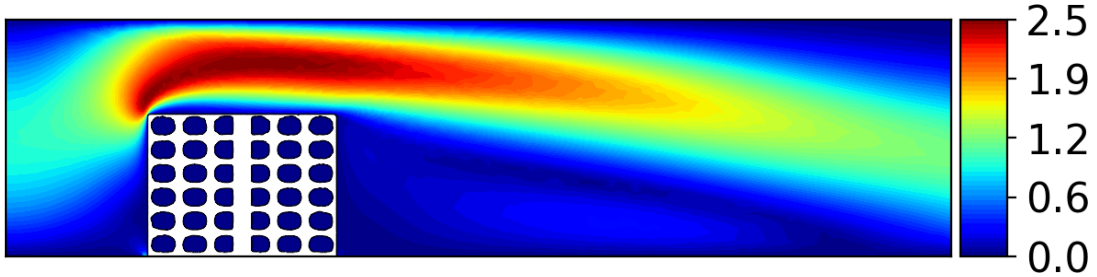
The initial and optimized shapes featuring the norm of the fluid velocity fields $\boldsymbol{v}$ are displayed on Figure 2.24. The corresponding elastic deformations induced by the fluid on the solid structure (obtained by moving mesh nodes along $\boldsymbol{u}$ for visualization purposes) are shown on Figure 2.25.Figure 2.26 shows parts of the initial and final meshes for the case using the surface expression of the shape derivative. Again, we point out that we used a feature of the library `mmg` to selectively enforce a fine mesh resolution near the interface Γ, while allowing larger triangles far from this boundary in order to save computational effort. Note that our final design are different from that in [319] because our Reynolds number is much larger and the location of the beam is different.

The evolution of the objective function and of the volume fraction are reported on Figure 2.28. Note that in the first part of the optimization, J increases sometimes substantially (especially in the case relying on the volume expression of the shape derivative) due to the fact that the volume constraint is not yet satisfied, or due to sudden discontinuities at topological changes.

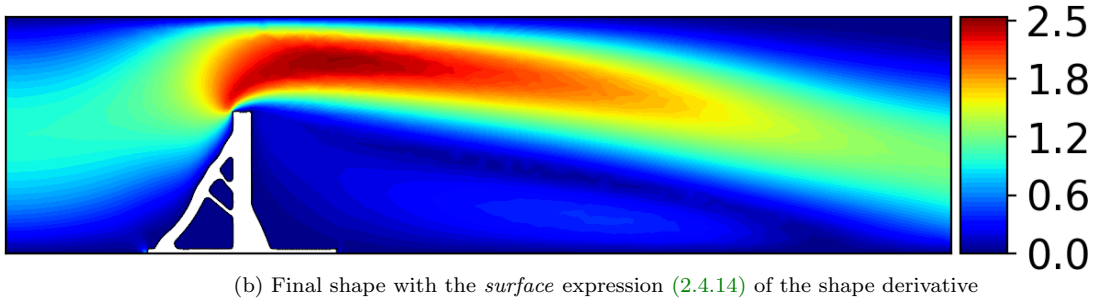
Several intermediate shapes are displayed on Figure 2.29 when using the surface expression of the shape derivative, and on Figure 2.30 when relying on the volume expression. The optimization path is surprisingly very different when using the volume expression; very thin parts of the structure exist for a longer time. Furthermore, although the final designs are somewhat similar, the one obtained with the volume expression has a slightly better performance for a very similar volume constraint. Note in particular the small solid bump near the top of the vertical beam which appears at the end of the optimization with the volume expression of the shape derivative. In our published work [153], we checked that this bump leads indeed to a better design; however it was obtained with the surface expression of the shape derivative at the cost of much more resolved meshes: the first design had a prescribed minimum edge length size $h_{\min} = 0.001$ with 28,000 vertices in [153] versus $h_{\min} = 0.003$ for approx. 11,000

vertices in the present work. The meshes of the optimized shape published in [153] are reproduced on Figure 2.27. Interestingly, in the present lower resolution case, only the volume expression of the shape derivative seems able to retrieve this bump.

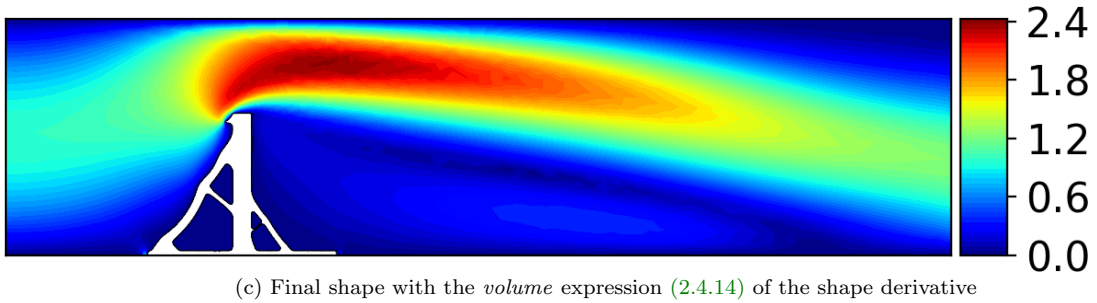
The optimization ran in approximately half an hour on a laptop equipped with Intel(R) Core(TM) i7-4702MQ @ 2.20 GHz. Note that for this example, no optimization of the implementation was performed for reducing the total computational time, e.g. with the use of preconditioners or parallelism when solving finite element problems that represent most of the computational effort for these 2-d examples. Such ingredients become unavoidable for 3-d examples and are discussed in chapter 6.



(a) Initial shape

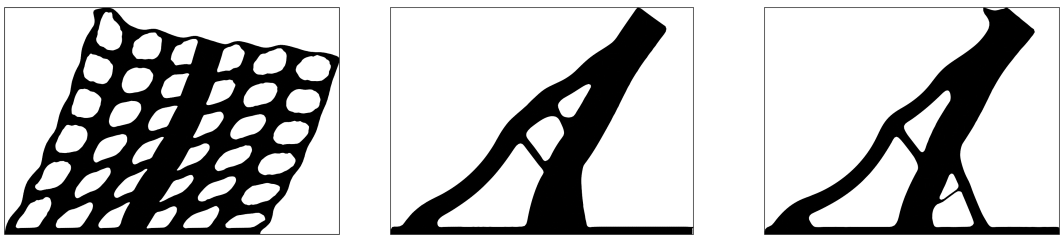


(b) Final shape with the *surface* expression (2.4.14) of the shape derivative



(c) Final shape with the *volume* expression (2.4.14) of the shape derivative

Figure 2.24: Initial and final shapes and velocity norm fields for the fluid-structure interaction test case of section 2.5.6. The optimized solid structure is depicted in white.



(a) Initial (b) Final (*surface* expression) (c) Final (*volume* expression)

Figure 2.25: Computed elastic deformations of the initial and optimized solid structures subjected to the pressure induced by an inlet flow.

2.5.7 Convective heat transfer

Our second example involving two different physics is concerned with a coupling of the flow and heat equations (2.2.1) and (2.2.2), i.e. the elastic behavior of the complement Ω_s of the optimized fluid phase

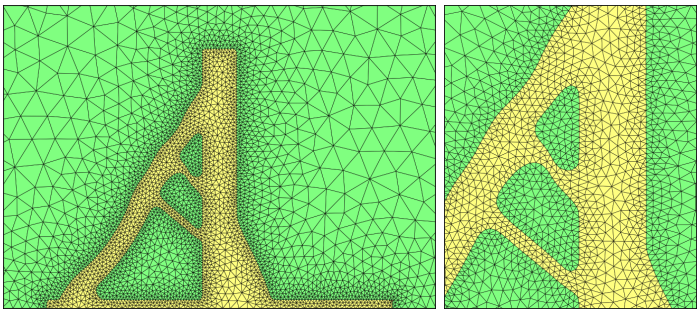


Figure 2.26: Zooms on the meshed solid structure of the final design obtained with the surface expression of the shape derivative. The minimum mesh size was set to $h_{\min} = 0.003$.

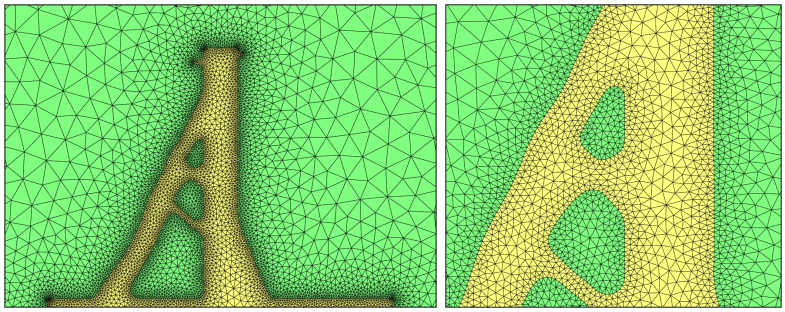
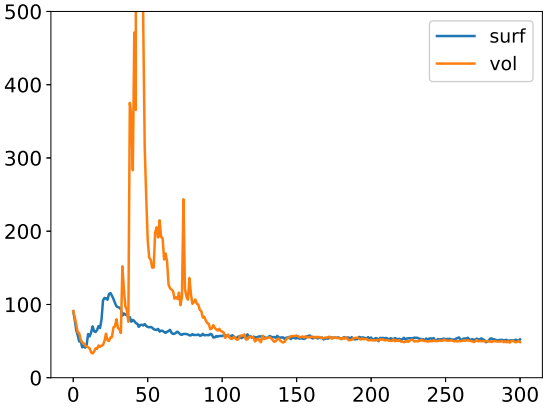
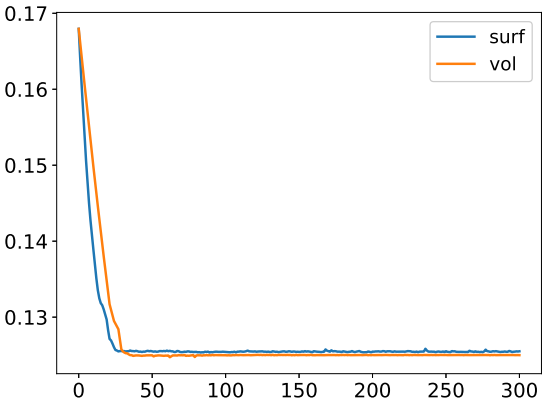


Figure 2.27: Zooms on the meshed solid structure of the final design from our published work [153] with the surface expression of the shape derivative. The minimum mesh size was set to $h_{\min} = 0.001$. The optimal shape is slightly different with the "bump" at the top left of the structure obtained with the volume expression.



(a) Objective function J. Final values: $J = 52$ (surface expression) and $J = 49$ (volume expression).



(b) Volume fraction constraint $\eta_f(\Omega_s) = \text{Vol}(\Omega_s)/\text{Vol}(D)$. Final value: $\eta_f(\Omega_s) = 0.125$ in both cases, up to an error lower than 5.10^{-4}.

Figure 2.28: Convergence history for the fluid-structure test case of section 2.5.6. The mentions "surf" and "vol" refer respectively to the use of the surface and volume expressions of the shape derivatives.

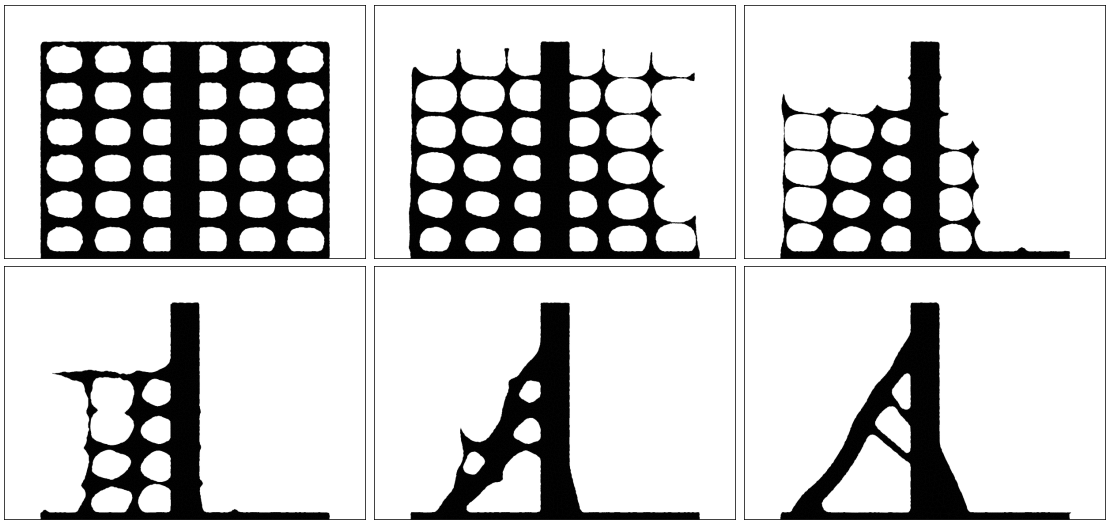


Figure 2.29: Intermediate iterations 0, 6, 15, 26, 56 and 300 for the fluid-structure test case of section 2.5.6 using the *surface* expression (2.4.14) of the shape derivative.

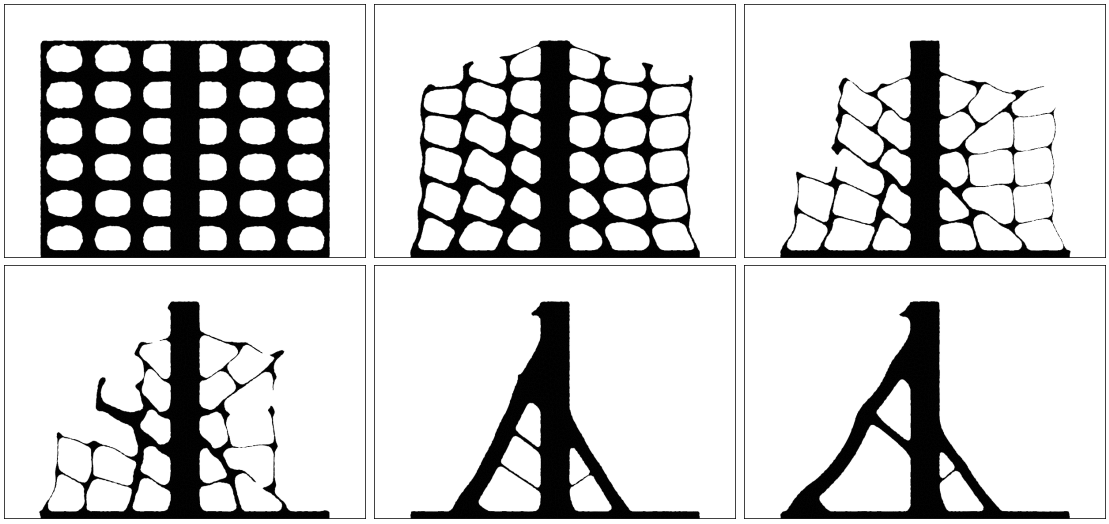


Figure 2.30: Intermediate iterations 0, 15, 40, 56, 150 and 300 for the fluid-structure test case of section 2.5.6 using the *volume* expression (2.4.14) of the shape derivative.

L	ρ	c_p	ν	$v_{\max}$	k_s	k_f	ω	T_{in}	T_{low}	T_{up} (Ex. 1)	T_{up} (Ex. 2)
0.1	10	100	0.005	1	10	1	0.4	0	10	-5	10

Table 2.5: Numerical values of the physical parameters in the convective heat transfer problem of section 2.5.7.

Ω_f is not taken into account (equation (2.2.3) is ignored). This test case features a cavity where a fluid is entering with a parabolic profile (with maximal velocity $v_{\max} = 1$) and an inlet temperature T_{in}. The setting is similar to that in [225] (where a density based method is used), although we rely on different parameters values; namely, a higher Reynold number (Re=$\frac{\rho L v_{max}}{5\nu} = 40$ while Re=3 in [225]) and different prescribed temperatures T_{up} and T_{low} for the lower and upper walls respectively. The other regions of the boundary of the cavity are insulated from the outside, i.e. zero normal fluxes boundary conditions hold for the temperature. The setting is represented on Figure 2.31 and numerical values of the parameters involved are reported in Table 2.5.

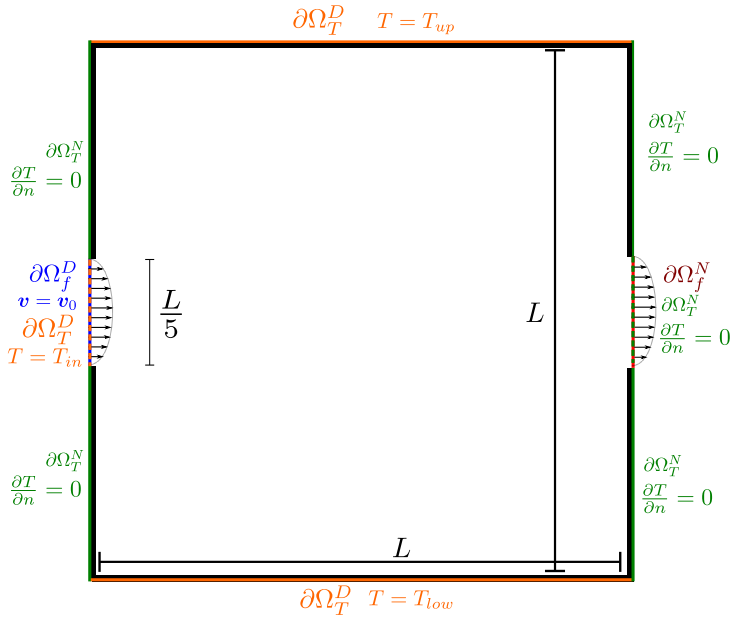


Figure 2.31: Setting of the convective heat transfer test case of section 2.5.7. The black layers at the walls stand for solid, non optimizable boundaries.

Our aim is maximize the heat transferred by the fluid subject to an upper bound on the output pressure drop and a volume constraint:

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{v}(\Gamma), T(\Gamma)) := -\int_{\Omega_f} \rho c_p \boldsymbol{v} \cdot \nabla T \mathrm{d}x \\ s.t. \quad & \begin{cases} \texttt{DP}(p(\Gamma)) := \int_{\partial\Omega_f^D} p \mathrm{d}s - \int_{\partial\Omega_f^N} p \mathrm{d}s \leq \texttt{DP}_{static} \\ \qquad\qquad \mathrm{Vol}(\Omega_f) = V_{target}. \end{cases} \end{aligned} \tag{2.5.14}$$

Note that the objective function J rewrites indeed as the opposite of heat transferred from the inlet $\partial\Omega_f^D$ to the outlet $\partial\Omega_f^N$ upon integration by parts:

$$\int_{\Omega_f} \rho c_p \boldsymbol{v} \cdot \nabla T \mathrm{d}x = \int_{\partial\Omega_f} \rho c_p T (\boldsymbol{v} \cdot \boldsymbol{n}) \mathrm{d}s = \int_{\partial\Omega_f^N} \rho c_p T (\boldsymbol{v} \cdot \boldsymbol{n}) \mathrm{d}s + \int_{\partial\Omega_f^D} \rho c_p T_0 (\boldsymbol{v}_0 \cdot \boldsymbol{n}) \mathrm{d}s$$

where the second term is a constant depending on the inlet data. The upper bound constraint on the static pressure drop is set to $\texttt{DP}_{static} = 11.4$ and the volume target to $V_{target} = 0.2|D|$.

Remark 2.16. This test case improves the one treated in our published work [153], sec. 6.3.,in that we now handle a fully constrained problem (2.5.14) instead of a penalized version of it. The viscous energy

dissipation (as considered in the drag minimization test case of section 2.5.4 and in the test case of [153]) has also been replaced by the static pressure drop DP, which has a more straightforward interpretation in industrial applications and also considered in topology optimization works, e.g. [255].

Remark 2.17. The constraint function DP does not *a priori* make sense because the pressure p belongs a priori to $L^2(\Omega_f)/\mathbb{R}$ and has no well defined trace on the boundary, unless possibly under higher regularity assumptions. However, it is believed interesting to observe that our shape derivative formulas seem to work despite the assumptions violation.

For this example, the formulas needed for the calculation of the shape sensitivity of the objective function are:

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\Omega_f} [-\rho c_p \boldsymbol{v}\cdot\nabla T)\mathrm{div}(\boldsymbol{\theta}) + \rho c_p \boldsymbol{v}\cdot\nabla\boldsymbol{\theta}^T\nabla T]\mathrm{d}x, \quad \overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = 0 \tag{2.5.15}$$

$$\frac{\partial \mathfrak{J}}{\partial(\widehat{\boldsymbol{v}},\widehat{p})}(\boldsymbol{w}',q') = -\int_{\Omega_f} \rho c_p \boldsymbol{w}'\cdot\nabla T\mathrm{d}x, \quad \frac{\partial \mathfrak{J}}{\partial \widehat{T}}(S') = -\int_{\Omega_f} \rho c_p \boldsymbol{v}\cdot\nabla S'\mathrm{d}x. \tag{2.5.16}$$

Furthermore, the shape derivative of the pressure drop constraint DP (remark 2.17) can be calculated (at least formally) with the following partial derivatives:

$$\frac{\partial \mathtt{DP}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = 0, \quad \overline{\frac{\partial \mathtt{DP}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = 0, \tag{2.5.17}$$

$$\frac{\partial \mathtt{DP}}{\partial(\widehat{\boldsymbol{v}},\widehat{p})}(\boldsymbol{w}',q') = \int_{\partial\Omega_f^D} q'\mathrm{d}s - \int_{\partial\Omega_f^N} q'\mathrm{d}s. \tag{2.5.18}$$

We consider two possible configurations for the applied temperature T_{up} at the upper wall of the system. The first example corresponds to the configuration considered in [224]; the upper and lower wall temperatures are equal and higher than the inlet temperature: $T_{\mathrm{in}} < T_{\mathrm{up}} = T_{\mathrm{low}}$. In a second example (Ex.2), we consider the case where the upper wall temperature is lower than the inlet temperature: $T_{\mathrm{up}} < T_{\mathrm{in}} < T_{\mathrm{low}}$.

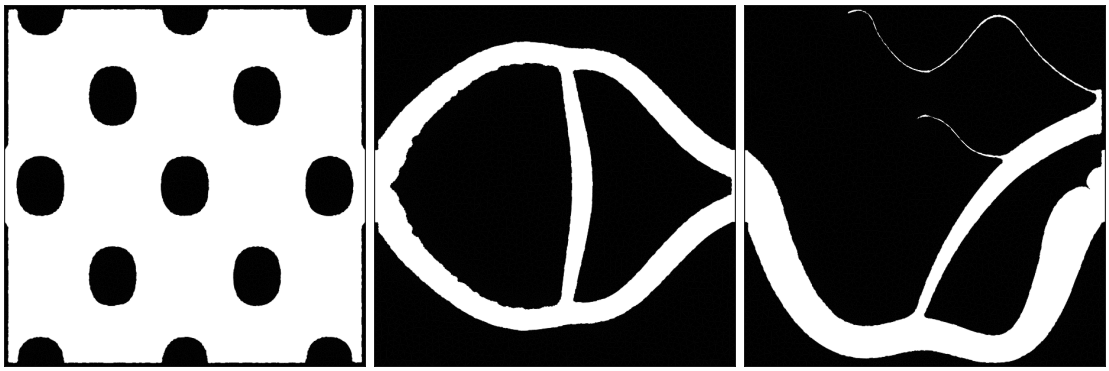
The optimized shapes, associated with temperature fields and fluid velocity fields are represented respectively on Figs. 2.32 to 2.33. Several intermediate shapes arising in the course of the optimization process are depicted on Figure 2.36. In the first case featuring $T_{\mathrm{up}} = T_{\mathrm{low}}$, we retrieve an optimized shape analogous to those presented in [224]. It consists of two main pipes connecting the inlet to the outlet with an additional vertical tubular bar of fluid domain where the velocity is almost zero. As already noticed in [224], this vertical inclusion takes advantage of the low diffusivity of the fluid material in order to insulate thermally the main pipes from the cold input left boundary T_{in}. Note that in contrast with the results presented in [153], the final shape is symmetric despite the use of non-symmetric meshes because the symmetry of the level set function was enforced at every iteration (see further details in chapter 6 about this point).

In the second case featuring $T_{\mathrm{up}} < T_{\mathrm{in}}$, an analogous behavior is observed: the optimized shape consists of a main pipe accumulating the hot temperature from the bottom wall with an additional outgrowth of the fluid domain where the velocity is almost zero which insulates the main pipe from the cold upper wall.

The convergence curves for the objective and constraint functions are depicted on Figure 2.37. Note that the slight increase of objective function at the end in case of the second configuration $T_{\mathrm{low}} < T_{\mathrm{in}} < T_{\mathrm{up}}$ is related to the disappearance of the very thin outgrowths of fluid at the top of the domain: these are beneficial for the performance of the shape but may tend to disappear because they are thinner than the prescribed mesh resolution.

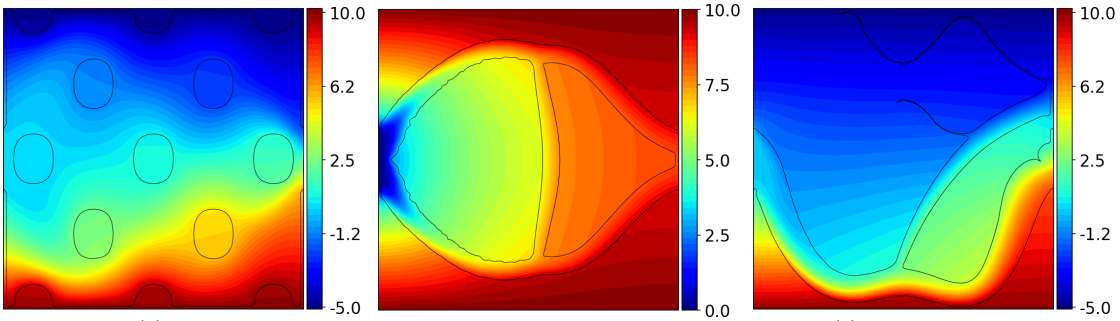
2.5.8 Optimization of a compliant thermoelastic solid with fluid-structure interaction

We finally turn to a shape optimization example in the full three-physic setting presented in section 2.2. A fluid is flowing from the left to the right of a two-dimensional pipe; at the center of this pipe, a solid body is attached to the boundary of a small non optimizable square ω of side length c. The flow is entering the pipe at the inlet with a parabolic profile (with maximal velocity $v_{max} = 1$), and a prescribed temperature T_{in} and the solid body receives a thermal flux h applied at the boundary $\partial\omega$ of the square. The reference temperature of the solid material is equal to the fluid inlet temperature: $T_{\mathrm{ext}} = T_{\mathrm{in}}$. All the other boundaries in this device are insulated from the outside: zero Neumann boundary conditions $\frac{\partial T}{\partial \boldsymbol{n}} = 0$ hold for the temperature; see Figure 2.38 for a schematic of the problem.



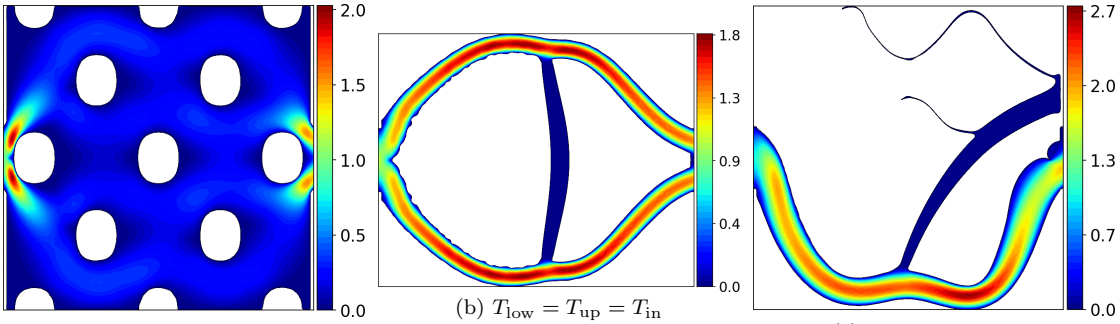
(a) Initial (b) $T_{\text{low}} = T_{\text{up}} = T_{\text{in}}$ (c) $T_{\text{up}} < T_{\text{in}} < T_{\text{low}}$

Figure 2.32: Initial and final configurations for the convective heat transfer test case of section 2.5.7.



(a) Initial (b) $T_{\text{low}} = T_{\text{up}} = T_{\text{in}}$ (c) $T_{\text{up}} < T_{\text{in}} < T_{\text{low}}$

Figure 2.33: Initial and final temperature fields for the convective heat transfer test case of section 2.5.7.



(a) Initial (b) $T_{\text{low}} = T_{\text{up}} = T_{\text{in}}$ (c) $T_{\text{up}} < T_{\text{in}} < T_{\text{low}}$

Figure 2.34: Initial and final norm fields of the velocity $(\boldsymbol{v}, p)$ for the convective heat transfer test case of section 2.5.7.

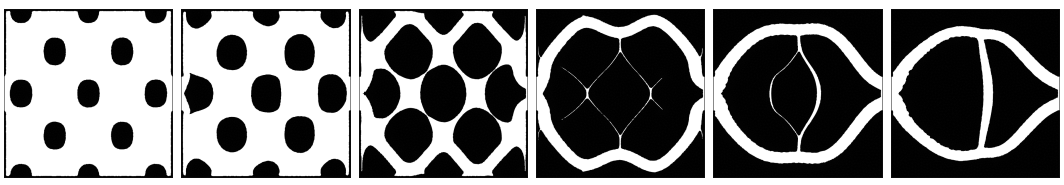


Figure 2.35: Intermediate iterations 0, 10, 30, 55, 188 and 300 for the convective heat transfer test case of section 2.5.7 with $T_{\text{low}} = T_{\text{up}} = T_{\text{in}}$.

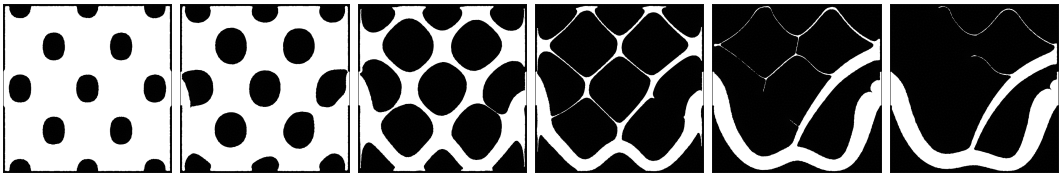
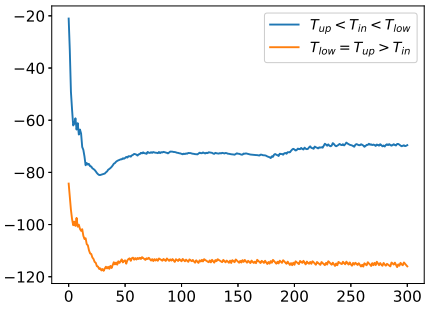
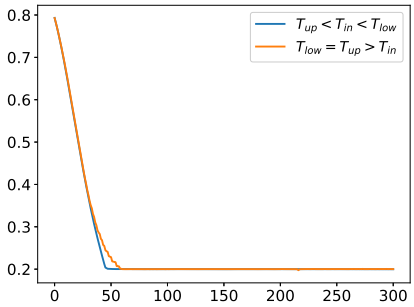


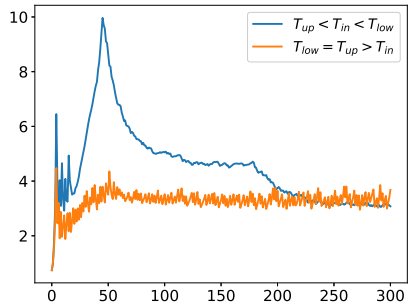
Figure 2.36: Intermediate iterations 0, 10, 30, 55, 188 and 300 for the convective heat transfer test case of section 2.5.7 with $T_{\text{up}} < T_{\text{in}} < T_{\text{low}}$.



(a) Objective function J. Final values: $J = -1.2e+02$ ($T_{\text{low}} = T_{\text{up}} < T_{\text{in}}$) and $J = -70$ ($T_{\text{up}} < T_{\text{in}} < T_{\text{low}}$).



(b) Volume fraction constraint $\eta_f(\Omega_f) := \text{Vol}(\Omega_f)/|D|$. Final value: $\eta_f(\Omega_f) = 0.2$ (in both cases).



(c) Static pressure drop constraint DP. Final values: DP = 3.3 (in both cases).

Figure 2.37: Convergence history for the convective heat transfer test case of section 2.5.7.

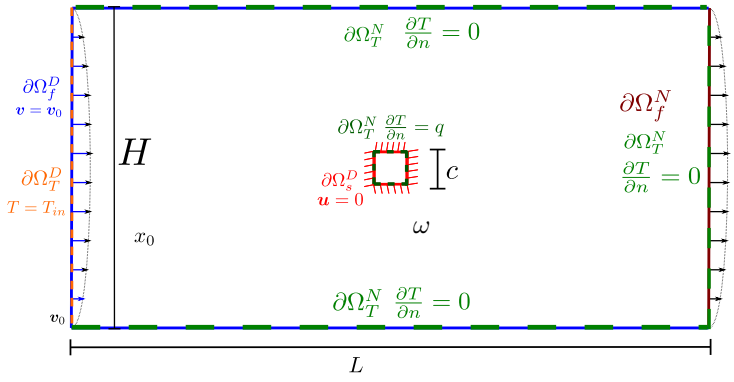


Figure 2.38: Setting of the thermoelastic fluid-structure problem of section 2.5.8.

L	H	c	ρ	c_p	ν	$v_{\max}$	k_s	k_f	T_{in}	T_{ref}	λ	μ	α	h
2	1	0.1	1	0.5	0.01	1	10	1	0	0	12.96	5.55	3	$\pm$ 250

Table 2.6: Parameter values for the thermoelastic fluid-structure test case of section 2.5.8.

We consider in total four different physical configurations corresponding to two possible signs of the thermal flux—either $h > 0$ or $h < 0$—and to two different systems for the physical behavior of the flow: either the Navier-Stokes system (2.2.1), or its linear Stokes counterpart. In the case where $h > 0$, the square boundary $\partial\omega$ plays the role of a thermal source: the high temperature in the solid body induces thermal expansion. In the latter case, where $h < 0$, $\partial\omega$ plays the role of a thermal sink: the lower temperature in the solid body induces thermal contraction. In both cases, the role of the fluid is to mitigate the temperature variations induced in Ω_s by the thermal source term h.

The Reynolds number is set to $\text{Re} = \frac{\rho L v_{max}}{\nu} = 200$ and the Péclet number is $\text{Pe} = \frac{L v_{\max}}{k_f} = 1000$. The volume of the solid phase is imposed to be equal to $\text{Vol}(\Omega_s) = 0.03\text{Vol}(D)$ where $\text{Vol}(D) = 2$ is the volume of the total domain. In all the considered regimes, a sufficiently high value of the thermal dilation coefficient α is used so as to make the thermoelastic effect dominant. The various numerical values for the physical parameters of the problem are summarized in Table 2.6.

Our aim is to minimize the mechanical efforts induced in the solid structure Ω_s by thermal dilation effects and the stress imposed by the fluid subject to the volume constraint:

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{u}(\Gamma)) := \int_{\Omega_s} \sigma_s(\boldsymbol{u}, T_s) : \nabla \boldsymbol{u} \mathrm{d}x = \int_{\Omega_s} (Ae(\boldsymbol{u}) : e(\boldsymbol{u}) - \alpha(T_s - T_{ref})\mathrm{div}(\boldsymbol{u}))\mathrm{d}x, \\ s.t. \quad & \text{Vol}(\Omega_s) = V_{target}. \end{aligned} \tag{2.5.19}$$

The objective functional $J(\Gamma, \boldsymbol{u}(\Gamma))$ corresponds to the internal energy stored inside the structure. Its sensitivities with respect to the shape Γ are calculated thanks to the following formulas:

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) = \int_{\Omega_s} (\sigma_s(\boldsymbol{u}, T_s) : \nabla \boldsymbol{u} \mathrm{div}(\boldsymbol{\theta}) - 2Ae(\boldsymbol{u}) : \nabla \boldsymbol{u} \nabla \boldsymbol{\theta} + \alpha(T_s - T_{ref})\mathrm{Tr}(\nabla \boldsymbol{u} \nabla \boldsymbol{\theta}))\mathrm{d}x, \tag{2.5.20}$$

$$\overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) = -\int_{\Gamma} [\sigma_s(\boldsymbol{u}, T_s) : \nabla \boldsymbol{u} - 2\boldsymbol{n} \cdot Ae(\boldsymbol{u}) \nabla \boldsymbol{u} \cdot \boldsymbol{n} + \alpha(T_s - T_{ref})\boldsymbol{n} \cdot \nabla \boldsymbol{u} \cdot \boldsymbol{n}](\boldsymbol{\theta} \cdot \boldsymbol{n})\mathrm{d}s, \tag{2.5.21}$$

$$\frac{\partial \mathfrak{J}}{\partial \widehat{\boldsymbol{u}}}(\boldsymbol{r}') = \int_{\Omega_s} (2Ae(\boldsymbol{u}) : e(\boldsymbol{r}') - \alpha(T_s - T_{ref})\mathrm{div}(\boldsymbol{r}'))\mathrm{d}x, \tag{2.5.22}$$

$$\frac{\partial \mathfrak{J}}{\partial \widehat{T}}(S') = -\int_{\Omega_s} \alpha S' \mathrm{div}(\boldsymbol{u})\mathrm{d}x. \tag{2.5.23}$$

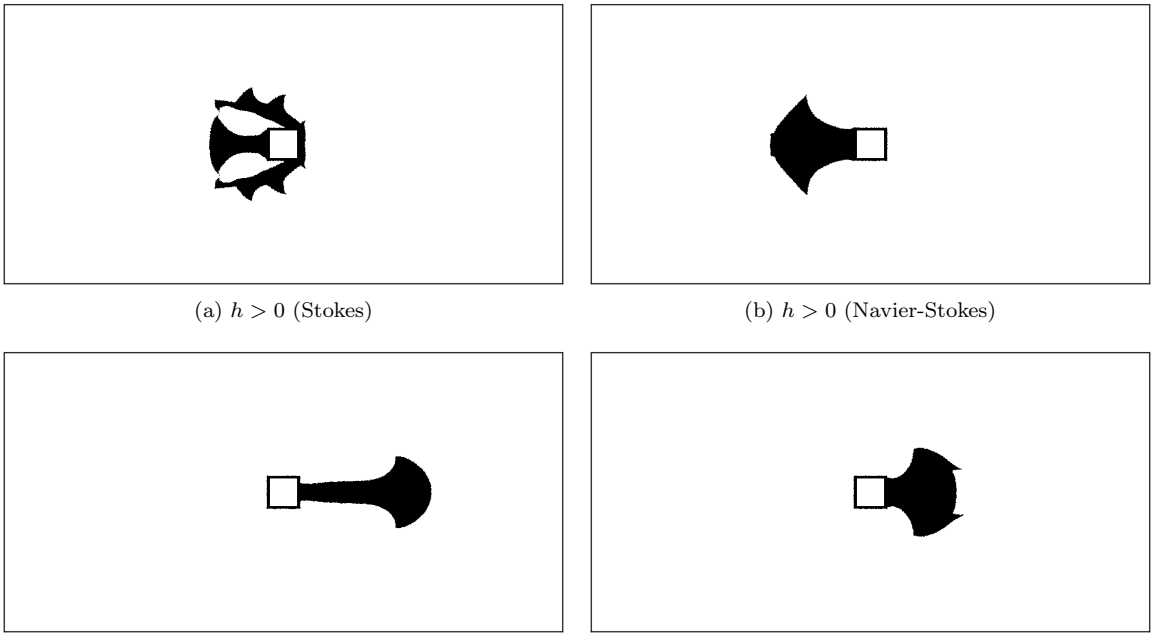
The optimized shapes in the four situations are displayed in Figure 2.39. Note that for the Stokes flow with $h > 0$, we used an initial shape perforated with holes in order to improve the convergence. For the other test cases, we used a disk shape initial domain. The convergence histories for the objective function and deviation to the volume constraint are shown on Figure 2.40. Notably, our optimization algorithm is able to decrease the objective function while keeping constant the volume fraction in the solid phase. Several intermediate shapes are represented in Figure 2.41 in the situation where $h > 0$ and the fluid behavior is driven either by the Stokes or the full Navier-Stokes equations (2.2.1). For this latter case, the state variables $\boldsymbol{v}, T$ and $\boldsymbol{u}$ are additionnally depicted on Figure 2.42. In all four cases, the solid part Ω_s tends to have a large contact surface with the fluid, so as to mitigate the effect of the thermal source (recall that $T_{\text{in}} = T_{\text{ref}}$). The optimized shapes in the cases $h > 0$ and $h < 0$ are dramatically different, and are quite unintuitive from the mechanical viewpoint. Finally, the optimized shapes for a common value of h are noticeably different between the Stokes and Navier-Stokes cases, which could be expected due to the non negligible value of the Reynolds number. We have indeed checked that the optimized shape in the case of a Stokes flow has worse performance when evaluated in the context of a Navier-Stokes flow than the optimized shape in this setting (and vice-versa).

In this example, the objective function J turns out to be very sensitive with respect to very small variations of the shape. Recall that we do not resort to any upwinding scheme in our implementation. Therefore, we used a very fine mesh resolution (the minimum edge length is `hmin`=0.001) as well as the

volume expression (2.3.27) of the shape derivative, which both enhanced the quality of the optimization process; see Figure 2.43.

On average, each intermediate mesh of D has approximately 20,000 vertices. A typical 300 iteration run of any of the aforementioned test cases (including Newton iterations for the numerical resolution of the Navier-Stokes equations) took approximately 6 hours on a 2.50 GHz Intel(R) Xeon(R) CPU.

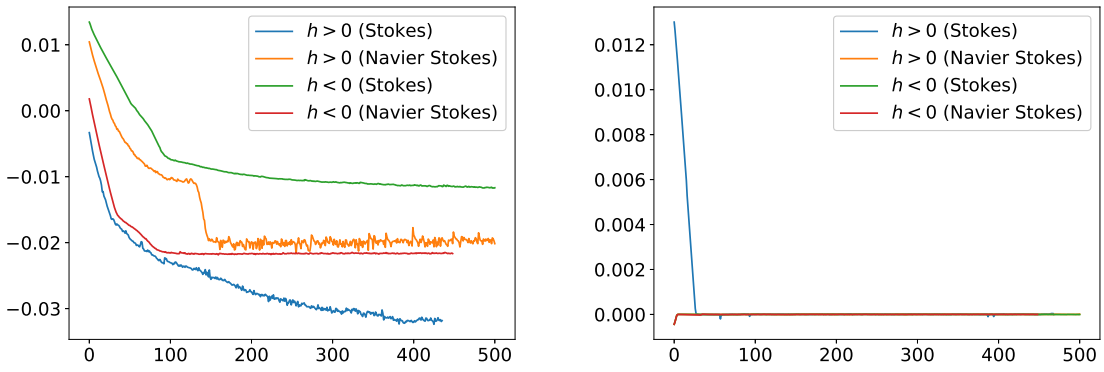
As we have already mentioned, these are preliminary results. The ongoing work presented in chapter 6 focuses on more realistic 3-d test cases.



(a) $h > 0$ (Stokes) (b) $h > 0$ (Navier-Stokes)

(c) $h < 0$ (Stokes) (d) $h < 0$ (Navier-Stokes)

Figure 2.39: Optimized shapes for the three-physic test case of section 2.5.8 in the four considered physical situations.



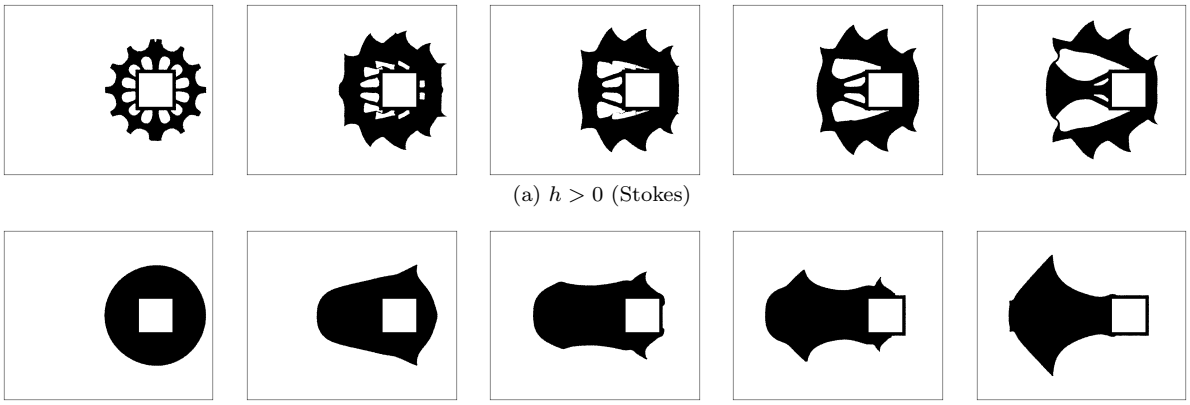
(a) Objective functions J (b) Deviations $\eta_s - 0.03$ to the volume constraint.

Figure 2.40: Convergence histories of the objective and constraint functions in the three-physic test case of section 2.5.8.

2.6 APPENDIX

2.6.1 Proof of propositions 2.3 and 2.4

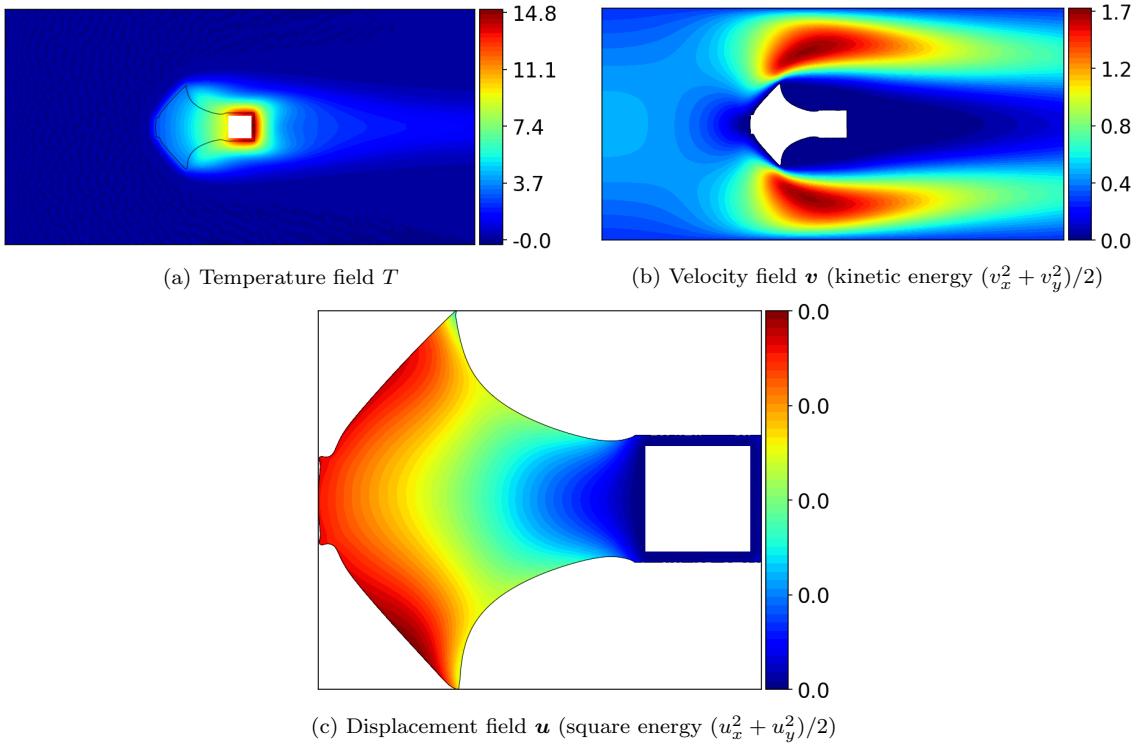
We provide in this appendix a proof of propositions 2.3 and 2.4, or equivalently of (2.4.13) and (2.4.14), which is a mere adaptation of the arguments involved in section 2.3. Using classical arguments based on the implicit function theorem (see e.g. [184]), one proves that under the condition that the linearized



(a) $h > 0$ (Stokes)

(b) $h > 0$ (Navier-Stokes)

Figure 2.41: From left to right and top to bottom: iterations 1, 35, 80, 120 and 300 of the optimization process in the three-physic context of section 2.5.8 for Stokes and Navier Stokes flow where $h > 0$.



(a) Temperature field T

(b) Velocity field $\boldsymbol{v}$ (kinetic energy $(v_x^2 + v_y^2)/2$)

(c) Displacement field $\boldsymbol{u}$ (square energy $(u_x^2 + u_y^2)/2$)

Figure 2.42: State variables $\boldsymbol{v}$, T and $\boldsymbol{u}$ for the optimized configuration of the three-physic shape optimization problem of section 2.5.8, in the situation $h > 0$ and solved with the Navier-Stokes equations.

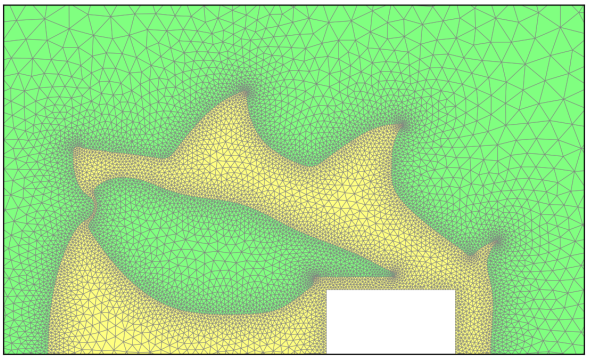


Figure 2.43: Zoom on the mesh for the final configuration of the Stokes case of Ex. 1.

version of the state equations (2.2.1) to (2.2.3) are well posed (see remark 2.12), the mappings $\boldsymbol{v}(\Gamma_{\boldsymbol{\theta}})\circ(I+\boldsymbol{\theta})$, $p(\Gamma_{\boldsymbol{\theta}})\circ(I+\boldsymbol{\theta})$, $T(\Gamma_{\boldsymbol{\theta}})\circ(I+\boldsymbol{\theta})$, and $\boldsymbol{u}(\Gamma_{\boldsymbol{\theta}})\circ(I+\boldsymbol{\theta})$ are differentiable with respect to $\boldsymbol{\theta}$. Differentiating the variational formulations (2.4.2) to (2.4.4), one finds that the Fréchet derivatives $\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta}), \dot{T}(\boldsymbol{\theta})$ and $\dot{\boldsymbol{u}}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}=0$ solve the following variational problems:

Find $(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \in V_{\boldsymbol{v},p}(\Gamma)$ such that $\forall(\boldsymbol{w}', q') \in V_{\boldsymbol{v},p}(\Gamma)$,

$$\begin{aligned}\int_{\Omega_f} [\sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) : \nabla \boldsymbol{w}' + \rho \boldsymbol{w}' \cdot \nabla \boldsymbol{v} \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) + \rho \boldsymbol{w}' \cdot \nabla \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \boldsymbol{v} - q' \mathrm{div}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}))] \mathrm{d}x \\ = \int_{\Omega_f} [\boldsymbol{w}' \cdot \mathrm{div}(\boldsymbol{f}_f \otimes \boldsymbol{\theta}) - (\sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{w}' + \rho \boldsymbol{w}' \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{v}) \mathrm{div}(\boldsymbol{\theta})] \mathrm{d}x \\ + \int_{\Omega_f} (\sigma_f(\boldsymbol{v}, p) : (\nabla \boldsymbol{w}' \nabla \boldsymbol{\theta}) + \sigma_f(\boldsymbol{w}', q') : (\nabla \boldsymbol{v} \nabla \boldsymbol{\theta}) + \rho \boldsymbol{w}' \cdot \nabla \boldsymbol{v} \nabla \boldsymbol{\theta} \cdot \boldsymbol{v}) \mathrm{d}x, \quad (2.6.1)\end{aligned}$$

Find $\dot{T}(\boldsymbol{\theta}) \in V_T(\Gamma)$ such that $\forall S' \in V_T(\Gamma)$,

$$\begin{aligned}\int_{\Omega_s} k_s \nabla \dot{T}(\boldsymbol{\theta}) \cdot \nabla S' \mathrm{d}x + \int_{\Omega_f} (k_f \nabla \dot{T}(\boldsymbol{\theta}) \cdot \nabla S + \rho c_p S' \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \nabla T) \mathrm{d}x = -\int_{\Omega_f} \rho c_p S' \boldsymbol{v} \cdot \nabla \dot{T}(\boldsymbol{\theta}) \mathrm{d}x \\ + \int_{\Omega_s} [\mathrm{div}(Q_s \boldsymbol{\theta}) S' + k_s (\nabla \boldsymbol{\theta} + \nabla \boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta}) I) \nabla T \cdot \nabla S'] \mathrm{d}x \\ + \int_{\Omega_f} [\mathrm{div}(Q_f \boldsymbol{\theta}) S' + k_f (\nabla \boldsymbol{\theta} + \nabla \boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta}) I) \nabla T \cdot \nabla S'] \mathrm{d}x \\ + \int_{\Omega_f} (-\rho c_p S' \boldsymbol{v} \cdot \nabla T \mathrm{div}(\boldsymbol{\theta}) + \rho c_p S' \boldsymbol{v} \cdot \nabla \boldsymbol{\theta}^T \nabla T) \mathrm{d}x, \quad (2.6.2)\end{aligned}$$

Find $\dot{\boldsymbol{u}}(\boldsymbol{\theta}) \in V_{\boldsymbol{u}}(\Gamma)$ such that $\forall \boldsymbol{r}' \in V_{\boldsymbol{u}}(\Gamma)$,

$$\begin{aligned}\int_{\Omega_s} Ae(\dot{\boldsymbol{u}}(\boldsymbol{\theta})) : \nabla \boldsymbol{r}' \mathrm{d}x = \int_{\Omega_s} \alpha \dot{T}(\boldsymbol{\theta}) \mathrm{div}(\boldsymbol{r}') \mathrm{d}x + \int_{\Omega_s} [-\mathrm{div}(\boldsymbol{\theta}) \sigma_s(\boldsymbol{u}, T_s) : \nabla \boldsymbol{r}' + \mathrm{div}(\boldsymbol{f}_s \otimes \boldsymbol{\theta}) \cdot \boldsymbol{r}'] \mathrm{d}x \\ + \int_{\Omega_s} (\sigma_s(\boldsymbol{u}, T_s) : (\nabla \boldsymbol{r} \nabla \boldsymbol{\theta}) + Ae(\boldsymbol{r}) : (\nabla \boldsymbol{u} \nabla \boldsymbol{\theta})) \mathrm{d}x - \int_{\Gamma} \boldsymbol{r}' \cdot \sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \cdot \boldsymbol{n} \mathrm{d}s, \quad (2.6.3)\end{aligned}$$

where $\sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \cdot \boldsymbol{n}$ is an element of the dual space $H_{00}^{-1/2}(\Gamma, \mathbb{R}^d)$ of $H_{00}^{1/2}(\Gamma, \mathbb{R}^d)$ whose action is given by (differentiating (2.4.5) with respect to $\boldsymbol{\theta}$):

$$\begin{aligned}\forall \boldsymbol{r}' \in H_{00}^{1/2}(\Gamma, \mathbb{R}^d), -\int_{\Gamma} \boldsymbol{r}' \cdot \sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \cdot \boldsymbol{n} \mathrm{d}s \\ = \int_{\Omega_f} (\mathrm{div}(\boldsymbol{f}_f \otimes \boldsymbol{\theta}) \cdot \tilde{\boldsymbol{r}} - (\rho \tilde{\boldsymbol{r}} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{v} + \sigma_f(\boldsymbol{v}, p) : \nabla \tilde{\boldsymbol{r}}) \mathrm{div}(\boldsymbol{\theta})) \mathrm{d}x \\ + \int_{\Omega_f} (\rho \tilde{\boldsymbol{r}} \cdot \nabla \boldsymbol{v} \nabla \boldsymbol{\theta} \cdot \boldsymbol{v} + \sigma_f(\boldsymbol{v}, p) : (\nabla \tilde{\boldsymbol{r}} \nabla \boldsymbol{\theta}) + \sigma_f(\tilde{\boldsymbol{r}}, \tilde{q}) : (\nabla \boldsymbol{v} \nabla \boldsymbol{\theta})) \mathrm{d}x \\ - \int_{\Omega_f} (\sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) : \nabla \tilde{\boldsymbol{r}} + \rho \tilde{\boldsymbol{r}} \cdot \nabla \boldsymbol{v} \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) + \rho \tilde{\boldsymbol{r}} \cdot \nabla \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \boldsymbol{v} - \tilde{q} \mathrm{div}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}))) \mathrm{d}x, \quad (2.6.4)\end{aligned}$$

for any extension $(\tilde{\boldsymbol{r}}, \tilde{q}) \in V_{\boldsymbol{v},p}(\Gamma)$ satisfying $\tilde{\boldsymbol{r}} = \boldsymbol{r}'$ on Γ. Note that the above expression is independent of the chosen extension because of (2.6.1) with $\boldsymbol{w}' = \tilde{\boldsymbol{r}}$ and $q' = \tilde{q}$. Then by the definition (2.4.7) of $\mathfrak{J}$:

$$\begin{aligned}J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}})) \\ = \mathfrak{J}(\boldsymbol{\theta}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}) \circ (I+\boldsymbol{\theta}), p(\Gamma_{\boldsymbol{\theta}}) \circ (I+\boldsymbol{\theta}), T(\Gamma_{\boldsymbol{\theta}}) \circ (I+\boldsymbol{\theta}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}) \circ (I+\boldsymbol{\theta})), \quad (2.6.5)\end{aligned}$$

whence the chain rule yields:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \Big[J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{u}(\Gamma_{\boldsymbol{\theta}})) \Big](\boldsymbol{\theta}) = \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) + \frac{\partial \mathfrak{J}}{\partial (\boldsymbol{v}, p)}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) + \frac{\partial \mathfrak{J}}{\partial T}(\dot{T}(\boldsymbol{\theta})) + \frac{\partial \mathfrak{J}}{\partial \boldsymbol{u}}(\boldsymbol{u}(\boldsymbol{\theta})). \quad (2.6.6)$$

One then uses the adjoint equations (2.4.8) to (2.4.10) with $\boldsymbol{r}' = \dot{\boldsymbol{u}}(\boldsymbol{\theta})$, $S' = \dot{T}(\boldsymbol{\theta}), \boldsymbol{w}' = \dot{\boldsymbol{v}}(\boldsymbol{\theta}), q' = \dot{p}(\boldsymbol{\theta})$ as test functions to obtain:

$$\frac{\partial \mathfrak{J}}{\partial \boldsymbol{u}}(\dot{\boldsymbol{u}}(\boldsymbol{\theta})) = \int_{\Omega_s} Ae(\boldsymbol{r}) : \nabla \dot{\boldsymbol{u}}(\boldsymbol{\theta}) \mathrm{d}x = \int_{\Omega_s} Ae(\dot{\boldsymbol{u}}(\boldsymbol{\theta})) : \nabla \boldsymbol{r} \mathrm{d}x, \tag{2.6.7}$$

$$\frac{\partial \mathfrak{J}}{\partial T}(\dot{T}(\boldsymbol{\theta})) = \int_{\Omega_s} k_s \nabla S \cdot \nabla \dot{T}(\boldsymbol{\theta}) \mathrm{d}x + \int_{\Omega_f} (k_f \nabla S \cdot \nabla \dot{T}(\boldsymbol{\theta}) + \rho c_p S \boldsymbol{v} \cdot \nabla \dot{T}(\boldsymbol{\theta})) \mathrm{d}x - \int_{\Omega_s} \alpha \dot{T}(\boldsymbol{\theta}) \mathrm{div}(\boldsymbol{r}) \mathrm{d}x, \tag{2.6.8}$$

$$\begin{aligned}
&\frac{\partial \mathfrak{J}}{\partial (\boldsymbol{v}, p)}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \\
&= \int_{\Omega_f} (\sigma_f(\boldsymbol{w}, q) : \nabla \dot{\boldsymbol{v}}(\boldsymbol{\theta}) + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) + \rho \boldsymbol{w} \cdot \nabla \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \boldsymbol{v} - \dot{p}(\boldsymbol{\theta}) \mathrm{div}(\boldsymbol{w})) \mathrm{d}x + \int_{\Omega_f} \rho c_p S \nabla T \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \mathrm{d}x \\
&= \int_{\Omega_f} (\sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) : \nabla \boldsymbol{w} + \rho \boldsymbol{w} \cdot \nabla \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \boldsymbol{v} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) - q \mathrm{div}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}))) \mathrm{d}x + \int_{\Omega_f} \rho c_p S \nabla T \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \mathrm{d}x.
\end{aligned} \tag{2.6.9}$$

Using now equations (2.6.2) and (2.6.3) with $\boldsymbol{r}' = \boldsymbol{r}, S' = S$ as test functions and (2.6.4) with $(\tilde{\boldsymbol{r}}, \tilde{q}) = (\boldsymbol{w}, q)$ as an extension of $\boldsymbol{r}' = \boldsymbol{r} \in H^{1/2}_{00}(\Gamma, \mathbb{R}^d)$ to eliminate the bilinear terms, the above three equations rewrite:

$$\begin{aligned}
\frac{\partial \mathfrak{J}}{\partial \boldsymbol{u}}(\dot{\boldsymbol{u}}(\boldsymbol{\theta})) = &- \int_{\Gamma} \boldsymbol{r} \cdot \sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \cdot \boldsymbol{n} \mathrm{d}s + \int_{\Omega_s} \alpha \dot{T}(\boldsymbol{\theta}) \mathrm{div}(\boldsymbol{r}) \mathrm{d}x \\
&+ \int_{\Omega_s} [-\mathrm{div}(\boldsymbol{\theta}) \sigma_s(\boldsymbol{u}, T_s) : \nabla \boldsymbol{r} + \mathrm{div}(\boldsymbol{f}_s \otimes \boldsymbol{\theta}) \cdot \boldsymbol{r}] \mathrm{d}x \\
&+ \int_{\Omega_s} (\sigma_s(\boldsymbol{u}, T_s) : (\nabla \boldsymbol{r} \nabla \boldsymbol{\theta}) + Ae(\boldsymbol{r}) : (\nabla \boldsymbol{u} \nabla \boldsymbol{\theta})) \mathrm{d}x,
\end{aligned} \tag{2.6.10}$$

$$\begin{aligned}
\frac{\partial \mathfrak{J}}{\partial T}(\dot{T}(\boldsymbol{\theta})) = &- \int_{\Omega_s} \alpha \dot{T}(\boldsymbol{\theta}) \mathrm{div}(\boldsymbol{r}) \mathrm{d}x - \int_{\Omega_f} \rho c_p S \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \cdot \nabla T \mathrm{d}x \\
&+ \int_{\Omega_s} [\mathrm{div}(Q_s \boldsymbol{\theta}) S + k_s (\nabla \boldsymbol{\theta} + \nabla \boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta}) I) \nabla T \cdot \nabla S] \mathrm{d}x \\
&+ \int_{\Omega_f} [\mathrm{div}(Q_f \boldsymbol{\theta}) S + k_f (\nabla \boldsymbol{\theta} + \nabla \boldsymbol{\theta}^T - \mathrm{div}(\boldsymbol{\theta}) I) \nabla T \cdot \nabla S] \mathrm{d}x \\
&+ \int_{\Omega_f} (-\rho c_p S \boldsymbol{v} \cdot \nabla T \mathrm{div}(\boldsymbol{\theta}) + \rho c_p S \boldsymbol{v} \cdot \nabla \boldsymbol{\theta}^T \nabla T) \mathrm{d}x,
\end{aligned} \tag{2.6.11}$$

$$\begin{aligned}
\frac{\partial \mathfrak{J}}{\partial (\boldsymbol{v}, p)}(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) = &\int_{\Gamma} \boldsymbol{r} \cdot \sigma_f(\dot{\boldsymbol{v}}(\boldsymbol{\theta}), \dot{p}(\boldsymbol{\theta})) \cdot \boldsymbol{n} \mathrm{d}s + \int_{\Omega_f} \rho c_p S \nabla T \cdot \dot{\boldsymbol{v}}(\boldsymbol{\theta}) \mathrm{d}x \\
&+ \int_{\Omega_f} (\mathrm{div}(\boldsymbol{f}_f \otimes \boldsymbol{\theta}) \cdot \boldsymbol{w} - (\rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{v} + \sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{w}) \mathrm{div}(\boldsymbol{\theta})) \mathrm{d}x \\
&+ \int_{\Omega_f} (\rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \nabla \boldsymbol{\theta} \cdot \boldsymbol{v} + \sigma_f(\boldsymbol{v}, p) : (\nabla \boldsymbol{w} \nabla \boldsymbol{\theta}) + \sigma_f(\boldsymbol{w}, q) : (\nabla \boldsymbol{v} \nabla \boldsymbol{\theta})) \mathrm{d}x.
\end{aligned} \tag{2.6.12}$$

Formula (2.4.13) follows by summing up the above three equations. If H^2 regularity holds for $\boldsymbol{v}, \boldsymbol{u}, T$ and H^1 regularity holds for p on their respective domains of definition, then an integration by parts allows

to rewrite (2.4.13) as;

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}},\boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}),p(\Gamma_{\boldsymbol{\theta}}),T(\Gamma_{\boldsymbol{\theta}}),\boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta})\\
&\qquad=\int_\Gamma \boldsymbol{g}_{\mathfrak{J}}\cdot\boldsymbol{\theta}\mathrm{d}s+\int_\Gamma(\boldsymbol{f}_f\cdot\boldsymbol{w}-\sigma_f(\boldsymbol{v},p):\nabla\boldsymbol{w}-\rho\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v})(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s\\
&\qquad+\int_\Gamma[\boldsymbol{n}\cdot\sigma_f(\boldsymbol{v},p)\nabla\boldsymbol{w}\cdot\boldsymbol{\theta}+\boldsymbol{n}\cdot\sigma_f(\boldsymbol{w},q)\nabla\boldsymbol{v}\cdot\boldsymbol{\theta}+\rho(\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{\theta}]\mathrm{d}s\\
&\qquad+\int_\Gamma\left(k_s\nabla T_s\cdot\nabla S_s-k_f\nabla T_f\cdot\nabla S_f+Q_fS-Q_sS_s\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s\\
&\qquad+\int_\Gamma\left(-k_s(\nabla T_s\cdot\boldsymbol{\theta})(\nabla S_s\cdot\boldsymbol{n})-k_s(\nabla S_s\cdot\boldsymbol{\theta})(\nabla T_s\cdot\boldsymbol{n})\right)\mathrm{d}s\\
&\qquad+\int_\Gamma\left(k_f(\nabla T_f\cdot\boldsymbol{\theta})(\nabla S_f\cdot\boldsymbol{n})+k_f(\nabla S_f\cdot\boldsymbol{\theta})(\nabla T_f\cdot\boldsymbol{n})\right)\mathrm{d}s\\
&+\int_\Gamma\left[(\sigma_s(\boldsymbol{u},T_s):\nabla\boldsymbol{r}-\boldsymbol{f}_s\cdot\boldsymbol{r})(\boldsymbol{\theta}\cdot\boldsymbol{n})-\boldsymbol{n}\cdot\sigma_s(\boldsymbol{u},T_s)\nabla\boldsymbol{r}\cdot\boldsymbol{\theta}-\boldsymbol{n}\cdot Ae(\boldsymbol{r})\nabla\boldsymbol{u}\cdot\boldsymbol{\theta}\right]\mathrm{d}s+\int_\Gamma\boldsymbol{\Lambda}\cdot\boldsymbol{\theta}\mathrm{d}x,
\end{aligned}\tag{2.6.13}$$

where $\boldsymbol{\Lambda}$ is a $L^1(D,\mathbb{R}^d)$ function obtained from Green's identity. The Hadamard structure theorem implies that (2.6.13) vanishes for compactly supported fields $\boldsymbol{\theta}$ or for fields $\boldsymbol{\theta}$ tangent to Γ. This implies that in fact, $\boldsymbol{\Lambda}=0$, and (2.4.14) follows by removing the terms depending on the tangential component of $\boldsymbol{\theta}$ on Γ.

2.6.2 Calculating the shape derivative of a particular objective functional and its adjoint system with Céa's method

In this appendix, we consider the simplified setting of section 2.3, and we show that the adjoint boundary condition $p_s=p_f$ corresponding to the equality of normal derivatives for the primal variables in (2.3.2) can be retrieved formally with the classical Céa's Lagrangian method [84] also reviewed in chapter 1, section 1.2.3. We shall as well illustrate once again the calculation of the shape derivative with this method on a particular type of objective functional J.

We restrict ourselves in this part to objective functionals J of the form:

$$J(\Gamma,u_s(\Gamma),u_f(\Gamma))=\int_{\Omega_f}j_f(u_f(\Gamma))\mathrm{d}x+\int_{\Omega_s}j_s(u_s(\Gamma))\mathrm{d}x,\tag{2.6.14}$$

for two C^2 functions $j_s,j_f:\mathbb{R}\to\mathbb{R}$ with bounded second-order derivatives:

$$||j_s''||_{L^\infty(\mathbb{R})}<\infty,\ ||j_f''||_{L^\infty(\mathbb{R})}<\infty.$$

Proposition 2.5. *The functional $\boldsymbol{\theta}\mapsto J(\Gamma_{\boldsymbol{\theta}},u_s(\Gamma_{\boldsymbol{\theta}}),u_f(\Gamma_{\boldsymbol{\theta}}))$, from $W_0^{1,\infty}(D,\mathbb{R}^d)$ into $\mathbb{R}$, as defined in (2.6.14), is differentiable at $\boldsymbol{\theta}=0$ and, under H^2 regularity of the variables u_f,u_s,p_f,p_s, the shape derivative reads:*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}},u_f(\Gamma_{\boldsymbol{\theta}}),u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta})=\\
&\qquad\int_\Gamma\left[j_f(u_f)-j_s(u_s)+f_fp_f-f_sp_s+\mu\nabla u_s\cdot\nabla p_s+\nu\frac{\partial u_f}{\partial\boldsymbol{n}}\frac{\partial p_f}{\partial\boldsymbol{n}}\right](\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s.
\end{aligned}\tag{2.6.15}$$

Remark 2.18. Formula (2.6.15) is equivalent to the following one, which was obtained in (2.3.34) by the Lagrangian method in section 2.3.2,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}},u_f(\Gamma_{\boldsymbol{\theta}}),u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta})=\\
&\int_\Gamma\left[j_f(u_f)-j_s(u_s)+f_fp_f-f_sp_s-\nu\nabla u_f\cdot\nabla p_f+\mu\nabla u_s\cdot\nabla p_s+2\nu\frac{\partial u_f}{\partial\boldsymbol{n}}\frac{\partial p_f}{\partial\boldsymbol{n}}\right](\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s,
\end{aligned}\tag{2.6.16}$$

because the boundary condition $u_f=0$ on Γ implies that $\nu\frac{\partial u_f}{\partial\boldsymbol{n}}\frac{\partial p_f}{\partial\boldsymbol{n}}=\nu\nabla u_f\cdot\nabla p_f$.

Proof. We introduce the following Lagrangian

$$\mathcal{L}(\Gamma,\widehat{u}_f,\widehat{u}_s,\widehat{p}_f,\widehat{p}_s,\widehat{\lambda}) = \int_{\Omega_f} j_f(\widehat{u}_f)\mathrm{d}x + \int_{\Omega_s} j_s(\widehat{u}_s)\mathrm{d}x$$
$$- \int_{\Omega_f}(-\nu\Delta\widehat{u}_f - f_f)\widehat{p}_f\mathrm{d}x - \int_{\Omega_s}(\mu\nabla\widehat{u}_s\cdot\nabla\widehat{p}_s - f_s\widehat{p}_s)\mathrm{d}x - \int_\Gamma \nu\frac{\partial\widehat{u}_f}{\partial\boldsymbol{n}}\widehat{p}_s\mathrm{d}s - \int_\Gamma\widehat{\lambda}\widehat{u}_f\mathrm{d}s, \quad (2.6.17)$$

where the above 'hat' functions all belong to the Sobolev space $H_0^1(D)$ which is independent of the position of the interface Γ. Following the methodology described in [17, 84, 253], Lagrange multipliers $\widehat{p}_f, \widehat{p}_s, \widehat{\lambda}$ are introduced in (2.6.17) to enforce the state equations (2.3.1) and (2.3.2) and the Dirichlet boundary condition $u_f = 0$ on the moving interface Γ. The main idea of Céa's method consists in finding the equations for the values p_f, p_s of the adjoint states $\widehat{p}_f$ and $\widehat{p}_s$ and the value λ of the Lagrange multiplier $\widehat{\lambda}$ by requiring that the partial derivatives $\frac{\partial\mathcal{L}}{\partial\widehat{u}_f}$, $\frac{\partial\mathcal{L}}{\partial\widehat{u}_s}$ at $\widehat{u}_f = u_f$, $\widehat{u}_s = u_s$ vanish. The latter partial derivatives read, for arbitrary v_f, $v_s \in H_0^1(D)$:

$$\frac{\partial\mathcal{L}}{\partial\widehat{u}_f}(v_f) = \int_{\Omega_f} j_f'(u_f)v_f\mathrm{d}x + \int_{\Omega_f}\nu\Delta v_f p_f\mathrm{d}x - \int_\Gamma\nu\frac{\partial v_f}{\partial\boldsymbol{n}}p_s\mathrm{d}s - \int_\Gamma\lambda v_f\mathrm{d}s$$
$$= \int_{\Omega_f}(j_f'(u_f)+\nu\Delta p_f)v_f\mathrm{d}x + \int_\Gamma\nu\left(\frac{\partial v_f}{\partial\boldsymbol{n}}p_f - \frac{\partial p_f}{\partial\boldsymbol{n}}v_f\right)\mathrm{d}s - \int_\Gamma\nu\frac{\partial v_f}{\partial\boldsymbol{n}}p_s\mathrm{d}s - \int_\Gamma\lambda v_f\mathrm{d}s \quad (2.6.18)$$

$$\frac{\partial\mathcal{L}}{\partial\widehat{u}_s}(v_s) = \int_{\Omega_s} j_s'(u_s)v_s\mathrm{d}x - \int_{\Omega_s}\mu\nabla v_s\cdot\nabla p_s\mathrm{d}x$$

Now requiring $\frac{\partial\mathcal{L}}{\partial\widehat{u}_s}(v_s)$ vanish for any $v_s \in H_0^1(D)$ yields the adjoint equation (and the attached boundary conditions) (2.3.23) for p_s.

Likewise, requiring $\frac{\partial\mathcal{L}}{\partial\widehat{u}_f}(v_f)$ to vanish for any $v_f \in H_0^1(D)$ should lead to the other adjoint system (2.3.24) but the derivation is a bit more subtle. First, choosing arbitrary v_f with compact support in Ω_f in (2.6.18) yields:

$$-\nu\Delta p_f = j_f'(u_f).$$

Second, choosing smooth v_f such that $v_f = 0$ on Γ and that v_f has an arbitrary trace $\frac{\partial v_f}{\partial\boldsymbol{n}}$ yields the Dirichlet interface condition $p_f = p_s$ on Γ. Thus, the adjoint system (2.3.24) for p_f is completely recovered. Finally, choosing v_f in (2.6.18) with arbitrary trace on Γ leads to the optimal value of the Lagrange multiplier $\lambda = -\nu\frac{\partial p_f}{\partial\boldsymbol{n}}$ on Γ.

Assuming that the solutions u_s and u_f to (2.3.1) and (2.3.2) are differentiable with respect to Γ (which is where Céa's method is only formal), and using that the partial derivatives of $\mathcal{L}$ with respect to $\widehat{u}_f$ and $\widehat{u}_s$ vanish at $(\boldsymbol{\theta},\widehat{u}_f,\widehat{u}_s,\widehat{p}_f,\widehat{p}_s,\widehat{\lambda}) = (0,u_f,u_s,p_f,p_s,\lambda)$, a simple use of the chain rule produces:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}},u_f(\Gamma_{\boldsymbol{\theta}}),u_s(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) = \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta})$$
$$= \int_\Gamma\left(j_f(u_f) - j_s(u_s) + \mu\nabla u_s\cdot\nabla p_s - f_s p_s - \nu\mathrm{div}(p_s\cdot\nabla u_f) - \lambda\frac{\partial u_f}{\partial\boldsymbol{n}}\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s$$
$$= \int_\Gamma\left(j_f(u_f) - j_s(u_s) + \mu\nabla u_s\cdot\nabla p_s - f_s p_s - \nu\nabla p_s\cdot\nabla u_f + f_f p_s + \nu\frac{\partial u_f}{\partial\boldsymbol{n}}\frac{\partial p_f}{\partial\boldsymbol{n}}\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s, \quad (2.6.19)$$

where we have used the technical proposition 1.7 of chapter 1 when differentiating $\int_\Gamma\nu\frac{\partial u_f}{\partial\boldsymbol{n}}p_s\mathrm{d}s$ with the normal vector depending on Γ. Using that $u_f = 0$ and $\frac{\partial p_s}{\partial\boldsymbol{n}} = 0$ on Γ, it follows that $\nabla u_f\cdot\nabla p_s = 0$ on Γ, and since $p_f = p_s$ on Γ, we retrieve expression (2.6.16). □

CHAPTER 3

NULL SPACE GRADIENT FLOWS FOR CONSTRAINED SHAPE OPTIMIZATION

Contents

Note : most of the content of this chapter has been submitted in the preprint [155]:

F. Feppon, G. Allaire, and C. Dapogny, *Null space gradient flows for constrained optimization with applications to shape optimization*, Submitted, (2019).

However, several changes have been made in the introduction and in section 3.4 for a better understanding, as well as a new section 3.5 providing additional discussions and comparisons of our method with other algorithms. Some redundant introductory parts with the previous chapters have also been removed.

3.1 Introduction

The purpose of this chapter is to introduce efficient and reliable algorithmic methodologies for the resolution of general shape optimization problems featuring an arbitrary number of equality and inequality constraints; for multiphysics applications in the context of chapter 2 , these problems can be generically formulated as

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \\ s.t. \quad & \begin{cases} g_i(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = 0, & 1 \le i \le p, \\ h_j(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \le 0, & 1 \le j \le q. \end{cases} \end{aligned} \tag{3.1.1}$$

The ability to account for equality or inequality constraints g_i or h_j in the resolution of such optimal design problems is a *sine qua non* condition towards the application of shape and topology optimization methods to realistic industrial applications. Indeed, industrial designs are in most cases devised with

respect to a variety of load specifications, such as minimum thickness, maximum curvature radius, upper bound limits on the stress in the case of a mechanical structure or on the pressure drop between the outlet and inlet in the case of a fluid device.

Over the past decades, many iterative algorithms have been proposed to solve generic constrained optimization problems of the form:

$$\min_{x\in V} \quad J(x) \qquad \text{s.t.} \begin{cases} \boldsymbol{g}(x) = 0 \\ \boldsymbol{h}(x) \leq 0, \end{cases} \tag{3.1.2}$$

where V is the optimization set, $J : V \to \mathbb{R}$ is a differentiable objective function, $\boldsymbol{g} : V \to \mathbb{R}^p$ and $\boldsymbol{h} : V \to \mathbb{R}^q$ are differentiable functions accounting for p equality and q inequality differentiable constraints, respectively. Classical gradient-based algorithms for the numerical resolution of (3.1.2) include, e.g., Penalty, Lagrangian, Interior Point and Trust Region Methods, Sequential Quadratic or Linear Programming (SQP or SLP) [70, 244, 321, 161, 310], the Method of Moving Asymptotes (MMA) [298], the Method of Feasible Directions [327, 308].

A major difficulty related to the practical use of these algorithms for topology optimization lies in that all the aforementioned techniques require fine tuning of the algorithm parameters in order to actually solve the minimization problem. These parameters are e.g. the penalty coefficients in the Augmented Lagrangian and Interior Point methods, the size of the trust region in SLP algorithms, the strategy for approximating the Hessian matrix in SQP, the bounds on the asymptotes in MMA and the Topkis parameters in MFD. The correct determination of these parameters is strongly case-dependent and often unintuitive: for instance, the penalty coefficients must be neither 'too large' nor 'too small' in Lagrangian methods, the SLP trust region size—which acts as a step length—cannot be chosen too small (otherwise the involved quadratic subproblems may not have a solution). For shape and topology optimization applications, a fair amount of trials and errors is often required in order to obtain satisfying minimizing sequences of shapes. Since every optimization step depends on the resolution of partial differential equations, such tunings are very tedious, time consuming for 2-d cases, and simply not affordable for realistic 3-d applications.

As a matter of fact, advanced mathematical programming methods are not frequently described for shape optimization based on Hadamard's method. Rather, for simplicity of implementation, Penalty and Augmented Lagrangian Methods are often used, all the more when only one constraint is considered [32, 113]. Morin et. al. introduced a variant of SQP in [237] but they treated a volume constraint with a Lagrange Multiplier method. For more complex applications, some authors have proposed adapted variants of Sequential Linear Programming [135] or of the Method of Feasible Directions [150]. Let us remark that in shape optimization based on the method of Hadamard, an additional difficulty comes into play due to the classical identification and regularization step of the descent direction. So far, this matter has not been treated explicitly for constrained optimization problems: common approaches rather compute a descent direction *first*, before performing a regularization (see e.g. [135, 150]), which may alter the decrease property of the updating step.

In this chapter, we propose a novel method for constrained optimization which is rather easy to implement and reliable in the sense that it allows to solve (3.1.2) without the need for tuning non physical parameters; this makes it particularly well adapted to the specificities of shape and topology optimization applications. The essence of our method is a modification of the celebrated gradient flow so as to make it able to 'see the constraints': optimization trajectories $x(t)$ are obtained by solving an Ordinary Differential Equation (ODE):

$$\dot{x}(t) = -\alpha_J \boldsymbol{\xi}_J(x(t)) - \alpha_C \boldsymbol{\xi}_C(x(t)). \tag{3.1.3}$$

An admissible local minimizer to (3.1.1) is then reached as the stationary point x^* of the continuous trajectory $x(t)$ starting from any (feasible or not) initialization $x(0)$ and whatever the value of $\alpha_J, \alpha_C > 0$. This property is retrieved at the discrete level provided (3.1.3) is discretized with a sufficiently small Euler step size Δt.

The descent direction $\dot{x}$ is a combination of a so-called 'null space' direction $\boldsymbol{\xi}_J(x)$ and a 'range space' direction $\boldsymbol{\xi}_C(x)$, lying respectively in the null space of the constraints and in its orthogonal complement (for this reason, we call the ODE (3.1.3) a 'null space' gradient flow). The null space direction $\boldsymbol{\xi}_J(x)$ is the projection of the gradient $\nabla J(x)$ onto the cone of feasible directions. Our approach relies on a suitable dual program for the resolution of the combinatorial character of the inequality constraints: we

precisely identify those active inequality constraints that the optimization path is allowed to 'unstick' from (thus re-entering into the feasible domain) and, conversely, those inequality constraints to which it must remain tangent (see Figure 3.1 below). As a result, $-\boldsymbol{\xi}_J(x)$ is always the best possible descent direction respecting locally both equality and inequality constraints. The range space direction $\boldsymbol{\xi}_C(x)$ is a Gauss-Newton direction $\boldsymbol{\xi}_C(x)$ which is aimed to smoothly lead the optimization path toward the feasible domain. Finally, $\alpha_J, \alpha_C > 0$ are two (optional) parameters introduced which scale the relative decrease rates of the objective function and of the violation of the constraints; we shall see in particular that the latter quantity decreases along trajectories $x(t)$ at least as fast as $e^{-\alpha_C t}$.

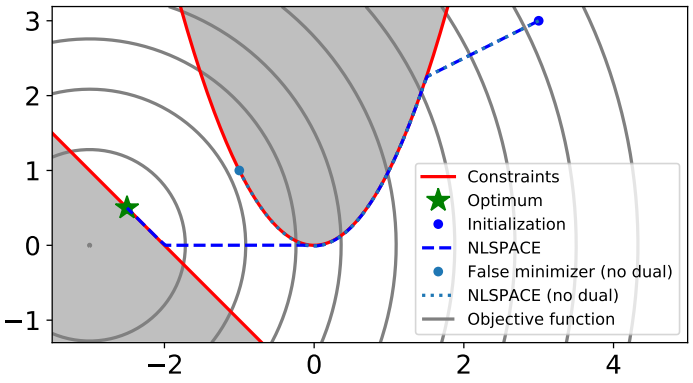


Figure 3.1: An example of optimization trajectory produced by our null space gradient flow (3.1.3). Trajectories travel tangentially to an optimal subset $\widehat{I}(x) \subset \widetilde{I}(x)$ of the active constraints $\widetilde{I}(x)$, which is determined by a dual problem (see section 3.3). A less optimal trajectory is obtained without the identification of the set $\widehat{I}(x)$, because it is unable to escape the tangent space to the constraints labeled by $\widetilde{I}(x)$.

More specifically, for a given subset of indices $I \subset \{1, \dots, q\}$, denote by $\boldsymbol{h}_I(x) := (h_i(x))_{i \in I}$ the collection of inequality constraints and by $\boldsymbol{C}_I(x)$ the matrix

$$\boldsymbol{C}_I(x) := \begin{bmatrix} \boldsymbol{g}(x) \\ \boldsymbol{h}_I(x) \end{bmatrix}. \tag{3.1.4}$$

Then, for inequality constrained problems, the directions $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ in (3.1.3) are defined as follows:

$$\boldsymbol{\xi}_J(x) = (I - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)}(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)})\nabla J(x), \tag{3.1.5}$$

$$\boldsymbol{\xi}_C(x) = \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)})^{-1}\boldsymbol{C}_{\widetilde{I}(x)}(x), \tag{3.1.6}$$

where I is the identity matrix and $(\mathrm{D}\boldsymbol{C}(x))_{ij} = \partial_j C_i(x)$ denotes the Jacobian matrix of a vector function $\boldsymbol{C}(x) = (C_i(x))_i$ (the dependence with respect to x is omitted when the context is clear). The symbol $\mathcal{T}$ denotes the transposition operator; it may differ from the usual transpose T if the optimization set V is infinite dimensional (see below and section 3.2). Formulas (3.1.5) and (3.1.6) involve two different subsets $\widehat{I}(x) \subset \widetilde{I}(x)$ of indices of the inequality constraints $\{1, ..., q\}$: the first one $\widetilde{I}(x)$ is the set of all saturated or violated constraints, defined by

$$\widetilde{I}(x) = \{i \in \{1, \dots, q\} \,|\, h_i(x) \geq 0\}. \tag{3.1.7}$$

The set $\widehat{I}(x) \subset \widetilde{I}(x)$ is an optimal subset which is characterized by the following 'dual' quadratic optimization subproblem:

$$(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x)) := \arg \min_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^p \\ \boldsymbol{\mu} \in \mathbb{R}^{\widetilde{q}(x)}, \, \boldsymbol{\mu} \geq 0}} ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}||_V. \tag{3.1.8}$$

The problem (3.1.8) amounts to compute the projection of $\nabla J(x)$ onto the cone of feasible directions. It allows to determine the optimal set $\widehat{I}(x)$ from the positive components of the optimal Lagrange multiplier $\boldsymbol{\mu}^*(x)$:

$$\widehat{I}(x) := \{i \in \widetilde{I}(x) \,|\, \mu_i^*(x) > 0\}. \tag{3.1.9}$$

In proposition 3.4, we show that this definition of $\widehat{I}(x)$ ensures that $-\boldsymbol{\xi}_J(x)$ is the 'best' descent direction respecting locally both equality and inequality constraints. This key idea to use these two different sets of indices $\widehat{I}(x)$ and $\widetilde{I}(x)$ ensures that $\boldsymbol{\xi}_J(x)$ is obtained by projection of $\nabla J(x)$ on the largest possible subspace staying tangential to the set of constraints. This approach turns out to be very efficient efficient in the case of a large number of (possibly violated) inequality constraints (Figure 3.1). As a result of the flexibility of ODE approaches, our method depends truly only on the discretization step Δt for (3.1.3), and on the physically interpretable, dimensionless parameters α_J, α_C, which makes them relatively easy to tune for the user.

It turns out that our ODE (3.1.3) is a generalization of rather classical dynamical system approaches to nonlinear constrained optimization, which are maybe less known in the topology optimization community. When the problem (3.1.2) features no constraint, (3.1.5) and (3.1.6) become $\boldsymbol{\xi}_J(x) = \nabla J(x)$ and $\boldsymbol{\xi}_C(x) = 0$, so that the ODE (3.1.3) reduces to the standard gradient flow

$$\dot{x}(t) = -\nabla J(x(t)). \tag{3.1.10}$$

When (3.1.2) also features equality constraints $\boldsymbol{g}(x) = 0$, but no inequality constraint (in that case $\boldsymbol{C}_{\widetilde{I}(x)}(x) = \boldsymbol{C}_{\widehat{I}(x)}(x) = \boldsymbol{g}(x)$), the same ODE as ours/(3.1.3) was previously derived and studied in the early 1980s by Tanabe [300] (without the Gauss-Newton direction $\boldsymbol{\xi}_C(x(t))$) and by Yamashita [317] (with both $\boldsymbol{\xi}_J(x(t))$ and $\boldsymbol{\xi}_C(x(t))$). In this particular case, the solution to the dual problem (3.1.8) admits a closed-form expression and (3.1.3) reads with our notation

$$\dot{x} = -\alpha_J(I - \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g})\nabla J(x) - \alpha_C \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x). \tag{3.1.11}$$

In the general situation where (3.1.2) features both inequality and equality constraints, variants of the ODE (3.1.11) have been considered by different authors, however with a different method from ours [277, 200, 201, 282]. The most common approach in the literature consists in introducing q slack variables $\{z_i\}_{1\leq i\leq q} \in \mathbb{R}^q$ so as to convert the q inequalities $h_i(x) \leq 0$ for $1 \leq i \leq q$ into as many equality constraints $h_i(x) + z_i^2 = 0$, before then solving the ODE (3.1.3) in the augmented space $(x, z) \in V \times \mathbb{R}^q$. This approach offers convergence guarantees [277] and could also be beneficial for shape optimization, however this is not the strategy we have retained. Indeed, our method does not need to resort to slack variables for handling inequality constraints, and it present small additional advantages described in section 3.5.1.

Another important contribution of this chapter is the exposure of our dynamical system strategy in a setting compatible with the inherently infinite dimensional aspect of shape optimization based on the method of Hadamard. In such context where the optimization variable x belongs to a Hilbert space V (if not a more general set, see below), this is achieved by making a clear distinction made between the Fréchet derivative $\mathrm{D}J(x(t))$ (which is an element of the dual space V') and the gradient $\nabla J(x(t))$ (which is an element of V): the gradient $\nabla J(x(t))$ is obtained by identification of $\mathrm{D}J(x(t))$ with an element of V thanks to the Riesz representation theorem. The same distinction is also needed between the differential of a vector valued function $\mathrm{D}\boldsymbol{C}(x(t))$ and its transposition $\mathrm{D}\boldsymbol{C}(x(t))^{\mathcal{T}}$.

Several works in the field of shape and topology optimization can be related to ours. In fact, our method is very close in spirit to the recent work of Barbarosie et. al. [60], who derived an iterative algorithm for equality constrained optimization which turns out to be a discretization of (3.1.3) with a variable scaling for the parameter α_C. For inequality constraints, the authors proposed (without convergence results) an active set strategy also based on the extraction of an appropriate subset of the active constraints. However their method relies on a different algorithm from ours, that yields generally a different (suboptimal) set than $\widehat{I}(x)$, see remark 3.7 below for more details. Yulin and Xiaoming [322] also suggested to project the gradient of the objective function onto the convex cone of feasible directions; nevertheless, they remained elusive regarding how the projection problem should be solved or how violated constraints should be tackled.

Finally, let us mention as well that the discretization of the flow (3.1.3) can be related to non dynamical system approaches for constrained optimization such as SQP methods (see [277] for a comparison with the ODE approach) and to null space iterative methods [69, 70, 244].

The present chapter is organized as follows. section 3.2 introduces useful notation for distinguishing gradient and differentials for constrained optimization on Hilbert spaces V. We review as well a few motivations at the origin of the definitions of $\boldsymbol{\xi}_J(x(t))$ and $\boldsymbol{\xi}_C(x(t))$ (in (3.1.5) and (3.1.6)) in the case where inequality constraints are absent. In this context, the properties of the flow (3.1.3), classical when

the minimization set V is a finite dimensional vector space, are reviewed for the more general context where V is a Hilbert space. We detail then in section 3.3 the necessary adaptations to account for inequality constraints and in particular the introduction of the dual subproblem allowing to determine the null space direction $\boldsymbol{\xi}_J(x)$. Under some non restrictive technical assumptions, we prove in proposition 3.5 decreasing properties for the trajectories of the "null space" gradient flow (3.1.3) towards points satisfying the Karush, Kuhn and Tucker (KKT) condition. Numerical implementation and discretization aspects are detailed in section 3.4. Section 3.5 provides pedagogical illustrations of our method on simple academic test cases, and comparisons to a few other optimization methods of the literature including the method of slack variables for inequality constraints. Shape optimization applications are eventually considered in section 3.6. After clarifying the necessary adaptations required to extend the discretization of (3.1.3) to sets featuring a rather generic manifold structure, we explain how our algorithm can be integrated within the level set method for shape optimization [311, 32, 24]. The ease of implementation and efficiency of our method is demonstrated numerically for the optimal design of a bridge structure subject to multiple loads, which involves up to ten constraints.

3.2 Null space gradient flows for equality-constrained optimization in Hilbert spaces

In this section, we consider the case where the optimization takes place on a Hilbert space V with inner product $\langle\cdot,\cdot\rangle_V$ and relative norm $||\cdot||_V = \langle\cdot,\cdot\rangle_V^{1/2}$; see section 3.6 for the description of the more general situation associated to our shape optimization applications. In this part only, we consider the problem (3.1.2) where only equality constraints are present, namely:

$$\begin{aligned} &\min_{x\in V} \quad J(x) \\ &\text{s.t. } \boldsymbol{g}(x) = 0, \end{aligned} \tag{3.2.1}$$

where $J : V \to \mathbb{R}$ and $\boldsymbol{g} : V \to \mathbb{R}^p$ are Fréchet differentiable functions. The purpose of this section is to motivate the introduction of the ODE (3.1.3) for equality constrained optimization, and to review its properties in the present Hilbertian setting. Let us emphasize that, although this section is elementary and not completely new, it is not easily found as is in the literature. Since it is key in understanding our technique for handling inequality constraints in section 3.3, the present context is thoroughly detailed for the reader's convenience.

The section is organized as follows. Section 3.2.1 recalls definitions for the differential, the gradient, and the transpose operation in the Hilbertian context. We then sketch briefly in section 3.2.2 how the formulas (3.1.5) and (3.1.6) can be formally obtained, before stating the properties of the null space step $\boldsymbol{\xi}_J(x)$ and its relation to Lagrange multipliers by means of a dual problem in lemma 3.1. Finally, the decrease properties of the obtained dynamical system are reviewed in section 3.2.3.

3.2.1 Notation and first-order optimality conditions

The following definition sets the notation conventions about differentiability and gradients in Hilbert spaces used throughout this chapter. Note that they may differ from those used by other authors because a clear distinction is needed for our shape optimization purposes between gradient and Fréchet derivatives.

Definition 3.1. 1. A vector-valued function $\boldsymbol{g} : V \to \mathbb{R}^p$ is *differentiable* at a point $x \in V$ if there exists a continuous linear mapping $\mathrm{D}\boldsymbol{g}(x) : V \to \mathbb{R}^p$ such that

$$\boldsymbol{g}(x+h) = \boldsymbol{g}(x) + \mathrm{D}\boldsymbol{g}(x)h + o(h) \text{ with } \frac{o(h)}{||h||_V} \xrightarrow{h\to 0} 0. \tag{3.2.2}$$

$\mathrm{D}\boldsymbol{g}(x)$ is called the Fréchet derivative of $\boldsymbol{g}$ at x.

2. If $\boldsymbol{g} : V \to \mathbb{R}^p$ is differentiable, for any $\boldsymbol{\mu} \in \mathbb{R}^p$, the Riesz representation theorem [76] ensures the existence of a unique vector $\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\mu} \in V$ satisfying

$$\forall \boldsymbol{\mu} \in \mathbb{R}^p, \forall \boldsymbol{\xi} \in V, \ \langle \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\mu}, \boldsymbol{\xi}\rangle_V = \boldsymbol{\mu}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}, \tag{3.2.3}$$

where the superscript ^{T} stands for the usual transpose of a vector in the Euclidean space $\mathbb{R}^p$. The linear operator $\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} : \mathbb{R}^p \to V$ thus defined is called the *transpose* of $\mathrm{D}\boldsymbol{g}(x)$.

3. If $J : V \to \mathbb{R}$ is a scalar function differentiable at $x \in V$, the Riesz representation theorem ensures the existence of a unique vector $\nabla J(x) \in V$ satisfying

$$\forall \boldsymbol{\xi} \in V,\ \langle \nabla J(x), \boldsymbol{\xi} \rangle_V = \mathrm{D}J(x)\boldsymbol{\xi}. \tag{3.2.4}$$

This vector $\nabla J(x)$ is called the *gradient* of J at x.

Throughout the chapter, the explicit mention to x is sometimes omitted in the notation for differentials or gradients when the considered point $x \in V$ is clear.

Remark 3.1.

1. If V is the (finite-dimensional) Euclidean space $\mathbb{R}^k$, equipped with the standard inner product, the Fréchet derivative and the transpose of the differential of a vector valued function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^p$ are respectively given by the Jacobian matrix $(\mathrm{D}\boldsymbol{g})_{ij} = \partial_j g_i$ and its transpose $(\mathrm{D}\boldsymbol{g}^{\mathcal{T}})_{ij} = (\mathrm{D}\boldsymbol{g}^T)_{ij} = \partial_i g_j$. In the literature, the differential matrix $\mathrm{D}\boldsymbol{g}$ is often denoted with the nabla symbol $\nabla \boldsymbol{g}$. For the sake of clarity, we reserve the ∇ symbol to denote the gradient of scalar functions $J : V \to \mathbb{R}$, and it holds that $\nabla J(x) = \mathrm{D}J(x)^{\mathcal{T}}1$. The calligraphic transpose notation $\mathcal{T}$ appearing in the objects $\mathrm{D}J(x)^{\mathcal{T}}$ or $\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}$ encodes at the same time the operator transposition (reversing the input and range spaces) and the Riesz identifications.

2. Still in the case where $V = \mathbb{R}^k$ is finite-dimensional and a is given by a symmetric definite positive matrix A (that is $\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_V = \boldsymbol{\xi}^T A \boldsymbol{\xi}$), the transpose of a $p \times k$ matrix $M : \mathbb{R}^k \to \mathbb{R}^p$ with respect to a is $M^{\mathcal{T}} = A^{-1} M^T$. As we shall see in section 3.6, in shape optimization applications, $\langle , \rangle_V$ often stands for the bilinear form associated to an elliptic operator, hence the calligraphic transpose $\mathcal{T}$ encompasses the extension and regularization step of the shape derivative outlined in section 1.4.1. If V is the tangent space to some Riemannian manifold, the inner product $\langle , \rangle_V$ can be interpreted as a metric and $\nabla J(x)$, as given by (3.2.4), is the covariant gradient with respect to this metric.

3. When V is a general Hilbert space, for a vector-valued function $\boldsymbol{g} : V \to \mathbb{R}^p$ with coordinates $\boldsymbol{g}(x) = (g_i(x))_{1 \leq i \leq p}$, $\mathrm{D}\boldsymbol{g} : V \to \mathbb{R}^p$ is the 'row' matrix whose entries are the p linear forms $\mathrm{D}g_i(x) : V \to \mathbb{R}$. The transpose $\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}$ is the 'column' matrix gathering the p vectors $(\nabla g_i(x))_{1 \leq i \leq p}$ obtained by solving the p identification problems:

$$\forall \boldsymbol{\xi} \in V,\ \langle \nabla g_i(x), \boldsymbol{\xi} \rangle_V = \mathrm{D}g_i(x)\boldsymbol{\xi}; \tag{3.2.5}$$

more precisely:

$$\forall \boldsymbol{\mu} \in \mathbb{R}^d,\ \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\mu} = \sum_{i=1}^{p} \mu_i \nabla g_i(x).$$

In particular, the $p \times p$ matrix $\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}} \in \mathbb{R}^{p \times p}$ has entries

$$(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})_{ij} = \langle \nabla g_i, \nabla g_j \rangle_V = \mathrm{D}g_i(x)(\nabla g_j(x)).$$

Throughout this section, the equality constraints defined by the function $\boldsymbol{g}$ are said to be qualified at a point $x \in V$ if

$$\mathrm{rank}(\mathrm{D}\boldsymbol{g}(x)) = p, \text{ or equivalently } \mathrm{D}\boldsymbol{g}(x)\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \text{ is an invertible } p\text{-by-}p \text{ matrix.} \tag{3.2.6}$$

Note that (3.2.6) makes sense even at points $x \in V$ where $\boldsymbol{g}(x) \neq 0$, a fact that we shall use extensively in the sequel.

Let us then recall the classical definitions of critical points in terms of the first-order necessary optimality conditions (KKT) for the equality-constrained problem (3.2.1):

Definition 3.2 (see [70, 244])**.** A point x^* is said to satisfy the Karush, Kuhn and Tucker conditions (KKT) for the equality constrained minimization problem (3.2.1) if and only if there exists $\boldsymbol{\lambda}(x^*) \in \mathbb{R}^p$ such that:

$$\begin{cases} \nabla J(x^*) + \mathrm{D}\boldsymbol{g}(x^*)^{\mathcal{T}}\boldsymbol{\lambda}(x^*) = 0, \\ \boldsymbol{g}(x^*) = 0. \end{cases} \tag{3.2.7}$$

3.2.2 Definitions and properties of the null space and range space steps $\boldsymbol{\xi}_J$ and $\boldsymbol{\xi}_C$

We are now in position to define the null space and range space steps $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ featured in the ODE (3.1.11) for equality constrained problems in the present Hilbert space setting, owing to the definition (3.2.3) of the transpose:

Definition 3.3. For any point $x \in V$ satisfying the constraint qualification condition (3.2.6), the *null space* and *range space* directions $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ associated to (3.2.1) are respectively defined by:

$$\boldsymbol{\xi}_J(x) := (I - \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g})\nabla J(x), \tag{3.2.8}$$

$$\boldsymbol{\xi}_C(x) := \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x). \tag{3.2.9}$$

Let us provide a formal intuition motivating the expressions (3.2.8), (3.2.9) and the dynamical system (3.1.3). A classical practice in Lagrange multiplier methods for optimization which is inspired from the KKT optimality conditions (3.2.7) is to search for an iterative optimization scheme (indexed by the iteration number $_n$) of the form (see also [59])

$$x_{n+1} = x_n - \Delta t(\alpha_J \nabla J(x_n) + \mathrm{D}\boldsymbol{g}(x_n)^{\mathcal{T}}\boldsymbol{\lambda}_n), \tag{3.2.10}$$

where $\boldsymbol{\lambda}_n \in \mathbb{R}^p$ is a tentative value for the Lagrange multiplier $\boldsymbol{\lambda}$ in (3.2.7), α_J is a user-defined coefficient and Δt is the step increment between successive iterations. We determine the value of $\boldsymbol{\lambda}_n$ by imposing that the value $\boldsymbol{g}(x_{n+1})$ of the constraint decrease by a factor $(1 - \alpha_C \Delta t)$, up to some lower order term in Δt. Since

$$\boldsymbol{g}(x_{n+1}) = \boldsymbol{g}(x_n) - \Delta t \mathrm{D}\boldsymbol{g}(x_n)(\alpha_J \nabla J(x_n) + \mathrm{D}\boldsymbol{g}(x_n)^{\mathcal{T}}\boldsymbol{\lambda}_n) + o(\Delta t),$$

the requirement that $\boldsymbol{g}(x_{n+1}) \simeq (1 - \alpha_C \Delta t)\boldsymbol{g}(x_n)$ suggests the rule:

$$\boldsymbol{\lambda}_n = (\mathrm{D}\boldsymbol{g}(x_n)\mathrm{D}\boldsymbol{g}(x_n)^{\mathcal{T}})^{-1}\left(\alpha_C \boldsymbol{g}(x_n) - \alpha_J \mathrm{D}\boldsymbol{g}(x_n)\nabla J(x_n)\right). \tag{3.2.11}$$

We recognize then a time discretization scheme of the ODE (3.1.11) by replacing $\boldsymbol{\lambda}_n$ with the above value (3.2.11) in (3.2.10).

Properties of the null space step $\boldsymbol{\xi}_J$

In the finite-dimensional case where $V = \mathbb{R}^k$, it is well-known that the null space step $\boldsymbol{\xi}_J(x)$ defined by (3.2.8) is the orthogonal projection of the gradient $\nabla J(x)$ of the objective function onto the null space of the constraints

$$\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x)) = \{\boldsymbol{\xi} \in V \,|\, \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = 0\},$$

which is also the tangent space at x to the manifold $\{y \in V \,|\, \boldsymbol{g}(y) = \boldsymbol{g}(x)\}$. Of course, this is still true when V is a Hilbert space, as recalled in the next lemma.

Lemma 3.1. *Let $x \in V$ be a point satisfying the qualification condition (3.2.6) The following properties hold:*

1. *The space V has the following orthogonal decomposition:*

$$V = \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x)) \oplus \mathrm{Ran}(\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}),$$

where we have introduced the range $\mathrm{Ran}(\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}) := \{\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda} \,|\, \boldsymbol{\lambda} \in \mathbb{R}^p\}$ of $\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}$.

Moreover, the operator $\Pi_{g(x)} : V \to V$ defined by

$$\Pi_{g(x)} = I - \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g}(x) \tag{3.2.12}$$

is the orthogonal projection onto $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$.

2. *When $\Pi_{g(x)}(\nabla J(x)) \neq 0$, $-\boldsymbol{\xi}_J(x) = -\Pi_{g(x)}(\nabla J(x))$ is the best normalized feasible descent direction for J in the sense that*

$$-\frac{\boldsymbol{\xi}_J(x)}{||\boldsymbol{\xi}_J(x)||_V} = \arg\min_{\boldsymbol{\xi} \in V} \quad \mathrm{D}J(x)\boldsymbol{\xi} \quad \text{s.t.} \begin{cases} \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = 0 \\ \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_V \leq 1. \end{cases} \tag{3.2.13}$$

3. *The null space direction* $\boldsymbol{\xi}_J(x) = \Pi_{g(x)}(\nabla J(x))$ *is the closest least squares approximation to* $\nabla J(x)$ *within the space* $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$. *It alternatively reads*

$$\boldsymbol{\xi}_J(x) = \nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}^*(x), \tag{3.2.14}$$

where the Lagrange multiplier $\boldsymbol{\lambda}^*(x) := -(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g}\nabla J(x)$ *is the unique solution to the following least squares problem that is the dual of* (3.2.13)*:*

$$\boldsymbol{\lambda}^*(x) = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}||_V. \tag{3.2.15}$$

Remark 3.2. The qualification condition (3.2.6) is essentially used in the computation of the orthogonal projection onto $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$ from formula (3.2.12). Most of the statements of this chapter could circumvent this hypothesis by relying on a Singular Value Decomposition (SVD) of $\mathrm{D}\boldsymbol{g}(x)$.

Proof. 1. Any $\boldsymbol{\xi} \in V$ may be decomposed as $\boldsymbol{\xi} = \Pi_{g(x)}(\boldsymbol{\xi}) + (I - \Pi_{g(x)})(\boldsymbol{\xi})$, where it is straightforward to verify that $\Pi_{g(x)}(\boldsymbol{\xi}) \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$, and $(I - \Pi_{g(x)})(\boldsymbol{\xi}) \in \mathrm{Ran}(\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}})$. In addition, $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$ and $\mathrm{Ran}(\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}})$ are orthogonal for the inner product $\langle\cdot,\cdot\rangle_V$ since from (3.2.3), one has,

$$\forall \boldsymbol{\zeta} \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x)),\ \forall \boldsymbol{\lambda} \in \mathbb{R}^p,\ \ \langle \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}, \boldsymbol{\zeta}\rangle_V = \boldsymbol{\lambda}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\zeta} = 0.$$

2. It follows from the first point that for any $\boldsymbol{\xi} \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$ such that $||\boldsymbol{\xi}||_V \leq 1$,

$$\mathrm{D}J(x)\boldsymbol{\xi} = \langle \nabla J(x), \boldsymbol{\xi}\rangle_V = \langle \Pi_{g(x)}(\nabla J(x)), \boldsymbol{\xi}\rangle_V \geq -||\Pi_{g(x)}(\nabla J(x))||_V,$$

whence we easily infer that $\boldsymbol{\xi} := -\Pi_{g(x)}(\nabla J(x))/||\Pi_{g(x)}(\nabla J(x))||_V$ is the global minimizer of (3.2.13).

3. The Pythagore identity yields, for any $\boldsymbol{\xi} \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$,

$$||\nabla J(x) - \boldsymbol{\xi}||_V^2 = ||(I - \Pi_{g(x)})\nabla J(x)||_V^2 + ||\Pi_{g(x)}\nabla J(x) - \boldsymbol{\xi}||_V^2 \geq ||\nabla J(x) - \Pi_{g(x)}\nabla J(x)||_V^2.$$

Hence the orthogonal projection $\Pi_{g(x)}(\nabla J(x))$ is the best approximation of $\nabla J(x)$ within the space $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$. Recalling from the first point that the range $\mathrm{Ran}(\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}})$ is the orthogonal complement of $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$, we obtain also, for any $\boldsymbol{\lambda} \in \mathbb{R}^p$,

$$||\Pi_{g(x)}(\nabla J(x))||_V = ||\nabla J(x) - (I - \Pi_{g(x)})(\nabla J(x))||_V \leq ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}||_V,$$

whence the expression (3.2.14) and the minimization property (3.2.15) follow. Note that the uniqueness of the solution $\boldsymbol{\lambda}^*(x)$ to (3.2.15) results from the qualification condition (3.2.6).

Finally, the optimization problem (3.2.13) can be rewritten as

$$\min_{\substack{\boldsymbol{\xi}\in V \\ \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_V \leq 1}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^p} \ \mathrm{D}J(x)\boldsymbol{\xi} + \boldsymbol{\lambda}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}.$$

Hence the (formal) dual problem of (3.2.13) reads:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^p} \min_{\substack{\boldsymbol{\xi}\in V \\ \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_V \leq 1}} \mathrm{D}J(x)\boldsymbol{\xi} + \boldsymbol{\lambda}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}.$$

According to the definitions (3.2.3) and (3.2.4) of the gradient and of the Hilbertian transpose, the latter problem rewrites:

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^p} \min_{\substack{\boldsymbol{\xi}\in V \\ \langle\boldsymbol{\xi},\boldsymbol{\xi}\rangle_V \leq 1}} \langle \nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}, \boldsymbol{\xi}\rangle_V = -\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} ||\nabla J + \mathrm{D}\boldsymbol{g}^{\mathcal{T}}\boldsymbol{\lambda}||_V,$$

where for given $\boldsymbol{\lambda} \in \mathbb{R}^p$, the value

$$\boldsymbol{\xi}^* := -\frac{\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}}{||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}||_V}$$

is that achieving the minimum in the minimization problem at the left-hand side of the above identity. This shows that (3.2.15) is the dual problem of (3.2.13).

□

Properties of the range space step $\boldsymbol{\xi}_C$

The next lemma characterizes the range space step $\boldsymbol{\xi}_C(x)$, defined by (3.2.9), as the unique Gauss-Newton direction for the minimization of the constraint function $\boldsymbol{g}(x)$ which is tangential to the (linearized) set of constraints:

Lemma 3.2. *Let $x \in V$ satisfy the qualification condition (3.2.6); then:*

1. *The range space step $\boldsymbol{\xi}_C(x) = \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x)$ is orthogonal to $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$:*

$$\forall \boldsymbol{\xi} \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x)),\ \ \langle \boldsymbol{\xi}_C(x), \boldsymbol{\xi} \rangle_V = 0.$$

2. *$-\boldsymbol{\xi}_C(x)$ is a descent direction for the violation of the constraints:*

$$\mathrm{D}\boldsymbol{g}(x)(-\boldsymbol{\xi}_C(x)) = -\boldsymbol{g}(x). \tag{3.2.16}$$

3. *The set of solutions to the Gauss-Newton program*

$$\min_{\boldsymbol{\xi} \in V} ||\boldsymbol{g}(x) + \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}||^2 \tag{3.2.17}$$

is the affine subspace $\{-\boldsymbol{\xi}_C(x) + \boldsymbol{\zeta} \mid \boldsymbol{\zeta} \in \mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))\}$ of V.

Proof. 1. This easily follows from the point 1. of lemma 3.1.

2. This is an immediate consequence of the definition (3.2.9) of $\boldsymbol{\xi}_C(x)$. Note that (3.2.16) means that $-\boldsymbol{\xi}_C(x)$ is a descent direction for the violation of the constraints in the sense that it ensures that any coordinate $g_i(x)$, $i = 1, ..., p$, decreases along $-\boldsymbol{\xi}_C(x)$ if $g_i(x) \geq 0$ and increases if $g_i(x) \leq 0$.

3. Since (3.2.17) is a convex optimization problem, a necessary and sufficient condition for $\boldsymbol{\xi} \in V$ to be one solution is given by the usual first-order condition:

$$\forall \boldsymbol{\zeta} \in V,\ \ (\boldsymbol{g}(x) + \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi})^T(\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\zeta}) = \langle \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}(\boldsymbol{g}(x) + \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}), \boldsymbol{\zeta} \rangle_V = 0,$$

which rewrites:

$$\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = -\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{g}(x).$$

Since the matrix $(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})$ is invertible (as a consequence of the qualification condition (3.2.6)), this is in turn equivalent to:

$$\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = -\boldsymbol{g}(x).$$

Finally, (3.2.16) states that $-\boldsymbol{\xi}_C(x)$ is one particular solution to the above equation; therefore, any two solutions of this problem differ by some $\boldsymbol{\zeta}$ such that $\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\zeta} = 0$.

□

3.2.3 Decrease properties of the equality constrained gradient flow

The main features of the definitions of $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ are the facts that $\boldsymbol{\xi}_J$ is orthogonal to the set of constraints, i.e. $\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}_J(x) = 0$, and that $-\boldsymbol{\xi}_C(x)$ decreases the violation of the constraints while being orthogonal to $\boldsymbol{\xi}_J(x)$. These ensure that the values of the constraint functional $\boldsymbol{g}(x(t))$ decrease to zero along the trajectories of the ODE (3.1.3), independently of the behavior of $\boldsymbol{\xi}_J(x)$. Then, as soon as the violation of the constraint becomes sufficiently small, the objective function J decreases without affecting the asymptotic vanishing of $\boldsymbol{g}(x(t))$. We review these properties in the next proposition, which was also observed in [317] in the finite-dimensional context.

Proposition 3.1. *Assume that the trajectories $x(t)$ of the flow*

$$\begin{cases} \dot{x} = -\alpha_J(I - \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g}(x))\nabla J(x) - \alpha_C \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x) \\ x(0) = x_0 \end{cases} \tag{3.2.18}$$

exist on some time interval $[0, T]$ for $T > 0$, and that the qualification condition (3.2.6) holds at any point $x(t)$, $t \in [0, T]$. Then the following properties hold true:

1. *The violation of the constraints decreases exponentially:*

$$\forall t \in [0, T],\ \boldsymbol{g}(x(t)) = e^{-\alpha_C t}\boldsymbol{g}(x_0). \tag{3.2.19}$$

2. *$J(x(t))$ decreases 'as soon as the violation (3.2.19) of the constraints is sufficiently small' in the following sense: assume that* $\mathrm{rank}(\mathrm{D}\boldsymbol{g}) = p$ *on* $K = \{x \in V \,|\, ||\boldsymbol{g}(x)||_\infty \leq ||\boldsymbol{g}(x_0)||_\infty\}$ *and that*

$$\sup_{x\in K} ||\nabla J(x)||_V |\sigma_p^{-1}(x)| < +\infty, \tag{3.2.20}$$

where $\sigma_p(x)$ is the smallest singular value of $\mathrm{D}\boldsymbol{g}(x)$. Then there exists a constant $C > 0$ such that

$$\forall t \in [0,T],\ ||\Pi_{g(x)}(\nabla J(x(t)))||_V^2 > Ce^{-\alpha_C t} \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t} J(x(t)) < 0. \tag{3.2.21}$$

3. *Any stationary point x^* of (3.2.18) satisfies the first-order KKT conditions (3.2.7) of the optimization program (3.2.1), that is:*

$$\begin{cases} \boldsymbol{g}(x^*) = 0 \\ \exists \boldsymbol{\lambda}^* \in \mathbb{R}^p,\ \nabla J(x^*) + \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(x^*)\boldsymbol{\lambda}^* = \Pi_{g(x^*)}(\nabla J(x^*)) = 0. \end{cases} \tag{3.2.22}$$

Proof. 1. Using the definition (3.2.18), the decreasing property (3.2.16) together with the fact that $\boldsymbol{\xi}_J(x)$ is orthogonal to $\mathrm{Ker}(\mathrm{D}\boldsymbol{g}(x))$, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{g}(x(t))) = -\alpha_C \boldsymbol{g}(x(t)),$$

whence (3.2.19) follows easily.

2. Let us introduce the eigenvalue decomposition

$$\mathrm{D}\boldsymbol{g}(x)\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} = \sum_{i=1}^{p} \sigma_i(x)^2 \boldsymbol{u}_i(x)\boldsymbol{u}_i(x)^T, \text{ where } \sigma_1(x) \geq \ldots \geq \sigma_p(x) > 0,\ \boldsymbol{u}_i(x)^T\boldsymbol{u}_j(x) = \delta_{ij},$$

of the symmetric, positive definite $p \times p$ matrix $\mathrm{D}\boldsymbol{g}(x)\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}$. Let then $\boldsymbol{v}_i(x)^\dagger : V \to \mathbb{R}$ be the linear form defined for any $\boldsymbol{\xi} \in V$ by $\boldsymbol{v}_i(x)^\dagger\boldsymbol{\xi} = \sigma_i(x)^{-1}\boldsymbol{u}_i(x)^T\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}$ and let $\boldsymbol{v}_i(x)$ be the vector in V such that $\forall \boldsymbol{\xi} \in V$, $\langle \boldsymbol{v}_i(x), \boldsymbol{\xi}\rangle_V = \boldsymbol{v}_i(x)^\dagger\boldsymbol{\xi}$; more explicitly, $\boldsymbol{v}_i(x) = \sigma_i(x)^{-1}\mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{u}_i(x)$. These definitions allow to write a singular value decomposition for $\mathrm{D}\boldsymbol{g}(x)$; it is indeed easily verified from the definitions of $\boldsymbol{u}_i(x)$ and $\boldsymbol{v}_i(x)$ that:

$$\mathrm{D}\boldsymbol{g}(x) = \sum_{i=1}^{p} \sigma_i(x)\boldsymbol{u}_i(x)\boldsymbol{v}_i(x)^\dagger, \text{ and } \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} = \sum_{i=1}^{p} \sigma_i(x)\boldsymbol{v}_i(x)\boldsymbol{u}_i(x)^T$$

with $\langle \boldsymbol{v}_i(x), \boldsymbol{v}_j(x)\rangle_V = \boldsymbol{v}_i(x)^\dagger\boldsymbol{v}_j(x) = \delta_{ij}$. We now calculate:

$$\mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x) = \sum_{i=1}^{p} \sigma_i^{-1}(x)(\boldsymbol{u}_i(x)^T\boldsymbol{g}(x))\boldsymbol{v}_i(x),$$

whence we obtain the following inequality:

$$\forall x \in V,\ |\mathrm{D}J(x)\mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}| \leq \sigma_p^{-1}(x)||\nabla J(x)||_V\, ||\boldsymbol{g}(x)||. \tag{3.2.23}$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}t} J(x(t)) = -\alpha_J \mathrm{D}J(x(t))\boldsymbol{\xi}_J(x(t)) - \alpha_C \mathrm{D}J(x(t))\boldsymbol{\xi}_C(x(t)),$$

it follows that $\frac{\mathrm{d}}{\mathrm{d}t}J(x(t)) < 0$ as soon as $\alpha_J|\mathrm{D}J(x(t))\boldsymbol{\xi}_J(x(t))| > \alpha_C|\mathrm{D}J(x(t))\boldsymbol{\xi}_C(x(t))|$. Thus, from (3.2.19) and (3.2.23), the constant C in (3.2.21) can be selected as

$$C = p\frac{\alpha_C}{\alpha_J}||\boldsymbol{g}(x_0)|| \sup_{x\in K}\left[\sigma_p^{-1}(x)||\nabla J(x)||_V\right]. \tag{3.2.24}$$

3. Since the vectors $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ are orthogonal for any point $x \in V$, a stationary point x^* of (3.2.18) must satisfy

$$\Pi_{g(x^*)}(\nabla J(x^*)) = 0, \text{ and } \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{g}(x^*) = 0, \tag{3.2.25}$$

and so the first KKT condition in (3.2.7) is satisfied with the value $\boldsymbol{\lambda} = -(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{g}(x^*)\nabla J(x^*)$ of the Lagrange multiplier. Then left multiplication by $\mathrm{D}\boldsymbol{g}$ in the second identity in (3.2.25) implies $\boldsymbol{g}(x^*) = 0$, which completes the proof.

□

Remark 3.3. The solutions to the dynamical system (3.2.18) are defined for small time if $\boldsymbol{\xi}_J$ and $\boldsymbol{\xi}_C$ are locally Lipschitz vector fields, which is the case if e.g. J and $\boldsymbol{g}$ are of class $\mathcal{C}^2$ [127]. In the case where V is finite-dimensional, the assumption (3.2.20) is satisfied if the set $K = \{x \in V | \boldsymbol{g}(x) \leq \boldsymbol{g}(x_0)\}$ is bounded and the functions J and $\boldsymbol{g}$ are $\mathcal{C}^1$ functions. It is worth noting that even if the present regularity of J and $\boldsymbol{g}$ is not strong enough to ensure the existence of solutions to (3.2.18), similar properties to those of proposition 3.1 hold for the discretized scheme

$$x_{n+1} = x_n - \Delta t(\alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n)), \tag{3.2.26}$$

which is sufficient for optimization. One can indeed verify that:

1. At first order, the constraints decrease with a geometric rate: $\boldsymbol{g}(x_{n+1}) = (1 - \alpha_C \Delta t)\boldsymbol{g}(x_n) + o(\Delta t)$.
2. An accumulation point x^* of the sequence $(x_n)_{n\in\mathbb{N}}$ satisfies $\boldsymbol{g}(x^*) = 0$ and is a KKT point of the problem (3.2.1), satisfying (3.2.7).

Remark 3.4. It is possible to control more accurately the pace at which each of the constraints decreases in (3.2.19): consider a diagonal matrix of positive coefficients $\boldsymbol{K} = \mathrm{diag}(\kappa_i)_{1\leq i\leq p}$ and replace the definition (3.2.9) of $\boldsymbol{\xi}_C(x)$ by

$$\boldsymbol{\xi}_C(x) := \mathrm{D}\boldsymbol{g}^{\mathcal{T}}(\mathrm{D}\boldsymbol{g}\mathrm{D}\boldsymbol{g}^{\mathcal{T}})^{-1}\boldsymbol{K}\boldsymbol{g}(x).$$

Then each constraint function g_i decreases at its own rate $\kappa_i\alpha_C$ along the solution $x(t)$ of (3.2.18):

$$\forall t \in [0, T],\ g_i(x(t)) = e^{-\kappa_i\alpha_C t} g_i(x_0).$$

3.3 Extension to equality and inequality constraints

We now proceed to extend the dynamical system (3.1.3) or (3.2.18) so as to handle inequality constraints as well. We return to the full optimization problem (3.1.2), still posed in a Hilbert space V with inner product $\langle\cdot,\cdot\rangle_V$, and where the objective $J : V \to \mathbb{R}$, equality constraints $\boldsymbol{g} : V \to \mathbb{R}^p$ and inequality constraints $\boldsymbol{h} : V \to \mathbb{R}^q$ are differentiable functions.

Inspired by the methodology developed in section 3.2, we still propose to solve the equality and inequality constrained problem (3.1.2) thanks to a dynamical system of the form:

$$\begin{cases} \dot{x}(t) = -\alpha_J \boldsymbol{\xi}_J(x(t)) - \alpha_C \boldsymbol{\xi}_C(x(t)) \\ x(0) = x_0, \end{cases} \tag{3.3.1}$$

whose discretized version reads:

$$x_{n+1} = x_n - \Delta t(\alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n)). \tag{3.3.2}$$

In what follows, notation conventions related to index sets associated to inequality constraints are introduced in section 3.3.1. The range space step $\boldsymbol{\xi}_C(x)$ is defined in section 3.3.2 from a formula analogous to (3.1.6). The definition of the null space step $\boldsymbol{\xi}_J(x)$ is examined in details in section 3.3.3; it involves a procedure discriminating a relevant subset $\widehat{I}(x) \subset \widetilde{I}(x)$ of the saturated or violated constraints, which relies on the introduction of the dual problem (3.1.8). Finally, the properties of the flow (3.3.1) are outlined in section 3.3.4.

3.3.1 Notation and preliminaries

For the convenience of the reader, a few notation conventions from the introduction are now made more precise. The set of indices of saturated or violated inequality constraints at $x \in V$ is denoted by $\widetilde{I}(x)$:

$$\widetilde{I}(x) = \{i \in \{1, \dots, q\} \,|\, h_i(x) \geq 0\}, \tag{3.3.3}$$

and $\widetilde{q}(x) := \mathrm{Card}(\widetilde{I}(x))$ is the number of such constraints. For a subset $I \subset \{1, ..., q\}$, the vector $\boldsymbol{h}_I(x) = (h_i(x))_{i\in I}$ collects the inequality constraints indexed by I and $\boldsymbol{C}_I(x)$, defined by (3.1.4), collects all equality constraints $\boldsymbol{g}(x)$ and those selected inequality constraints $\boldsymbol{h}_I(x)$.

In the present context, the constraints are said to be qualified at $x \in V$ if the linearized saturated or violated constraints are independent, that is,

$$\mathrm{rank}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}(x)) = p + \widetilde{q}(x). \tag{3.3.4}$$

If the point x satisfies the constraints, (3.3.4) is one usual qualification condition (of course, there are other possible qualification conditions, see [70, 244]), but let us insist again that we leave the room for this definition to apply to points x where constraints are not satisfied. Define $\Pi_{\boldsymbol{C}_I} : V \to V$, the orthogonal projection operator onto $\mathrm{Ker}(\mathrm{D}\boldsymbol{C}_I(x))$, by

$$\Pi_{\boldsymbol{C}_I} = I - \mathrm{D}\boldsymbol{C}_I(x)^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_I(x)\mathrm{D}\boldsymbol{C}_I(x)^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}_I(x), \tag{3.3.5}$$

and let $(\boldsymbol{\lambda}_I(x), \boldsymbol{\mu}_I(x)) \in \mathbb{R}^p \times \mathbb{R}_+^{\mathrm{Card}(I)}$ be the corresponding Lagrange multipliers:

$$\begin{bmatrix} \boldsymbol{\lambda}_I(x) \\ \boldsymbol{\mu}_I(x) \end{bmatrix} := -(\mathrm{D}\boldsymbol{C}_I\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}_I(x)\nabla J(x). \tag{3.3.6}$$

Last but not least, let us recall the necessary first-order optimality conditions (the KKT conditions) for equality and inequality constrained problems:

Definition 3.4 (KKT conditions, [70, 244])**.** A point $x^* \in V$ is said to satisfy the Karush, Kuhn and Tucker conditions for (3.1.2) if and only if there exist $\boldsymbol{\lambda}(x^*) \in \mathbb{R}^p$ and $\boldsymbol{\mu}(x^*) \in \mathbb{R}_+^q$ such that:

$$\begin{cases} \nabla J(x^*) + \mathrm{D}\boldsymbol{g}(x^*)^{\mathcal{T}}\boldsymbol{\lambda}(x^*) + \mathrm{D}\boldsymbol{h}(x^*)^{\mathcal{T}}\boldsymbol{\mu}(x^*) = 0, \\ \boldsymbol{g}(x^*) = 0, \;\; \boldsymbol{h}(x^*) \leq 0, \\ \forall i = 1, ..., q, \;\; \mu_i h_i(x^*) = 0. \end{cases} \tag{3.3.7}$$

3.3.2 Definition of the range space step

Definition 3.5 (range space step)**.** The range step $\boldsymbol{\xi}_C(x)$ associated with the optimization problem (3.1.2) is defined by

$$\boldsymbol{\xi}_C(x) := \mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}^{\mathcal{T}})^{-1}\boldsymbol{C}_{\widetilde{I}(x)}(x), \tag{3.3.8}$$

where $\widetilde{I}(x)$ is the subset of saturated or violated constraints, defined by (3.3.3).

The purpose of the range space step $\boldsymbol{\xi}_C(x)$ is to decrease the violation of the constraints as we shall see in proposition 3.5 below. The counterpart of lemma 3.2 holds exactly in this context, in particular:

1. $\boldsymbol{\xi}_C(x)$ is orthogonal to $\mathrm{Ker}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)})$.
2. $-\boldsymbol{\xi}_C(x)$ is a Gauss-Newton direction for the violation of the constraints:

$$\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}(-\boldsymbol{\xi}_C(x)) = -\boldsymbol{C}_{\widetilde{I}(x)}(x).$$

3.3.3 Definition and properties of the null space step

The definition of the null space direction $\boldsymbol{\xi}_J(x)$ is slightly more involved than in the equality constrained case since it is not obtained by replacing $\mathrm{D}\boldsymbol{g}(x)$ by $\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}$ in (3.2.8). It requires the introduction of a different subset $\widehat{I}(x) \subset \widetilde{I}(x)$, which is now detailed.

The null space step $\boldsymbol{\xi}_J(x)$ is sought, up to a change of sign, as a best normalized descent direction diminishing violated or saturated inequality constraints. Following the characterization lemma 3.1 for equality constrained problems, $-\boldsymbol{\xi}_J(x)$ shall be set positively proportional to the solution $\boldsymbol{\xi}^*(x)$ of the following minimization problem:

$$\begin{aligned} \boldsymbol{\xi}^*(x) = \arg\min_{\boldsymbol{\xi} \in V} \quad & \mathrm{D}J(x)\boldsymbol{\xi} \\ \text{s.t.} \quad & \begin{cases} \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = 0 \\ \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi} \leq 0 \\ ||\boldsymbol{\xi}||_V \leq 1. \end{cases} \end{aligned} \tag{3.3.9}$$

The problem (3.3.9) could be solved directly with standard quadratic programming algorithms. However, it is convenient to characterize explicitly the minimizer $\boldsymbol{\xi}^*(x)$ of (3.3.9) by examining the dual problem. This will allow us to obtain in definition 3.6 an explicit formula for the null space direction $\boldsymbol{\xi}_J(x)$, in the form of (3.1.5).

Proposition 3.2. *Let $x \in V$ satisfy the qualification condition (3.3.4). There exists a unique couple of multipliers $\boldsymbol{\lambda}^*(x) \in \mathbb{R}^p$ and $\boldsymbol{\mu}^*(x) \in \mathbb{R}_+^{\widetilde{q}(x)}$ solution to the following quadratic optimization problem which is the dual of (3.3.9):*

$$(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x)) := \arg \min_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^p \\ \boldsymbol{\mu} \in \mathbb{R}^{\widetilde{q}(x)},\, \boldsymbol{\mu} \geq 0}} ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}||_V. \tag{3.3.10}$$

Proof. Problem (3.3.9) is equivalent to the following min-max formulation:

$$\min_{\substack{\boldsymbol{\xi} \in V \\ \langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_V \leq 1}} \max_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^p \\ \boldsymbol{\mu} \in \mathbb{R}_+^{\widetilde{q}(x)}}} \mathrm{D}J(x)\boldsymbol{\xi} + \boldsymbol{\lambda}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} + \boldsymbol{\mu}^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi}.$$

Exchanging formally the min and the max and performing the maximization with respect to $\boldsymbol{\xi}$ as in the proof of lemma 3.1 yields that (3.3.10) is the dual problem of (3.3.9) up to a change of sign (the duality gap between (3.3.10) and (3.3.9) will be shown to vanish in proposition 3.3). The program (3.3.10) brings into play the closed convex set $\mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$ and the least squares functional

$$(\boldsymbol{\lambda}, \boldsymbol{\mu}) \mapsto \left|\left| \nabla J(x) + \mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix} \right|\right|_V .$$

The latter is strictly convex over $\mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$ by virtue of (3.3.4). Hence, (3.3.10) has a unique solution. □

The optimization problem (3.3.10) belongs to the class of non negative least squares problems; it can be solved efficiently with a number of dedicated solvers, such as `cvxopt` [41] or `IPOPT` [310]. One nice feature of (3.3.10) lies in that its dimension is the number $p + \widetilde{q}(x)$ of saturated or violated constraints, which is small for many practical cases, e.g. in all the shape optimization applications considered in this thesis. It is also possible to exploit the 'sparsity' of the constraints when $p + \widetilde{q}(x)$ is large, see remark 3.8 below.

The next proposition relates the optimal values and the solutions $\boldsymbol{\xi}^*(x)$ and $(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x))$ of the primal and dual problems (3.3.9) and (3.3.10). In essence, we show that the optimal feasible descent direction $\boldsymbol{\xi}^*(x)$ of (3.3.12) is the projection of the gradient $\nabla J(x)$ onto the cone of feasible directions. The proof follows classical arguments of linear programming duality theory and it is detailed for the convenience of the reader.

Proposition 3.3. *Let $x \in V$ satisfy the qualification condition (3.3.4) and denote*

$$m^*(x) := ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda}^*(x) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}^*(x)||_V$$

the value of the dual problem (3.3.10). Then the value of the primal problem (3.3.9) is $p^*(x) = -m^*(x)$ and the following alternative holds:

1. *$m^*(x) = 0$: the first line of the KKT conditions (3.3.7) for the minimization problem (3.1.2) holds with (necessarily unique) Lagrange multipliers $(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x)) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$:*

$$\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda}^*(x) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}^*(x) = 0. \tag{3.3.11}$$

   One particular minimizer of (3.3.9) is $\boldsymbol{\xi}^*(x) = 0$.

2. *$m^*(x) > 0$: (3.3.11) does not hold and there exists a unique minimizer $\boldsymbol{\xi}^*(x)$ to (3.3.9), given by*

$$\boldsymbol{\xi}^*(x) = -\frac{\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda}^*(x) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}^*(x)}{||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda}^*(x) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}^*(x)||_V}. \tag{3.3.12}$$

Proof. Let $\boldsymbol{\xi} \in V$ be a feasible direction for the problem (3.3.9), i.e. $\mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} = 0$, $\mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi} \leq 0$ and $||\boldsymbol{\xi}||_V \leq 1$. Then for any $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$, it holds

$$\begin{aligned} \mathrm{D}J(x)\boldsymbol{\xi} \geq & \quad \mathrm{D}J(x)\boldsymbol{\xi} + \boldsymbol{\lambda}^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi} + \boldsymbol{\mu}^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi} \\ = & \quad \langle \nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}, \boldsymbol{\xi} \rangle_V \\ \geq & \quad -||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}||_V \end{aligned} \tag{3.3.13}$$

Since (3.3.13) holds for any feasible direction $\boldsymbol{\xi}$ for (3.3.9), and for any $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$, it follows:

$$\min_{\substack{\boldsymbol{\xi} \in V \\ \boldsymbol{\xi} \text{ feasible for (3.3.9)}}} \mathrm{D}J(x)\boldsymbol{\xi} \geq - \min_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^p \\ \boldsymbol{\mu} \in \mathbb{R}^{\widetilde{q}(x)}, \boldsymbol{\mu} \geq 0}} ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}||_V. \tag{3.3.14}$$

Therefore, we have proven that $p^*(x) \geq -m^*(x)$. We now examine the alternative $m^*(x) = 0$ or $m^*(x) > 0$:

1. If $m^*(x) = 0$, then (3.3.14) implies $p^*(x) \geq 0$. Therefore, the value of (3.3.9) is $p^*(x) = -m^*(x) = 0$, attained in particular at $\boldsymbol{\xi}^* = 0$, and more generally at any feasible $\boldsymbol{\xi}^* \in V$ satisfying $\boldsymbol{\mu}^*(x)^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi}^* = 0$, as follows readily from the KKT conditions for (3.3.9). Furthermore, the KKT equation (3.3.11) is satisfied by definition of $m^*(x) = 0$.

2. Assume now $m^*(x) > 0$. The KKT condition for (3.3.9) states that for any local optimum $\boldsymbol{\xi}'$, there exists $(\boldsymbol{\lambda}', \boldsymbol{\mu}') \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$ and $\alpha \geq 0$ such that,

$$\forall \boldsymbol{\xi} \in V,\ (\mathrm{D}J(x) + \boldsymbol{\lambda}'^T \mathrm{D}\boldsymbol{g}(x) + \boldsymbol{\mu}'^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x))\boldsymbol{\xi} = -\alpha \langle \boldsymbol{\xi}', \boldsymbol{\xi} \rangle_V. \tag{3.3.15}$$

   Using Riesz identifications of the gradient and the differentials, we obtain

$$\alpha \boldsymbol{\xi}' = -(\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}' + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}'),$$

   and since $m^*(x) > 0$, it is necessary that $\alpha > 0$. The complementarity condition $\alpha(\langle \boldsymbol{\xi}', \boldsymbol{\xi}' \rangle_V - 1) = 0$ yields then $||\boldsymbol{\xi}'||_V = 1$, which readily implies:

$$\boldsymbol{\xi}' = -\frac{\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}' + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}'}{||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}' + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}'||_V}.$$

   Then the complementarity condition for (3.3.9) implies $\boldsymbol{\mu}'^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi}' = 0$. Therefore it holds that

$$\begin{aligned} \mathrm{D}J(x)\boldsymbol{\xi}' &= \mathrm{D}J(x)\boldsymbol{\xi}' + \boldsymbol{\lambda}'^T \mathrm{D}\boldsymbol{g}(x)\boldsymbol{\xi}' + \boldsymbol{\mu}'^T \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\boldsymbol{\xi}' \\ &= \langle \nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}' + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}', \boldsymbol{\xi}' \rangle_V \\ &= -||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}' + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}'||_V. \end{aligned} \tag{3.3.16}$$

   The previous equation together with the inequality (3.3.13) with $\boldsymbol{\xi} = \boldsymbol{\xi}'$ then implies that $(\boldsymbol{\lambda}', \boldsymbol{\mu}')$ achieves the minimum of (3.3.10). By uniqueness, this implies $\boldsymbol{\lambda}' = \boldsymbol{\lambda}^*(x)$ and $\boldsymbol{\mu}' = \boldsymbol{\mu}^*(x)$, hence $\boldsymbol{\xi}' = \boldsymbol{\xi}^*(x)$. Furthermore, $p^*(x) = \mathrm{D}J(x)\boldsymbol{\xi}^*(x) = \mathrm{D}J(x)\boldsymbol{\xi}' = -m^*(x)$.

□

Finally, the next proposition characterizes explicitly the expression of the optimal descent direction $\boldsymbol{\xi}^*(x)$ from the signs of the multiplier $\boldsymbol{\mu}^*(x)$, and highlights in which sense the problem (3.3.9) is combinatorial. Let us recall the definitions (3.3.5) and (3.3.6) for the projection operator Π_{C_I} and the multipliers $(\boldsymbol{\lambda}_I(x), \boldsymbol{\mu}_I(x))$ which are used in the result.

Proposition 3.4. *In the context of point (2) in proposition 3.3, let $\boldsymbol{\xi}^*(x)$ and $(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x))$ be the minimizers of the primal and dual problems (3.3.9) and (3.3.10). Define the subset $\widehat{I}(x) \subset \widetilde{I}(x)$ by*

$$\widehat{I}(x) := \{i \in \widetilde{I}(x) \,|\, \mu_i^*(x) > 0\}. \tag{3.3.17}$$

1. $(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x))$ and $\boldsymbol{\xi}^*(x)$ are explicitly given in terms of $\widehat{I}(x)$ by:

$$\begin{bmatrix} \boldsymbol{\lambda}^*(x) \\ \widehat{\boldsymbol{\mu}}^*(x) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_{\widehat{I}(x)}(x) \\ \boldsymbol{\mu}_{\widehat{I}(x)}(x) \end{bmatrix} = -(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)} \mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}^{\mathcal{T}})^{-1} \mathrm{D}\boldsymbol{C}_{\widehat{I}(x)} \nabla J(x), \tag{3.3.18}$$

$$\boldsymbol{\xi}^*(x) = -\frac{\Pi_{\boldsymbol{C}_{\widehat{I}(x)}}(\nabla J(x))}{||\Pi_{\boldsymbol{C}_{\widehat{I}(x)}}(\nabla J(x))||_V}, \tag{3.3.19}$$

where $\widehat{\boldsymbol{\mu}}^*(x) := (\mu_i^*(x))_{i \in \widehat{I}(x)}$ is the vector collecting all positive components of $\boldsymbol{\mu}^*(x)$.

2. *$\widehat{I}(x)$ is equivalently the unique solution to either of the following discrete optimization problems:*

$$\begin{aligned} \widehat{I}(x) = \arg \max_{I \subset \widetilde{I}(x)} \quad & ||\Pi_{\boldsymbol{C}_I}(\nabla J(x))||_V \\ s.t. \quad & \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\Pi_{C_I}(\nabla J(x)) \geq 0, \end{aligned} \tag{3.3.20}$$

$$\begin{aligned} \widehat{I}(x) = \arg \min_{I \subset \widetilde{I}(x)} \quad & ||\Pi_{C_I}(\nabla J(x))||_V \\ s.t. \quad & \boldsymbol{\mu}_I(x) \geq 0. \end{aligned} \tag{3.3.21}$$

In particular, $\widehat{I}(x)$ is the unique subset $I \subset \widetilde{I}(x)$ satisfying simultaneously both feasibility conditions

$$\mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\Pi_{C_I}(\nabla J(x)) \geq 0 \text{ and } \boldsymbol{\mu}_I(x) \geq 0.$$

Proof. 1. The complementary condition for the primal and dual problems (3.3.9) and (3.3.10) reads

$$\forall i \in \widetilde{I}(x), \quad \mu_i^*(x)\mathrm{D}h_i(x)\boldsymbol{\xi}^*(x) = 0. \tag{3.3.22}$$

Therefore, $\mathrm{D}h_i(x)\boldsymbol{\xi}^*(x) = 0$ for all indices $i \in \widehat{I}(x)$, which implies that $\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}(x)\boldsymbol{\xi}^*(x) = 0$. Then, after left multiplication of (3.3.12) by $(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}$, we obtain (3.3.18), whence (3.3.19) follows.

2. Let $I \subset \widetilde{I}(x)$ a subset satisfying $\mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)\Pi_{C_I}(\nabla J(x)) \geq 0$. This implies that

$$\boldsymbol{\xi} = -\Pi_{\boldsymbol{C}_I}(\nabla J(x))/||\Pi_{\boldsymbol{C}_I}(\nabla J(x))||_V$$

is feasible for the primal problem (3.3.9), and we obtain by definition of $\boldsymbol{\xi}^*(x)$ that

$$-||\Pi_{\boldsymbol{C}_{\widehat{I}(x)}}(\nabla J(x))||_V = \mathrm{D}J(x)\boldsymbol{\xi}^*(x) \leq \mathrm{D}J(x)\boldsymbol{\xi} = -||\Pi_{\boldsymbol{C}_I}(\nabla J(x))||_V, \tag{3.3.23}$$

whence the maximization property (3.3.20).

For $I \subset \widetilde{I}(x)$ satisfying $\boldsymbol{\mu}_I(x) \geq 0$, we obtain feasible multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ for the dual problem (3.3.10) by taking $\boldsymbol{\mu}$ to be equal to $\boldsymbol{\mu}_I$ on the indices of I and extended by 0 in the complementary subset $\widetilde{I}(x) \setminus I$. Then the optimality of $(\boldsymbol{\lambda}^*(x), \boldsymbol{\mu}^*(x))$ for this dual problem reads:

$$\begin{aligned} ||\Pi_{\boldsymbol{C}_{\widehat{I}(x)}}(\nabla J(x))||_V &= ||\nabla J + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda}^*(x) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^{\mathcal{T}}\boldsymbol{\mu}^*(x)||_V \\ &\leq ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}}\boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}^{\mathcal{T}}\boldsymbol{\mu}||_V = ||\Pi_{\boldsymbol{C}_I}(\nabla J(x))||_V, \end{aligned} \tag{3.3.24}$$

whence the minimization property (3.3.21).

□

Remark 3.5. In view of (3.3.17), the optimal multiplier $\boldsymbol{\mu}^*(x)$ can be interpreted as an indicator variable specifying which constraints of $\widetilde{I}(x)$ are 'not aligned' with the gradient $\nabla J(x)$ and should be kept in the subset $\widehat{I}(x)$. The best descent direction (in the sense of (3.3.9)) is obtained by projecting the gradient $\nabla J(x)$ onto the tangent space of the constraint subset $\widehat{I}(x)$ rather than onto the full set of violated or saturated constraints $\widetilde{I}(x)$. Indeed, the descent direction $\boldsymbol{\xi} = -\Pi_{C_{\widetilde{I}(x)}}\nabla J(x)$ that would be obtained by projecting $\nabla J(x)$ on the whole set $\widetilde{I}(x)$ would only keep them all constant at first order, i.e. $\mathrm{D}h_i(x)\boldsymbol{\xi} = 0$, (see remark 3.9 for more details). It is therefore more efficient to project $\nabla J(x)$ only on those constraints associated to the indices $i \in \widehat{I}(x)$, thus allowing the remaining ones (associated to $i \in \widetilde{I}(x) \setminus \widehat{I}(x)$, indicating vanishing multipliers $\mu_i^*(x) = 0$) to decrease since the calculated descent direction ensures that $\mathrm{D}h_i(x)\boldsymbol{\xi}^*(x) \leq 0$ holds for all $i = 1, ..., q$.

Remark 3.6. The use of a dual problem such as (3.3.10) in order to obtain information about which constraints should remain active is classical in active sets methods, see e.g. [77, 187, 244]. In principle, the optimal subset $\widehat{I}(x)$ could be found by solving the discrete problems (3.3.20) or (3.3.21). However, we expect that in practice, it is more efficient to rely on iterative solvers relying on gradient descent for solving the dual problem (3.3.10), e.g. a cone programming solver or a non negative least squares algorithm such as [77]. This is what we do in the sequel.

Having introduced the subset $\widehat{I}(x)$ (defined in (3.3.17)), we are now able to define the null space direction $\boldsymbol{\xi}_J(x)$ in the present context: $-\boldsymbol{\xi}_J(x)$ is set to be a positive multiple of the optimal descent direction $\boldsymbol{\xi}^*(x)$ supplied by (3.3.19).

Definition 3.6. For any point $x \in V$ satisfying the constraint qualification (3.3.4), the null space direction $\boldsymbol{\xi}_J(x)$ at x for the optimization problem (3.1.2) is defined by:

$$\boldsymbol{\xi}_J(x) := \Pi_{\boldsymbol{C}_{\widehat{I}(x)}}(\nabla J(x)) = (I - \mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}(x)^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)})\nabla J(x), \tag{3.3.25}$$

where $\widehat{I}(x)$ is the optimal set defined by (3.3.17).

The main point in definition 3.6 is that, while all violated and saturated constraints are taken into account in the Gauss-Newton direction $\boldsymbol{\xi}_C(x)$ defined by (3.3.8), only those constraints in $\widehat{I}(x)$, not aligned with the gradient $\nabla J(x)$, occur in the definition of $\boldsymbol{\xi}_J(x)$.

Remark 3.7. With our notation, the optimization scheme proposed by Barbarosie et. al. [59, 60] reads

$$\begin{cases} x_{n+1} = x_n - \Delta t \nabla J(x_n) - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I(x_n)}\boldsymbol{\nu}_n \\ \quad \boldsymbol{\nu}_n = -\Delta t(\mathrm{D}\boldsymbol{C}_{I(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I(x_n)})^{-1}\mathrm{D}\boldsymbol{C}_{I(x_n)}\nabla J(x_n) + \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I(x_n)}(\mathrm{D}\boldsymbol{C}_{I(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I(x_n)})^{-1}\boldsymbol{C}_{I(x_n)}, \end{cases} \tag{3.3.26}$$

where the set $I(x_n)$ is obtained by removing indices from $\widetilde{I}(x_n)$ one by one, starting from the index i_0 associated with the most negative multiplier $\nu_{n,i_0} < 0$, until all of them become non negative. Therefore, the set $I(x_n)$ used in this strategy and that $\widehat{I}(x_n)$ featured in our strategy, given by (3.3.17), do not coincide in general; one could think of configurations where the procedure of [60] would fail to find the optimal set $\widehat{I}(x_n)$ (for example if $i_0 \in \widehat{I}(x_n)$) and would project the gradient on a less optimal subset of constraints. We note that no convergence result is given by the authors about this procedure.

Remark 3.8. Let us discuss two extreme cases related to the involved computational effort in the numerical implementation of (3.3.25). Upon discretization, we may assume that $V = \mathbb{R}^k$ is a finite-dimensional space.

1. If the total number $p+\widetilde{q}$ of saturated or violated constraints is small compared to the dimension k of V, it is best, for numerical efficiency, to assemble the small square matrix $(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)})$ and to invert it by a direct method.

2. If $V = \mathbb{R}^k$ is equipped with an inner product encoded by a matrix A, and if $p+\widetilde{q}$ is of the order of k or larger, the computation of the inverse of $(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)})$ can be expensive. However, it can be still tractable if both $\mathrm{D}\boldsymbol{C}$ and A are sparse matrices. For instance, this occurs in the case of bound constraints on the optimization variable $x = (x_1, ..., x_k)$, e.g. constraints of the form $\alpha_i \leq x_i \leq \beta_i$, $i = 1, ..., k$. Recalling from remark 3.1 that in this setting, $\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}(x)} = A^{-1}\mathrm{D}\boldsymbol{C}^{T}_{\widehat{I}(x)}$, it can be verified that the vector
$$X := A^{-1}\mathrm{D}\boldsymbol{C}^{T}_{\widehat{I}(x)}(\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}A^{-1}\mathrm{D}\boldsymbol{C}^{T}_{\widehat{I}(x)})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\nabla J(x)$$
can be computed as the solution to the sparse linear system
$$\begin{bmatrix} A & -\mathrm{D}\boldsymbol{C}^{T}_{\widehat{I}(x)} \\ \mathrm{D}\boldsymbol{C}_{\widehat{I}(x)} & 0 \end{bmatrix}\begin{bmatrix} X \\ \Lambda \end{bmatrix} = \begin{bmatrix} 0 \\ \mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}\nabla J(x) \end{bmatrix},$$
where $\Lambda \in \mathbb{R}^{p+\mathrm{Card}(\widehat{I}(x))}$ is an extra slack variable, which yields the null space directions $\boldsymbol{\xi}_J(x) = \nabla J(x) - X$. A similar strategy based on the sparsity of A and $\mathrm{D}\boldsymbol{C}_{\widehat{I}(x)}$ can be used to compute the range space direction $\boldsymbol{\xi}_C(x)$ of (3.3.8), or to solve the dual quadratic subproblem (3.3.10).

Remark 3.9. As we have already mentioned, the Lagrange multiplier $\boldsymbol{\mu}^*(x)$ given by (3.3.18) may be understood as an indicator of which inequality constraints are aligned with the gradient of J. To further highlight this, it is instructive to consider the particular situation where the gradients of the constraint functions are orthogonal, i.e.:

$$\begin{aligned} &\langle \nabla g_i(x), \nabla g_j(x) \rangle_V = 0, \text{ for } i, j = 1, ..., p, \ i \neq j, \\ &\langle \nabla h_i(x), \nabla h_j(x) \rangle_V = 0, \text{ for } i, j = 1, ..., q, \ i \neq j, \\ &\langle \nabla g_i(x), \nabla h_j(x) \rangle_V = 0, \text{ for } i = 1, ..., p, \ j = 1, ..., q. \end{aligned}$$

In this case, it easily follows from the Pythagore theorem that for any $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$,

$$||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^\mathcal{T}\boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x)}(x)^\mathcal{T}\boldsymbol{\mu}||_V^2 = \left|\left|\nabla J(x) + \sum_{i=1}^p \lambda_i \nabla g_i(x) + \sum_{j\in\widetilde{I}(x)} \mu_j \nabla h_j(x)\right|\right|_V^2$$
$$= ||\nabla J(x)||_V^2 + \sum_{i=1}^p \left(\lambda_i^2 ||\nabla g_i(x)||_V^2 + 2\lambda_i \langle \nabla J(x), \nabla g_i(x)\rangle_V\right)$$
$$+ \sum_{j\in\widetilde{I}(x)} \left(\mu_j^2 ||\nabla h_j(x)||_V^2 + 2\mu_j \langle \nabla J(x), \nabla h_j(x)\rangle_V\right).$$

Therefore the minimization problem (3.3.10) is separable with respect to the variables $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x)}$: $(\lambda_i^*(x))_{1\leq i\leq p}$ and $(\mu_i^*(x))_{i\in\widetilde{I}(x)}$ are the respective solutions to the minimization problems:

$$\forall i \in 1\dots p,\ \ \lambda_i^*(x) = \arg\min_{t\in\mathbb{R}} \left(t^2 ||\nabla g_i(x)||_V^2 + 2t\langle \nabla J(x), \nabla g_i(x)\rangle_V\right),$$

$$\forall i \in \widetilde{I}(x),\ \ \mu_i^*(x) = \arg\min_{\substack{t\in\mathbb{R}\\ t\geq 0}} \left(t^2 ||\nabla h_i(x)||_V^2 + 2t\langle \nabla J(x), \nabla h_i(x)\rangle_V\right),$$

which yields eventually:

$$\lambda_i^*(x) = -\frac{\langle \nabla J(x), \nabla g_i(x)\rangle_V)}{||\nabla g_i(x)||_V^2}, \qquad \mu_i^*(x) = \begin{cases} 0 & \text{if } \langle \nabla J(x), \nabla h_i(x)\rangle_V \geq 0, \\ -\frac{\langle \nabla J(x), \nabla h_i(x)\rangle_V}{||\nabla h_i(x)||_V^2} & \text{otherwise.} \end{cases}$$

Hence, $\mu_i^*(x)$ is positive if and only if the direction $-\nabla J(x)$ leads to an increase (i.e. a violation) of the i^{th} inequality constraint.

In the general case where all the constraint gradients are not mutually orthogonal, the interpretation of $\boldsymbol{\mu}^*(x)$ is similar, up to the additional complication that (3.3.10) accounts for the combinatorics behind the possible alignments between different constraint gradients. In the following, with a slight abuse of language, we shall nevertheless refer to the indices $i \in \widetilde{I}(x) \setminus \widehat{I}(x)$ as those associated to constraints which are 'aligned' with $\nabla J(x)$.

3.3.4 Decrease properties of the trajectories of the null space ODE

The final result of this section is the counterpart of proposition 3.1 in the case of the equality and inequality constrained optimization problem (3.1.2).

Proposition 3.5. *Assume that the trajectories $x(t)$ of the flow*

$$\begin{cases} \dot{x}(t) = -\alpha_J \boldsymbol{\xi}_J(x(t)) - \alpha_C \boldsymbol{\xi}_C(x(t)) \\ x(0) = x_0, \end{cases} \tag{3.3.27}$$

with $\boldsymbol{\xi}_J$ and $\boldsymbol{\xi}_C$ given by (3.3.8) and (3.3.25) exist on some interval $[0, T]$ for $T > 0$ and are such that:

(a) the set $\widetilde{I}(x(t))$ defined in (3.3.3) is constant over $[0, T]$:

$$\forall t \in [0, T],\ \widetilde{I}(x(t)) = \widetilde{I}(x_0);$$

(b) the constraints remain qualified along the flow $x(t)$, in the sense of (3.3.4).

Then the following properties hold true:

1. *The violation of the constraints decreases exponentially:*

$$\forall t \in [0, T],\ \boldsymbol{g}(x(t)) = e^{-\alpha_C t}\boldsymbol{g}(x_0) \text{ and } \boldsymbol{h}_{\widetilde{I}(x(t))}(x(t)) \leq e^{-\alpha_C t}\boldsymbol{h}_{\widetilde{I}(x_0))}(x_0). \tag{3.3.28}$$

2. *$J(x(t))$ decreases 'as soon as the violation (3.3.28) of the constraints is sufficiently small' in the following sense. Assume that* $\mathrm{rank}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x_0)}(x))$ *is maximal for all x in $K = \{x \in V \,|\, ||\boldsymbol{C}_{\widetilde{I}(x_0)}(x)||_\infty \leq ||\boldsymbol{C}_{\widetilde{I}(x_0)}(x_0)||_\infty\}$ and*

$$\sup_{x\in K} ||\nabla J(x)||_V |\sigma_p^{-1}(x)| < +\infty. \tag{3.3.29}$$

where $\sigma_p(x)$ is the smallest singular value of $\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}(x)$. Then there exists a constant $C > 0$ such that

$$\forall t \in [0,T],\ ||\Pi_{C_{\widehat{I}(x(t))}}(\nabla J(x(t)))||_V^2 > Ce^{-\alpha_C t} \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t}J(x(t)) < 0. \tag{3.3.30}$$

3. *Any stationary point x^* of the flow (3.3.27) satisfies the KKT optimality conditions (3.3.7) which equivalently rewrite:*

$$\begin{cases} \nabla J(x^*) + \mathrm{D}\boldsymbol{g}(x^*)^\mathcal{T}\boldsymbol{\lambda}^*(x^*) + \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x^*)}(x^*)^\mathcal{T}\boldsymbol{\mu}^*(x^*) = 0, \\ (\boldsymbol{g}(x^*) = 0 \text{ and } \boldsymbol{h}_{\widetilde{I}(x^*)}(x^*) = 0) \Leftrightarrow \boldsymbol{C}_{\widetilde{I}(x^*)}(x^*) = 0, \end{cases} \tag{3.3.31}$$

where $(\boldsymbol{\lambda}^*(x^*), \boldsymbol{\mu}^*(x^*)) \in \mathbb{R}^p \times \mathbb{R}_+^{\widetilde{q}(x^*)}$ are defined in (3.3.10) or (3.3.18).

Proof. 1. The definition (3.3.8) of $\boldsymbol{\xi}_C(x(t))$ implies $\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x(t))}\boldsymbol{\xi}_C(x(t)) = \boldsymbol{C}_{\widetilde{I}(x(t))}(x(t))$, and since $-\boldsymbol{\xi}_J(x(t))$ is positively proportional to $\boldsymbol{\xi}^*(x(t))$ (proposition 3.3), it holds

$$\mathrm{D}\boldsymbol{C}_{\widehat{I}(x(t))}\boldsymbol{\xi}_J(x(t)) = 0,\ -\mathrm{D}\boldsymbol{h}_{\widetilde{I}(x(t))\setminus\widehat{I}(x(t))}(x(t))\boldsymbol{\xi}_J(x(t)) \leq 0.$$

Therefore we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{C}_{\widehat{I}(x(t))}(x(t)) = -\alpha_C\boldsymbol{C}_{\widehat{I}(x(t))}(x(t)) \text{ and } \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{h}_{\widetilde{I}(x(t))\setminus\widehat{I}(x(t))}(x(t)) \leq -\alpha_C\boldsymbol{h}_{\widetilde{I}(x(t))\setminus\widehat{I}(x(t))}(x(t)) \tag{3.3.32}$$

from which (3.3.28) follows by application of Gronwall's lemma.

2. The proof is identical to that of proposition 3.1.

3. A stationary point x^* of (3.3.27) satisfies by definition $-\alpha_J\boldsymbol{\xi}_J(x^*) - \alpha_C\boldsymbol{\xi}_C(x^*) = 0$. Left multiplication of this identity by $\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x^*)}(x^*)$ yields:

$$-\alpha_J\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x^*)}(x^*)\boldsymbol{\xi}_J(x^*) - \alpha_C\boldsymbol{C}_{\widetilde{I}(x^*)}(x^*) = 0. \tag{3.3.33}$$

Remembering now that from definition (3.3.9),

$$-\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x^*)}\boldsymbol{\xi}_J(x^*) \leq 0 \text{ and } \boldsymbol{C}_{\widetilde{I}(x^*)}(x^*) \geq 0,$$

equality in (3.3.33) can hold only if both terms vanish. In particular, we infer that $\boldsymbol{C}_{\widetilde{I}(x^*)}(x^*) = 0$, a fact which implies $\boldsymbol{\xi}_C(x^*) = 0$ and which encompasses the last two lines of the KKT conditions (3.3.7). Returning to the fact that $-\alpha_J\boldsymbol{\xi}_J(x^*) - \alpha_C\boldsymbol{\xi}_C(x^*) = 0$, we obtain that $\boldsymbol{\xi}_J(x^*) = 0$, which corresponds to the first line in (3.3.7). This completes the proof.

□

Remark 3.10. Since the sets $\widetilde{I}(x)$ or $\widehat{I}(x)$ are subject to change as soon as inequality constraints become active or inactive, or if not enough regularity holds, the ODE (3.3.27) has in general a discontinuous right-hand side and is defined only formally (note that a rigorous mathematical meaning could still be provided with the theory of non smooth ODEs, see [126, 158]). However and as discussed further on in the next remarks, its discretization makes sense and exhibits the same decrease properties as its continuous counterpart for sufficiently small steps Δt.

Remark 3.11. The assumption (a) in proposition 3.5, whereby the index set $\widetilde{I}(x(t))$ remains constant is essentially made to ensure that the right-hand side of the flow (3.3.27) is continuous. Indeed, in such a case, the range space direction $\boldsymbol{\xi}_C(x(t))$ is continuous by its definition (3.3.8), while the null space step $\boldsymbol{\xi}_J(x(t))$ is continuous because

$$\boldsymbol{\xi}_J(x(t)) = \nabla J(x(t)) + \mathrm{D}\boldsymbol{C}_{\widetilde{I}(x(t))}\begin{bmatrix}\boldsymbol{\lambda}^*(x(t)) \\ \boldsymbol{\mu}^*(x(t))\end{bmatrix}$$

and it can be shown that the multipliers $(\boldsymbol{\lambda}^*(x(t)), \boldsymbol{\mu}^*(x(t)))$ defined by (3.3.10) are continuous functions. At a time T corresponding to a sudden change of the index set $\widetilde{I}(x(t))$, we assume that the solution $x(t)$ can be extended by restarting the ODE (3.3.27) with the new index set $\widetilde{I}(x(T))$. From (1) in proposition 3.5, the bound $\boldsymbol{h}(x(t)) \leq e^{-\alpha_C t}\boldsymbol{h}(x(0))$ still holds after the time T for all constraints $i \in \{1, \ldots, q\}$:

constraints are asymptotically satisfied. Properties (2) and (3) remain true, up to an adjustment of the constant C in (3.3.30) (which can be taken global since there are finitely many possible sets $\widetilde{I}(x(t))$). There may exist situations where the set of asymptotically saturated constraints $\widetilde{I}(x(t))$ could oscillate indefinitely. However (2) states that $x(t)$ always keeps improving (in the sense of (3.3.30)), and (3) states that if $x(t)$ eventually converges, it is necessarily towards a KKT point.

Remark 3.12. In practice, the analysis of proposition 3.5 is sufficient because, similarly to the conclusions of remark 3.3, analogous properties hold for the discrete scheme

$$x_{n+1} = x_n - \Delta t\big(\alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n)\big). \tag{3.3.34}$$

Indeed, one can easily check that:

1. Up to first order, the violation of the constraints decreases at a geometric rate:

$$\boldsymbol{C}(x_{n+1}) = (1 - \alpha_C \Delta t)\boldsymbol{C}(x_n) + o(\Delta t). \tag{3.3.35}$$

   This suggests that in order to obtain a stable scheme, one must a priori select α_C and Δt such that $0 < \alpha_C \Delta t < 2$.

2. An accumulation point x^* of the sequence $(x_n)_{n\in\mathbb{N}}$ is feasible, i.e. $\boldsymbol{C}_{\widetilde{I}(x^*)}(x^*) = 0$ and is a KKT point.

Finally, note that one flexibility of this ODE approach is that at the continuous level, the results of proposition 3.5 do not depend on the values of the parameters $\alpha_J > 0$ and $\alpha_C > 0$. Therefore the convergence of the discrete scheme towards the continuous trajectory should hold as soon as the discretization step size $\Delta t > 0$ is sufficiently small.

3.4 Numerical discretization and time-stepping schemes for the null space ODE

This short section describes practical implementation details for the discretization of the ODE (3.1.3) by an explicit Euler scheme. Two important issues are discussed respectively in sections 3.4.1 and 3.4.2. First, we propose small adaptations in the computation of $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ in order to account for the discontinuous changes of the right-hand side $-(\alpha_J \boldsymbol{\xi}_J + \alpha_C \boldsymbol{\xi}_C)$. Then, a merit function is proposed for adapting the time step Δt. The complete implementation of the algorithm is summarized in section 3.4.3 below.

3.4.1 Accounting for discontinuities near the inequality constraint barriers

A potential issue when implementing directly the Euler time-stepping scheme (3.3.2) comes from the fact that the vector fields $\boldsymbol{\xi}_J$ and $\boldsymbol{\xi}_C$ given by (3.3.8) and (3.3.25) are characterized by the same discontinuities as the discrete index mapping $x \mapsto \widetilde{I}(x)$. As a result, abrupt oscillations of the discrete optimization path (x_n) may occur near the boundary of the feasible set: if $h_i(x_n) = 0$ and $i \in \widehat{I}(x_n)$ for some index $i \in \{1, \dots, q\}$, then in the definition (3.3.25) of $\boldsymbol{\xi}_J(x_n)$, the gradient $\nabla J(x_n)$ is projected tangentially to the constraint h_i, but it is not projected after any slight deviation (e.g. due to the discretization) making this constraint inactive ($h_i(x_{n+1}) < 0$). This kind of issue is very classical in the discretization of ODEs with discontinuous vector fields and can be tackled by various methods, see e.g. [126] for a review.

In this section, we suggest a simple alternative that works well in practice to stabilize trajectories near these boundaries: inequality constraints are felt from a short distance by replacing the set $\widetilde{I}(x_n)$ in (3.1.7) with the set $\widetilde{I}_\varepsilon(x_n)$ of inequality constraints violated "up to ε_i":

$$\widetilde{I}_\varepsilon(x_n) = \{i \in \{1, \dots, q\} \,|\, h_i(x_n) \geq -\varepsilon_i\}. \tag{3.4.1}$$

The tolerances $\varepsilon_i > 0$ can be estimated in an automatic fashion, independent of an arbitrary rescaling of the constraints, thanks to an posteriori bound we now detail. Let $\mathtt{h}$ be a user-defined parameter accounting for the distance from the optimization path at which the constraints should be felt.

Assume now that the current point x_n satisfies the constraint h_i up to the uncertainty $\mathtt{h}$ on its location: by this we mean that there exists some unknown point x_n^* such that $||x_n^* - x_n|| \leq \mathtt{h}$, $h_i(x_n) > 0$ and $h_i(x_n^*) = 0$. Then the error $\mathtt{h}$ for the location of x_n propagates to the constraint values $h_i(x_n)$ according to the following inequality:

$$h_i(x_n) = |h_i(x_n) - h_i(x_n^*)| \simeq |\mathrm{D}h_i(x_n)(x_n^* - x_n)| \leq ||\nabla h_i(x_n)||_V \mathtt{h}. \tag{3.4.2}$$

It is therefore natural to set

$$\varepsilon_i := ||\nabla h_i(x_n)||_V \mathtt{h} \tag{3.4.3}$$

for the value of ε_i in (3.4.1). In our implementation relying on the discretization of (3.1.3), the parameter $\mathtt{h}$ is set proportional to the time step Δt: $\mathtt{h} = \mathtt{K}\Delta t$ for a constant $\mathtt{K}$ to be defined by the user. The parameter $\mathtt{h}$ should be defined in accordance with the typical distance $||\Delta x||_V = ||x_{n+1} - x_n||_V$ between two successive iterations; in the academic examples in section 3.5 below considering optimization in $\mathbb{R}^k$, we may set e.g. $\mathtt{h} = 0.01$ for a typical increment size $||\Delta x||_V \simeq 0.1$. For our shape optimization applications in section 3.6, $\mathtt{h}$ is typically of the order of the discretization mesh size, see section 3.6.2 below.

Note that more generally, the a posteriori bound (3.4.2) allows to assert whether a constraint $C_i(x_n)$ can be considered as satisfied or not with respect to the numerical discretization.

The dual problem (3.3.10) is then solved with $\widetilde{I}_\varepsilon(x_n)$ instead of $\widetilde{I}(x_n)$ in order to obtain a new subset of indices $\widehat{I}_\varepsilon(x_n)$ which indicates which constraints are likely to be not aligned with the gradient $\nabla J(x_n)$ when crossing the barrier $\{\boldsymbol{h} = 0\}$. The null space and range space steps $\boldsymbol{\xi}_J(x_n)$ and $\boldsymbol{\xi}_C(x_n)$ in step 4 of algorithm 3.1 are finally replaced with $\boldsymbol{\xi}_{J,\varepsilon}(x_n)$ and $\boldsymbol{\xi}_{C,\varepsilon}(x_n)$ computed as follows:

$$\boldsymbol{\xi}_{J,\varepsilon}(x_n) := (I - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}_\varepsilon(x_n)}(\mathrm{D}\boldsymbol{C}_{\widehat{I}_\varepsilon(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}_\varepsilon(x_n)})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}_\varepsilon(x_n)})\nabla J(x_n), \tag{3.4.4}$$

$$\boldsymbol{\xi}_{C,\varepsilon}(x_n) := \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I^*_\varepsilon(x_n)}(\mathrm{D}\boldsymbol{C}_{I^*_\varepsilon(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I^*_\varepsilon(x_n)})^{-1}\boldsymbol{C}_{I^*_\varepsilon(x_n)}(x_n), \tag{3.4.5}$$

where $I^*_\varepsilon(x_n) = \widetilde{I}(x_n) \cup \widehat{I}_\varepsilon(x_n)$ is the set of constraints that are either violated, saturated or not aligned with the gradient $\nabla J(x_n)$ at $\boldsymbol{h} = -(\varepsilon_1, ..., \varepsilon_q)^T$. The use of $\widehat{I}_\varepsilon(x_n)$ in the definition of $\boldsymbol{\xi}_{J,\varepsilon}(x_n)$ ensures that the gradient $\nabla J(x_n)$ is being projected tangentially to the constraint on a small layer near the boundary of the feasible set. As a result, no abrupt discontinuity occurs anymore for $\boldsymbol{\xi}_{J,\varepsilon}$ and $\boldsymbol{\xi}_{C,\varepsilon}$ when crossing the boundary of the feasible domain while remaining in this layer. Including constraints $i \in \widehat{I}_\varepsilon(x_n)$ in the Gauss-Newton direction $\boldsymbol{\xi}_{C,\varepsilon}(x_n)$ even if they are satisfied (i.e. if $-\varepsilon_i \leq h_i(x_n) \leq 0$) further allows to stabilize the values of these constraints closer to zero.

3.4.2 Time step adaptation based on a merit function.

The ODE (3.1.3) is discretized by an explicit scheme of the form:

$$x_{n+1} = x_n - \Delta t_n\big(\alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n)\big), \tag{3.4.6}$$

with a variable time step $\Delta t_n > 0$. The practical implementation of such a strategy is often guided by a merit function, i.e. an indicator allowing to detect that a step has been too large, a situation where a choice has to be made regarding whether the step should be reduced or accepted. For our null space algorithm, a merit function which resembles very much that of the Augmented Lagrangian Method is readily available, however with a specific choice of multipliers:

Lemma 3.3. *For a given $x_n \in V$, let $\mathtt{merit}_{x_n} : V \to \mathbb{R}$ be the function defined by*

$$\mathtt{merit}_{x_n}(x) := \alpha_J \left(J(x) + \boldsymbol{\Lambda}(x_n)^T \boldsymbol{C}_{\widetilde{I}(x_n)}(x)\right) + \frac{\alpha_C}{2} \boldsymbol{C}_{\widetilde{I}(x_n)}(x)^T \boldsymbol{S}(x_n) \boldsymbol{C}_{\widetilde{I}(x_n)}(x) \tag{3.4.7}$$

where $\boldsymbol{\Lambda}(x_n) = \begin{bmatrix} \boldsymbol{\lambda}^*(x_n)^T & \boldsymbol{\mu}^*(x_n)^T \end{bmatrix}^T$ is the vector of multipliers defined as the solution to the dual problem (3.3.10) (see (3.3.18)) and $\boldsymbol{S}(x_n) = (\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x_n)}(x_n)\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x_n)}(x_n)^{\mathcal{T}})^{-1}$ is symmetric positive definite. Then (3.4.6) is a gradient step for decreasing the function $\mathtt{merit}_{x_n}$, namely:

$$\nabla \mathtt{merit}_{x_n}(x_n) = \alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n).$$

Proof. It is a straightforward computation of the gradient of (3.4.7). □

A possible implementation of an optimization strategy of the form (3.4.6) based on this merit function is summarized in algorithm 3.1, which requires the introduction of a few extra parameters:

- `time_step`: choose a fixed time step $\Delta t > 0$.
- `maxtrials`: the optimization time step is decreased up to `maxtrials` times until the value of the merit function has decreased. If the merit function has not decreased after all `maxtrials` steps, the smallest step is accepted.

- `tolLag`: a small threshold for the values of the Lagrange multipliers μ_i^* under which these are considered to be 0 (we took `tolLag=1e-8`). This value is used for the identification of the set $\widehat{I}(x_n)$ according to the definition (3.3.17); it should be set in accordance with the machine precision and that of the quadratic programming solver for the dual problem (3.3.10).

Let us emphasize that these parameters have a quite intuitive and physical meaning, so that the task of assigning their values does not involve fine tunings in practice.

Importantly, the rescaling induced by the inverse of the correlation matrix $(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x_n)})^{-1}$ normalizes all the constraints; in particular, the whole algorithm 3.1 is invariant under multiplication of the constraints by arbitrary positive constants (up to the machine precision for the step 3); a preliminary rescaling of the constraints is therefore not required from the user.

3.4.3 Overall algorithm pseudo code

The resulting algorithmic implementation of the null space gradient flow taking into account both adaptations of section 3.4.1 and (3.4.2) is summarized in algorithm 3.1 below.

3.5 Comparisons with other methods and illustrations on academic test cases

This section is purely pedagogical and serves to illustrate and compare our null space gradient flow to other classical first order methods for constrained optimization problems. The method of slack variables for treating inequality constraints with the equality constrained gradient flow (3.1.11) is reviewed and compared with our method in section 3.5.1. The comparison of our gradient flow algorithm with the more classical SLP and Augmented Lagrangian methods are presented in section 3.5.2.

3.5.1 Comparison with the method of slack variables for inequality constraints

It is classical to introduce slack variables so as to turn inequality constraints in the problem (3.1.2) into equality constraints of an augmented problem including these additional variables. This section briefly reviews the method investigated by [277] in the context of dynamical system approaches to constrained optimization and compares it with our method based on the dual problem (3.3.10).

The method of slack variables consists in replacing the problem (3.1.2) with the following equivalent one, involving q extra variables $(z_1, \dots, z_q) \in \mathbb{R}^q$:

$$\begin{aligned} \min_{\substack{x\in V\\ z\in\mathbb{R}^q}} \quad & J(x) \\ \text{s.t.} \quad & \boldsymbol{C}(x,z) = 0, \end{aligned} \tag{3.5.1}$$

where the augmented vector of constraints $\boldsymbol{C}(x,z)$ reads:

$$\boldsymbol{C}(x,z) := \begin{bmatrix} \boldsymbol{g}(x) \\ h_1(x) + \frac{1}{2}z_1^2 \\ \vdots \\ h_q(x) + \frac{1}{2}z_q^2 \end{bmatrix} \in \mathbb{R}^{p+q}.$$

Problem (3.5.1) is an equality constrained optimization problem of the form (3.2.1), set over the Hilbert space

$$\widetilde{V} := V \times \mathbb{R}^q, \text{ with inner product } \langle (x,z),(x',z')\rangle_{\widetilde{V}} := \langle x, x'\rangle_V + z^T z'.$$

It can be solved thanks to the proposed algorithm in section 3.2 for equality-constrained problems. The

Algorithm 3.1 Discretization of the null space gradient flow (3.3.27), based on a merit function.

for $n = 1 \dots \texttt{maxiter}$ **do**

1. Compute the gradients $\nabla J(x_n)$, $\nabla g_i(x_n)$ and $\nabla h_j(x_n)$ for $1 \leq i \leq p$, $1 \leq j \leq q$ by solving, if necessary, the identification problem (3.2.3) and (3.2.4).
2. For all inequality constraints $1 \leq i \leq q$, compute the tolerance

$$\varepsilon_i := ||\nabla h_i(x_n)||_V \, \texttt{h}.$$

3. Determine the set $\widetilde{I}(x_n)$ of active or violated constraints and the set $\widetilde{I}_\varepsilon(x_n)$ of constraints violated "up to ε_i":

$$\widetilde{I}(x_n) = \{i \in \{1, \dots, q\} \,|\, h_i(x_n) \geq 0\}$$
$$\widetilde{I}_\varepsilon(x_n) = \{i \in \{1, \dots, q\} \,|\, h_i(x_n) \geq -\varepsilon_i\}.$$

4. Denote by $\widetilde{q}_\varepsilon := \text{Card}(\widetilde{I}_\varepsilon)$. Solve the dual problem

$$(\boldsymbol{\lambda}^*_\varepsilon(x_n), \boldsymbol{\mu}^*_\varepsilon(x_n)) := \arg \min_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^p \\ \boldsymbol{\mu} \in \mathbb{R}^{\widetilde{q}_\varepsilon(x)}, \, \boldsymbol{\mu} \geq 0}} ||\nabla J(x) + \mathrm{D}\boldsymbol{g}(x)^{\mathcal{T}} \boldsymbol{\lambda} + \mathrm{D}\boldsymbol{h}_{\widetilde{I}_\varepsilon(x)}(x)^{\mathcal{T}} \boldsymbol{\mu}||_V$$

to obtain the optimal Lagrange multiplier $\boldsymbol{\mu}^*(x_n)$. Infer the subset $\widehat{I}_\varepsilon(x_n) \subset \widetilde{I}_\varepsilon(x_n)$ indicating which constraints must remain active (proposition 3.4) :

$$\widehat{I}_\varepsilon(x_n) = \{i \in \widetilde{I}_\varepsilon(x_n) \,|\, \mu^*_{\varepsilon,i}(x_n) > \texttt{tolLag}\}. \tag{3.4.8}$$

5. Let $I^*_\varepsilon(x_n) := \widetilde{I}(x_n) \cup \widehat{I}_\varepsilon(x_n)$. Extract the vectors $\boldsymbol{C}_{\widehat{I}_\varepsilon(x_n)}(x_n)$ and $\boldsymbol{C}_{I^*_\varepsilon(x_n)}(x_n)$ (defined by (3.1.4)) and compute

$$\begin{aligned} \boldsymbol{\xi}_J(x_n) &= (I - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}_\varepsilon(x_n)}(\mathrm{D}\boldsymbol{C}_{\widehat{I}_\varepsilon(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widehat{I}_\varepsilon(x_n)})^{-1}\mathrm{D}\boldsymbol{C}_{\widehat{I}_\varepsilon(x_n)})\nabla J(x_n), \\ \boldsymbol{\xi}_C(x_n) &= \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I^*_\varepsilon(x_n)}(\mathrm{D}\boldsymbol{C}_{I^*_\varepsilon(x_n)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{I^*_\varepsilon(x_n)})^{-1}\boldsymbol{C}_{I^*_\varepsilon(x_n)}. \end{aligned} \tag{3.4.9}$$

for $k = 1 \dots \texttt{maxtrials}$ **do**

Compute the step

$$x_{n+1} = x_n - \frac{\Delta t}{2^{k-1}}(\alpha_J \boldsymbol{\xi}_J(x_n) + \alpha_C \boldsymbol{\xi}_C(x_n)).$$

if $\texttt{merit}_{x_n}(x_{n+1}) < \texttt{merit}_{x_n}(x_n)$ **then**

break

end if

end for

end for

associated gradient flow for (3.5.1) reads:

$$\begin{cases} \begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = -\alpha_J (I - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}\mathrm{D}\boldsymbol{C}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}) \begin{bmatrix} \nabla J(x(t)) \\ 0 \end{bmatrix} - \alpha_C \mathrm{D}\boldsymbol{C}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}\mathrm{D}\boldsymbol{C}^{\mathcal{T}})^{-1}\boldsymbol{C}(x(t), z(t)), \\ \begin{bmatrix} x(0) \\ z(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ z_0 \end{bmatrix}, \end{cases} \tag{3.5.2}$$

where $(x_0, z_0) \in \widetilde{V}$ is a suitable initialization. In our implementation, for comparison purposes, the variable z is initialized with a value $z_0 \in \mathbb{R}^q$ in such a way that the inequality constraints of (3.1.2) which are inactive for x_0 (i.e. $h_i(x_0) < 0$) are associated with satisfied equality constraints $C_{p+i}(x_0, z_0) = 0$ in (3.5.1):

$$\forall i \in \{1, \dots, q\},\ z_{0,i} = \sqrt{2|h_i(x_0)|}.$$

In the finite-dimensional setting $V = \mathbb{R}^k$ and when J, $\boldsymbol{g}$ and $\boldsymbol{h}$ are $\mathcal{C}^2$ functions, Schropp and Singer proved in [277] that:

(i) stationary points of the extended flow (3.5.2) are exactly critical points of (3.1.2), that is points x^* satisfying (3.3.7) but with $\boldsymbol{\mu}(x^*) \in \mathbb{R}^q$ possibly negative;

(ii) among all possible critical points, only KKT points (fulfilling all three conditions (3.3.7) with $\boldsymbol{\mu}(x^*) \in \mathbb{R}^q_+$) are asymptotically stable equilibria.

As a consequence, the solution vector $x(t)$ to (3.5.2) converges in practice to a KKT point for problem (3.1.2).

The main differences between the slack variable approach and our proposed flow (3.3.27) for dealing with equality and inequality constrained problems can be summarized as follows.

1. Any point x^{crit} satisfying the constraints ($\boldsymbol{C}_{\widetilde{I}(x^{crit})}(x^{crit}) = 0$) and $\Pi_{\boldsymbol{C}_{\widetilde{I}(x^{crit})}}(\nabla J(x^{crit})) = 0$ is a stationary point of the extended dynamical system (3.5.2), although it might violate the full KKT condition (because (3.3.6) may yield possible negative values of the multiplier $\boldsymbol{\mu}_{\widetilde{I}(x^{crit})}(x^{crit})$). In contrast, x^* is a stationary points of the flow (3.3.27) if and only if it is a true feasible KKT point; see proposition 3.5.

2. The computation of $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$ in our flow (3.3.27) requires to invert a matrix of size at most $(p+\widetilde{q}(x))$-by-$(p+\widetilde{q}(x))$ with $\widetilde{q}(x)$ the number of active or violated constraints at x. The process of equalizing inequality constraints as in [277, 282] rather requires to invert the full $(p+q)$-by-$(p+q)$ matrix $\mathrm{D}\boldsymbol{C}(x,z)\mathrm{D}\boldsymbol{C}(x,z)^{\mathcal{T}}$. Our method is therefore more efficient if $\widetilde{q}(x) \ll q$, that is if a lot of inequality constraints are inactive.

3. At feasible points, our ODE (3.3.27) follows the best locally admissible descent direction (with respect to the norm of V). This is not the case for the extended ODE (3.5.2). Therefore, from a common feasible point x, the ODE (3.3.27) always decreases the objective function with a steeper slope $\mathrm{d}J/\mathrm{d}s$ with respect to the parameterization induced by the path length s, defined as a function of the time t by

$$s(t) = \int_0^t ||\dot{x}(\alpha)||_V \mathrm{d}\alpha. \tag{3.5.3}$$

This property is illustrated in the academic examples of section 3.5, and in particular on Figure 3.9a below.

All in all, our observations based on the simple numerical examples of the next section 3.5.3 tend to illustrate that both flows (3.3.27) and (3.5.2) may have equivalent performances for solving the non linear optimization problem (3.1.2), this performance being measured in term of the total length

$$S = \int_0^{+\infty} ||\dot{x}||_V \mathrm{d}t$$

covered by the optimization path to reach the optimum. However, the two ODEs (3.3.27) and (3.5.2) yield optimization paths of essentially different natures. Our null space flow (3.3.27) ignores inactive constraints and those aligned with the gradient of the objective function. As a result, it produces non smooth paths that are more likely to reach quickly the saturation of the constraint. The extended flow (3.5.2) yields smoother trajectories that more likely stay away from the constraints, at the cost of inverting at every step the full matrix $\mathrm{D}\boldsymbol{C}(x,z)\mathrm{D}\boldsymbol{C}(x,z)^{\mathcal{T}}$ of the size of the total number of constraints (active and inactive).

3.5.2 Comparisons with 'iterative' optimization algorithms

This part compares the null space gradient flow with a few classical first order methods for constrained optimization which are not directly interpretable as the discretization of some dynamical system. We consider in the next paragraphs the Augmented Lagrangian Method and Sequential Linear Programming (SLP). Many other algorithms exist and could have been examined, such as the Method of Feasible Directions (MFD) [327, 308] or the Method of Moving Asymptotes (MMA) [298]. These methods can be considered as more or less sophisticated variant of SLP: they both rely on the resolution of linearized subproblems over a trust region that must be determined by the user. These methods present therefore similar characteristics to the SLP method. The reader is referred to the textbooks [244, 70] for further material on iterative methods for constrained optimization.

The Augmented Lagrangian method

The Augmented Lagrangian Method is an optimization algorithm which is very popular partly because of its implementation simplicity: it consists in replacing (3.1.2) with a sequence of unconstrained minimization problems

$$\min_{x\in V} \quad \mathcal{L}_k(x) = J(x) + \boldsymbol{\lambda}_k^T \boldsymbol{g}(x) + \boldsymbol{\mu}_k^T \boldsymbol{h}(x) + \frac{\alpha_k}{2}||\boldsymbol{g}(x)||^2 + \frac{\alpha_k}{2}||\boldsymbol{h}_{\widetilde{I}(x)}(x)||^2. \tag{3.5.4}$$

Here, the tentative Lagrange multipliers have been denoted $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ in order to be consistent with our previous notation. The parameter α_k serves to penalize the violation of the constraints.

A sequence x_n which *hopefully* converges to a solution to the original problem (3.1.2) is typically computed by alternating gradient steps for the unconstrained minimization of (3.5.4),

$$x_{n+1} = x_n - \Delta t\Big(\nabla J(x_n) + \mathrm{D}\boldsymbol{g}(x_n)^{\mathcal{T}}\boldsymbol{\lambda}_k + \mathrm{D}\boldsymbol{h}(x_n)^{\mathcal{T}}\boldsymbol{\mu}_k + \alpha_k \mathrm{D}\boldsymbol{g}(x_n)^{\mathcal{T}}\boldsymbol{g}(x_n) + \alpha_k \mathrm{D}\boldsymbol{h}_{\widetilde{I}(x_n)}(x_n)^{\mathcal{T}}\boldsymbol{h}_{\widetilde{I}(x_n)}\Big) \tag{3.5.5}$$

with updates of the Lagrange multipliers $\boldsymbol{\lambda}_k$, $\boldsymbol{\mu}_k$, and possibly (but this is not compulsory) of the penalization parameter α_k. Note that the iteration indices n and k may be independent: one major difficulty of the method lies in the correct initialization and updates of the parameters $\boldsymbol{\lambda}_k$, $\boldsymbol{\mu}_k$, α_k. Generally, the parameter α_k is assumed to be constant, and an update rule is commonly assumed for updating $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ (see [244]):

$$\lambda_{k+1,i} := \lambda_{k,i} + \alpha_k \boldsymbol{g}_i(x_k^*), 1 \le i \le p, \qquad \mu_{k+1,j} := \begin{cases} \mu_{k,j} + \alpha_k h_j(x_k^*) \text{ if } h_j(x_k^*) > -\dfrac{\mu_{k,j}}{\alpha_k} \\ 0 \text{ otherwise,} \end{cases}, 1 \le j \le q \tag{3.5.6}$$

where x_k^* is a minimizer for (3.5.4). Of course, in practice, (3.5.6) is not computed with the true minimizer x_k^* but with a current iterate x_n obtained after one or more iterations of (3.5.5).

Besides being very difficult to use as soon as the problem involves more than one constraint, this algorithm is rather slow, because it uses *tentative* values for the Lagrange multipliers $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ rather than the optimal multipliers $\boldsymbol{\lambda}^*(x_n), \boldsymbol{\mu}^*(x_n)$ of (3.3.10). As a result, it can take a lot of iterations for $\boldsymbol{\lambda}_k$ and $\boldsymbol{\mu}_k$ to converge to the correct multipliers, often resulting in oscillations around the constraints at convergence. Recall that actually, lemma 3.3 highlights that our null space algorithm can be seen as an Augmented Lagrangian Method with a "clever" choice of $\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k$, and α_k (in fact, α_k is replaced by the symmetric positive definite matrix $\boldsymbol{S}(x_n)$).

The examples of section 3.5.3 rely on the implementation of this algorithm available in [150]. The step α_k is kept constant equal to 1.1, and (3.5.6) is performed after every single iteration of (3.5.5), which means $n = k$.

Sequential Linear Programming

The SLP algorithm is a very popular first order method for equality and inequality constrained problems. The descent direction given by the SLP algorithm is of the form (3.3.34). The main idea of the method is to obtain the next iterated x_{n+1} by solving a linearized version of the original problem (3.1.2) around the current guess x_n:

$$x_{n+1} := \arg\min \quad J(x_n) + \mathrm{D}J(x_n)\cdot(x_{n+1}-x_n) \qquad s.t. \begin{cases} \boldsymbol{g}(x_n) + \mathrm{D}\boldsymbol{g}(x_n)\cdot(x_{n+1}-x_n) = 0 \\ \boldsymbol{h}(x_n) + \mathrm{D}\boldsymbol{h}(x_n)\cdot(x_{n+1}-x_n) \leq 0 \\ ||x_{n+1}-x_n||_V \leq \delta. \end{cases} \tag{3.5.7}$$

The constraint $||x_{n+1}-x_n||_V \leq \delta$ requires that x_{n+1} should be found in a trust region of size $\delta > 0$ around x_n (otherwise the value of the subproblem (3.5.7) could be $-\infty$ due to unbounded minimizing sequences). Note that when the original problem (3.1.2) is set in $\mathbb{R}^k$, this constraint is more often expressed as a bound constraint

$$-\delta \leq x_{n+1,i} - x_{n,i} \leq \delta, \quad \forall 1 \leq i \leq k,$$

because this allows to use gradient projection methods [244].

Actually, there is a connection between null space iterates (3.3.34) and SLP iterates. Indeed, assume that the constraint $||x_{n+1}-x_n||_V \leq \delta$ is saturated and that the inequality constraints

$$h_i(x_n) + \mathrm{D}h_i(x_n)\cdot(x_{n+1}-x_n) \leq 0$$

get saturated for i belonging to some index set I_n. Then it is readily verified (in the line of the point 2. of lemma 3.1, or in the line of proposition 3.4) that the solution of (3.5.7) is explicitly given by

$$x_{n+1} = x_n - \alpha_n(I - \mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_{I_n}\mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}_{I_n})\nabla J(x_n) - \mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_{I_n}\mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}})^{-1}\boldsymbol{C}_{I_n}(x_n) \tag{3.5.8}$$

where $\boldsymbol{C}_I = \begin{bmatrix} \boldsymbol{g}(x_n) & \boldsymbol{h}_I \end{bmatrix}^T$ is the vector of saturated linearized constraints, and $\alpha_n > 0$ is a scaling coefficient given by

$$\alpha_n = \frac{\sqrt{\delta^2 - ||\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_I\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}})^{-1}\boldsymbol{C}_I(x_n)||_V^2}}{||(I - \mathrm{D}\boldsymbol{C}_I^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_I\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}})^{-1}\mathrm{D}\boldsymbol{C}_I)\nabla J(x_n)||_V}.$$

Note that if the trust region size δ is too small, i.e. if

$$\delta < ||\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_I\mathrm{D}\boldsymbol{C}_I^{\mathcal{T}})^{-1}\boldsymbol{C}_I(x_n)||_V$$

then the problem (3.5.7) does not have a solution.

The scheme (3.5.8) is a variant of the discretization (3.3.34) of the null space gradient flow, up to the redefinition of α_C and α_J. The index set I_n may also be different from $\widehat{I}(x_n)$, because the subproblem (3.5.7) sees all the constraints while the null space method sees only the active or violated ones.

As is underlined by the expression of α_n, the major problem of the SLP method is the fact that the SLP subproblem (3.5.7) may not have a solution if the trust region parameter δ is too small. Some additional tuning operations are usually done to face this issue, see e.g. [174]. This issue is partly due to the fact that the update scheme (3.5.8) is not the discretization of some ODE, because the range space step $\mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}}(\mathrm{D}\boldsymbol{C}_{I_n}\mathrm{D}\boldsymbol{C}_{I_n}^{\mathcal{T}})^{-1}\boldsymbol{C}_{I_n}(x_n)$ has not been scaled by some small parameter (e.g. by δ in our case). The SLP method directly imposes constraints to be satisfied at first order, whereas our null space method requires rather constraints to decrease by a factor $\alpha_C\Delta t$ at the next iteration (see the discussion at the beginning of section 3.2.2 and remark 3.3). Note that in contrast with the SLP method, the primal and dual subproblems (3.3.9) and (3.3.10) always admit a solution.

In the following examples of section 3.5.3, the SLP method is implemented with the code available in [150]. Let us note that this implementation of the SLP method does not solve the subproblem (3.5.7) in the unfeasible domain, but a variant in order to deal with violated constraints.

3.5.3 Comparative academic test cases in the euclidean space $\mathbb{R}^k$

In this section, we consider simple and illustrative academic examples in order to compare qualitatively the above strategies for dealing with inequality constraints in optimization problems; we consider:

- the method of section 3.5.1 for equalizing inequality constraints by means of slack variables; see (3.3.27). This strategy is hereafter labeled as 'SLACK';
- the proposed null space flow (3.3.27) in section 3.3.2, based on the dual problem (3.3.10) for solving the combinatorial character of the constraints. This method is labeled as 'NLSPACE';
- an alternative, naive version of (3.3.27) which does not take advantage of the use of a dual problem, and simply projects $\nabla J(x(t))$ on all the violated constraints:

$$\begin{cases} \dot{x} = -\alpha_J \widetilde{\boldsymbol{\xi}}_J(x(t)) - \alpha_C \boldsymbol{\xi}_C(x(t)) \\ \widetilde{\boldsymbol{\xi}}_J(x) := (I - \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)})^{-1}\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)})\nabla J(x) \\ \boldsymbol{\xi}_C(x) := \mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)}(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x)}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x)})^{-1}\boldsymbol{C}_{\widetilde{I}(x)}(x). \end{cases} \tag{3.5.9}$$

  In other words, all the violated or saturated constraints are taken into account in the computation of both the null space and range space directions $\boldsymbol{\xi}_J(x)$ and $\boldsymbol{\xi}_C(x)$. This strategy is labeled as 'NLSPACE (no dual)';
- the Augmented Lagrangian Method of section 3.5.2, labeled as 'AULG';
- the Sequential Linear Programming algorithm of section 3.5.2, labeled as 'SLP'.

To achieve our comparison purpose, algorithm 3.1 is implemented for the discretization of (3.3.27), (3.5.2) and (3.5.9), with straightforward adaptations for equalizing slack variables or disabling the resolution of the dual problem. In all the considered cases, we have set the values of α_J and α_C such that $\alpha_J/\alpha_C = 5/3$. The step size Δt for the discretization of the ODE (3.1.3) was chosen sufficiently small to compute continuous paths with satisfying accuracy. Our discussion is exclusively focused on the continuous trajectories of the considered ODEs. In particular, we do neither discuss the issue of the selection of the time step, nor the efficiency of these methods in terms of the needed number of iterations required to achieve convergence.

In order to compare the three methods without bias, we consider the arc length $s(t)$ (defined in (3.5.3)) as the common reference time for the three ODEs (3.3.27), (3.5.2) and (3.5.9); recall indeed that this quantity is invariant under any monotone parameterization change of the time t. In the convergence figures below, optimized quantities are then plotted with respect to the pseudo time $s(t)$ in abscissa; for example we plot the graph $t \mapsto (s(t), J(x(s(t))))$ in order to account for the evolution of the objective function J. The SLP and Augmented Lagrangian methods are not dynamical system approaches: we decided not to include them when plotting convergence histories for the objective function and constraints because the arc length $s(t)$ would not make sense for these.

We shall also plot the evolution of the Lagrange multipliers $s \mapsto \boldsymbol{\mu}(x(s))$ associated with $\boldsymbol{\xi}_J(x(s))$ or $\widetilde{\boldsymbol{\xi}}_J(x(s))$ for the ODEs (3.3.27) and (3.5.9). For that purpose, these Lagrange multipliers are defined on the violated indices $i \in \widetilde{I}(x(s))$ by (3.3.18) for the null space flow (3.3.27), and by

$$\boldsymbol{\mu}(x(s)) := -(\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x(s))}\mathrm{D}\boldsymbol{C}^{\mathcal{T}}_{\widetilde{I}(x(s))})^{-1}\mathrm{D}\boldsymbol{C}_{\widetilde{I}(x(s))}\nabla J(x(s)) \tag{3.5.10}$$

for the flow (3.5.9) that does not use the dual problem (3.3.10). For the indices $i \in \{1, \dots, q\} \setminus \widetilde{I}(x(s))$, the value of the Lagrange multiplier is set to $\mu_i(x(s)) := 0$ by convention. We do not plot Lagrange multipliers for the ODE (3.5.2) using slack variables because these are defined with respect to the extended variables $(x(s), z(s))$.

The examples of this section take place in the optimization set $V = \mathbb{R}^2$, which is equipped with the usual euclidean inner product; the Hilbert transposition $\mathcal{T} = T$ coincides with the usual transposition operator (see definition 3.1). For simplicity, these examples only involve inequality constraints; we consider the following three scenarios:

Test case 1: the initial point is unfeasible and the gradient of the objective function $\nabla J(x)$ is always aligned with the directions of the constraints;

Test case 2: the initial point is unfeasible, but the gradient $\nabla J(x)$ may not be aligned with the direction of constraints;

Test case 3: one of the constraints becomes inactive in the course of the optimization path.

Test case 1 : unfeasible initialization with initial gradient aligned with the constraints.

Our first example features the following problem, reproduced from [150]:

$$\begin{aligned} \min_{(x_1,x_2)\in\mathbb{R}^2} \quad & J(x_1,x_2) := x_2 + 0.3x_1 \\ \text{s.t.} \quad & \begin{cases} h_1(x_1,x_2) := -x_2 + \dfrac{1}{x_1} & \leq 0, \\ h_2(x_1,x_2) := x_1 + x_2 - 3 & \leq 0. \end{cases} \end{aligned} \tag{3.5.11}$$

This test case is designed so that for the chosen initial point $x_0 = (1.5, 2.25)$, the gradient of the objective function $\nabla J(x)$ is 'aligned' with the linear constraint h_2, in the sense that

$$-\nabla h_2(x_0) \cdot \nabla J(x_0) < 0.$$

Hence at least for small times (in fact during the whole optimization path), the constraint h_2 can be ignored since the minimization of J is naturally concurrent with a decrease of the value of h_2.

The optimization paths taken by the solutions of the three ODEs (3.3.27), (3.5.2) and (3.5.9) are plotted on Figure 3.2. The associated convergence histories for the values of the objective and constraint functions are displayed on Figure 3.3. Note the oscillations characterizing the Augmented Lagrangian Method to travel tangentially to the constraint boundary. The SLP algorithm is able to compute the tangential projection of $\nabla J(x(t))$ onto the constraint set; it finds a path that is similar to the one of the null space method.

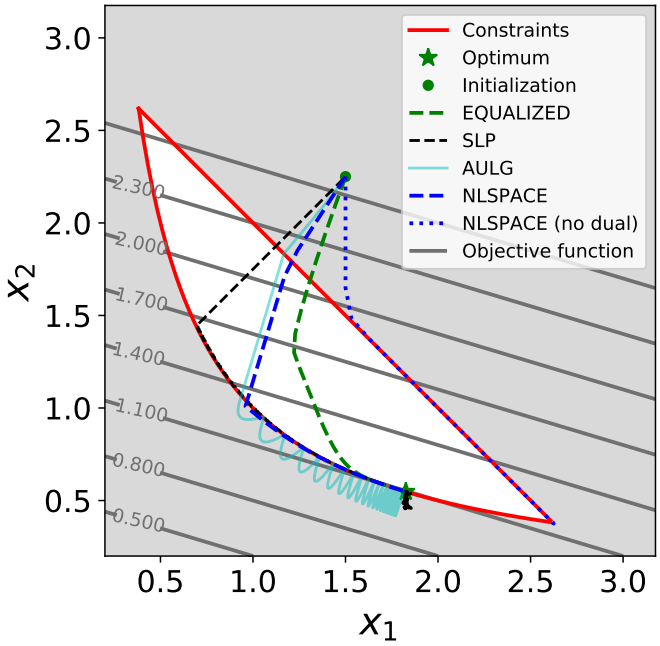
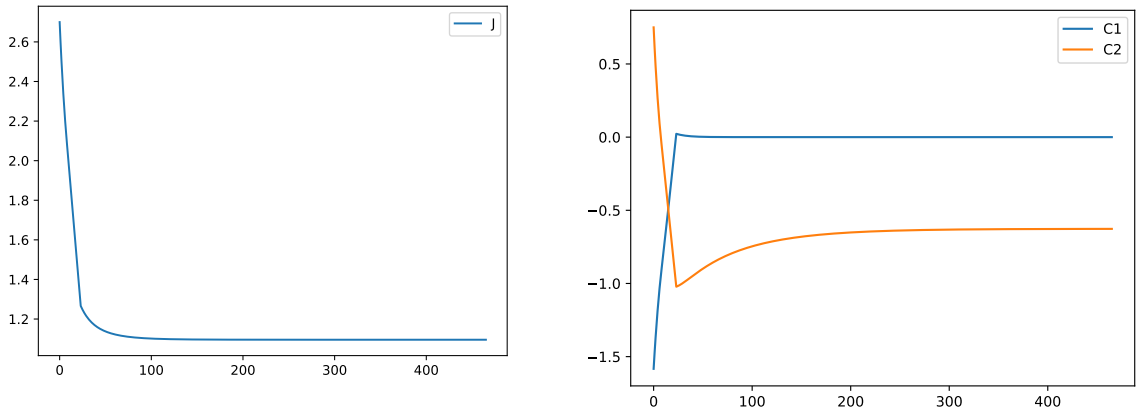


Figure 3.2: Optimization test case 1: optimization paths for an unfeasible initialization x_0 with $\nabla J(x_0)$ aligned in the direction of the constraints.

Let us comment the trajectory followed by the null space flow (3.3.27) in details. The gradient of the objective function remains aligned with the constraint h_2, which is associated with a zero Lagrange multiplier $\mu_2(x(s))$ (see Figure 3.4). During the first part of the optimization, the first constraint h_1 is not violated, hence the multiplier $\mu_1(x(s))$ is also set to zero. As a result, both constraints are ignored when computing the null space direction, which is set equal to the gradient: $\boldsymbol{\xi}_J(x(s)) = \nabla J(x(s))$. The optimization path $x(s)$ follows then almost the direction of the gradient $\nabla J(x(s))$ (without projection), up to a small deviation induced by the non zero Gauss-Newton direction $\boldsymbol{\xi}_C(x(s))$ in the unfeasible domain. When the hyperbolic constraint represented by h_1 becomes violated, the gradient is not aligned anymore with this constraint and the dual problem (3.3.10) yields a non-zero Lagrange multiplier $\mu_1(x(s))$ (near $s = 1.4$). From this point, the gradient $\nabla J(x(s))$ is then projected tangentially to the constraint h_1 till the optimum is attained.

In contrast, the path of the ODE (3.5.9) (which does not involve the dual problem (3.3.10)) fails to find the optimum as it is unable to unstick from the first saturated constraint. Notably, the gradient $\nabla J(x)$ is kept being projected tangentially to the violated constraint h_2 while it should not, which could have been detected from the negativity of the computed Lagrange multiplier μ_2 (see Figure 3.4).

Finally, the extended ODE (3.5.2) making use of slack variables feels the constraint h_1 from some distance, inducing a deviation of the trajectory in the direction of the optimum before reaching the saturation of the constraint. This allows the trajectory $x(s)$ to find globally a slightly shorter path than our null space flow (3.3.27).



(a) Objective function J (b) Constraints $\boldsymbol{h}$

Figure 3.3: History curves for the optimization test case 1.

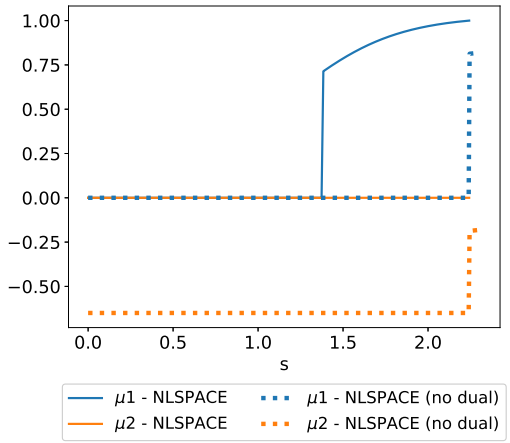


Figure 3.4: Evolution of the Lagrange multipliers $\mu_1(x(s)), \mu_2(x(s))$ for the optimization test case 1.

Test case 2 : unfeasible initialization with initial gradient not aligned with the constraints.

We now devise a test case where the gradient of the initialization is not aligned with the constraints. The feasible domain is the same as in the previous test case but the objective function is different:

$$\min_{(x_1,x_2)\in\mathbb{R}^2} \quad J(x_1,x_2) := (x_1-2)^2 + (x_2-2)^2$$
$$s.t. \begin{cases} h_1(x_1,x_2) := -x_2 + \dfrac{1}{x_1} \leq 0 \\ h_1(x_1,x_2) := x_1 + x_2 - 3 \leq 0. \end{cases} \tag{3.5.12}$$

We keep the same initialization $x_0 = (1.5, 2.25)$. Corresponding optimization paths and convergence curves are displayed on Figs. 3.5 and 3.6.

For this example, the linear constraint h_2 is not aligned with the gradient along the optimization path of the null space gradient flow (3.3.27). This is associated with a non-zero Lagrange multiplier $\mu_2(x(s)) > 0$ (see Figure 3.7): the gradient $\nabla J(x(s))$ is kept being projected tangentially to the constraint h_2 when computing $\boldsymbol{\xi}_J(x(s))$. Here, the combination with the Gauss-Newton direction $\boldsymbol{\xi}_C(x(s))$ allows to decrease simultaneously the objective function and the violation of the constraints, which enables the optimization path to reach directly the optimum when hitting the feasible set. Note that the convergence curve Figure 3.6a depicts a monotonically increasing objective function J after $s \geq 0.5$ (although we are minimizing J); this is of course due to the fact that constraints are never satisfied.

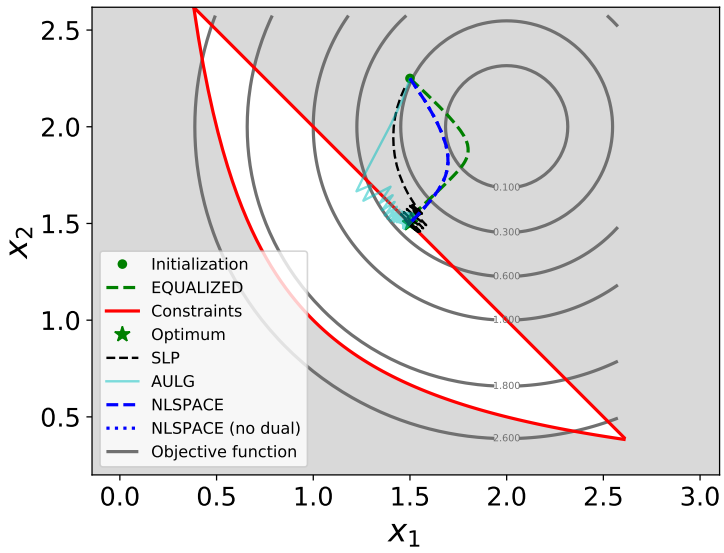
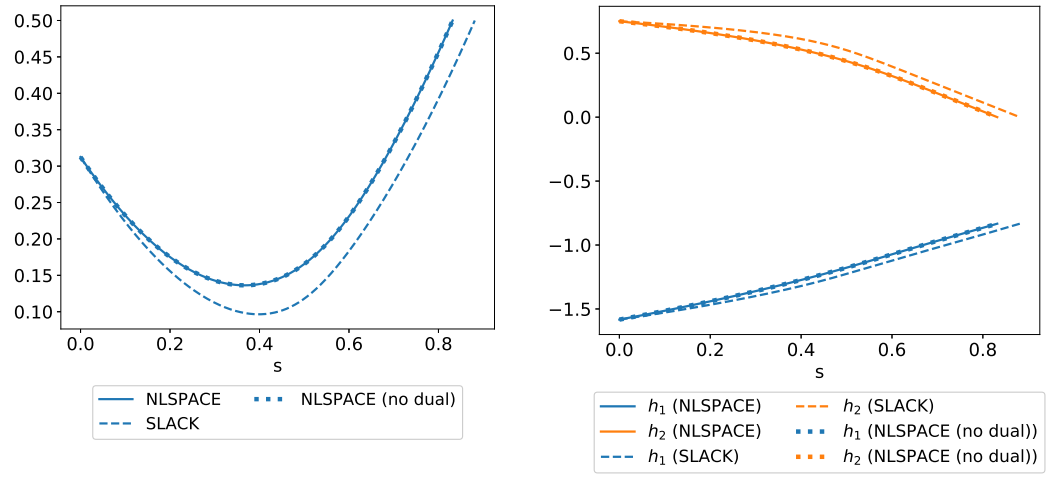


Figure 3.5: Optimization problem of section 3.5.3: unfeasible initialization x_0 with $\nabla J(x_0)$ not aligned in the direction of the constraints.



(a) Objective function J

(b) Constraints $\boldsymbol{h}$

Figure 3.6: History curves for the optimization test case 2.

Since the constraint h_2 remains active from initialization to the optimum, the optimization path is unchanged when disabling the dual problem (ODE (3.5.9)). Finally, the path selected by the extended flow (3.5.2) to reach the optimum is very similar to the one of our method, although slightly longer.

Test case 3: a saturated inequality constraint becoming inactive along the optimization path

This last optimization test case is designed to illustrate the relevance of the dual problem for detecting when a saturated inequality constraint ceases to be saturated. We consider a disconnected unfeasible

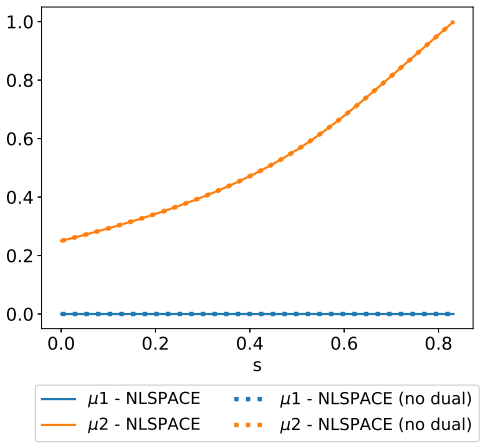


Figure 3.7: Evolution of the Lagrange multipliers $\mu_1(x(s)), \mu_2(x(s))$ for the optimization test case 2.

domain made from the reunion of a half-space and the interior region of a parabola:

$$\begin{aligned} \min_{(x_1,x_2)\in\mathbb{R}^2} \quad & J(x_1,x_2) = x_1^2 + (x_2+3)^2 \\ \text{s.t.} \quad & \begin{cases} h_1(x_1,x_2) = -x_1^2 + x_2 & \leq 0 \\ h_2(x_1,x_2) = -x_1 - x_2 - 2 & \leq 0. \end{cases} \end{aligned} \tag{3.5.13}$$

The feasible domain and optimization paths starting from the initialization $x_0 = (3,3)$ are displayed on Figure 3.8. Associated convergence curves are reported on Figure 3.9. The optimization paths obtained with the Augmented Lagrangian Method and the SLP method are also depicted. We observed oscillations of the SLP method in the region close to both constraints. We note as well, again, the tendency of the Augmented Lagrangian Method iterates to oscillate around the constraints.

For the null space flow (3.3.27), four different stages occur as is visible on the evolution of the Lagrange multipliers $\mu_1(x(s)), \mu_2(x(s))$ reported on Figure 3.10:

1. From $s = 0$ to $s = 1.73$, the trajectory $x(s)$ remains in the feasible domain. Lagrange multipliers $\mu_1(x(s)) = \mu_2(x(s)) = 0$ are set to 0 and the null space direction $\boldsymbol{\xi}_J(x(s)) = \nabla J(x(s))$ coincides with the gradient of the objective function, until $x(s)$ hits the parabolic barrier, which corresponds to the saturation of the first constraint h_1.

2. From $s = 1.73$ to $s = 4.4$, the resolution of the dual problem yields a non zero multiplier $\mu_1(x(s)) > 0$. The optimization trajectory $x(s)$ remains tangent to the first constraint, until reaching a limit point such that $\nabla J(x(s)) \cdot \nabla h_1(x(s)) = 0$. At this moment, it is not necessary to project the gradient tangentially to this constraint any more, and the values of both Lagrange multipliers $\mu_1(x(s)) = \mu_2(x(s))$ are equal to 0.

3. From $s = 4.4$ to $s = 6.5$, both constraints h_1 and h_2 are ignored and the trajectory $x(s)$ follows the gradient $-\nabla J(x(s))$, till the saturation of h_2.

4. From $s = 6.5$ to $s = 7.1$, the second Lagrange multiplier $\mu_2(x(s)) > 0$ has a positive value; $x(s)$ evolves then tangentially to this constraint till the optimum is attained.

As illustrated on Figure 3.8, the use of the dual problem is key in the detection of the moment when the optimization trajectory $x(s)$ needs to be released from active inequality constraints. Because of the discrete nature of the time stepping, the path followed by the ODE (3.5.9) necessarily enters slightly the violated parabolic domain. Since it does not use the information provided by the dual problem (3.3.10), the gradient $\nabla J(x(s))$ is kept being projected tangentially to the constraint h_1 till $x(s)$ converges to some stationary point (which is not a KKT point). As can be seen on Figure 3.10, the optimization trajectory followed by this ODE coincides with the one of the flow (3.3.27), till the instant $s = 4.4$ at which the Lagrange multiplier $\mu_1(x(s))$ becomes negative (which violates the feasibility condition of the dual problem (3.3.10)). Note that using larger steps could have allowed the trajectory $x(s)$ to exit "by chance" the unfeasible domain, and in that case convergence to the optimum would have been obtained. However this would not reflect the actual behavior of the continuous solutions of (3.5.9).

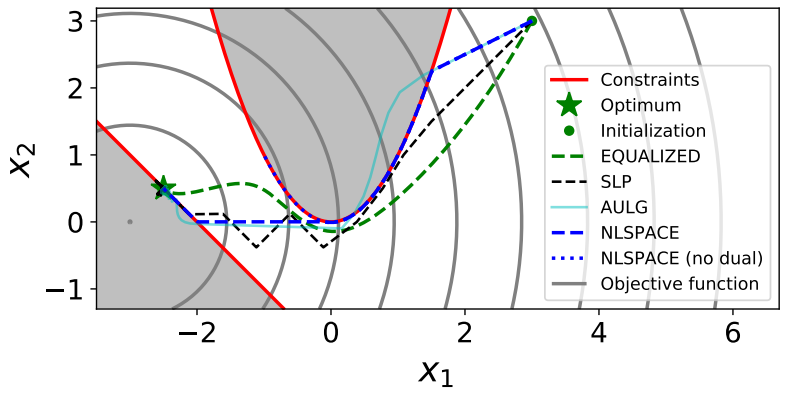
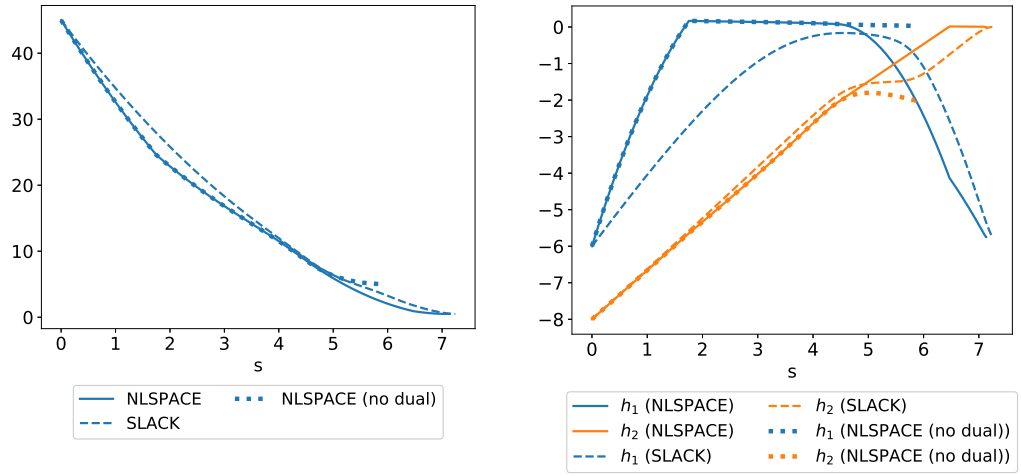


Figure 3.8: Optimization problem of section 3.5.3 : feasible initialization x_0 but the optimization has to find a path across the parabolic domain.



(a) Objective function J

(b) Constraints $\boldsymbol{h}$

Figure 3.9: History curves of the null space algorithm for the optimization problem of section 3.5.3.

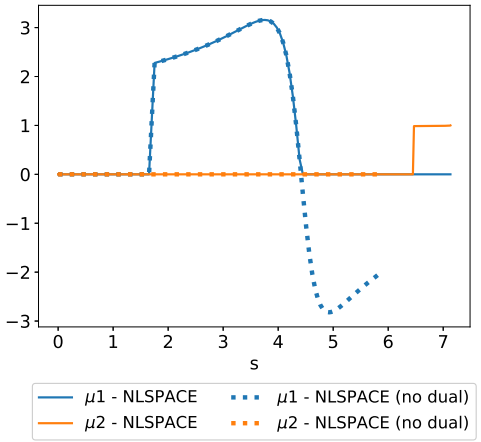


Figure 3.10: Evolution of Lagrange multipliers $\mu_1(x(s)), \mu_2(x(s))$ for the optimization problem of section 3.5.3.

Finally, the extended ODE (3.5.2) using slack variables finds a smooth path to the optimum. Since inactive constraints are felt from distance, the trajectory $x(s)$ is able to remain more strictly in the feasible domain for all times. The total length of the optimization path is almost the same than the one of the null space flow (3.3.27) (note the steeper descending slopes $\mathrm{d}J/\mathrm{d}s$ for the latter at intersection points of the two trajectories, see the point (3) in the discussion of section 3.5.1).

3.6 Optimization within the set of Lipschitz subdomains: applications to shape optimization

Our ultimate goal is to apply our optimization strategy to shape and topology optimization problems. In such applications, the optimization set is not a vector space V as in (3.1.2) but a more general set $\mathcal{X}$ of shapes in $\mathbb{R}^d$ ($d = 2$ or 3 in standard applications):

$$\mathcal{X} = \{\Omega \subset D \,|\, \Omega \text{ Lipschitz}\}, \tag{3.6.1}$$

where $D \subset \mathbb{R}^d$ is an enclosing 'hold-all' domain. Since $\mathcal{X}$ is not a Hilbert space, the present context does not fall into the optimization framework described in sections 3.2 and 3.3. However, $\mathcal{X}$ may be endowed with a manifold structure, which makes it possible to extend our dynamical system (3.3.1) to this context, up to small adaptations that we now describe.

In the whole section (and in the next chapters of the thesis), we consider a generic shape optimization problem

$$\begin{aligned} &\min_{\Omega \in \mathcal{X}} \quad J(\Omega) \\ &s.t. \begin{cases} \boldsymbol{g}(\Omega) = 0 \\ \boldsymbol{h}(\Omega) \leq 0, \end{cases} \end{aligned} \tag{3.6.2}$$

for shape differentiable objective and constraint functions $J : \mathcal{X} \to \mathbb{R}$, $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^p$ and $\boldsymbol{h} : \mathcal{X} \to \mathbb{R}^q$ (in the sense of chapter 1, definition 1.1). In section 3.6.1, we outline the analogy between the resolution of (3.6.2) with the method of Hadamard and classical methods for the optimization on smooth manifolds. Then, we explain how to extend the null space algorithm 3.1 to this setting, and in particular how the classical extension and regularization step of shape derivatives (outlined in chapter 1, section 1.4.1) is naturally included in our method when using the definition (3.2.3) of the Hilbertian transposition $\mathcal{T}$. A few implementation details regarding the normalization of the descent direction with respect to the mesh size are discussed in section 3.6.2. Finally, section 3.6.3 concludes this chapter with some further numerical illustrations of our constrained optimization algorithm in the context of shape optimization: we consider the model example of shape optimization of a bridge structure subjected to multiple load cases and featuring 10 constraints.

3.6.1 Manifold structures for shape optimization

Our extension of the previous material to the shape optimization context is inspired by 'classical' optimization strategies on a smooth embedded manifold $\mathscr{M} \subset \mathbb{R}^k$. In this context, a descent direction at a point $x_n \in \mathscr{M}$ for some objective functional is typically sought as an element $\boldsymbol{\xi}_n \in T_{x_n}\mathscr{M}$ of the tangent space $T_{x_n}\mathscr{M}$ to $\mathscr{M}$ at x_n; see e.g. [141, 5]. Then one relies on a *retraction* ρ_{x_n}, that is a mapping

$$\rho_{x_n} : T_{x_n}\mathscr{M} \to \mathscr{M}$$

satisfying the following two consistency conditions:

$$\begin{cases} \rho_{x_n}(0) = x_n \\ \forall \boldsymbol{\xi} \in T_{x_n}\mathscr{M}, \ \dfrac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \rho_{x_n}(t\boldsymbol{\xi}) = \boldsymbol{\xi}. \end{cases}$$

The mapping ρ_{x_n} then allows to convert $\boldsymbol{\xi}_n$ into a practical update of the actual point x_n on $\mathscr{M}$:

$$x_{n+1} := \rho_{x_n}(\Delta t \boldsymbol{\xi}_n), \tag{3.6.3}$$

where $\Delta t > 0$ is the descent step; see [6] and Figure 3.11. Since the new point x_{n+1} belongs to $\mathscr{M}$, this procedure can be repeated iteratively.

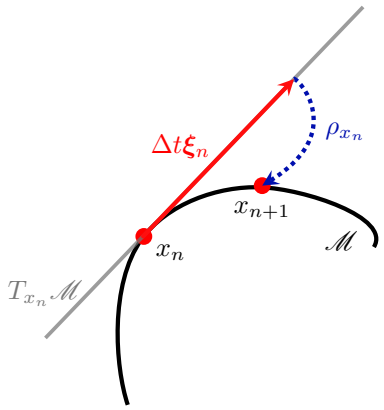


Figure 3.11: Optimization on a manifold $\mathscr{M}$: a retraction map ρ_{x_n} is used to project a tangential motion $\Delta t\boldsymbol{\xi}_n \in T_{x_n}\mathscr{M}$ from $x_n \in \mathscr{M}$ back onto the optimization domain $\mathscr{M}$.

The same idea can be used to apply the methods of sections 3.2 and 3.3 to the optimization problem (3.6.2), posed over the set of shapes $\mathcal{X}$. To this end, we rely on Hadamard's method (reviewed in chapter 1), which considers variations of a shape $\Omega \in \mathcal{X}$ of the form

$$\rho_\Omega(\boldsymbol{\theta}) := (I + \boldsymbol{\theta})(\Omega), \text{ for } \boldsymbol{\theta} \in W^{1,\infty}(D, \mathbb{R}^d) \text{ with } ||\boldsymbol{\theta}||_{W^{1,\infty}(D,\mathbb{R}^d)} < 1, \tag{3.6.4}$$

Formally, the set $W^{1,\infty}(D, \mathbb{R}^d)$ may be interpreted as the tangent space to $\mathcal{X}$ at Ω and the mapping ρ_Ω, which is defined by (3.6.4) on a neighborhood of $\boldsymbol{\theta} = 0$ in $W^{1,\infty}(D, \mathbb{R}^d)$, plays the role of a retraction. Other definitions are possible for such a transformation dictating how a shape should evolve according to a vector field $\boldsymbol{\theta}$, see the discussion of chapter 1, section 1.4.2. Note also that more rigorous manifold structures on shape spaces can be formulated, see e.g. [44, 278].

Usually, in the context of a general embedded manifold $\mathscr{M} \subset \mathbb{R}^k$, a differential structure on $\mathscr{M}$ is defined first (inducing a notion of derivative on $\mathscr{M}$), and the definition of a suitable retraction is inferred accordingly. In the framework of Hadamard's method however, it is the retraction ρ_Ω itself, that is the parametrization (3.6.4) by deformation fields $\boldsymbol{\theta}$, that is used to define the notion of derivative. Indeed, the definition of shape derivative of chapter 1, definition 1.1 is equivalent to saying that a functional $J(\Omega)$ is shape differentiable if and only if $J \circ \rho_\Omega$ is differentiable.

The key ingredient for applying the null space algorithm algorithm 3.1 to this context is the computation of the gradient $\nabla J(\Omega)$ and transposes $\mathrm{D}\boldsymbol{g}(\Omega)^\mathcal{T}$ and $\mathrm{D}\boldsymbol{h}(\Omega)^\mathcal{T}$ of the shape derivatives of the constraints. Following definition 3.1, these are computed by introducing a Hilbert space $V \subset W^{1,\infty}(D, \mathbb{R}^d)$ with scalar product $\langle \cdot, \cdot \rangle_V$ and by solving identification problems (3.2.3) and (3.2.4). Our practical implementation follows the strategy described in chapter 1, section 1.4.1: we set either

$$V = H^1(D, \mathbb{R}^d) \text{ with } \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in V,\ \langle \boldsymbol{\theta}, \boldsymbol{\theta}' \rangle_V := \int_D (\gamma^2 \nabla \boldsymbol{\theta} : \nabla \boldsymbol{\theta}' + \boldsymbol{\theta} \cdot \boldsymbol{\theta}') \mathrm{d}x \tag{3.6.5}$$

or

$$V = \{ v \nabla d_\Omega \,|\, v \in H^1(D) \} \text{ with } \forall v, w \in V,\ \langle v \nabla d_\Omega, \boldsymbol{w} \nabla d_\Omega \rangle_V := \int_D (\gamma^2 \nabla v \cdot \nabla w + vw) \mathrm{d}x. \tag{3.6.6}$$

The implicit dependence of V with respect to Ω is omitted. As mentioned in chapter 1, the parameter γ is set proportional to the minimum mesh element size `hmin`.

In light of the previous discussion, the proposed dynamical system (3.3.1) for tackling shape optimization problems of the form (3.6.2) is extended and discretized as follows.

1. The null space and range space directions $\boldsymbol{\xi}_J(\Omega)$ and $\boldsymbol{\xi}_C(\Omega)$ are computed as elements of V thanks to the formulas (3.3.8) and (3.3.25). This requires the computation of the gradient $\nabla J(\Omega)$ and of the transposes $\mathrm{D}\boldsymbol{g}^\mathcal{T}(\Omega)$, $\mathrm{D}\boldsymbol{h}^\mathcal{T}(\Omega)$ via the resolution of identification problems such as (3.2.3) and (3.2.4). In particular, steps 1 to 5 of algorithm 3.1 including the resolution of the dual problem (3.3.10) are achieved from the knowledge of the Fréchet derivatives and of their transposes.

2. The update (3.3.2) of the design from one iteration of the process to the next is performed by using the retraction map ρ_Ω as in (3.6.3):

$$\Omega_{n+1} := \rho_{\Omega_n}(-\Delta t(\alpha_J \boldsymbol{\xi}_J(\Omega_n) + \alpha_C \boldsymbol{\xi}_C(\Omega_n))); \tag{3.6.7}$$

in practice, the step 6 of algorithm 3.1 is adapted accordingly.

In our applications, let us recall that (3.6.7) is carried out by relying on the mesh evolution technique reviewed in chapter 1, section 1.4.2: at every iteration n, the current shape Ω_n is explicitly discretized as a submesh of a triangulated mesh $\mathcal{T}_n$ of D (see e.g. Figure 3.15 below). This also means that one does not use the retraction map ρ_Ω of (3.6.4) naturally considered by the method of Hadamard but rather the one $\widetilde{\rho}_\Omega$ associated with the use of the level set method described in eqn. (1.4.12) of chapter 1.

3.6.2 Adaptive normalizations for the null space and range space directions

We rely on algorithm 3.1 for our implementation of the null space flow (3.3.1) for numerical shape optimization. A few comments are in order regarding the appropriate scaling of the null and range space steps in relation with the size of the mesh discretization; we define accordingly variable coefficients $\alpha_{J,n}$ and $\alpha_{C,n}$ for the descent direction $\boldsymbol{\theta}_n$ in (3.6.8).

For stability reasons, the vertices of the current mesh $\mathcal{T}_n$ discretizing Ω_n should move by a distance which equals at most a few mesh elements in order to produce the subsequent shape Ω_{n+1}. Hence, the minimum edge length `hmin` of the computational mesh is a natural candidate for the limiting step size value `h` of the discussion in section 3.4.1. In our practical implementation, we set $\Delta t = 1$ and a descent direction $\boldsymbol{\theta}_n(x)$ is computed by estimating

$$\boldsymbol{\theta}_n := -(\alpha_{J,n}\boldsymbol{\xi}_J(\Omega_n) + \alpha_{C,n}\boldsymbol{\xi}_C(\Omega_n)), \tag{3.6.8}$$

where α_J and α_C of the update (3.6.7) have been replaced by dynamic coefficients $\alpha_{J,n}$ and $\alpha_{C,n}$.

The parameters $\alpha_{J,n}$ and $\alpha_{C,n}$ scaling the null space and range space steps $\boldsymbol{\xi}_J(\Omega_n)$ and $\boldsymbol{\xi}_C(\Omega_n)$ are updated dynamically in order to control the step size $||\boldsymbol{\theta}_n||_{L^\infty(D,\mathbb{R}^d)}$. Note that the infinity norm is considered because *all* values of the displacement $\boldsymbol{\theta}_n$ should be of the order of the mesh size. We consider A_J and A_C two user-defined parameters, which are expressed in terms of the minimum edge length `hmin` for a clearer intuitive meaning. The coefficients $\alpha_{J,n}$ and $\alpha_{C,n}$ are updated at every iteration according to the following rules:

$$\alpha_{J,n} := \begin{cases} \dfrac{A_J\texttt{hmin}}{||\boldsymbol{\xi}_J(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)}} & \text{if } n < \texttt{n}_0 \\ \\ \dfrac{A_J\texttt{hmin}}{\max(||\boldsymbol{\xi}_J(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)}, ||\boldsymbol{\xi}_J(\Omega_{\texttt{n}_0})||_{L^\infty(D,\mathbb{R}^d)})} & \text{if } n \geq \texttt{n}_0 \end{cases} \tag{3.6.9}$$

$$\alpha_{C,n} := \min\left(0.9, \frac{A_C\texttt{hmin}}{\max\left(\texttt{1e-9}, ||\boldsymbol{\xi}_C(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)}\right)}\right). \tag{3.6.10}$$

These normalizations ensure that the null space and range space steps always remain smaller than A_J and A_C times the mesh size:

$$\forall n \geq 0,\ ||\alpha_{J,n}\boldsymbol{\xi}_J(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)} \leq A_J\texttt{hmin} \text{ and } ||\alpha_{C,n}\boldsymbol{\xi}_C(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)} \leq \min(0.9, A_C\texttt{hmin}).$$

Actually, the null space component $\alpha_{J,n}\boldsymbol{\xi}_J(\Omega_n)$ of $\boldsymbol{\theta}_n$ is scaled to be exactly of the size $A_J\texttt{hmin}$ for the first $\texttt{n}_0$ iterations:

$$\forall 1 \leq n \leq \texttt{n}_0,\ ||\alpha_{J,n}\boldsymbol{\xi}_J(\Omega_n)||_{L^\infty(D,\mathbb{R}^d)} = A_J\texttt{hmin}.$$

Then, $\boldsymbol{\xi}_J(\Omega_n)$ is allowed to converge to 0 as $n \to \infty$.

The range step $\alpha_{C,n}\boldsymbol{\xi}_C(\Omega_n)$ is also set to remain smaller than the constant 0.9, in view of the stability condition $0 < \alpha_C\Delta t < 2$ (see remark 3.12). The role of the constant `1e-9` is only to avoid division by 0 when no constraint is active.

Remark 3.13. Since we measure step sizes with the infinity norm $||\boldsymbol{\theta}_n||_{L^\infty(D)}$ rather than with the Hilbertian norm $||\boldsymbol{\theta}_n||_V = ||\boldsymbol{\theta}_n||_{H^1(D,\mathbb{R}^d)}$, the tolerance bounds (3.4.3) need to be updated with respect to this norm as follows:

$$\varepsilon_i := \texttt{hmin}\int_{\partial\Omega} |v_{C_i}(\Omega_n)|\mathrm{d}s,$$

where it is assumed that the shape derivative of each constraint functional $C_i(\Omega_n)$ can be written as a boundary integral featuring the scalar field $v_{C_i}(\Omega_n)$:

$$\mathrm{D}C_i(\Omega_n)(\boldsymbol{\theta}) := \int_{\partial\Omega} v_{C_i}(\Omega_n)\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s.$$

3.6.3 Illustrations on a multiple load structural shape optimization test case

In this final section, we illustrate the efficiency of our optimization strategy on a practical structural design problem. Two possible configurations are investigated for the shape optimization of a bridge structure subjected to multiple loads, featuring either multiple objective criteria or multiple constraint functions.

Shape optimization setting

We consider the shape optimization of a bridge-like structure Ω contained in a two-dimensional rectangular hold-all domain $D \subset \mathbb{R}^2$ with size 10×2. The boundary of $\partial\Omega$ is divided into disjoint regions as:

$$\partial\Omega = \Gamma \cup \Gamma_D \cup \bigcup_{i=0}^{8} \Gamma_i,$$

where

- Γ_D is a non-optimizable part of the boundary on which the structure Ω is clamped, made of two segments with unit length at the lower extremities of D.
- For $i = 0, \dots, 8$, Γ_i is a non-optimizable subset of the upper side of D given by

$$\Gamma_i := \left[i\frac{10}{9}, (i+1)\frac{10}{9} \right] \times \{2\}, \quad \forall 0 \leq i \leq 8;$$

  Γ_i is subjected to a unit, vertical downward traction load $\boldsymbol{g}_i = (0, -1)$.
- The remaining region Γ is traction-free and it is the only region of $\partial\Omega$ which is subject to optimization.

Non-optimizable material layers of width 0.1 are additionally imposed on the upper part of the domain D and above each component of Γ_D; see Figure 3.12. We consider nine different load cases, that are

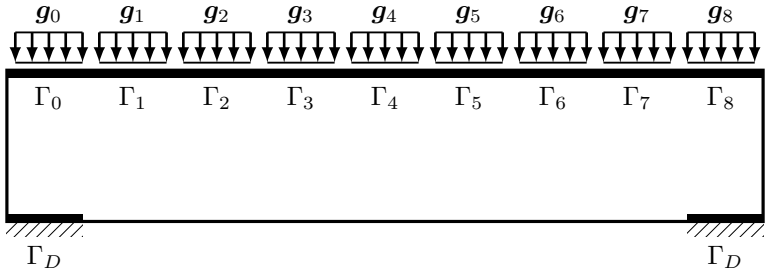


Figure 3.12: Geometric setting for the multiple load case test case

obtained by applying successively and exclusively each of the loads $\boldsymbol{g}_i$ on the region Γ_i. In each situation, the corresponding elastic displacement $\boldsymbol{u}_i$ is the unique solution in $H^1(\Omega, \mathbb{R}^d)$ to the linearized elasticity system:

$$\begin{cases} -\mathrm{div}(Ae(\boldsymbol{u}_i)) = 0 & \text{in } \Omega \\ Ae(\boldsymbol{u}_i)\boldsymbol{n} = 0 & \text{on } \Gamma \\ Ae(\boldsymbol{u}_i)\boldsymbol{n} = \boldsymbol{g}_i & \text{on } \Gamma_i \\ Ae(\boldsymbol{u}_i)\boldsymbol{n} = 0 & \text{on } \Gamma_j \text{ for } j \neq i \\ \boldsymbol{u}_i = 0 & \text{on } \Gamma_D, \end{cases} \tag{3.6.11}$$

where we recall that $e(\boldsymbol{u}) = (\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T)/2$ is the strain tensor associated to the displacement $\boldsymbol{u}$ and $Ae(\boldsymbol{u}) = 2\mu e(\boldsymbol{u}) + \lambda \mathrm{Tr}(e(\boldsymbol{u}))I$ is the corresponding stress tensor, involving the Hooke's law A. The Young modulus and the Poisson ratio are set to $E = 15$ and $\nu = 0.35$, which corresponds to $\lambda = 12.96$ and $\mu = 5.56$.

Starting from the initial structure Ω_0 depicted in Figure 3.13, we are interested in the simultaneous minimization of the volume $\mathrm{Vol}(\Omega)$ of Ω and in the maximization of all the compliances $C_i(\Omega)$ associated

with each load case $\boldsymbol{g}_i$. These quantities are defined by:

$$\mathrm{Vol}(\Omega) := \int_\Omega \mathrm{d}x, \quad C_i(\Omega) := \int_\Omega Ae(\boldsymbol{u}_i) : e(\boldsymbol{u}_i)\mathrm{d}x, \tag{3.6.12}$$

and their shape derivatives read (see chapter 2):

$$\mathrm{DVol}(\Omega)(\boldsymbol{\theta}) = \int_\Gamma \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s, \quad \mathrm{D}C_i(\Omega)(\boldsymbol{\theta}) = -\int_\Gamma Ae(\boldsymbol{u}_i) : e(\boldsymbol{u}_i)\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s. \tag{3.6.13}$$

Volume minimization with maximum compliance constraint

A first possible way to address this case featuring multiple concurrent objectives is to minimize the volume $\mathrm{Vol}(\Omega)$ subject to an upper bound constraint on the individual compliances $C_i(\Omega)$:

$$\begin{aligned} \min_{\Omega\in\mathcal{X}} \quad & \mathrm{Vol}(\Omega) \\ \text{s.t.} \quad & C_i(\Omega) \le C \quad \text{for all } i \in I \end{aligned} \tag{3.6.14}$$

where $I \subset \{0,1,\dots,8\}$ is a set of indices for the considered load cases. The value of the threshold C in (3.6.14) is set to a fraction of the maximum of the compliances $C_i(\Omega_0)$ of the initial design Ω_0:

$$C = 0.7 \max_{i=0,\dots,8} \int_{\Omega_0} Ae(\boldsymbol{u}_i) : e(\boldsymbol{u}_i)\mathrm{d}x \tag{3.6.15}$$

We solve (3.6.14) in the following three configurations:

1. *Case 1: single load case:* $I = \{4\}$ (only the central load $\boldsymbol{g}_4$ is applied)
2. *Case 2: three load cases:* $I = \{0,4,8\}$ (only the central load $\boldsymbol{g}_4$ and the two extreme loads $\boldsymbol{g}_0$ and $\boldsymbol{g}_8$ are applied).
3. *Case 3: all load cases:* $I = \{0,1,\dots,8\}$ (all nine loads are considered).

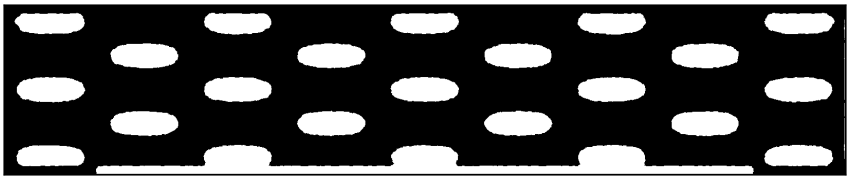
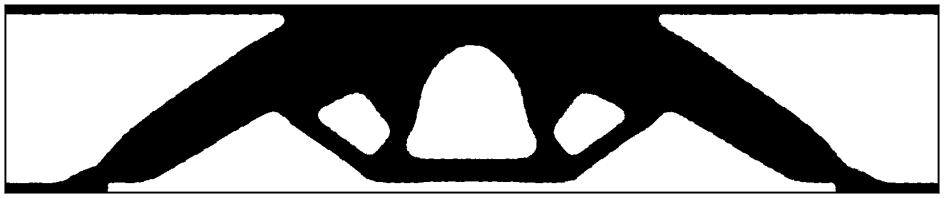


Figure 3.13: Initialisation Ω_0 (solid in *black*) for the shape optimization examples of section 3.6. The thin white layer at the bottom is a non optimizable part of the domain.

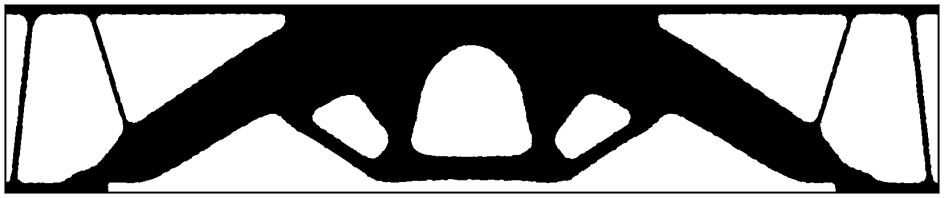
Let us emphasize that for this example (and the next ones), no fine tuning of the algorithm parameters A_J and A_C (determining the update of the values of $\alpha_{J,n}$ and $\alpha_{C,n}$ in (3.6.8)) of section 3.6.2 is required. The only intuition guiding our choice for this particular test case is that the value of A_J should be set lower than A_C. Indeed, a too high value of A_J might entail a too quick decrease of the volume, which would incur dramatic topological changes violating the rigidity constraints. Therefore these parameters were set to $A_J = 1$ and $A_C = 2$ for this test case. The minimum mesh size is $\texttt{hmin} = 0.03$.

The optimized shapes obtained in the three aforementioned situations are shown on Figure 3.14. The meshes of the initial and final designs, as well as several intermediate shapes corresponding to the nine load test-case are shown on Figs. 3.15 and 3.16. The convergence histories in the three situations are reported on Figs. 3.17 to 3.19. They allow to verify the decrease of the objective function even after the saturation of the constraints. Note that for this example and the one to follow, we observed that $\widehat{I}(\Omega_n)$ coincides with $\widetilde{I}(\Omega_n)$ at every iteration, however this situation is very specific to this test case and does not reproduce in generality[1]. As expected, the optimal value found for the volume of the solid distribution increases with the number of imposed constraints. The major structural change between the different situations is the addition of extra vertical bars of material near the extremities of the structure.

[1] for instance the thermoelasticity test cases of chapter 2, section 2.5.5 for which the volume constraint does not saturate featured $\widetilde{I}(\Omega_n) \neq \widehat{I}(\Omega_n)$



(a) One load (only $\boldsymbol{g}_4$ is considered).



(b) Three loads (only $\boldsymbol{g}_0, \boldsymbol{g}_4, \boldsymbol{g}_8$ are considered).

(c) All nine loads.

Figure 3.14: Optimized shapes for three possible configurations of the volume minimization problem (3.6.14) subject to maximum compliance constraint.

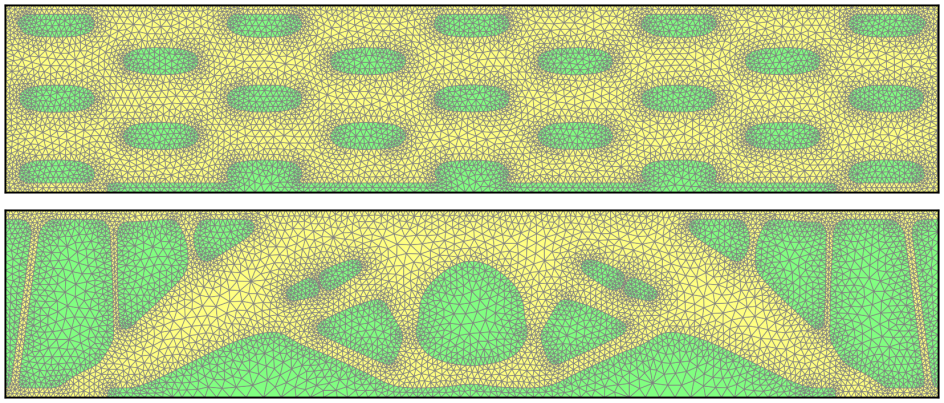


Figure 3.15: Meshes of the initialization and final shapes for the nine load case of Figure 3.14c ((3.6.14)).

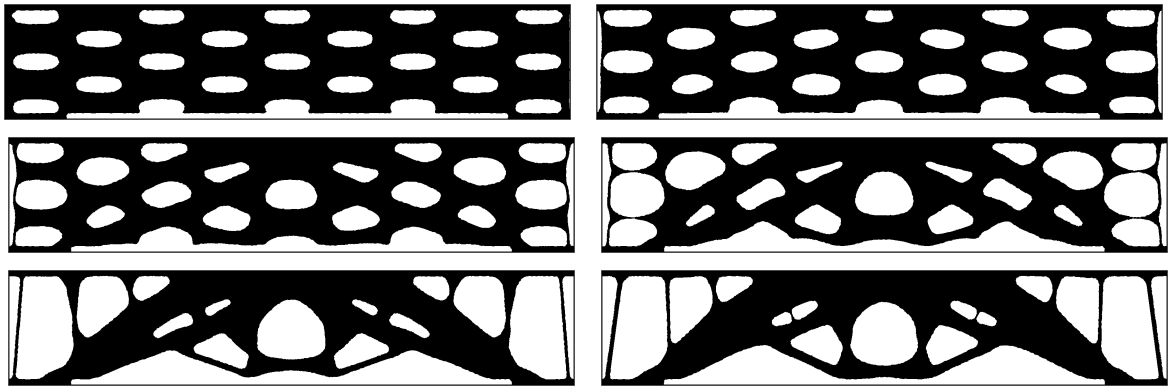
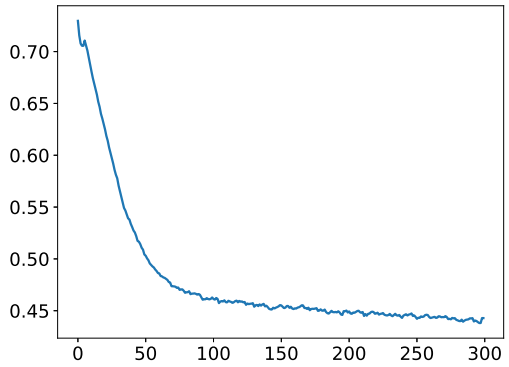
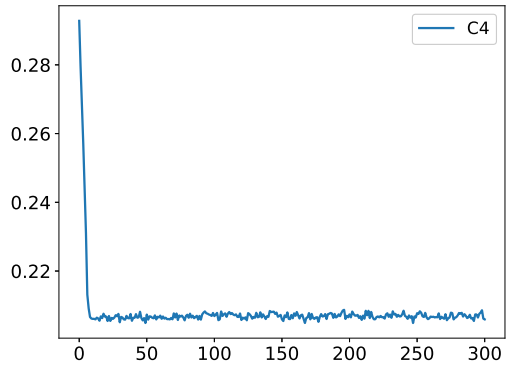


Figure 3.16: Intermediate minimizing shapes for the nine load case of the volume minimization problem of (3.6.14) (iterations 0, 5, 10, 20, 80, and 300).

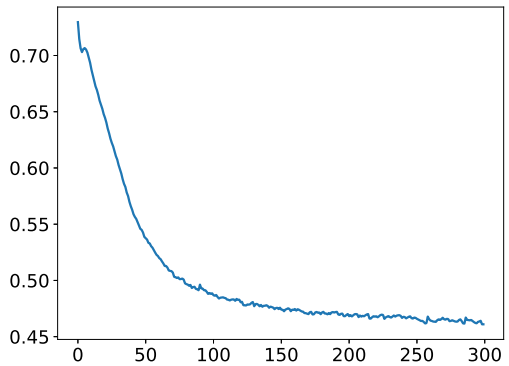


(a) Objective function $J(\Omega) = \text{Vol}(\Omega)$. Final value: $\text{Vol}(\Omega) = 0.44\text{Vol}(D)$.

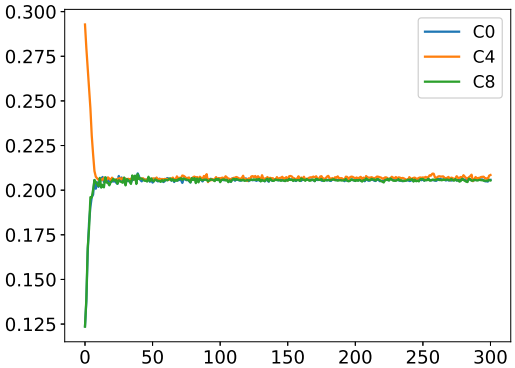


(b) Constraints (compliance values C_i). Final values: $C_4 = 1.29$.

Figure 3.17: Convergence history curves for the single load case of (3.6.14).

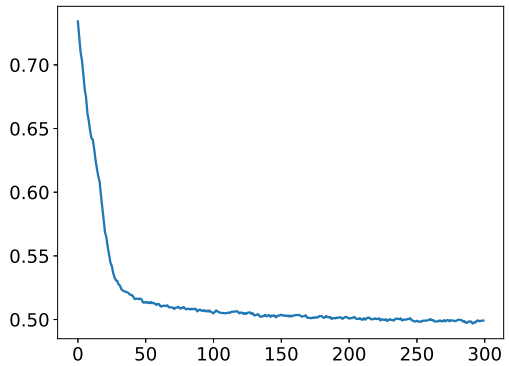


(a) Objective function $J(\Omega) = \text{Vol}(\Omega)$. Final value: $\text{Vol}(\Omega) = 0.46\text{Vol}(D)$.

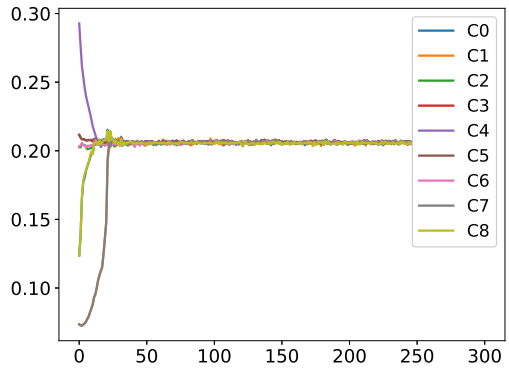


(b) Constraints (compliance values C_i). Final values: $C_0 = 1.29$, $C_4 = 1.30$, $C_8 = 1.28$.

Figure 3.18: Convergence history curves for the three load case of (3.6.14).



(a) Objective function $J(\Omega) = \text{Vol}(\Omega)$. Final value: $\text{Vol}(\Omega) = 0.50\text{Vol}(D)$.



(b) Constraints (compliance values C_i). Final values: $C_0 = 1.29$, $C_1 = 1.28$, $C_2 = 1.28$, $C_3 = 1.29$, $C_4 = 1.29$, $C_5 = 1.29$, $C_6 = 1.29$, $C_7 = 1.30$, $C_8 = 1.29$.

Figure 3.19: Convergence history curves for the nine load case of (3.6.14).

Min/Max compliance optimization with a volume constraint

We now consider a different approach where the maximum value of the compliances $C_i(\Omega)$ is minimized with an equality volume constraint:

$$\begin{aligned} \min_{\Omega\in\mathcal{X}} \quad & \max_{i\in I} \quad C_i(\Omega) \\ \text{s.t.} \quad & \mathrm{Vol}(\Omega) = \rho_0 \mathrm{Vol}(D) \end{aligned} \tag{3.6.16}$$

for a target volume fraction $\rho_0 = 0.5$ of elastic material and for the three load sets I introduced in the previous subsection. This problem may be given the form (3.1.2) after introducing a slack variable m:

$$\begin{aligned} \min_{(\Omega,m)\in\mathcal{X}\times\mathbb{R}} \quad & m \\ \text{s.t.} \quad & \begin{cases} \mathrm{Vol}(\Omega) = \rho_0 \mathrm{Vol}(D) \\ C_i(\Omega) \leq m \quad \text{for all } i \in I. \end{cases} \end{aligned} \tag{3.6.17}$$

The optimization must now be performed with respect to both the slack variable m and the domain geometry Ω, which demands minor adaptations of our optimization algorithm (similar e.g. to those in section 3.5.1): the optimization set $\mathcal{X}\times\mathbb{R}$ is equipped with the tensorized tangent space $\widetilde{V} = V \times \mathbb{R}$ and differentials are identified to gradients thanks to the inner product $\langle\cdot,\cdot\rangle_{\widetilde{V}}$ defined by

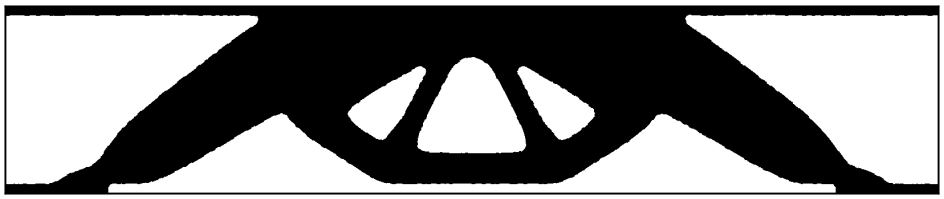
$$\forall (v,w) \in H^1(D,\mathbb{R}) \times H^1(D,\mathbb{R}),\ (l,m) \in \mathbb{R}\times\mathbb{R}, \quad \langle (v,l),(w,m)\rangle_{\widetilde{V}} := \langle v,w\rangle_V + lm, \tag{3.6.18}$$

where $\langle\cdot,\cdot\rangle_V$ is the scalar product of (3.6.5) or (3.6.6). The slack variable m is initialized with the maximum value of the compliance of the initial structure Ω_0 over all the considered loads:

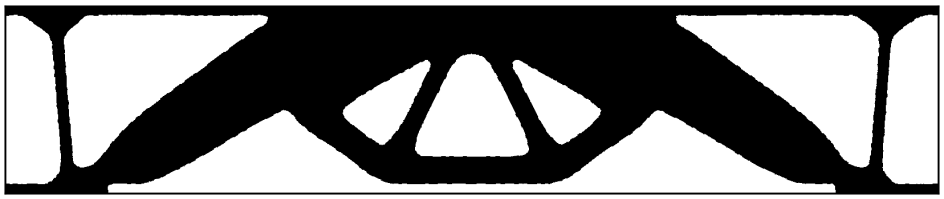
$$m_0 := \max_{i\in I} \quad C_i(\Omega_0), \tag{3.6.19}$$

and its values m_n are then updated along with the shape Ω_n according to algorithm 3.1.

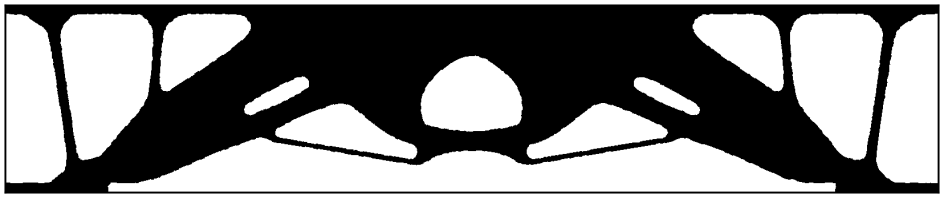
The resulting optimized structures are shown on Figure 3.20 for each of the three considered configurations and the associated convergence histories are displayed on Figs. 3.21 to 3.23 for the single, triple and nine load cases respectively. Note that sudden, abrupt peaks on the constraint curves correspond to topological changes (e.g. at iteration 38 for the nine load case) for which the displacements corresponding to the extremal loads $\boldsymbol{g}_0$ and $\boldsymbol{g}_8$ are especially sensitive. The decrease of all compliance functions is observed even after all the inequality constraints are saturated, which occurs as soon as where all compliances achieve a common value. As expected, the optimal design found for the nine load minimum compliance case (Figure 3.14c) is similar (up to a few bars) to the corresponding one found for the volume minimization (Figure 3.20c): indeed, both cases reach at convergence a volume fraction $\mathrm{Vol}(\Omega) = 0.5\mathrm{Vol}(D)$ and a maximum compliance $\max C_i(\Omega) \simeq 1.30$.



(a) One load (only $\boldsymbol{g}_4$ is considered).

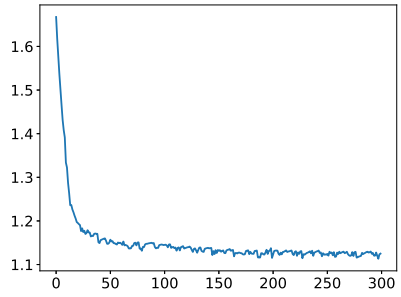


(b) Three loads (only $\boldsymbol{g}_0, \boldsymbol{g}_4, \boldsymbol{g}_8$ are considered).

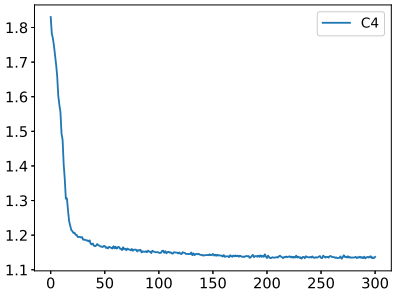


(c) All nine loads.

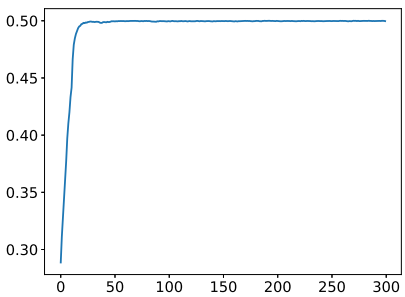
Figure 3.20: Optimized shapes for three possible configurations of the min/max optimization problem (3.6.17).



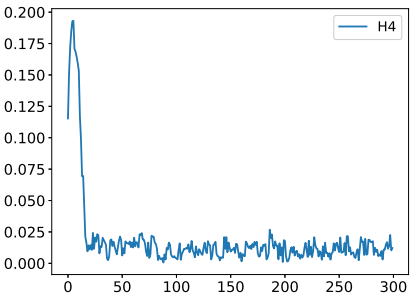
(a) Objective function m. Final value $m = 1.13$.



(b) Compliance $C_i(\Omega)$. Final value: $C_4 = 1.14$.

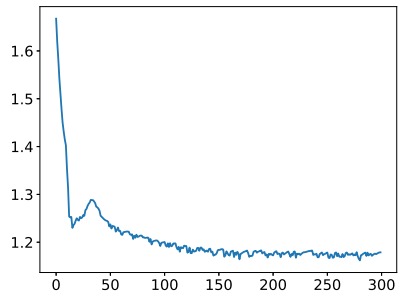


(c) Volume fraction Vol(Ω)/Vol(D). Final value: Vol(Ω)/Vol(D)=0.50.

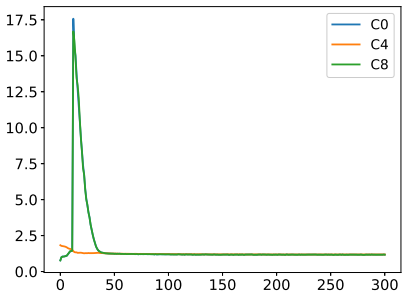


(d) Inequality constraint $H_i = C_i - m$

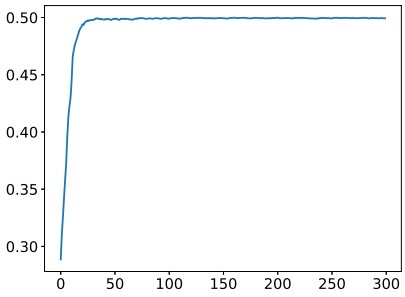
Figure 3.21: Convergence history curves for one load case of (3.6.17).



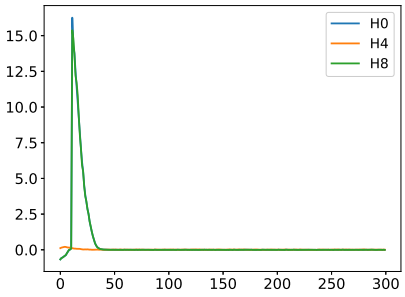
(a) Objective function m. Final value $m = 1.18$.



(b) Compliance $C_i(\Omega)$. Final values: $C_0 = 1.17$, $C_4 = 1.19$, $C_8 = 1.17$.

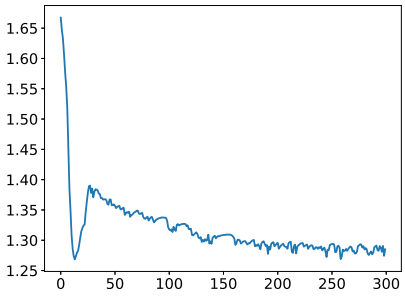


(c) Volume fraction Vol(Ω)/Vol(D). Final value: Vol(Ω)/Vol(D)=0.50.

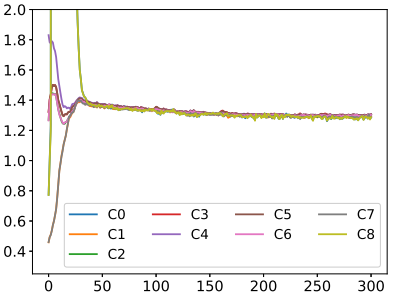


(d) Inequality constraints values: $H_i = C_i - m$.

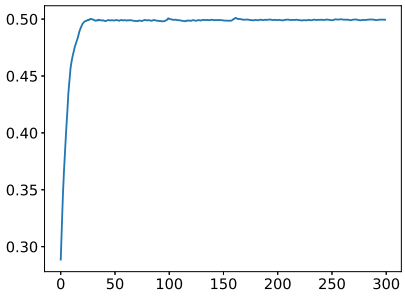
Figure 3.22: Convergence history curves for three load case of (3.6.17).



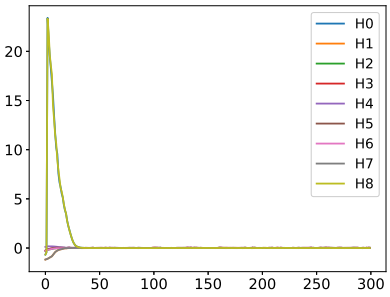
(a) Objective function m (eqn. (3.6.17))



(b) Compliance values $C_i(\Omega)$. Final values: $C_0 = 1.29$, $C_1 = 1.29$, $C_2 = 1.30$, $C_3 = 1.31$, $C_4 = 1.31$, $C_5 = 1.31$, $C_6 = 1.30$, $C_7 = 1.29$, $C_8 = 1.29$.



(c) Volume fraction Vol(Ω)/Vol(D). Final value: Vol(Ω)/Vol(D)=0.50.



(d) Inequality constraints $H_i = C_i - m$

Figure 3.23: Convergence history curves for nine load case of (3.6.17).

CHAPTER 4

A VARIATIONAL METHOD FOR COMPUTING SHAPE DERIVATIVES OF GEOMETRIC CONSTRAINTS ALONG RAYS

Contents

Note : most of the content of this chapter has been published in [154]. F. Feppon, G. Allaire, and C. Dapogny, *A variational formulation for computing shape derivatives of geometric constraints along rays*, To appear in M2AN, (2019).

The introduction has been partly modified in regards with this new section and the original motivations of the thesis. Some background material redundant with the previous chapters has also been removed.

4.1 Introduction

One of the key challenges regarding the systematic integration of topology optimization methods in industrial applications is the need for taking into account geometric constraints. These come very often from the final manufacturing process which generally brings into play constraints on the geometry of shapes; for instance minimum thickness constraints, maximum thickness constraints, minimum distance between members, casting constraints; see [320, 88, 178, 234] or the recent survey [220].

For our multiphysics shape optimization purposes, the starting motivation for the present chapter originates from the need for tackling non penetration constraints in liquid liquid heat exchangers design. In the next chapter 6 (section 5.1), we consider the heat exchanger design problem of Figure 4.1 featuring two fluid phases $\Omega_{f,hot} \subset D$ and $\Omega_{f,cold} \subset D$. A very natural specification of the problem is that these two phases should not interpenetrate during the optimization process of the union $\Omega = \Omega_{f,cold} \cup \Omega_{f,cold}$ of the two phases. Very few works have investigated the enforcement of this non-mixing condition; we are actually only aware of the thesis of Papazoglou [255] in the context of density based topology optimization.

Among the variety of methods proposed in the literature to enforce geometric constraints, several works [29, 30, 109, 110, 234] have proposed to formulate them by means of integral functionals involving the signed distance function d_Ω (see chapter 1, section 1.3) to the optimized domain Ω. More precisely, the shape Ω is sought among all possible subsets of a fixed 'hold-all' domain $D \subset \mathbb{R}^d$ (with still $d = 2, 3$ in applications) as the solution of a constrained minimization problem of the form:

$$\min_{\Omega \subset D} J(\Omega), \text{ s.t. } P(\Omega) \leq 0, \text{ where } P(\Omega) := \int_D j(d_\Omega(x))\mathrm{d}x, \tag{4.1.1}$$

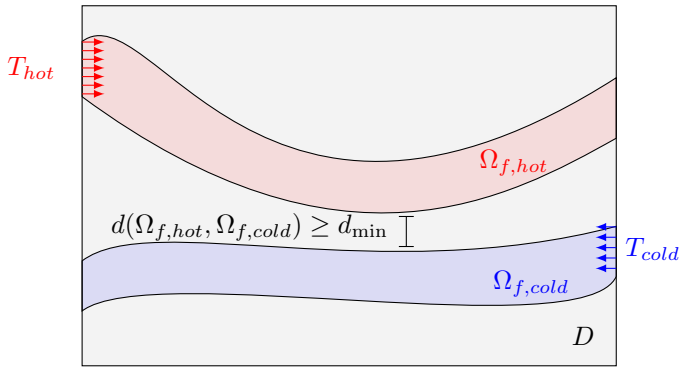


Figure 4.1: Setting for a 2-d fluid fluid heat exchanger design problem (treated in chapter 6, section 5.1). A hot fluid phase $\Omega_{f,1} \subset D$ is entering at the top left with a temperature T_{hot}, and a hot fluid phase $\Omega_{f,2}$ is entering in the reverse direction at the bottom right inlet. Both phases should remain at a distance $d(\Omega_{f,1}, \Omega_{f,2}) > d_{\min}$ from one another.

where $J(\Omega)$ is the objective function, $j : \mathbb{R} \to \mathbb{R}$ is a given, smooth function, and $P(\Omega)$ is a geometric constraint. The above framework is quite appealing insofar as the signed distance function d_Ω naturally lends itself to clear mathematical formulations of maximum, minimum thickness constraint functionals (and the other aforementioned geometric criteria). For instance, the non penetration constraint of the 2-d heat exchanger problem of Figure 4.1 can be conveniently formulated with the signed distance function of either of the phases $\Omega_{f,hot}$ or $\Omega_{f,cold}$, because it is sufficient to require the two phases to remain at a distance $d_{\min} > 0$ from one another:

$$\forall x \in \Omega_{f,hot},\ d_{\Omega_{f,cold}}(x) \geq d_{\min}.$$

This pointwise constraint can then be captured by an integral criterion $P(\Omega_{f,hot})$ as in (4.1.1), see section 5.1.

Allaire, Jouve and Michailidis [30] have shown that the problem (4.1.1) is amenable to numerical treatments by means of standard (e.g. steepest-descent) optimization algorithms. However, several technical stages in its numerical implementation pose difficulties. The leading motivation of the present chapter is to introduce a new variational method that make geometric constraints substantially simpler to implement in the framework of the method of Hadamard.

Let us provide a little more details about the main numerical issues raised by the implementation of (4.1.1), while staying at an explanatory level; see sections 4.3.1 and 4.4 below for full details. The use of traditional optimization methods for the program (4.1.1) relies (in particular) on the knowledge of the *shape derivative* of the considered constraint functional $\Omega \mapsto P(\Omega)$ (chapter 1, definition 1.1). It was shown in [23] that the shape derivative of $P(\Omega)$ as in (4.1.1) has the form

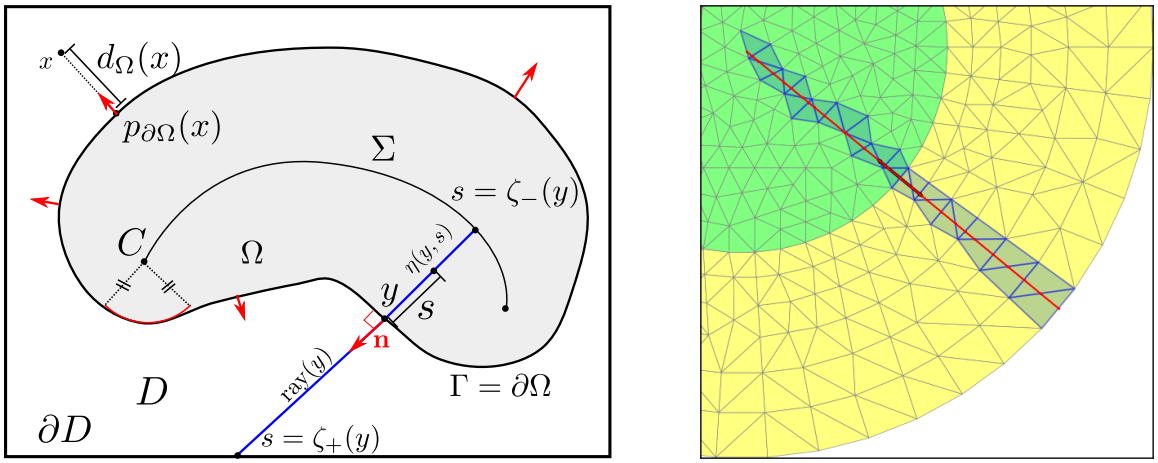
$$\mathrm{D}P(\Omega)(\boldsymbol{\theta}) = \int_{\partial\Omega} u\, \boldsymbol{\theta} \cdot \boldsymbol{n}\, \mathrm{d}y, \tag{4.1.2}$$

where $\mathrm{d}y$ refers to the surface measure on $\partial\Omega$, and the function $u \in L^1(\partial\Omega)$ is given by the formula

$$u(y) = -\int_{x \in \mathrm{ray}(y)} j'(d_\Omega(x)) \prod_{1 \leq i \leq n-1} (1 + \kappa_i(y) d_\Omega(x)) \mathrm{d}x, \qquad \forall y \in \partial\Omega. \tag{4.1.3}$$

In (4.1.2) and (4.1.3), $\boldsymbol{n}$ and $\kappa_i(y)$ stand for the unit normal vector to $\partial\Omega$ pointing outward Ω and the $(n-1)$ principal curvatures of $\partial\Omega$ respectively; the set $\mathrm{ray}(y)$ is the normal ray that originates from $y \in \partial\Omega$ and that stops at either the boundary ∂D of the hold-all domain, or at the skeleton (or medial axis) Σ of Ω; see Figure 4.2a for an illustration. The normal velocity $\boldsymbol{\theta} = -u\boldsymbol{n}$ is then exploited by optimization algorithms (such as our null space method of chapter 3, algorithm 3.1) as a constraint gradient to build a descent update for the shape Ω. The numerical component of this program raises the need to calculate the quantity $u : \partial\Omega \to \mathbb{R}$ defined in (4.1.3).

In principle, the formula (4.1.3) featuring integration along the normal rays to $\partial\Omega$ can be implemented efficiently. In particular, its numerical evaluations at several points $y \in \partial\Omega$ can be performed in parallel.



(a) The shape optimization setting: shape $\Omega \subset D$, skeleton Σ with normal rays $(\mathrm{ray}(y))_{y\in\partial\Omega}$, local coordinate change η, outward normal vector $\boldsymbol{n}$, center of curvature C, orthogonal projection $p_{\partial\Omega}(x)$ of a point $x \in D$ onto $\partial\Omega$. (see chapter 1, section 1.3 for the definitions).

(b) Integration along rays requires the detection of the skeleton of the shape and the computation of the triangle path at the mesh level, that is the intersection points of each triangle with the ray.

Figure 4.2: Definition of geometric quantities and integration along normal rays.

However, the practical implementation when all the functions involved in the integrand of (4.1.3) are discretized on a computational mesh is not trivial: it requires the computation of (i) the ray trajectories on the discrete mesh (for instance by some variant of the method of characteristics), as exemplified on Figure 4.2b, and (ii) the principal curvatures κ_i of the boundary $\partial\Omega$ the numerical discretization of Ω. As we shall illustrate in our numerical investigations, the accurate calculation of these quantities —even when $\partial\Omega$ is endowed with a discretization as an explicit triangulated surface—is not trivial and is still an active research direction, already in the case of two space dimensions, and especially in the three-dimensional context [233, 270, 143].

The objective of this chapter is to introduce a robust and easy to implement method that allows to compute integral quantities such as (4.1.3) by solving a variational problem for advection-like operators which alleviates the need for resorting to direct integration along rays. This yields an efficient numerical method to calculate shape derivatives of geometric constraints, which is very simple to implement using a standard finite element software, and which does not raise any additional algorithmic difficulty in 3-d than in 2-d.

Our numerical method for the calculation of (4.1.3) is based on the following variational problem, set over a suitable Hilbert space V_ω:

$$\text{Find } u \in V_\omega \text{ such that } \forall v \in V_\omega,\ \int_{\partial\Omega} uv\mathrm{d}s + \int_{D\setminus\overline{\Sigma}} \omega(\nabla d_\Omega \cdot \nabla u)(\nabla d_\Omega \cdot \nabla v)\mathrm{d}x = -\int_{D\setminus\overline{\Sigma}} j'(d_\Omega)v\mathrm{d}x, \tag{4.1.4}$$

where $\omega > 0$ is a rather arbitrary weight function (over which relevant assumptions shall be stated later on), and $\overline{\Sigma}$ is the closure of the skeleton set Σ of Ω (see Figure 4.2a and section 4.3.1 below). Indeed, our key result is to show that (4.1.4) is well-posed, and that the trace of its unique solution u on $\partial\Omega$ is exactly (4.1.3).

A formal insight that allows to see this is the following: assume that for any $v_0 \in \mathcal{C}^0(\partial\Omega)$, the function v satisfying $v = v_0$ on $\partial\Omega$ and taking constant values along the rays normal to $\partial\Omega$ (i.e. $\nabla d_\Omega \cdot \nabla v = 0$ in $D\setminus\overline{\Sigma}$) belongs to the space V_ω. Using v as a test function in the variational formulation (4.1.4) then yields:

$$\begin{aligned}\int_{\partial\Omega} uv_0\mathrm{d}s &= \int_{\partial\Omega} uv\mathrm{d}s + \int_{D\setminus\overline{\Sigma}} \omega(\nabla d_\Omega \cdot \nabla u)(\nabla d_\Omega \cdot \nabla v)\mathrm{d}x \\ &= -\int_{D\setminus\overline{\Sigma}} j'(d_\Omega)v\mathrm{d}x = -\int_{y\in\partial\Omega}\left(\int_{z\in\mathrm{ray}(y)} j'(d_\Omega(z)) \prod_{1\le i\le n-1}(1+\kappa_i(y)d_\Omega(z))\mathrm{d}s\right) v_0(y)\mathrm{d}y,\end{aligned} \tag{4.1.5}$$

where the last equality follows from a change of variables allowing to rewrite integration on $D \setminus \overline{\Sigma}$ as a nested integration on points $y \in \partial\Omega$ and on elements z in $\mathrm{ray}(y)$ (see (4.2.7) and (4.2.9) below). Since

$v_0 \in \mathcal{C}^0(\partial\Omega)$ can be chosen arbitrarily in (4.1.5), the identity (4.1.3) follows. These considerations are made rigorous in section 4.2.5.

The variational formulation (4.1.4) makes it possible to compute integrals (4.1.3) along the normal rays to $\partial\Omega$ without the need to calculate these rays or the curvatures κ_i explicitly on a discretization (i.e. a mesh) of the ambient space. This feature is especially convenient for shape optimization applications relying on the level set method, as described in section 4.4; there, the gradient of the signed distance function ∇d_Ω is easy to calculate on an unstructured mesh of the considered hold-all domain D from a $\mathbb{P}_1$ approximation of d_Ω. The variational formulation (4.1.4) can then be readily implemented in a finite element framework, even if the boundary $\Gamma = \partial\Omega$ is not meshed explicitly. Our method requires only a rough estimate of the location of the skeleton Σ, upon a judicious choice of the weight ω in (4.1.4); see the numerical examples in section 4.3. Let us emphasize that the variational approach (4.2.5) is equally simple to implement in any space dimension while the 3-d implementation of geometric integration along rays would require much more efforts than in 2-d.

Last, the previous arguments work identically when ∇d_Ω is replaced with an arbitrary $\mathcal{C}^1$ vector field β: we obtain that a variational formulation analogous to (4.1.4) (given in (4.2.5)) allows to compute integrals quantities along the characteristics curves of β, which is subject to offer wider applications than shape optimization.

With these perspectives in mind, the present chapter is organized as follows. Section 4.2 carefully discusses the mathematical setting that guarantees the existence and uniqueness of a solution to the variational problem (4.1.4) (in fact its generalization to arbitrary vector fields β), and the justification of the key identity (4.1.3) satisfied by the trace of its solution. For this purpose, we provide a variational theory for the advection operator $\beta \cdot \nabla$ associated to arbitrary $\mathcal{C}^1$ vector fields β on the weighted graph space V_ω; in particular, the existence of traces on Γ and an adapted Poincaré inequality for functions $v \in V_\omega$ are obtained. Note that many related works are available about these matters (e.g. [131, 144]), however they usually rely on strong boundedness assumptions on the divergence of β (typically $\mathrm{div}(\beta) \in L^\infty(D)$), which do not hold for our shape optimization applications (4.2.8) and (4.2.9). Indeed, in the latter situation where $\beta = \nabla d_\Omega$, the divergence $\mathrm{div}(\nabla d_\Omega)$ typically blows up near Σ; hence the need for our different approach. Our approach requires rather the existence of a flow η (see (4.2.1) below) associated with the vector field β.

In section 4.3, we investigate the numerical accuracy of our variational method for calculating integrals along characteristic curves in the shape optimization setting (4.2.8) where $\beta = \nabla d_\Omega$. After a short review of the properties of the signed distance function, we compare the direct numerical integration along rays with the use of our variational method on several numerical examples where the value of (4.1.3) is analytically known. We also consider "practical" cases where some of the regularity assumptions imposed by our framework are not fulfilled. Most importantly, we illustrate how the selection of an appropriate weight ω in the variational formulation (4.1.4) allows to deal with the presence of cracks in the working domain, for instance the skeleton Σ in shape optimization when it is not explicitly meshed.

The last two sections are dedicated on numerical applications. Section 4.4 illustrates the simplicity and effectiveness of our method on practical shape optimization applications. Section 4.4 elaborates on the works [25, 30, 110] concerned with manufacturing constraints: we demonstrate that our variational method allows for a convenient and efficient implementation of maximum and minimum thickness constraints in structural design. Finally, section 5.1 applies the method in order to enforce a non-mixing constraint for liquid-liquid heat exchanger problems of such as the one depicted in the above Figure 4.1.

4.2 Weighted graph space of the advection operator $\beta \cdot \nabla$ for velocity fields of class $\mathcal{C}^1$

This section is concerned with the mathematical analysis of a slightly more general variational problem than (4.1.4): ∇d_Ω is replaced by a rather arbitrary vector field β. This setting and some technical assumptions are described in section 4.2.1. In order to obtain the well-posedness of the corresponding variational problem and the trace identity (4.1.3), suitable functional spaces V_ω are introduced in section 4.2.2 in which the directional derivative $\beta \cdot \nabla$ naturally makes sense.

Then, section 4.2.3 investigates the density of $\mathcal{C}^1$ functions in the graph space V_ω. In section 4.2.4, we establish a trace theorem for functions in V_ω and we provide a Poincaré-type inequality. These two ingredients allow us to prove in section 4.2.5 the well posedness of the variational problem (4.1.4) and of the identity in this more general context involving arbitrary fields β.

4.2.1 Preliminaries, notation and assumptions

Preliminaries and notation

Let $U \subset \mathbb{R}^d$, $\Gamma \subset \overline{U}$ and $\beta : U \to \mathbb{R}^d$ be respectively a (possibly non smooth) bounded open set, a hypersurface and a vector field of class $\mathcal{C}^1$. Note we do not require Γ to be a compact manifold (Γ may differ from its closure $\overline{\Gamma}$); in this case, we require $\overline{\Gamma}$ to be a manifold with boundary, which in particular prevents Γ from showing spiralling patterns near its ends—an assumption which is needed for technical reasons (see e.g. the proof of lemma 4.3 below). Two examples of admissible settings are represented on Figure 4.3.

We assume that Γ is a Poincaré section or a *stream surface* for β, meaning that for any $y \in \Gamma$, there is a unique characteristic curve $s \mapsto \eta(y,s)$ passing through $y = \eta(y,0)$ at time $s = 0$, and that lives in the domain U on some maximal interval $s \in (\zeta_-(y), \zeta_+(y))$. In other words, for any $y \in \Gamma$, $(\zeta_-(y), \zeta_+(y))$ is the maximum existence interval such that the solution $s \mapsto \eta(y,s)$ of the ordinary differential equation

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}s}\eta(y,s) = \beta(\eta(y,s)), \\ \eta(y,0) = y, \end{cases} \tag{4.2.1}$$

remains in the domain U (note that by definition, $\zeta_- \leq 0 \leq \zeta_+$). We assume that ζ_- and ζ_+ are continuous, bounded functions on Γ, satisfying the following separation condition:

$$\exists \varepsilon > 0,\ \forall y \in \Gamma,\ \zeta_+(y) - \zeta_-(y) > \varepsilon. \tag{4.2.2}$$

The vector field β is required to be U-filling, in the sense that its related flow η realizes a $\mathcal{C}^1$ diffeomorphism from the tensor product set

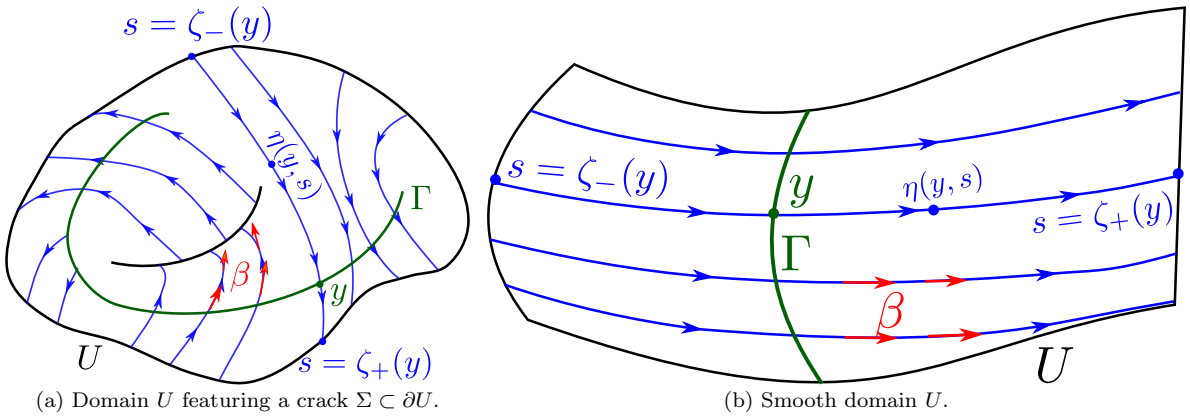
$$W = \{(y,s) \,|\, y \in \Gamma,\ s \in (\zeta_-(y), \zeta_+(y))\}, \tag{4.2.3}$$

onto U (see [212], Chap. IV), where W is a manifold obtained as an open subset of the tensor product of Γ with the real line $\mathbb{R}$ (note that the "open" character of W comes from the continuity assumption on ζ_- and ζ_+).

Finally, we denote by $|\mathrm{D}\eta|$ the Jacobian of the local coordinate change η:

$$\forall y \in \Gamma, \forall s \in (\zeta_-(y), \zeta_+(y)),\ \ |\mathrm{D}\eta|(y,s) = |\det(\nabla\eta)|(y,s), \tag{4.2.4}$$

where the Jacobian matrix of η reads $\nabla\eta = \left[\ \partial_y\eta \ \middle|\ \partial_s\eta\ \right]$ and ∂_y denotes the collection of derivatives with respect to the $(n-1)$ tangential coordinates of Γ.



(a) Domain U featuring a crack $\Sigma \subset \partial U$.

(b) Smooth domain U.

Figure 4.3: Two types of admissible domain U, with U-filling vector field β and stream surface Γ, in the framework of section 4.2.1.

Remark 4.1. Let us consider a particular case where U is a smooth bounded domain in $\mathbb{R}^d$, and $\Gamma \subset \partial U$ is defined as the inlet boundary of the flow field β, *i.e.*

$$\Gamma := \{x \in \partial U \,|\, \beta(x) \cdot \boldsymbol{n}(x) < 0\},$$

where $\boldsymbol{n}$ is the unit normal vector to ∂U, pointing outward U, the separation condition (4.2.2) exactly requires that the inflow and outflow boundaries be separated by a positive minimum distance, which is a rather classical assumption in the study of the advection operator, see e.g. Section 2.1.3 in [125]. In our case, assumption (4.2.2) is required in the proof of proposition 4.2.

The main point of this section is to mathematically justify that given rather arbitrary function $f : U \to \mathbb{R}$ and weight $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$, the trace of the solution u to the variational problem

$$\text{Find } u \in V_\omega \text{ such that } \forall v \in V_\omega,\ \int_\Gamma uv\mathrm{d}s + \int_U \omega(\beta\cdot\nabla u)(\beta\cdot\nabla v)\mathrm{d}x = \int_U fv\mathrm{d}x \tag{4.2.5}$$

is given explicitly in terms of line integrals by the formula

$$\forall y \in \Gamma,\ \ u(y) = \int_{\zeta_-(y)}^{\zeta_+(y)} (f\circ\eta\,|\mathrm{D}\eta|)(y,s)\mathrm{d}s, \tag{4.2.6}$$

which yields a numerical method for computing (4.2.6) by solving (4.2.5).

Throughout the chapter, the considered measure on W is that induced by the standard product measure of the surface measure $\mathrm{d}y$ on Γ and of the Lebesgue measure $\mathrm{d}s$ on $\mathbb{R}$. Thus, the space $L^1(W)$ of integrable functions on W is defined as the space of measurable functions $f : W \to \mathbb{R}$ such that $\int_\Gamma \int_{\zeta_-(y)}^{\zeta_+(y)} |f(y,s)|\mathrm{d}s\mathrm{d}y < +\infty$. Integration of f over W is then defined by:

$$\int_W f(y,s)\mathrm{d}s\mathrm{d}y := \int_\Gamma \int_{\zeta_-(y)}^{\zeta_+(y)} f(y,s)\mathrm{d}s\mathrm{d}y.$$

Under these notations, the classical change of variable formula between manifolds reads (see [212], Chap. XVI):

$$\forall f \in L^1(U),\ \int_U f\mathrm{d}x = \int_W f\circ\eta|\mathrm{D}\eta|\mathrm{d}s\mathrm{d}y, \tag{4.2.7}$$

where the Jacobian $|\mathrm{D}\eta|$ is defined by (4.2.4).

The shape optimization setting as a particular case

The shape optimization setting outlined in the introduction, which is exemplified on Figure 4.2a, reduces to the particular case

$$U = D\backslash\overline{\Sigma},\ \Gamma = \partial\Omega,\ \beta = \nabla d_\Omega. \tag{4.2.8}$$

where $\Omega \subset D$ are bounded Lipschitz domains of $\mathbb{R}^d$ (note that in order for $\beta = \nabla d_\Omega$ to be a $\mathcal{C}^1$ vector field, Ω is assumed to be in fact a $\mathcal{C}^2$ domain in all what follows).

The bound functions $\zeta_-(y)$ and $\zeta_+(y)$ are the distances at which $\mathrm{ray}(y)$ hits either the skeleton Σ of Ω or the boundary ∂D of the hold-all domain D.

In this situation, the local coordinate change η and its Jacobian $|\mathrm{D}\eta|$ are explicitly given by (see section 4.3.1)

$$\eta(y,s) = y + s\,\nabla d_\Omega(y),\ |\mathrm{D}\eta|(y,s) = \prod_{i=1}^{d-1}(1+\kappa_i(y)s), \qquad \forall y \in \partial\Omega,\ \forall s \in (\zeta_-(y),\zeta_+(y)), \tag{4.2.9}$$

so that the function u of (4.1.3) coincides with the expression (4.2.6) for $f = -j'(d_\Omega)$. Let us emphasize that, in this context, the open set U is not smooth because it features a "crack", namely the skeleton Σ (we call it a "cracked domain" in Figure 4.2a); in particular, U is not located on one side of its boundary. This "lack of smoothness" of U prevents from using many convenient results from functional analysis [302], and thus raises the need for several technical ingredients in the sequel, which otherwise would have been fairly classical.

Assumptions on weight functions ω

Let us now consider a weight function $\omega : U \to \mathbb{R}$, which will be one of the key ingredients of our variational formulation (4.2.5), satisfying the following assumptions:

(H1) $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ is a positive, continuous function on U.

(H2) The transported weight $\alpha := \omega \circ \eta |\mathrm{D}\eta| \in \mathcal{C}^0(W, \mathbb{R}_+^*)$ over W is such that the function

$$\begin{array}{rccc} g_\alpha & : & \Gamma & \longrightarrow & \mathbb{R}_+ \cup \{+\infty\} \\ & & y & \longmapsto & \displaystyle\int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y,s)\mathrm{d}s \end{array} \tag{4.2.10}$$

is uniformly bounded, i.e. $g_\alpha \in L^\infty(\Gamma)$.

(H3) The transported weight $\alpha = \omega \circ \eta |\mathrm{D}\eta|$ over W is such that the function

$$\begin{array}{rccc} h_\alpha & : & \Gamma & \longrightarrow & \mathbb{R}_+ \cup \{+\infty\} \\ & & y & \longmapsto & \displaystyle\int_{\zeta_-(y)}^{\zeta_+(y)} \left[\alpha(y,s) \int_0^s \alpha^{-1}(y,t)\mathrm{d}t\right] \mathrm{d}s \end{array} \tag{4.2.11}$$

is uniformly bounded: $h_\alpha \in L^\infty(\Gamma)$.

Note that, above and in the sequel, the notations $\alpha^{-1}(y,s) = 1/\alpha(y,s)$ and $\omega^{-1}(y,s) = 1/\omega(y,s)$ are used for the inverses of the scalar weights α and ω respectively, whereas the notation $\eta^{-1} : U \mapsto W$ shall stand for the reciprocal mapping of the diffeomorphism η.

Hypothesis (H1) is essentially a regularity assumption, positivity being no surprise for ω to be an admissible weight in the variational formulation (4.2.5). Notice that no assumption is made about the behavior of ω near the boundary of U; in particular, $\omega(x)$ may tend to 0 as x approaches ∂U. Hypothesis (H2) is an upper-boundedness assumption for α. The weights that we are going to consider in our practical applications in section 4.3.3 will satisfy (H1) and (H2) almost automatically. Finally, (H3) is roughly a monotonicity constraint for the decay of α towards 0 as $s \to \zeta_\pm(y)$. In practice, we will rely on the following lemma which provides a simple monotonicity condition under which the condition (H3) is fulfilled, indicating that the class of weights satisfying (H3) is large enough.

Lemma 4.1. *Let $\alpha \in \mathcal{C}^0(W, \mathbb{R}_+^*)$ be a weight of the product form:*

$$\forall (y,s) \in W,\ \alpha(y,s) = f(y,s)g(y,s) \tag{4.2.12}$$

where f and g are two positive functions on W satisfying:

(i) There exist two constants $g_-, g_+ \in \mathbb{R}$ such that for all $(y,s) \in W$, $0 < g_- \leq g(y,s) \leq g_+$.

(ii) For any $y \in \Gamma$, the real function $s \mapsto f(y,s)$ is non increasing on $(0, \zeta_+(y))$ and non decreasing on $(\zeta_-(y), 0)$.

Then α satisfies the condition (H3).

Proof. Under the above conditions, it holds, for any $y \in \Gamma$ and for $s \in (0, \zeta_+(y))$:

$$\left|\alpha(y,s) \int_0^s \alpha^{-1}(y,t)\mathrm{d}t\right| \leq \frac{g_+}{g_-} f(y,s) \int_0^s f^{-1}(y,t)\mathrm{d}t \leq \frac{g_+}{g_-}\zeta_+(y) \leq \frac{g_+}{g_-}||\zeta_+ - \zeta_-||_{L^\infty(\Gamma)}.$$

Arguing in the same fashion allows to prove a similar estimate when $s \in (\zeta_-(y), 0)$, which finally implies that $||h_\alpha||_{L^\infty(\Gamma)} \leq \frac{g_+}{g_-}||\zeta_+ - \zeta_-||^2_{L^\infty(\Gamma)}$; this allows to conclude. □

Remark 4.2. The statement of lemma 4.1 does not require f or g to be continuous on W.

Remark 4.3. From the Liouville theorem for ordinary differential equations [95], it holds

$$\forall (y,s) \in W,\ |\mathrm{D}\eta|(y,s) = \exp\left(\int_0^s \mathrm{div}(\beta) \circ \eta(y,t)\mathrm{d}t\right). \tag{4.2.13}$$

Therefore, it is straightforward to verify that (H1) to (H3) are satisfied for the constant weight $\omega = 1$ whenever $\mathrm{div}(\beta) \in L^\infty(U)$—a customary assumption in the study of advection operators; see *e.g.* [51, 195, 125, 257, 67].

Our setting, based on (H1) to (H3), allows to handle more general velocity fields β, with unbounded divergence, which leaves room for the Jacobian $|\mathrm{D}\eta|$ to vanish on the boundary ∂U. This feature is crucial

to deal with the shape optimization setting (4.2.8); in there, the divergence $\mathrm{div}(\beta)$ of $\beta = \nabla d_\Omega$ blows up near *centers of curvatures* $C \in \overline{\Sigma}$ (see Figure 4.2a and section 4.3.1 below) where one of the principal curvatures κ_i is such that $-\kappa_i(p_{\partial\Omega}(x))d_\Omega(x) \to 1$ as $x \to C$. For instance, consider the following very simple situation where $\Omega = D$ is the unit ball in two space dimensions. Its skeleton reduces to a single point, its center. Then,

$$\forall x \in D\backslash\{0\},\ d_\Omega(x) = ||x|| - 1,\ \Delta d_\Omega(x) = \frac{1}{||x||},$$

where $||x||$ is the euclidean norm in $\mathbb{R}^2$. Therefore $\Delta d_\Omega(x)$ blows up at the center $x = 0$ which implies $\mathrm{div}(\nabla d_\Omega) \notin L^\infty(D\backslash\overline{\Sigma})$. However, in this case, (H1) to (H3) hold for $\omega = 1$: from

$$\forall (y,s) \in \partial\Omega \times (-1,0),\ \eta(y,s) = (1+s)y,\ |\mathrm{D}\eta(y,s)| = 1 + s,$$

we obtain that $\alpha(y,s) = 1 + s$ and $\alpha(y,s)\int_0^s \alpha^{-1}(y,t)dt = (1+s)\log(1+s)$ are uniformly bounded on $W = \partial\Omega \times (-1,0)$.

In the following, for a weight ω satisfying (H1), we denote by $L^2_\omega(U)$ and $L^2_\alpha(W)$ the weighted spaces

$$L^2_\omega(U) = \left\{ v \text{ measurable } \mid \int_U \omega v^2 \mathrm{d}x < +\infty \right\}, \quad L^2_\alpha(W) = \left\{ \widetilde{v} \text{ measurable } \mid \int_W \alpha \tilde{v}^2 \mathrm{d}x < +\infty \right\}, \tag{4.2.14}$$

with respective corresponding L^2 norms $||v||_{L^2_\omega(U)} := \left(\int_U \omega v^2 \mathrm{d}x\right)^{1/2}$ and $||\widetilde{v}||_{L^2_\alpha(W)} = \left(\int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y\right)^{1/2}$.

4.2.2 Definition of the graph space V_ω of the advection operator $\beta \cdot \nabla$

We recall that the assumption $\beta \in \mathcal{C}^1(U, \mathbb{R}^d)$ implies that $\mathrm{div}(\beta)$ belongs to $L^\infty_{loc}(U)$.

Definition 4.1 (Derivative along characteristic curves)**.** Let $v \in L^1_{\mathrm{loc}}(U)$ be a locally integrable function on U. The directional derivative $\beta \cdot \nabla v \in \mathcal{D}'(U)$ is the distribution on U defined by

$$\forall \phi \in \mathcal{C}^\infty_c(U),\ \int_U (\beta \cdot \nabla v)\phi \mathrm{d}x = \int_U (-(\beta \cdot \nabla \phi)v - \mathrm{div}(\beta)\phi v)\mathrm{d}x. \tag{4.2.15}$$

Remark 4.4. The definition (4.2.15) of $\beta \cdot \nabla$ mimics the integration by part formula that holds for functions v of class $\mathcal{C}^1$. Considering test functions $\phi \in \mathcal{C}^\infty_c(U)$ allows to avoid imposing classical regularity requirements on the open domain U such as that being locally located on one side of its boundary [302].

Definition 4.2 (Graph space of the operator $\beta \cdot \nabla$)**.** Let $\omega \in \mathcal{C}^0(U, \mathbb{R}^*_+)$ be a positive weight on U. The weighted graph space V_ω of the advection operator $\beta \cdot \nabla$ is defined by:

$$V_\omega = \{v \in L^2_\omega(U) \,|\, \beta \cdot \nabla v \in L^2_\omega(U)\}. \tag{4.2.16}$$

V_ω is a Hilbert space when it is endowed with the norm

$$||v||_{V_\omega} := \left(\int_U \omega v^2 \mathrm{d}x + \int_U \omega |\beta \cdot \nabla v|^2 \mathrm{d}x\right)^{1/2}.$$

Remark 4.5. The completeness of V_ω follows from very standard arguments, see e.g. [306, 125]. It is also easy to see from the assumption $\omega \in \mathcal{C}^0(U, \mathbb{R}^*_+)$ that the inclusion $L^2_\omega(U) \subset L^1_{loc}(U)$ holds.

Let us then introduce corresponding notions of derivative with respect to the s variable and of graph space on the set W:

Definition 4.3 (Graph space of the derivative ∂_s)**.** Let $\alpha \in \mathcal{C}^0(W, \mathbb{R}^*_+)$ be a positive weight on W. The weak derivative $\partial_s \tilde{v} \in \mathcal{D}'(W)$ of a function $\tilde{v} \in L^1_{\mathrm{loc}}(W)$ is the distribution defined by

$$\forall \phi \in \mathcal{C}^\infty_c(W),\ \int_W \partial_s \tilde{v}\phi \mathrm{d}y\mathrm{d}s = -\int_W \tilde{v}\partial_s \phi \mathrm{d}y\mathrm{d}s. \tag{4.2.17}$$

The weighted graph space of the operator ∂_s is defined by:

$$\widetilde{V}_\alpha = \{\tilde{v} \in L^2_\alpha(W) \,|\, \partial_s \tilde{v} \in L^2_\alpha(W)\}. \tag{4.2.18}$$

This space is a Hilbert space when it is equipped with the norm

$$||\tilde{v}||_{\widetilde{V}_\alpha} := \left(\int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y + \int_W \alpha |\partial_s \tilde{v}|^2 \mathrm{d}s\mathrm{d}y\right)^{1/2}.$$

Remark 4.6. The defining identity (4.2.15) of the weak derivative $\beta \cdot \nabla v$ also holds for test functions $\phi \in C_c^1(U)$, as follows from a standard density argument. Likewise, (4.2.17) holds for test functions ϕ which are only continuously differentiable with respect to the s variable.

The motivation behind the introduction of the spaces $\widetilde{V}_\alpha$ is that they allow to transfer the difficulty of studying the "curvilinear" directional derivative $\beta \cdot \nabla$ over the domain U, to that of the more standard, "flat" derivative ∂_s over W. The price to pay is the need to account for the weight of the Jacobian $|\mathrm{D}\eta|$ resulting from the change of variables. This point is made precise by the following technical lemma.

Lemma 4.2. *Let $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ be a positive weight on U, and let $\alpha = \omega \circ \eta |\mathrm{D}\eta|$. The following equivalence holds true:*

$$v \in V_\omega \Leftrightarrow v \circ \eta \in \widetilde{V}_\alpha.$$

Additionally, for any $v \in V_\omega$, the function $\tilde{v} = v \circ \eta$ satisfies:

(i) $\partial_s \tilde{v} = (\beta \cdot \nabla v) \circ \eta$,

(ii) $||v||_{V_\omega} = ||v \circ \eta||_{\widetilde{V}_\alpha}$.

Proof. Let $v \in V_\omega$; it follows from the change of variables (4.2.7) that $\tilde{v} := v \circ \eta$ belongs to $L^2_\alpha(W)$. Moreover, we observe that $\tilde{v}$ has a weak derivative $\partial_s \tilde{v}$ given by:

$$\partial_s \tilde{v} = (\beta \cdot \nabla v) \circ \eta. \tag{4.2.19}$$

Indeed, using the definition (4.2.15) of the weak derivative along β with a test function $\phi \in \mathcal{C}_c^1(U)$ of the form $\phi = \tilde{\phi} \circ \eta^{-1}$ for an arbitrary $\tilde{\phi} \in \mathcal{C}_c^1(W)$ (see remark 4.6), then using (4.2.7), we obtain:

$$\forall \tilde{\phi} \in \mathcal{C}_c^1(W), \quad \int_W (\beta \cdot \nabla v) \circ \eta \, \tilde{\phi} |\mathrm{D}\eta| \mathrm{d}s\mathrm{d}y = \int_W (-\partial_s \tilde{\phi} \, \tilde{v} - \mathrm{div}(\beta) \circ \eta \, \tilde{\phi} \tilde{v}) |\mathrm{D}\eta| \mathrm{d}s\mathrm{d}y, \tag{4.2.20}$$

where we have used the equality $\partial_s \tilde{\phi} = (\beta \cdot \nabla \phi) \circ \eta$. Now using that $|\mathrm{D}\eta|$ is continuously differentiable with respect to s with $\partial_s |\mathrm{D}\eta| = \mathrm{div}(\beta) \circ \eta |\mathrm{D}\eta|$, as a consequence of the Liouville formula (4.2.13), (4.2.20) rewrites:

$$\forall \tilde{\phi} \in \mathcal{C}_c^1(W), \int_W (\beta \cdot \nabla v) \circ \eta \, \tilde{\phi} |\mathrm{D}\eta| \mathrm{d}s\mathrm{d}y = \int_W (-\partial_s \tilde{\phi} \, \tilde{v} |\mathrm{D}\eta| - \partial_s |\mathrm{D}\eta| \tilde{\phi} \tilde{v}) \mathrm{d}s\mathrm{d}y = -\int_W \tilde{v} \partial_s (\tilde{\phi} |\mathrm{D}\eta|) \mathrm{d}s\mathrm{d}y.$$

By a standard density argument, the above equality holds more generally for functions $\tilde{\phi}$ that are only continuously differentiable with respect to s on W. Therefore, taking $\tilde{\phi} = \tilde{\psi}/|\mathrm{D}\eta|$ as a test function in the above equation for arbitrary $\tilde{\psi} \in \mathcal{C}_c^1(W)$ yields (4.2.19). The change of variables (4.2.7) now yields $||v||_{V_\omega} = ||v \circ \eta||_{\widetilde{V}_\alpha}$, and so $v \circ \eta \in \widetilde{V}_\alpha$.

Conversely, one proves in a similar way that if $\tilde{v}$ is in $\widetilde{V}_\alpha$, the function $v = \tilde{v} \circ \eta^{-1}$ belongs to V_ω, which terminates the proof. □

Remark 4.7. In the following we prove a density result, a trace theorem and a Poincaré inequality in $\widetilde{V}_\alpha$ and we then obtain the direct counterparts of these results in the setting of the graph space V_ω thanks to the above lemma 4.2. There exists actually a wide literature about weighted Sobolev spaces such as $\widetilde{V}_\alpha$ [148, 207, 209, 306], in which analogous results are proved for weights α in the so-called Mückenhoupt class A_2 (see e.g. [136] for a definition of the class A_p). Our working assumptions however are of a different nature: for instance, the hypothesis (H3) essentially requires that the inverse weight α^{-1} belong to the Mückenhoupt class A_1.

4.2.3 Density of functions of class $\mathcal{C}^1$ in the weighted space V_ω

We now examine the density of $\mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$ in $\widetilde{V}_\alpha$, whence we shall infer the density of $\mathcal{C}^1(U) \cap V_\omega$ in V_ω—the loss of regularity between both statements coming from the fact that the coordinate change η is only of class $\mathcal{C}^1$ on W.

Our study classically involves mollifying functions; since the space $\widetilde{V}_\alpha$ of interest contains functions defined on the manifold W, a little treatment is in order. In particular, we shall need the so-called *Ahlfors regularity* of Γ (see [115]); this is the purpose of the next lemma, whose proof is outlined for the convenience of the reader. Here and throughout the chapter, $B(y, h) \subset \mathbb{R}^d$ stands for the open ball with center y and radius h. The measure of a Lebesgue measurable set $A \subset \mathbb{R}^d$ is denoted by $|A|$.

Lemma 4.3 (Area covered by extrinsic balls on an embedded manifold)**.** *Let $\Gamma \subset \mathbb{R}^d$ be a $\mathcal{C}^1$ hypersurface such that either Γ is compact ($\Gamma = \overline{\Gamma}$) or $\overline{\Gamma}$ is a compact manifold with boundary. Then there exists $h_0 > 0$ and constants $m > 0$ and $M > 0$ depending only on Γ such that for any $0 < h < h_0$,*

$$\forall y \in \Gamma,\ \ mh^{d-1} \leq |\Gamma \cap B(y,h)| \leq Mh^{d-1}. \tag{4.2.21}$$

Proof. Since $\overline{\Gamma}$ is compact, there exist finitely many open subset $V_i \subset \mathbb{R}^d$, $i = 1, ..., N$, such that $\overline{\Gamma} \subset \bigcup_i V_i$, and for each $i = 1, ..., N$, there exists a local coordinate chart $\phi_i : U_i \subset \mathbb{R}^{d-1} \to \phi_i(U_i) \subset \overline{\Gamma}$ such that $\phi_i(U_i) = V_i \cap \overline{\Gamma}$. The set U_i is a convex open subset of $\mathbb{R}^{d-1}$ if $V_i \cap \partial\Gamma = \emptyset$; when the latter intersection is not empty, U_i is of the form $U_i = \widetilde{U}_i \cap \mathbb{R}^{d-1}_+$, where $\widetilde{U}_i \subset \mathbb{R}^{d-1}$ is a convex open subset, and $\mathbb{R}^{d-1}_+$ is the upper half-space of $\mathbb{R}^{d-1}$. Let $h_0 > 0$ be a Lebesgue's number associated with this cover, that is:

$$\forall y \in \Gamma,\ \exists i \in \{1, ..., N\},\ \ B(y, h_0) \subset\subset V_i. \tag{4.2.22}$$

Now, given $y \in V_i$, one has for any $0 < h < h_0$:

$$|\Gamma \cap B(y,h)| = \int_\Gamma \mathbf{1}_{B(y,h)} \mathrm{d}z = \int_{\phi_i(U_i)} \mathbf{1}_{B(y,h)} \mathrm{d}z = \int_{U_i} \mathbf{1}_{\phi_i^{-1}(B(y,h))} |\mathrm{D}\phi_i| \mathrm{d}x,$$

where i is the index supplied by (4.2.22) and $\mathbf{1}_A$ denotes the characteristic function of a subset $A \subset \mathbb{R}^d$, and $|\mathrm{D}\phi_i|$ is the Jacobian associated to the change of variables between manifolds induced by ϕ_i.

Let $\sigma_{d-1}(\nabla\phi_i(x)) \leq \sigma_1(\nabla\phi_i(x))$ be respectively the smallest and the largest singular values of the $n \times (n-1)$ Jacobian matrix $\nabla\phi_i(x)$, and let $0 < \sigma_- \leq \sigma_+$ be defined by:

$$\sigma_- = \min_{i=1,...,N} \inf_{x \in U_i} \sigma_{d-1}(\nabla\phi_i(x)),\ \ \sigma_+ = \max_{i=1.,...N} \sup_{x \in U_i} \sigma_1(\nabla\phi_i(x)).$$

Applying the Taylor formula to ϕ_i yields:

$$\forall i = 1, ..., N,\ \ \forall (x_0, x_1) \in U_i,\ \ \sigma_- ||x_1 - x_0|| \leq ||\phi_i(x_1) - \phi_i(x_0)|| \leq \sigma_+ ||x_1 - x_0||.$$

Therefore, possibly shrinking the value of the constant h_0 supplied by (4.2.22) and taking $x_0 = \phi_i^{-1}(y)$, we obtain:

$$\forall 0 < h < h_0,\ B\left(x_0, \frac{h}{\sigma_+}\right) \subset \phi_i^{-1}(B(y,h)) \subset B\left(x_0, \frac{h}{\sigma_-}\right) \subset U_i.$$

Finally, denoting by $|B_{d-1}|$ the volume of the unit ball in $\mathbb{R}^{d-1}$, we obtain:

$$\begin{aligned}
\frac{1}{2}|B_{d-1}| \left(\frac{\sigma_-}{\sigma_+}\right)^{d-1} h^{d-1} &\leq \int_{U_i} \mathbf{1}_{B(x_0, h/\sigma_+)} |\mathrm{D}\phi_i| \mathrm{d}x \\
&\leq \int_{U_i} \mathbf{1}_{\phi_i^{-1}(B(y,h))} |\mathrm{D}\phi_i| \mathrm{d}x = |\Gamma \cap B(y,h)| \\
&\leq \int_{U_i} \mathbf{1}_{B(x_0, h/\sigma_-)} |\mathrm{D}\phi_i| \mathrm{d}x \leq |B_{d-1}| \left(\frac{\sigma_+}{\sigma_-}\right)^{d-1} h^{d-1},
\end{aligned}$$

which completes the proof. □

In what follows, the hypersurface Γ is the one introduced by section 4.2.1. In the next lemma, we construct the kernels ρ_h and $\boldsymbol{\xi}_h$ which shall be used for mollification purposes in W.

Lemma 4.4. *For any $h > 0$, there exist two positive, smooth functions $\rho_h \in \mathcal{C}^\infty_c(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$ and $\boldsymbol{\xi}_h \in \mathcal{C}^\infty_c(\mathbb{R}, \mathbb{R})$ satisfying the conditions:*

(i) For all points $y \in \mathbb{R}^d$, $\mathrm{supp}(\rho_h(y, \cdot)) \subset B(y,h)$.

(ii) For any $y \in \mathbb{R}^d$, $\int_\Gamma \rho_h(y,z)\mathrm{d}z = 1$.

(iii) There exist constants $C > 0$ and $h_0 > 0$ depending only on Γ such that

$$\forall 0 < h < h_0,\ \ \forall z \in \Gamma, \int_\Gamma \rho_h(y,z)\mathrm{d}y \leq C. \tag{4.2.23}$$

(iv) $\mathrm{supp}(\boldsymbol{\xi}_h) \subset [-h, h]$ and $\int_{\mathbb{R}} \boldsymbol{\xi}_h(s)\mathrm{d}s = 1$.

Proof. Let $\rho \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R})$ be a smooth, positive function such that $\mathrm{supp}(\rho) \subset B(0,1)$, $\rho \leq 1$ on $\mathbb{R}^d$ and $\rho_{|B(0,1/2)} = 1$; we define ρ_h by:

$$\forall h > 0, \forall (y,z) \in \mathbb{R}^d \times \mathbb{R}^d, \ \ \rho_h(y,z) := \frac{\rho(\frac{y-z}{h})}{\displaystyle\int_\Gamma \rho\left(\frac{y-z}{h}\right) \mathrm{d}z}; \tag{4.2.24}$$

note that $\rho_h(y,z)$ is not a function of $(y-z)$. The conditions (i) and (ii) of the statement are obviously satisfied by (4.2.24). The condition (iii) is a consequence of lemma 4.3, which implies:

$$\forall 0 < h < h_0, \forall y \in \Gamma, \int_\Gamma \rho\left(\frac{y-z}{h}\right) \mathrm{d}z \geq \left|\Gamma \cap B\left(y, \frac{h}{2}\right)\right| \geq m\left(\frac{h}{2}\right)^{d-1},$$

$$\forall 0 < h < h_0, \forall z \in \Gamma, \int_\Gamma \rho\left(\frac{y-z}{h}\right) \mathrm{d}y \leq |\Gamma \cap B(z,h)| \leq M h^{d-1},$$

so that (4.2.23) holds with $C = 2^{d-1}M/m$.

Eventually, a function $\boldsymbol{\xi}_h$ satisfying (iv) is constructed from any positive function $\boldsymbol{\xi} \in \mathcal{C}_c^\infty(\mathbb{R}, \mathbb{R})$ with compact support inside $[-1,1]$ and unit integral over $\mathbb{R}$, by setting $\boldsymbol{\xi}_h = h^{-1}\boldsymbol{\xi}(\cdot/h)$. □

Definition 4.4 (Mollification on the tensor product manifold W)**.** For $h > 0$, let ρ_h and $\boldsymbol{\xi}_h$ be two functions as in the statement of lemma 4.4. For any $u \in L^1_{\mathrm{loc}}(W)$, $(y,s) \in W$ and $h > 0$ sufficiently small (depending on (y,s)), the mollification of u is the function $u_h = \rho_h \boldsymbol{\xi}_h * u$ defined by

$$(\rho_h \boldsymbol{\xi}_h * u)(y,s) = \int_\Gamma \int_{\mathbb{R}} \rho_h(y,z)\boldsymbol{\xi}_h(s-t)u(z,t)\mathrm{d}t\mathrm{d}z. \tag{4.2.25}$$

Note that for a given open subdomain $W_1 \subset\subset W$ and because $\overline{\Gamma}$ is compact, there exists $h_0 > 0$, depending on W_1, sufficiently small so that (4.2.25) makes sense for any $(y,s) \in W_1$ and $0 < h < h_0$.

Lemma 4.5. *The following properties hold true:*

(i) If $u \in L^1_{loc}(W)$, for any subdomain $W_1 \subset\subset W$ and for $h > 0$ sufficiently small, the convolution $\rho_h \boldsymbol{\xi}_h * u$ is of class $\mathcal{C}^\infty$ on W_1.

(ii) If $u \in \widetilde{V}_\alpha$, then for any subdomain $W_1 \subset\subset W$ and for h sufficiently small, $\partial_s(\rho_h\boldsymbol{\xi}_h * u) = \rho_h\boldsymbol{\xi}_h * \partial_s u$.

(iii) If $\phi \in \mathcal{C}^0(W, \mathbb{R})$ then $\rho_h\boldsymbol{\xi}_h * \phi \to \phi$ in $L^\infty_{loc}(W)$ as $h \to 0$.

Proof. (i) This results from the Lebesgue dominated convergence theorem and the smoothness of ρ_h and $\boldsymbol{\xi}_h$.

(ii) This is again a consequence of the Lebesgue dominated theorem and of an integration by parts:

$$\begin{aligned}\forall (y,s) \in W_1, (\rho_h\boldsymbol{\xi}_h * \partial_s u)(y,s) &= \int_\Gamma \int_{\mathbb{R}} \rho_h(y,z)\boldsymbol{\xi}_h(s-t)\partial_s u(z,t)\mathrm{d}t\mathrm{d}z \\ &= \int_\Gamma \int_{\mathbb{R}} \rho_h(y,z)\partial_s\boldsymbol{\xi}_h(s-t)u(z,t)\mathrm{d}t\mathrm{d}z = \partial_s(\rho_h\boldsymbol{\xi}_h * u).\end{aligned}$$

(iii) For a given subset $W_1 \subset\subset W$, let $\varepsilon > 0$ and $h_1 > 0$ be uniform continuity constants for ϕ on W_1, i.e.

$$\forall\ 0 < h < h_1,\ \ (y,s) \in W_1,\ (z,t) \in W_1, (||y-z|| < h \text{ and } |s-t| < h) \Rightarrow |\phi(z,t) - \phi(y,s)| < \varepsilon.$$

Then for any $0 < h < h_1$:

$$\begin{aligned}\forall (y,s) \in W_1,\ |\rho_h\boldsymbol{\xi}_h * \phi(y,s) - \phi(y,s)| &= \left|\int_\Gamma \int_{\mathbb{R}} \rho_h(y,z)\boldsymbol{\xi}_h(s-t)(\phi(z,t) - \phi(y,s))\mathrm{d}t\mathrm{d}z\right| \\ &\leq \varepsilon \int_\Gamma \int_{\mathbb{R}} \rho_h(y,z)\boldsymbol{\xi}_h(s-t)\mathrm{d}t\mathrm{d}z = \varepsilon.\end{aligned}$$

□

Proposition 4.1. *Let $\alpha \in \mathcal{C}^0(W, \mathbb{R}_+^*)$. Then for any function $u \in L^2_{\alpha,\mathrm{loc}}(W)$, $u_h := \rho_h \boldsymbol{\xi}_h * u$ belongs to $L^2_{\alpha,\mathrm{loc}}(W) \cap \mathcal{C}^\infty(W)$ and $u_h \to u$ in $L^2_{\alpha,\mathrm{loc}}(W)$.*

Proof. We adapt the proof of Prop. 1.18, p.30 in [268]. Let $W_1 \subset\subset W$ be an open subset of W and consider another open subset W_2 such that $W_1 \subset\subset W_2 \subset\subset W$. For $\varepsilon > 0$, let $\phi \in \mathcal{C}^0(W)$ be such that (see [269]):

$$||u - \phi||_{L^2_\alpha(W_2)} < \varepsilon. \tag{4.2.26}$$

From lemma 4.5, $\rho_h \boldsymbol{\xi}_h * \phi \to \phi$ in $L^\infty(W_1)$ as $h \to 0$ and so, for h small enough, $||\rho_h \boldsymbol{\xi}_h * \phi - \phi||_{L^2_\alpha(W_1)} \leq \varepsilon$. Then,

$$||u - u_h||_{L^2_\alpha(W_1)} \leq ||u - \phi||_{L^2_\alpha(W_1)} + ||\phi - \rho_h \boldsymbol{\xi}_h * \phi||_{L^2_\alpha(W_1)} + ||\rho_h \boldsymbol{\xi}_h * \phi - \rho_h \boldsymbol{\xi}_h * u||_{L^2_\alpha(W_1)}. \tag{4.2.27}$$

The first two terms in the right-hand side of (4.2.27) are controlled by ε owing to the previous discussion; as for the last term, we get from the Cauchy-Schwarz inequality:

$$\begin{aligned}
\forall (y,s) \in W_1,\ |\rho_h \boldsymbol{\xi}_h * (\phi - u)(y,s)|^2 &= \left| \int_\Gamma \int_{\mathbb{R}} \rho_h(y,z) \boldsymbol{\xi}_h(s-t)(\phi(z,t) - u(z,t)) \mathrm{d}t\mathrm{d}z \right|^2 \\
&\leq \left(\int_\Gamma \int_{\mathbb{R}} \rho_h(y,z) \boldsymbol{\xi}_h(s-t) \alpha^{-1}(z,t) \mathrm{d}t\mathrm{d}z \right) \left(\int_\Gamma \int_{\mathbb{R}} \rho_h(y,z) \boldsymbol{\xi}_h(s-t) \alpha(z,t) |\phi(z,t) - u(z,t)|^2 \mathrm{d}t\mathrm{d}z \right).
\end{aligned}$$

Multiplying both sides by α and integrating over W_1 yields:

$$\begin{aligned}
&||\rho_h \boldsymbol{\xi}_h * (\phi - u)||^2_{L^2_\alpha(W_1)} \\
&\quad \leq ||\alpha(\rho_h \boldsymbol{\xi}_h * \alpha^{-1})||_{L^\infty(W_1)} \int_{(y,s) \in W_1} \left(\int_\Gamma \int_{\mathbb{R}} \rho_h(y,z) \boldsymbol{\xi}_h(s-t) \alpha(z,t) |\phi(z,t) - u(z,t)|^2 \mathrm{d}t\mathrm{d}z \right) \mathrm{d}s\mathrm{d}y \\
&\quad \leq C ||\alpha(\rho_h \boldsymbol{\xi}_h * \alpha^{-1})||_{L^\infty(W_1)} ||\phi - u||^2_{L^2_\alpha(W_2)},
\end{aligned}$$

where C is the constant supplied by the condition (iii) in the statement of lemma 4.4. By assumption, α^{-1} is a continuous function on W_2, which implies by lemma 4.5 that $\alpha(\rho_h \boldsymbol{\xi}_h * \alpha^{-1}) \to 1$ in $L^\infty(W_1)$. In particular, $||\alpha(\rho_h \boldsymbol{\xi}_h * \alpha^{-1})||_{L^\infty(W_1)}$ is bounded by some constant C'. Finally, using (4.2.26), we obtain from (4.2.27) that for $h > 0$ small enough:

$$||u - u_h||_{L^2_\alpha(W_1)} \leq (CC' + 2)\varepsilon,$$

which is the desired result. □

We conclude this subsection with the desired density result of $\mathcal{C}^1$ functions in V_ω:

Corollary 4.1. *(i) Let $\alpha \in \mathcal{C}^0(W, \mathbb{R}_+^*)$ be a positive weight on W; the space $\mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$ is dense in $\widetilde{V}_\alpha$.*

(ii) Let $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ be a positive weight on U; the space $\mathcal{C}^1(U) \cap V_\omega$ is dense in V_ω.

Proof. The proof of the density (i) of $\mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$ in $\widetilde{V}_\alpha$ relies on a partition of unity argument and on the properties of lemma 4.5 and proposition 4.1, exactly along the lines of the proof of Theorem 5.15 in [268], to which the reader is referred for details.

The density (ii) of $\mathcal{C}^1(U)$ in V_ω follows then from the density of $\mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$ in $\widetilde{V}_\alpha$ with $\alpha = \omega \circ \eta |\mathrm{D}\eta|$ and by composition with the $\mathcal{C}^1$ diffeomorphism η. □

Remark 4.8. Corollary 4.1 does not imply the density of $\mathcal{C}^1(\overline{U})$ in V_ω. A result of this kind would require careful regularity assumptions on ∂U and on the behavior of ω near ∂U.

4.2.4 Trace theorem and Poincaré inequality in V_ω

In this section, the trace operator on Γ is defined and studied for functions in the weighted space V_ω, or equivalently for functions in $\widetilde{V}_\alpha$ on $\Gamma \times \{0\} = \{(y,0) \,|\, y \in \Gamma\}$. In the sequel, with a little abuse of notations, the latter set $\Gamma \times \{0\}$ is identified with Γ.

Proposition 4.2 (Trace theorem). *Let $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ be a positive weight on U. The trace operator*

$$\begin{array}{rccc} \gamma & : & \mathcal{C}^1(U) & \to & L^2(\Gamma) \\ & & v & \mapsto & \gamma(v) = v_{|\Gamma} \end{array} \tag{4.2.28}$$

induces a bounded operator $V_\omega \to L^2(\Gamma)$; there exists a constant $C > 0$ (possibly depending on the weight ω) such that

$$\forall v \in V_\omega,\ ||\gamma(v)||_{L^2(\Gamma)} \leq C||v||_{V_\omega}. \tag{4.2.29}$$

Proof. Introducing $\alpha = \omega \circ \eta|\mathrm{D}\eta|$, using the change of variables (4.2.7) and the density result of corollary 4.1, it is enough to prove that there exists a constant $C > 0$ such that:

$$\forall \tilde{v} \in \mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha,\ \ ||\gamma(\tilde{v})||_{L^2(\Gamma)} \leq C||\tilde{v}||_{\widetilde{V}_\alpha}. \tag{4.2.30}$$

Let us consider the following partition of Γ: $\Gamma = \Gamma_+ \cup \Gamma_-$, where $\Gamma_+ = \{y \in \Gamma \,|\, \zeta_-(y) \geq -\varepsilon/2\}$, $\Gamma_- = \Gamma \backslash \Gamma_+$, and ε is the parameter featured in the separation condition (4.2.2). Then

$$\forall y \in \Gamma_+,\ 0 < \varepsilon/2 < \zeta_+(y),$$

$$\forall y \in \Gamma_-,\ \zeta_-(y) < -\varepsilon/2 < 0.$$

Now, let $a = \varepsilon/2$ and $\boldsymbol{\xi} \in \mathcal{C}_c^\infty(\mathbb{R})$ be such that $\boldsymbol{\xi}(0) = 1$ and $\boldsymbol{\xi}(-a) = \boldsymbol{\xi}(a) = 0$. Let K be the bounded measurable set defined by $K := (\Gamma_- \times [-a, 0]) \cup (\Gamma_+ \times [0, a]) \subset W$. Then for any function $\tilde{v} \in \mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$, it holds:

$$\begin{aligned} \int_\Gamma |\tilde{v}(y,0)|^2 \mathrm{d}y &= \int_\Gamma |\tilde{v}(y,0)^2 \boldsymbol{\xi}(0)| \mathrm{d}y = \int_{\Gamma_-} \int_{-a}^0 \partial_s(\tilde{v}^2 \boldsymbol{\xi}) \mathrm{d}s\mathrm{d}y + \int_{\Gamma_+} \int_a^0 \partial_s(\tilde{v}^2 \boldsymbol{\xi}) \mathrm{d}s\mathrm{d}y \\ &\leq \int_K \left| 2\tilde{v}\partial_s\tilde{v}\boldsymbol{\xi} + \tilde{v}^2 \partial_s \boldsymbol{\xi} \right| \mathrm{d}s\mathrm{d}y \\ &\leq 2||\alpha^{-1}\boldsymbol{\xi}||_{L^\infty(K)} ||\tilde{v}||_{L^2_\alpha(W)} ||\partial_s \tilde{v}||_{L^2_\alpha(W)} + ||\alpha^{-1}\partial_s \boldsymbol{\xi}||_{L^\infty(K)} ||\tilde{v}||^2_{L^2_\alpha(W)} \\ &\leq (2||\alpha^{-1}\boldsymbol{\xi}||_{L^\infty(K)} + ||\alpha^{-1}\partial_s \boldsymbol{\xi}||_{L^\infty(K)}) ||\tilde{v}||^2_{\widetilde{V}_\alpha}, \end{aligned}$$

which implies (4.2.30) and therefore terminates the proof of proposition 4.2. □

Remark 4.9. The proof of proposition 4.2 supplies the existence of the trace on Γ of an arbitrary function $\tilde{v} \in \widetilde{V}_\alpha$, which we shall also denote by $\tilde{v}_{|\Gamma}$.

For later purposes (see section 4.2.5), we shall need the surjectivity of the above trace operator; this is the purpose of the next proposition.

Proposition 4.3 (Surjectivity of traces). *Let $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ be a positive continuous weight.*

(i) The trace operator defined by (4.2.28) is surjective from V_ω onto $L^2(\Gamma)$:

$$L^2(\Gamma) = \{v_{|\Gamma} \,|\, v \in V_\omega\}.$$

(ii) If ω additionally satisfies (H2), then any function $v_0 \in L^2(\Gamma)$ can be extended constantly along the characteristics of β: there exists $v \in V_\omega$ such that $v_{|\Gamma} = v_0$ and $\beta \cdot \nabla v = 0$.

Proof. (i) We rather prove that $L^2(\Gamma) = \{\tilde{v}_{|\Gamma} \,|\, \tilde{v} \in \widetilde{V}_\alpha\}$, where $\alpha = \omega \circ \eta|\mathrm{D}\eta|$; see remark 4.9. To this end, consider $\boldsymbol{\xi}$ and K be as in the proof of proposition 4.2. For an arbitrary function $v_0 \in L^2(\Gamma)$, we define $\tilde{v}$ by the formula

$$\tilde{v}(y, s) := v_0(y)\boldsymbol{\xi}(s) \text{ a.e. in } W.$$

Obviously, $\tilde{v}(y, 0) = v_0(y)$ and $\partial_s \tilde{v}(y, s) = v_0(y)\partial_s \boldsymbol{\xi}(s)$, whence the Cauchy-Schwarz inequality yields:

$$\int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y = \int_\Gamma \int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y, s)\boldsymbol{\xi}(s)^2 v_0(y)^2 \mathrm{d}s\mathrm{d}y < ||\alpha||_{L^\infty(K)} ||\boldsymbol{\xi}||_{L^2(\mathbb{R})} ||v_0||^2_{L^2(\Gamma)} < +\infty,$$

$$\int_W \alpha |\partial_s \tilde{v}|^2 \mathrm{d}s\mathrm{d}y = \int_\Gamma \int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y, s) v_0(y)^2 |\partial_s \boldsymbol{\xi}(s)|^2 \mathrm{d}s\mathrm{d}y \leq ||\alpha||_{L^\infty(K)} ||\partial_s \boldsymbol{\xi}||_{L^2(\mathbb{R})} ||v_0||^2_{L^2(\Gamma)} < +\infty.$$

Hence $\tilde{v}$ is a function in $\widetilde{V}_\alpha$ such that $\tilde{v}_{|\Gamma} = v_0$, which is the desired result.

(ii) If (H2) is satisfied, then for an arbitrary function $v_0 \in L^2(\Gamma)$, we simply define $\tilde{v}$ by the formula:

$$\tilde{v}(y,s) := v_0(y) \text{ a.e. in } W.$$

Then, clearly $\partial_s \tilde{v} = 0$; what's more, one has $\tilde{v} \in L^2_\alpha(W)$ as follows from the following estimate:

$$\int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y = \int_\Gamma \int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y,s) v_0(y)^2 \mathrm{d}s\mathrm{d}y \leq ||g_\alpha||_{L^\infty(\Gamma)} ||v_0||^2_{L^2(\Gamma)}, \tag{4.2.31}$$

where g_α is as in the statement of (H2). Hence the function $\tilde{v}$ belongs to $\widetilde{V}_\alpha$, and so $v := \tilde{v} \circ \eta^{-1}$ is an element of V_ω which has the desired properties owing to lemma 4.2.

□

We now prove a Poincaré-type inequality in the spaces V_ω under the additional assumptions (H2) and (H3) about the weight ω.

Proposition 4.4 (Poincaré type inequality on V_ω)**.** *Let $\omega \in \mathcal{C}^0(U, \mathbb{R}^*_+)$ be a weight satisfying the assumptions (H1) to (H3). Then there exists a constant $C > 0$ (depending on ω) such that:*

$$\forall v \in V_\omega, \int_U \omega\, v^2 \mathrm{d}x \leq C \left[\int_\Gamma v^2 \mathrm{d}y + \int_U \omega |\beta \cdot \nabla v|^2 \mathrm{d}x \right]. \tag{4.2.32}$$

Proof. Introducing again $\alpha = \omega \circ \eta |\mathrm{D}\eta| \in \mathcal{C}^0(W, \mathbb{R}^*_+)$ and using the change of variables (4.2.7), we rather prove the analogous Poincaré inequality in W, that is:

$$\forall \tilde{v} \in \widetilde{V}_\alpha, \int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y \leq C \left[\int_\Gamma \tilde{v}^2 \mathrm{d}s + \int_W \alpha |\partial_s \tilde{v}|^2 \mathrm{d}s\mathrm{d}y \right]. \tag{4.2.33}$$

Furthermore, since $\mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$ is dense in $\widetilde{V}_\alpha$, it is enough to prove that (4.2.33) holds for $\tilde{v} \in \mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$, which we now do. To this end, for arbitrary $\tilde{v} \in \mathcal{C}^\infty(W) \cap \widetilde{V}_\alpha$, a use of Taylor's formula yields:

$$\forall (y,s) \in W,\ \tilde{v}(y,s) = \tilde{v}(y,0) + \int_0^s \partial_s \tilde{v}(y,t) \mathrm{d}t. \tag{4.2.34}$$

In (4.2.34), the Cauchy-Schwarz inequality implies that, for $(y,s) \in W$:

$$\int_0^s |\partial_s \tilde{v}(y,t)| \mathrm{d}t \leq \left(\int_0^s \alpha^{-1}(y,t) \mathrm{d}t \right)^{1/2} \left(\int_{\zeta_-(y)}^{\zeta_+(y)} \alpha |\partial_s \tilde{v}|^2 (y,t) \mathrm{d}t \right)^{1/2}. \tag{4.2.35}$$

Now squaring (4.2.34), using the Young's inequality ($\forall a, b \in \mathbb{R}, (a+b)^2 \leq 2a^2 + 2b^2$) together with (4.2.35), then multiplying by $\alpha(y,s)$, we obtain:

$$\forall (y,s) \in W,\ \alpha(y,s) |\tilde{v}(y,s)|^2 \leq 2\alpha(y,s) |\tilde{v}(y,0)|^2 + 2\alpha(y,s) \int_0^s \alpha^{-1}(y,t) \mathrm{d}t \int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y,t) |\partial_s \tilde{v}|^2 (y,t) \mathrm{d}t. \tag{4.2.36}$$

Integrating (4.2.36) over W now results in:

$$\begin{aligned} \int_W \alpha \tilde{v}^2 \mathrm{d}s\mathrm{d}y &\leq 2 \int_\Gamma \tilde{v}^2(y,0) \left(\int_{\zeta_-(y)}^{\zeta_+(y)} \alpha(y,s) \mathrm{d}s \right) \mathrm{d}y + 2 \int_\Gamma h_\alpha(y) \left(\int_{\zeta_-(y)}^{\zeta_+(y)} \alpha |\partial_s \tilde{v}|^2 \mathrm{d}s \right) \mathrm{d}y \\ &\leq 2||g_\alpha||_{L^\infty(\Gamma)} \int_\Gamma \tilde{v}^2 \mathrm{d}y + 2||h_\alpha||_{L^\infty(\Gamma)} \int_W \alpha |\partial_s \tilde{v}|^2 \mathrm{d}s\mathrm{d}y, \end{aligned}$$

where g_α and h_α are the functions featured in (H2) and (H3). This completes the proof of (4.2.33), and so that of proposition 4.4. □

Remark 4.10. This Poincaré type inequality is close in spirit to the "curvilinear" Poincaré inequality of Azerad [50, 51], who considered the weight $\omega = 1$ and vector fields β satisfying $\mathrm{div}\beta = 0$.

4.2.5 Well-posedness of the variational problem (4.2.5)

We are now in a position to state and prove the main result of this section.

Proposition 4.5. *Let $\omega \in \mathcal{C}^0(U, \mathbb{R}_+^*)$ be a positive weight on U, satisfying the assumptions (H1) to (H3). Then:*

(i) For any function $f \in L^2_{\omega^{-1}}(U)$, there exists a unique solution $u \in V_\omega$ to the variational problem

$$\text{Find } u \in V_\omega \text{ such that } \forall v \in V_\omega, \int_\Gamma uv\mathrm{d}s + \int_U \omega(\beta\cdot\nabla u)(\beta\cdot\nabla v)\mathrm{d}x = \int_U fv\mathrm{d}x. \tag{4.2.37}$$

(ii) The trace $u_{|\Gamma} \in L^2(\Gamma)$ of the solution u to (4.2.37), is independent of the weight ω; it is given by:

$$u(y) = \int_{\zeta_-(y)}^{\zeta_+(y)} f\circ\eta|\mathrm{D}\eta|\mathrm{d}s \quad \text{a.e. on } \Gamma. \tag{4.2.38}$$

Proof. (i) The assumption $f \in L^2_{\omega^{-1}}(U)$ ensures that $v \mapsto \int_U fv\mathrm{d}x$ is a continuous linear form on V_ω owing to the Cauchy-Schwartz inequality:

$$\forall v \in V_\omega, \left|\int_U fv\mathrm{d}x\right| \leq ||f||_{L^2_{\omega^{-1}}(U)}||v||_{L^2_\omega(U)}.$$

Moreover, the Poincaré inequality of proposition 4.4 ensures the coercivity on V_ω of the bilinear form

$$(u,v) \mapsto \int_\Gamma uv\mathrm{d}y + \int_U \omega(\beta\cdot\nabla u)(\beta\cdot\nabla v)\mathrm{d}x.$$

Hence, the classical Lax-Milgram theorem (see e.g. [144]) yields the existence and uniqueness of a solution $u \in V_\omega$ to (4.2.37).

(ii) Since ω satisfies (H2), it follows from proposition 4.3 (ii) that for any $v_0 \in \mathcal{C}^0(\Gamma)$, there exists a function $v \in V_\omega$ such that $v_{|\Gamma} = v_0$ and $\beta\cdot\nabla v = 0$. Taking v as an admissible test function in (4.2.37) and using once again the change of variables (4.2.7) yields:

$$\forall v_0 \in \mathcal{C}^0(\Gamma), \int_\Gamma uv_0\mathrm{d}y = \int_U fv\mathrm{d}x = \int_\Gamma v_0(y)\left(\int_{\zeta_-(y)}^{\zeta_+(y)} f\circ\eta|\mathrm{D}\eta|\mathrm{d}s\right)\mathrm{d}y, \tag{4.2.39}$$

whence (4.2.38).

□

Remark 4.11. Problem (4.2.5) or (4.2.37) can be solved, for an arbitrary choice of weight ω satisfying (H1) to (H3), if the right hand side satisfies $f \in L^2_{\omega^{-1}}(U)$. This holds true, for example when f belongs to $L^\infty(U)$, as soon as the weight ω satisfies: $\omega^{-1} \in L^1(U)$. The latter property is not a consequence of (H1) to (H3) as explained in Remark 4.13.

Remark 4.12. If (H2) is not satisfied, which is the case if for example ω blows up "too fast" near some part of the boundary of U, functions $v \in V_\omega$ are expected to vanish near this part of ∂U and then V_ω may not contain all functions which are constant along the characteristic curves of β.

4.3 Numerical methods for integration along normal rays

This section provides recipes for the implementation of the variational formulation (4.2.5) with the finite element method, and numerical comparisons with the line integral formula (4.2.6). For simplicity and because we are motivated by our shape optimization applications, this comparison is considered in the context where these lines are normal rays to the shape.

The general setting of section 4.2 to the shape optimization context outlined in the introduction; in section 4.3.1. We then clarify, for comparison purposes, some of the algorithmic stages required by the direct integration of (4.2.6) along rays in section 4.3.2—such as the numerical computation of the principal curvatures κ_i of the considered shapes, and the delicate detection of their skeleton on unstructured meshes—. We discuss next in section 4.3.3 the construction of suitable weights ω that allow to to solve accurately the variational problem (4.2.5) by means of $\mathbb{P}_1$ conforming finite elements. Finally, in section 4.3.4, we compare on several 2-d analytical examples the numerical calculations of (4.2.6) by means of our variational formulation (4.2.5) with those produced by direct integration of this formula along rays.

4.3.1 Shape optimization context: normal rays and flow map of the signed distance function gradient

This section explicits the shape optimization context in the framework of the previous section 4.2 and in the perspective of applying the variational problem (4.2.5) for enforcing distance constraints. A fixed, bounded and Lipschitz 'hold-all' open domain $D \subset \mathbb{R}^d$ is considered, as well as a bounded $\mathcal{C}^2$ open subdomain $\Omega \subset D$.

The vector field β is given by the gradient of the signed distance function:

$$U = D\backslash\overline{\Sigma},\ \Gamma = \partial\Omega,\ \beta = \nabla d_\Omega. \tag{4.3.1}$$

The bound functions $\zeta_-(y)$ and $\zeta_+(y)$ correspond to the extremities of the normal rays as follows:

Definition 4.5 (Normal rays)**.** For $y \in \partial\Omega$, the ray emerging from y is the one-dimensional segment

$$\mathrm{ray}(y) := \left\{x \in D\backslash\overline{\Sigma},\ p_{\partial\Omega}(x) = y\right\} = \{y + s\nabla d_\Omega(y)\,|\,s \in (\zeta_-(y), \zeta_+(y))\}. \tag{4.3.2}$$

where $\zeta_-(y)$ and $\zeta_+(y)$ are the maximum distances at which $\mathrm{ray}(y)$ hits either the skeleton Σ or the boundary ∂D of the hold-all domain:

$$\forall y \in \partial\Omega,\ \zeta_+(y) = \sup\{s \geq 0\,|\,\{y + t\nabla d_\Omega(y)\,|\,t \in [0, s)\} \cap (\overline{\Sigma} \cup \partial D) = \emptyset\}, \tag{4.3.3}$$

$$\forall y \in \partial\Omega,\ \zeta_-(y) = \inf\{s \leq 0\,|\,\{y + t\nabla d_\Omega(y)\,|\,t \in (s, 0]\} \cap (\overline{\Sigma} \cup \partial D) = \emptyset\}. \tag{4.3.4}$$

The functions ζ_- and ζ_+ are continuous on $\partial\Omega$ (see [82] or [215]).

Finally, the next proposition explicits the local coordinate change η and its Jacobian $|\mathrm{D}\eta|$ featured in the identity (4.2.6) (see [64] and chapter 1, section 1.3). Because this will be useful when investigating the formulation of relevant weights ω in lemma 4.6, we also recall the value of Δd_Ω.

Proposition 4.6 (Shape optimization setting)**.** *Let $\Omega \subset D$ be a domain of class $\mathcal{C}^2$; then the signed distance function d_Ω is of class $\mathcal{C}^2$ on the open set $U := D\backslash\overline{\Sigma}$. Hence $\beta := \nabla d_\Omega$ is a vector field of class $\mathcal{C}^1$ on U; the associated flow map $\eta : W \to D\backslash\overline{\Sigma}$ is a diffeomorphism of class $\mathcal{C}^1$, whose expression reads:*

$$\forall (y, s) \in W,\ \eta(y, s) = y + s\nabla d_\Omega(y), \tag{4.3.5}$$

where W is the set defined by (4.2.3). The inverse flow mapping $\eta^{-1} : U \to W$ is given by

$$\forall x \in U,\ \eta^{-1}(x) = (p_{\partial\Omega}(x), d_\Omega(x)). \tag{4.3.6}$$

The divergence of the vector field $\beta = \nabla d_\Omega$ and the Jacobian $|\mathrm{D}\eta|$ of the flow map η are respectively given by

$$\forall x \in D\backslash\overline{\Sigma},\ \mathrm{div}(\nabla d_\Omega)(x) = \Delta d_\Omega(x) = \sum_{i=1}^{d-1} \frac{\kappa_i(p_{\partial\Omega}(x))}{1 + d_\Omega(x)\kappa_i(p_{\partial\Omega}(x))}, \tag{4.3.7}$$

$$\forall (y, s) \in W,\ |\mathrm{D}\eta|(y, s) = |\det(\nabla\eta)|(y, s) = \prod_{i=1}^{d-1}(1 + s\kappa_i(y)). \tag{4.3.8}$$

4.3.2 Computing curvatures and detecting the skeleton for direct integration along the rays

Before going to the numerical aspects of our variational method, we clarify important practical details which are required in the implementation of the direct integration along characteristics involved in the calculation of (4.2.6).

We first discuss the delicate issue of detecting mesh triangles crossing the skeleton Σ of Ω when traveling along the rays (note that this step is not required by our variational method). Then, we detail the method we used to compute the curvature κ (there is only one curvature $\kappa := \kappa_1$ in 2-d) required in the line integral formula (4.1.3). Note that these steps serve only for comparison purposes with our variational method. These are fairly classical numerical issues which could otherwise be addressed with more sophisticated techniques, see e.g. [270, 124].

For our present numerical applications, the hold-all domain D is equipped with a simplicial mesh $\mathcal{T}$ featuring a discretization of the domain Ω as a submesh. The only information we use about Ω is an accurate approximation d_h of the signed distance function d_Ω as an element of the space V_h of Lagrange $\mathbb{P}_1$ finite element, where h is the maximum mesh element size. In our context, this approximation d_h is obtained by the algorithm of [111] implemented in the software program `mshdist`.

Detection of the skeleton Σ of Ω and identification of normal rays

The numerical detection of Σ in the course of the identification of the set $\text{ray}(y)$ for some given point $y \in \partial\Omega$ is achieved by assessing the following criterion (independently of the dimension), holding at the continuous level (see [122]):

$$\forall y \in \partial\Omega,\ \forall z_0, z_1 \in \text{ray}(y),\ \frac{d_\Omega(z_1) - d_\Omega(z_0)}{(z_1 - z_0)\cdot \nabla d_\Omega(y)} = 1.$$

In our implementation, when computing the ray emerging at some point $y \in \partial\Omega$ (which is detected by the fact that $d_h(y) = 0$), we travel the triangles in the mesh $\mathcal{T}$ in the normal direction $\boldsymbol{n} = \nabla d_h(y)$, and we stop the calculation of the ray in the triangle $T \in \mathcal{T}$ where the entering and exiting points z_0 and z_1 satisfy:

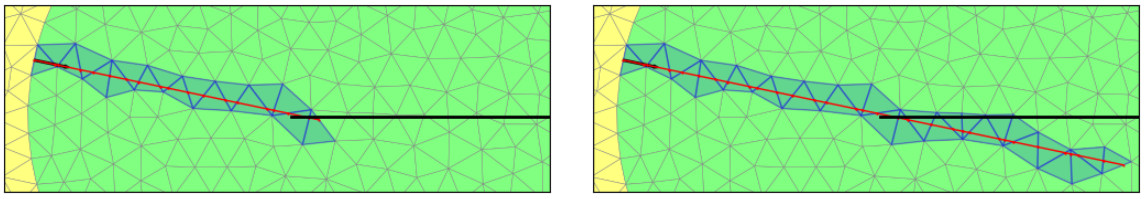
$$\left| \frac{d_h(z_1) - d_h(z_0)}{||z_1 - z_0||} - \text{sign}(d_h(z_0)) \right| \geq \texttt{tolRay}, \tag{4.3.9}$$

where `tolRay` a small tolerance (set to 0.3 in our implementation). This provides meanwhile an approximate location of the skeleton Σ, up to a tolerance of the order of the mesh size.

Our criterion (4.3.9) differs from that used in the related works [30, 234]. In there, the authors detect Σ by looking at changes in the monotonicity of the signed distance function d_h along the ray, i.e. they rely on the following property of the (continuous) signed distance function d_Ω:

$$\forall y \in \partial\Omega,\ \forall z_0, z_1 \in \text{ray}(y),\ (d_\Omega(z_1) - d_\Omega(z_0))((z_1 - z_0)\cdot \nabla d_\Omega(y)) \geq 0. \tag{4.3.10}$$

Our personal experiment with the above criterion suggests that it may sometimes fail to detect the skeleton Σ accurately, because such a change in monotonicity may simply not occur when the ray is supposed to cross Σ in the neighborhood of center of curvatures (see remark 1.10). Our criterion (4.3.9) may also fail depending on the chosen tolerance parameter, but it offered visible improvements (Figure 4.4) in our academic test-cases. Note that when integrating along the ray, the last triangle, where the skeleton is hit, is included in the integration.



(a) Criterion (4.3.9) (b) Criterion (4.3.10)

Figure 4.4: Comparison between the two criteria of section 4.3.2 for the detection of Σ when travelling along rays in an unstructured mesh (Skeleton displayed in the black line).

Estimating the curvature κ of a 2-d subdomain based on its signed distance function

In this part, we detail our method for the numerical approximation, in 2-d, of the principal curvature κ (there is only one in 2-d) from the knowledge of a $\mathbb{P}^1$ discretization d_h of the signed distance function d_Ω at the nodes of the mesh $\mathcal{T}$. We essentially rely on the fact that in 2-d, κ is given by the trace of Δd_Ω on the boundary $\partial\Omega$ (in view of (4.3.7). In 3-d, the estimation of Δd_Ω would not be sufficient to evaluate the values of both principal curvatures κ_1 and κ_2: these could e.g. be computed from the eigenvalues of the Hessian matrix $\nabla^2 d_\Omega$.

Our first step towards estimating Δd_Ω consists in calculating a $\mathbb{P}_1$ interpolation $g_h \in V_h \times V_h$ of the piecewise constant gradient ∇d_h by solving the following variational problem:

$$\forall \psi_h \in V_h \times V_h,\ \int_D g_h \cdot \psi_h \mathrm{d}x = \int_D \nabla d_h \cdot \psi_h \mathrm{d}x. \tag{4.3.11}$$

The approximation of the divergence $\text{div}(\nabla d_\Omega)$ is then calculated as the (piecewise constant) divergence of the reconstructed field g_h. Unfortunately, this procedure generally produces a very noisy approximation characterized by a lot of spurious oscillations when the mesh resolution increases (see Figure 4.5). In

order to overcome this difficulty, we calculate a regularization κ_h of this noisy estimation with a Laplace kernel, namely we solve:

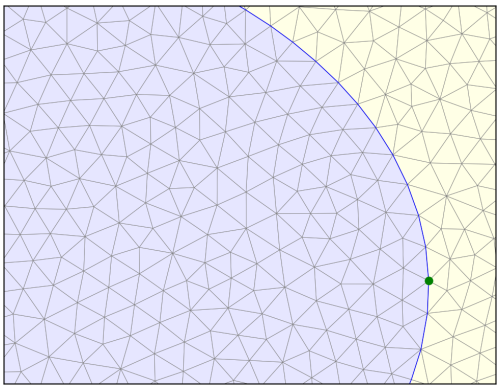
$$\text{Find } \kappa_h \in V_h \text{ such that } \forall \psi_h \in V_h,\ \int_D \left(\gamma_h^2 \nabla \kappa_h \cdot \nabla \psi_h + \kappa_h \psi_h\right) \mathrm{d}x = \int_D \mathrm{div}(g_h)\psi_h \mathrm{d}x, \tag{4.3.12}$$

where $\gamma_h > 0$ is a regularization length scale (equal to $3h_{\max}$ where $h_{\max}$ is the maximum edge length in the mesh). This procedure yields satisfying results in practice (even with shapes Ω characterized by discontinuous curvatures, up to some over smoothing near the discontinuities), although we do not have a proof of convergence of the approximation κ_h towards the exact function Δd_Ω.

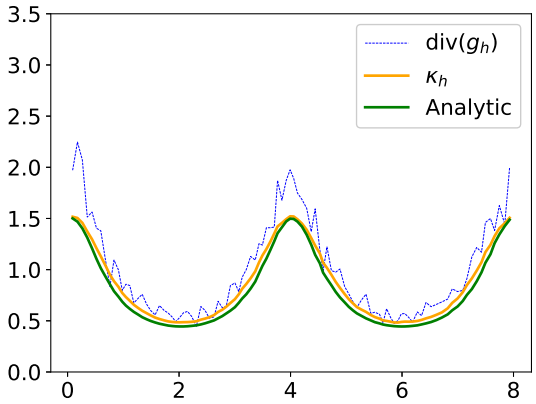
Let us illustrate our method by considering the example of an ellipse Ω inside a square-shaped hold-all domain D: let D and Ω be defined by

$$D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, |x_1| < 2 \text{ and } |x_2| < 2\},\ \ \Omega = \left\{(x_1, x_2) \in D \,|\, \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} < 1\right\}, \tag{4.3.13}$$

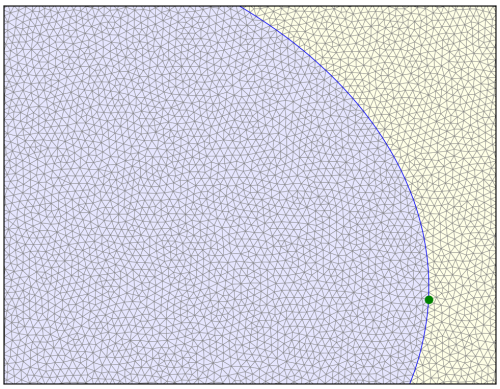
where $a = 1.5$ and $b = 1$. The skeleton of Ω is explicitly known in this case: $\Sigma = \{(x_1, 0) \,|\, |x_1| < a - b^2/a\}$ and the curvature κ of $\partial\Omega$ at a point $y = (y_1, y_2)$ is given by $\kappa(y) = ab/\gamma^3$ with $\gamma = \sqrt{\frac{b^2}{a^2}y_1^2 + \frac{a^2}{b^2}y_2^2}$. The difference between the exact curvature $\kappa(y)$ of $\partial\Omega$ and its reconstruction κ_h (at the boundary nodes discretizing $\partial\Omega$) using both procedures (4.3.11) and (4.3.12) is represented in Figure 4.5.



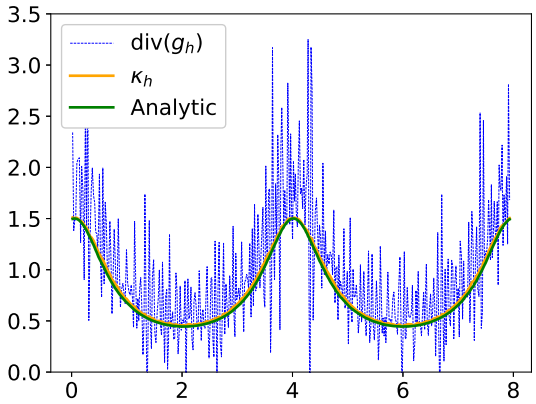
(a) Zoom on a mesh discretization of an ellipse $\Omega \subset D$ with a maximum edge length $h_{\max}$=0.08.



(b) Estimated values of Δd_Ω on $\partial\Omega$.



(c) Zoom on a finer mesh discretization of $\Omega \subset D$ with a maximum edge length $h_{\max}$=0.02.



(d) Estimated values of Δd_Ω on $\partial\Omega$.

Figure 4.5: Estimation of Δd_Ω on the mesh $\mathcal{T}$ for the shape Ω in (4.3.13) and for two different mesh resolutions. The x coordinate on the right-hand graphs represents the arc length coordinate on $\partial\Omega$ when the starting point is the green reference point. Estimates of the mean curvature of $\partial\Omega$, $\mathrm{div}(g_h)$ and κ_h (see (4.3.11) and (4.3.12) for the definitions) are compared to the analytical value $\kappa(y)$.

4.3.3 Admissible numerical weights built upon the signed distance function

In this section, we discuss the numerical resolution of the variational problem (4.2.5) in the shape optimization context (4.2.8) and (4.2.9) (see also proposition 4.6). The numerical setting is the same

as that of the previous section 4.3.2: the hold-all domain D is equipped with a simplicial mesh $\mathcal{T}$ (i.e. composed of triangles in 2-d, or tetrahedra in 3-d, although our method would work with other kinds of meshes) featuring a discretization of the domain Ω as a submesh, and we rely on the Lagrange $\mathbb{P}_1$ finite element method for the discretization of (4.2.5) on account of its robustness and ease of implementation. In other terms, the approximation u_h to the solution $u \in V_\omega$ of (4.2.5) is sought in the space V_h of continuous piecewise linear functions on each simplex of $\mathcal{T}$.

We start by introducing some motivations for the use in (4.2.5) of weights ω vanishing on the skeleton, and we provide (in lemma 4.6 below) a formula for analytical and admissible weights satisfying approximately this property. We then discuss the issue of computing numerically these weights. Finally, we perform a few numerical experiments where we show that weights vanishing on the skeleton make $\mathbb{P}_1$ finite elements able to capture discontinuous test functions across the skeleton, which allows to confirm numerically the latter motivations.

Motivations for weights vanishing on the skeleton

As already mentioned, our final target is to calculate the trace (4.2.6) on $\Gamma = \partial\Omega$ of u. In the continuous setting of section 4.2, this trace does not depend on the choice of the weight ω as long as it fulfills (H1) to (H3), so that in principle, any such weight could be used. However, when U is a cracked domain (typically, the skeleton Σ is a crack in the working domain, i.e., $U = D\backslash\overline{\Sigma}$) and the crack is not explicitly discretized in the mesh $\mathcal{T}$, then the most simple choice $\omega = 1$ (which is an admissible weight on account of lemma 4.6 below) might not work well in practice. Indeed and as we shall illustrate below in this section, test functions of (4.2.5) which belong to V_ω with $\omega = 1$ are in general discontinuous across Σ and not well captured by $\mathbb{P}_1$ finite elements. In many shape optimization applications, discretizing the skeleton Σ of the current domain Ω at every iteration of the optimization process or resorting to discontinuous finite elements (in order to be able to set $\omega = 1$) is very inconvenient, for instance if working with fixed structured meshes as it is performed in many applications built on the level set method [26, 311].

Therefore, we shall be interested in determining weights ω adapted to our commitment to use the space V_h of Lagrange $\mathbb{P}_1$ finite elements for the resolution of (4.2.5) without the need for an accurate discretization of the skeleton Σ (alternative approaches could be to use discontinuous finite elements close to the skeleton, or to have a zero weight on the degrees of freedom corresponding to modes close to the skeleton and to remove the null space in the corresponding linear system, but they seemed to be more complicated to implement, at least to us). The weight ω should be chosen in such a way that arbitrary functions $v \in V_\omega$ are well approximated (in the norm $||\cdot||_{V_\omega}$) by functions $v_h \in V_h$, as is reflected by the classical Céa's lemma (see e.g. [144]):

$$||u - u_h||_{V_\omega} \leq C \inf_{v_h \in V_h} ||u - v_h||_{V_\omega}, \tag{4.3.14}$$

for a constant $C > 0$ (which possibly depends on ω). The space V_h of Lagrange $\mathbb{P}_1$ elements is a conformal finite element space in the sense that the inclusion $V_h \subset V_\omega$ always holds (because functions of V_h are smoother than those of V_ω), however V_h may be "too small" to guarantee a correct approximation of discontinuous solutions $u \in V_\omega$ in the sense of (4.3.14). Heuristically, and without looking for a very precise statement, these considerations call for the choice of a weight ω almost vanishing on $\overline{\Sigma}$, so that the approximation error $||u - v_h||_{V_\omega}$ in (4.3.14) attributes a lesser weight to a neighborhood of $\overline{\Sigma}$ where u is expected to be discontinuous while the functions $v_h \in V_h$ are continuous.

We now provide explicit candidates for weights ω which fulfill the conditions (H1) to (H3) while taking small values near the skeleton, that we are going to use in our practical implementations.

Lemma 4.6. *For any real numbers $q \geq 0, r \geq 0$, the weight*

$$\omega = \frac{1}{1 + |d_\Omega|^q |\Delta d_\Omega|^r} \tag{4.3.15}$$

satisfies the conditions (H1) to (H3) (this includes in particular the constant weight $\omega = 1/2$ for $q = r = 0$).

Proof. At first, it follows readily from the definition that ω belongs to $\mathcal{C}^0(U, \mathbb{R}_+^*)$, and is uniformly bounded on U, so that (H1) and (H2) are trivially satisfied. We then define $\kappa_-(y)$ and $\kappa_+(y)$ by:

$$\kappa_-(y) = \min(0, \min_i \kappa_i(y)), \ \ \kappa_+(y) = \max(0, \max_i \kappa_i(y)), \tag{4.3.16}$$

as well as the corresponding multiplicities $m_-(y)$ and $m_+(y)$:

$$m_\pm(y) = \mathrm{Card}(\{i \in \{1,\dots,n-1\} \mid \kappa_i(y) = \kappa_\pm(y)\}).$$

Using formula (4.3.7) for Δd_Ω, the weight $\alpha = \omega \circ \eta |\mathrm{D}\eta|$ is decomposed as

$$\begin{aligned}\alpha(y,s) &= \frac{\prod_{i=1}^{d-1}(1+\kappa_i(y)s)}{1+|s|^q\left|\sum_i \frac{\kappa_i(y)}{1+\kappa_i(y)s}\right|^r} = \frac{\prod_{i=1}^{d-1}(1+\kappa_i(y)s)^{r+1}}{\prod_{i=1}^{d-1}(1+\kappa_i(y)s)^r + |s|^q\left|\sum_i \kappa_i(y)\prod_{j\neq i}(1+\kappa_j(y)s)\right|^r} \\ &= f(y,s)g(y,s)\end{aligned} \tag{4.3.17}$$

where f and g are the following functions:

$$f(y,s) = (1+s\kappa_-(y))^{m_-(y)(r+1)}\mathbf{1}_{s\geq 0}(s) + (1+s\kappa_+(y))^{m_+(y)(r+1)}\mathbf{1}_{s<0}(s), \tag{4.3.18}$$

$$g(y,s) = \frac{\prod_{i,\kappa_i\neq\kappa_-}(1+\kappa_i(y)s)^{r+1}\mathbf{1}_{s\geq 0}(s) + \prod_{i,\kappa_i\neq\kappa_+}(1+\kappa_i(y)s)^{r+1}\mathbf{1}_{s<0}(s)}{\prod_{i=1}^{d-1}(1+\kappa_i(y)s)^r + |s|^q\left|\sum_i \kappa_i(y)\prod_{j\neq i}(1+\kappa_j(y)s)\right|^r}. \tag{4.3.19}$$

Then, f satisfies clearly the monotonicity condition (ii) in the statement of lemma 4.1 and g is a continuous function on each of the domains $W_- = \{(y,s) \in W \mid s \leq 0\}$ and $W_+ = \{(y,s) \in W \mid s \geq 0\}$, that does not vanish on the compact sets $\overline{W_-}$ and $\overline{W_+}$. The assumptions of lemma 4.1 are therefore fulfilled, so that ω satisfies (H3). This terminates the proof. □

Remark 4.13.

- When it comes to solving (4.2.5) by relying on proposition 4.5 with some data $f \in L^\infty(U)$, it is useful to observe that f belongs to $L^2_{\omega^{-1}}(U)$ as soon as the weight ω satisfies $\omega^{-1} \in L^1(U)$ (see Remark 4.11), which is the case if it is of the form (4.3.15) with $r < 2$. In the following numerical experiments, we shall see however that using values for r which are larger than 2 still provides good results in practice: in general, taking higher values of q and r yields a faster decay of ω near the skeleton.

- Taking $r > 0$ ensures that the weight (4.3.15) will vanish at points $x \in \overline{\Sigma}$ that are centers of curvatures (for which Δd_Ω blows up). Taking $q > 0$ allows to make sure that $\omega = 1/2$ on $\partial\Omega$ whatever the value of r, and to accentuate the decay of ω near the skeleton. Not that in general, ω will not vanish on points $x \in \overline{\Sigma}$ that are not centers of curvatures. However, it still takes very small values on $\overline{\Sigma}$ and it is convenient to use in the implementation. There is no unicity of weights appropriate for the numerical computation and variants can easily be imagined.

- For the most general setting where U is an arbitrary open set, the methodology of this section extends naturally by considering weights ω which vanish on the cracked parts of ∂U that are not explicitly meshed.

Numerical computations of the weight based on the Laplacian of the signed distance function

For our numerical applications below, we shall use the weights of lemma 4.6 in the definition of our variational formulation (4.2.5). This requires the computation of the Laplacian Δd_Ω on the triangulated mesh $\mathcal{T}$ based on the $\mathbb{P}^1$ estimation of the signed distance function d_Ω. For this purpose, we use the same regularization method outlined in section 4.3.2 for the computation of the numerical curvature.

Importantly, from proposition 4.5, the variational formulation (4.2.5) is rather insensitive to the choice of the weight ω, and as a result the estimation of Δd_Ω does not need to be very accurate as long as it takes large values near the skeleton (as we shall illustrate below in section 4.3.3). In contrast, the estimation of the principal curvatures κ_i for the direct method would *need* to be accurate. From a numerical standpoint, the formula (4.3.15) featuring Δd_Ω at the denominator is convenient to obtain numerically vanishing weights near the skeleton (even if (4.3.15) truely vanish for $r > 0$ at centers of curvatures). Indeed, ∇d_Ω is discontinuous across the skeleton, which should reflect in high numerical values of Δd_Ω when computing numerically the divergence $\mathrm{div}(\nabla d_\Omega)$ with the method of section 4.3.2.

Assessing the choice of the weight ω when using a $\mathbb{P}_1$ discretization: generating numerical test functions constant along rays

In this section, we perform several numerical experiments about the influence of the choice of the weight function ω in the resolution of the variational problem (4.2.5) using the Lagrange $\mathbb{P}_1$ finite element method. With this perspective in mind, and in the 2-d numerical setting described in sections 4.3.2 and 4.3.3, we consider the issue of generating numerical functions $v \in V_\omega$ which are constant along the normal rays to the shape Ω. Namely, we solve the boundary-value problem

$$\begin{cases} \nabla d_\Omega \cdot \nabla v = 0 & \text{in } D\backslash\overline{\Sigma}, \\ v = v_0 & \text{on } \partial\Omega, \end{cases} \tag{4.3.20}$$

for given data $v_0 \in L^2(\partial\Omega)$.

Using the variational setting of section 4.2, we show that it is possible to obtain the solution v to (4.3.20) by solving a variational problem of the same nature of (4.2.5).

Proposition 4.7. *Let $\omega \in \mathcal{C}^0(D\backslash\overline{\Sigma}, \mathbb{R}_+^*)$ be a weight satisfying (H1) to (H3). There exists a unique solution $v \in V_\omega$ to the following variational problem:*

$$\text{Find } v \in V_\omega \text{ such that } \forall w \in V_\omega,\ \int_{\partial\Omega} vw\mathrm{d}s + \int_{D\backslash\overline{\Sigma}} \omega(\nabla d_\Omega \cdot \nabla v)(\nabla d_\Omega \cdot \nabla w)\mathrm{d}x = \int_{\partial\Omega} v_0 w\mathrm{d}s. \tag{4.3.21}$$

The solution v is independent of the choice of ω as long as (H1) to (H3) are satisfied, and it is given by the formula:

$$v(x) = v_0(p_{\partial\Omega}(x)),\ \text{a.e. } x \in D\backslash\overline{\Sigma}. \tag{4.3.22}$$

Proof. To see that (4.3.20) and (4.3.21) are equivalent, it is sufficient to take $v \in V_\omega$ constant along rays in (4.3.21) as in the proof of proposition 4.5, which yields $v = v_0$ on $\partial\Omega$, and then $\nabla d_\Omega \cdot \nabla v = 0$. □

The formulation (4.3.21) is to be compared to the so-called Galerkin Least Square formulation and SUPG methods for advection-reaction problems [144], with the difference that usual assumptions of uniformly bounded divergence (which do not hold in our applications) are replaced with the regularity assumptions of section 4.2.1.

Remark 4.14. The problem of building constant functions along normal rays of the form (4.3.20) may be solved on unstructured meshes owing to variants of the fast marching algorithm; see e.g. [97].

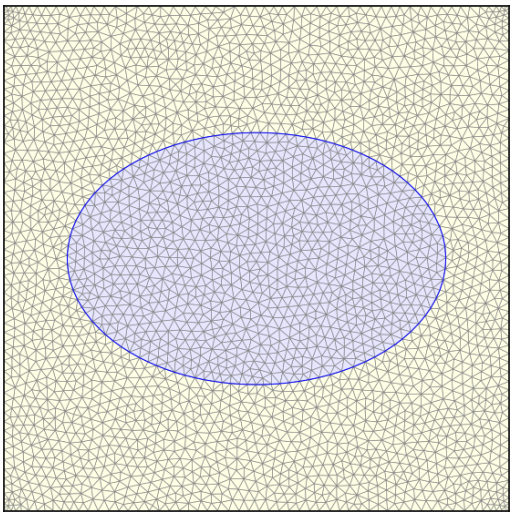
We now verify that a good approximation of the solution v to (4.3.20) (or more precisely (4.3.21)), which is in particular discontinuous across the skeleton Σ of Ω, can be obtained either by truncating manually the skeleton from the computational mesh $\mathcal{T}$ and taking the weight $\omega = 1$, or by selecting a weight ω taking "small" values near Σ. For the latter experiment, we use the weight $\omega = 1/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^2)$ (see also remark 4.13 about the "small" values of ω near Σ). Of course, removing the skeleton from the mesh is not a straightforward task in full generality but it is performed here for the sake of comparison.

Let us consider again the ellipse example of (4.3.13). We consider two different computational meshes $\mathcal{T}$ and $\mathcal{T}'$. The former is a triangular mesh of D, and the latter $\mathcal{T}'$ is a triangular mesh of $D\backslash\overline{\Sigma}$ (i.e. the skeleton $\overline{\Sigma}$ has been manually removed). In both $\mathcal{T}$ and $\mathcal{T}'$, the considered shape Ω is explicitly discretized as a submesh; see Figure 4.6 for an illustration. The variational problem (4.3.21) is numerically solved for a boundary datum $v_0 \in L^2(\partial\Omega)$ given by (see Figure 4.7d):

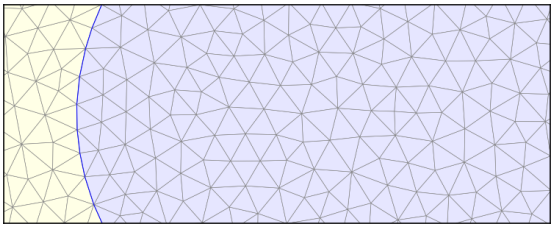
$$\forall (y_1, y_2) \in \partial\Omega,\ v_0(y_1, y_2) = \cos(3y_1)^2 + 20y_2, \tag{4.3.23}$$

and the computed finite element solution is plotted on Figure 4.7 in the following three situations.

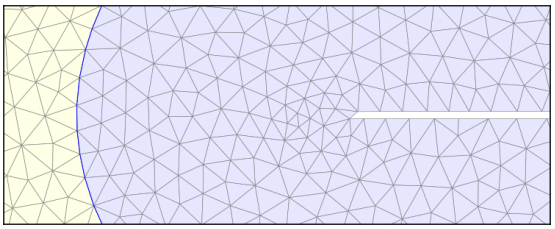
- The mesh $\mathcal{T}'$ is used, in which D and $\Omega \subset D$ are meshed explicitly, and where a thin layer around $\overline{\Sigma}$ has been manually removed (see Figure 4.6c). The solution v to (4.3.21) is computed using the constant weight $\omega = 1$ and the result is displayed on Figure 4.7a. As expected, the fact that $\overline{\Sigma}$ is absent from $\mathcal{T}'$ allows the numerical solution v to have very different values on either sides of $\overline{\Sigma}$.
- The mesh $\mathcal{T}$ of D (where $\overline{\Sigma}$ has not been removed) is used (see Figure 4.6b), and v is computed with the constant weight $\omega = 1$; the result is represented on Figure 4.7b. The formulation (4.3.21) proves numerically stable with the choice $\omega = 1$, but it tends to over smoothen the sharp discontinuities of v near the skeleton Σ, which results in a loss of accuracy for the extension problem (4.3.20).



(a) Mesh $\mathcal{T}$ of the hold-all domain D with $\Omega \subset D$ (*in blue*) discretized as a submesh.



(b) Zoom on the mesh $\mathcal{T}$.



(c) Zoom on the mesh $\mathcal{T}'$ (skeleton Σ truncated)

Figure 4.6: The two meshes $\mathcal{T}$ and $\mathcal{T}'$ used for the numerical example of section 4.3.3.

- The mesh $\mathcal{T}$ of D is used again, but the solution v to (4.3.21) is now computed by using the weight $\omega = 2/(1+|d_\Omega|^{3.5}|\Delta d_\Omega|^2)$; the result is represented on Figure 4.7c. The obtained numerical solution is much closer to the expected result (4.3.22): the values of the solution function v look constant along the normal rays up to a small neighborhood of the skeleton (of the size of the mesh element size), where sharp variations are observed, as expected. The numerical procedure in this case seems therefore to achieve the same order of accuracy than in the experiment using the truncated mesh $\mathcal{T}'$.

4.3.4 Numerical comparisons between the variational method and direct integration along rays

We now investigate the numerical evaluation of the function $u \in L^2(\partial\Omega)$ in (4.2.6) for several 2-d academic configurations of domains D, Ω and functions f. In particular, we compare the evaluation of u obtained by direct integration along rays (*i.e.* implementing directly the formula (4.2.6)) to that obtained by solving the variational formulation (4.2.5) on meshes $\mathcal{T}$ (resp. $\mathcal{T}'$) of D in which Ω is explicitly discretized and the skeleton Σ of Ω is not removed (resp. is removed).

A domain with trivial skeleton: the case of a circle

We first consider the case where Ω is a disk enclosed in a larger disk D:

$$D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1^2 + x_2^2 < 4\}, \text{ and } \Omega = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, x_1^2 + x_2^2 < 1\}; \tag{4.3.24}$$

see Figure 4.8. In this case, the skeleton Σ is reduced to the point 0. The considered function f is:

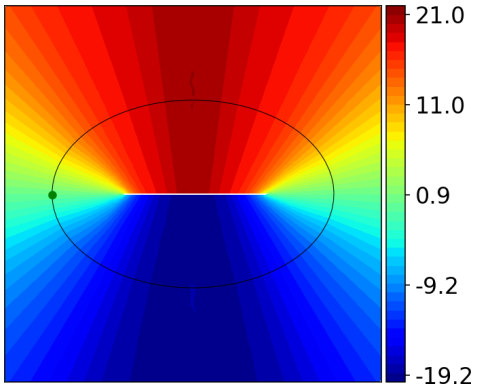
$$\forall x = (x_1, x_2) \in D, \;\; f(x_1, x_2) = x_2,$$

which belongs to the finite element space V_h, and is therefore amenable to an exact integration when the travel procedure along rays of section 4.3.2 is used.

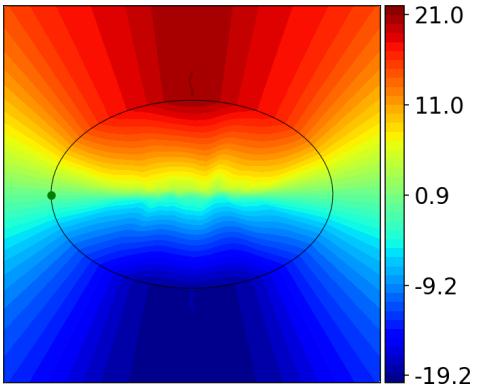
In this situation, the sought function u, given by (4.2.6), is known analytically; a calculation in polar coordinates indeed yields:

$$\forall (y_1, y_2) \in \partial\Omega, \;\; u(y_1, y_2) = \frac{8}{3} y_2. \tag{4.3.25}$$

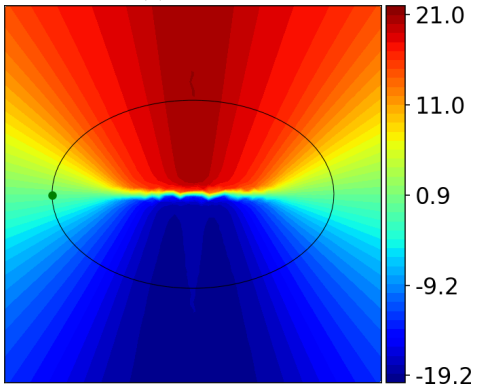
Comparisons are displayed on Figure 4.9 between the version of u obtained after direct integration along rays, and the numerical solutions of (4.2.5) obtained for various choices of weight functions ω on the mesh $\mathcal{T}$ of D (where Σ has not been removed). We observe a good match between the variational and the direct method. As expected, the solutions computed thanks to our variational method are less accurate when the constant weight $\omega = 1$ is chosen. A significant increase in accuracy is achieved by



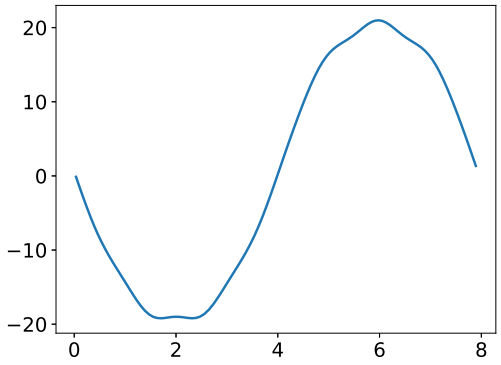
(a) Mesh $\mathcal{T}'$, $\omega = 1$.



(b) Mesh $\mathcal{T}$, $\omega = 1$



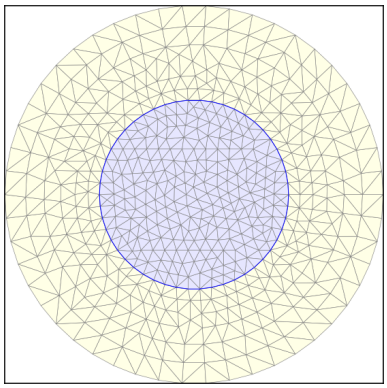
(c) Mesh $\mathcal{T}$, $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^2)$



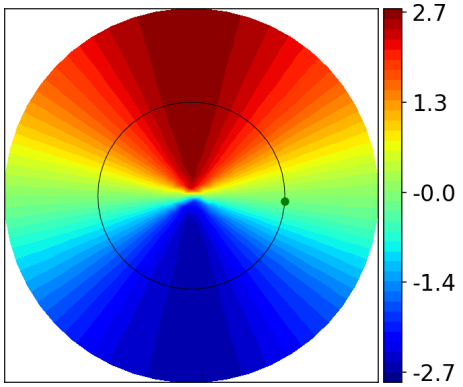
(d) Input function $v_0(y_1, y_2) = \cos(3y_1)^2 + 20y_2$ for the problem (4.3.21). The horizontal coordinate is the arc length with the green point as reference point.

Figure 4.7: Numerical resolution of the problem (4.3.20) using the variational problem (4.3.21) for various weights, with or without removing the skeleton Σ from the computational mesh.

selecting a weight ω vanishing at the center of the circle, i.e. $\omega = 1/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$. Note that this weight does not fulfill the condition $\omega \in L^1_{\omega^{-1}}(U)$ (see remark 4.12), but works well in numerical practice.

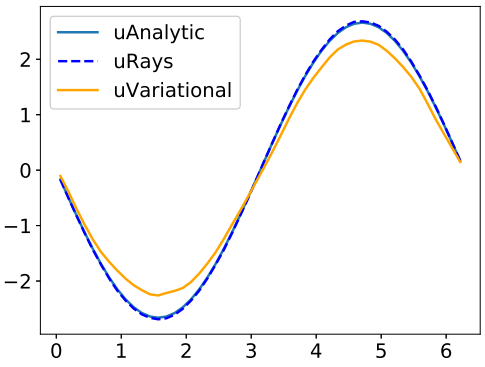


(a) The mesh $\mathcal{T}$ for D (part corresponding to Ω in blue)

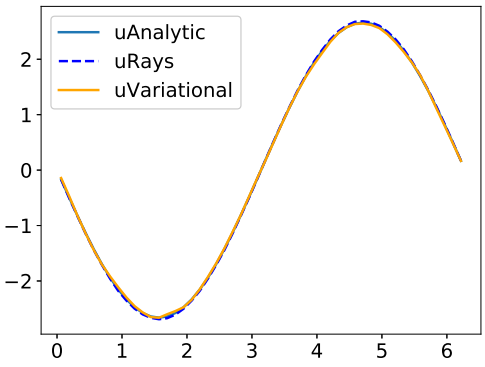


(b) The numerical solution u extended constantly along rays. The green point provides the 0 reference for the horizontal axes of Figure 4.9.

Figure 4.8: Setting of the disk test case of (4.3.24) in section 4.3.4.



(a) $\omega = 1$



(b) $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$

Figure 4.9: Comparison between direct integration along rays, analytical formula (4.1.3), and finite element solution of (4.2.5) for two choices of weights ω for the example of the test case of (4.3.24) in section 4.3.4. In the legends of this figure and all those to follow, `uAnalytic`, `uRays` and `uVariational` refer respectively to the analytical value of u, its numerical estimation using (4.2.38), and the trace of the variational solution of (4.2.5).

A $\mathcal{C}^2$ domain with non trivial skeleton

We take on the example (4.3.13) where Ω is an ellipse inside a square-shaped hold-all domain D, reusing both meshes $\mathcal{T}$ and $\mathcal{T}'$ depicted on Figure 4.6. To evaluate the influence of the skeleton on our numerical method, we compute the quantity (4.2.5) for a function f which is supported in Ω only (see Figure 4.10):

$$f(x_1, x_2) = (1 + x_2)\mathbf{1}_\Omega(x_1, x_2)\,. \tag{4.3.26}$$

The presence of the constant 1 in (4.3.26) avoids the "simplification" of having f vanishing on Σ. After an explicit calculation based on (4.2.6), the exact solution is given by:

$$\forall y = (y_1, y_2) \in \partial\Omega,\ \ u(y) = (1 + y_2)\zeta_- - \frac{\zeta_-^2}{2}\left(\frac{y_2}{\zeta_-} + \kappa(1 + y_2)\right) + \frac{y_2}{3}\kappa\zeta_-^2, \tag{4.3.27}$$

with $\kappa = \frac{ab}{\gamma^3}$, $\zeta_- = \frac{\gamma b}{a}$, $\gamma = \sqrt{\frac{b^2}{a^2}y_1^2 + \frac{a^2}{b^2}y_2^2}$.

The numerical trace u (which is extended along rays using (4.3.21) to ease the visualization) obtained by solving the variational formulation (4.2.5) on $\mathcal{T}$ or $\mathcal{T}'$, and for various choices of weights, is represented

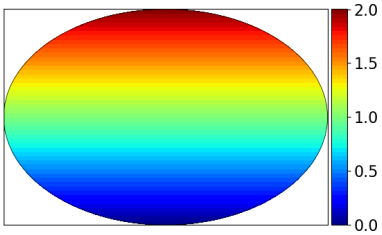
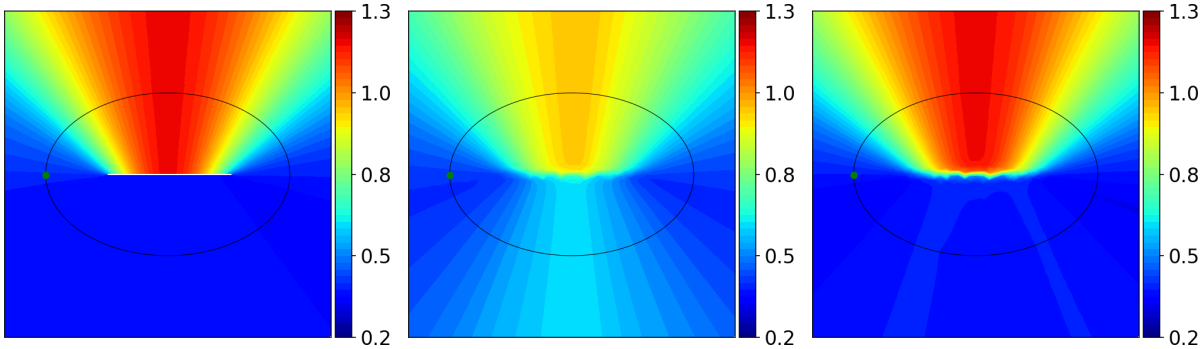


Figure 4.10: The function f for the identification problem (4.2.6) for the ellipse test case (4.3.13) of section 4.3.4.

on Figure 4.11. On Figure 4.12, we compare for each of these strategies the boundary values of the numerical trace u to that computed with (4.1.3) by direct integration along rays, and to the analytical expression (4.3.27).

We note that the solutions obtained by integrating along rays are generally characterized by small spurious oscillations. We explain this error by the fact that our criterion (4.3.9) detects the skeleton up to an error proportional to the mesh size (the actual location of the skeleton is not estimated within the last travelled triangle; see Figure 4.13). Such a procedure could be improved by using more accurate skeleton detection methods; see for instance the works [37, 45] in the field of computer graphics. We verify that the amplitude of these oscillations decreases with the size of the mesh, as shown on Figure 4.12d.



(a) Mesh $\mathcal{T}'$ (skeleton manually truncated), $\omega = 1$

(b) Mesh $\mathcal{T}$, $\omega = 1$.

(c) Mesh $\mathcal{T}$, $\omega = 2/(1+|d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$

Figure 4.11: The numerical solution u of the variational problem (4.2.5) extended constantly along rays (for visualization purpose). Inaccuracies occur if the skeleton Σ is not removed from the mesh or if the weight ω does not vanish in its vicinity.

A $\mathcal{C}^1$ domain, with discontinuous curvature

We now consider the case where Ω is a stadium, i.e. the reunion of a rectangle and two half-disks. Define (see Figure 4.14)

$$D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, |x_1| \leq 2 \text{ and } |x_2| < 2\} \tag{4.3.28}$$

and

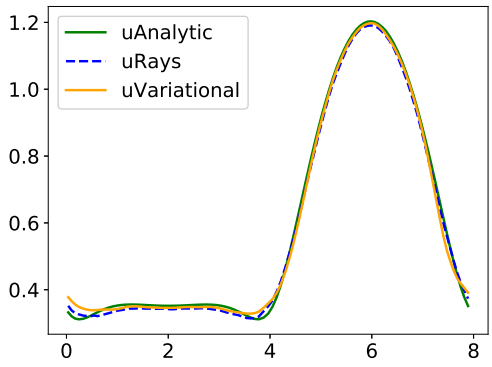
$$\Omega = \{(x_1, x_2) \in D \,|\, (|x_1| < 0.5 \text{ and } |x_2| < 0.5) \text{ or } (x_1 - 0.5)^2 + x_2^2 < 0.25 \text{ or } (x_1 + 0.5)^2 + x_2^2 < 0.25\}. \tag{4.3.29}$$

The domain Ω is not of class $\mathcal{C}^2$, and the curvature κ of the boundary $\partial\Omega$ is discontinuous at the points where $x_1 = \pm 0.5$. Hence, this example does not fall into the admissible setting of proposition 4.6, and there is, in principle, no guarantee that our variational method based on (4.2.5) should still work.

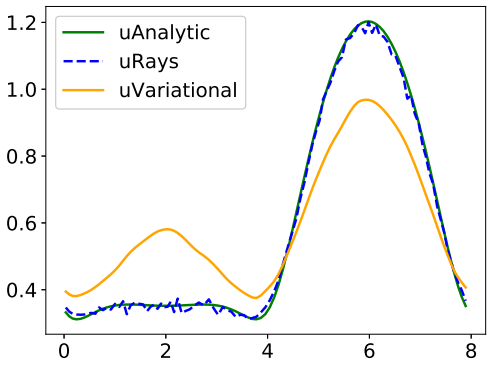
In this example, the skeleton Σ of Ω is explicitly given by $\Sigma = \{(x_1, 0) \,|\, |x_1| \leq 0.5\}$. We calculate the function u in (4.2.6) associated to the datum function f defined by (4.3.26); the analytical solution u for this problem is given by

$$u(y_1, y_2) = \begin{cases} \dfrac{1}{2} + \dfrac{1}{8}\text{sign}(y_1) \text{ if } |y_1| < 0.5, \\ \dfrac{1}{4} + \dfrac{y_2}{6} \text{ if } |y_1| > 0.5. \end{cases}$$

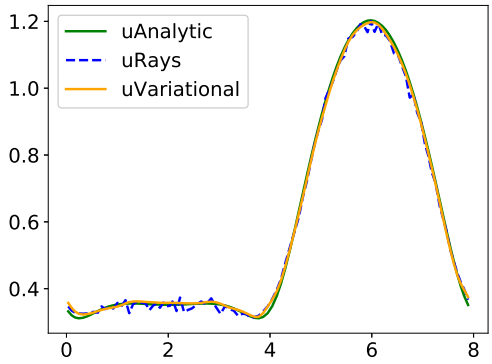
In particular, u is ill-defined at the points $x \in \partial\Omega$ where the curvature is discontinuous.



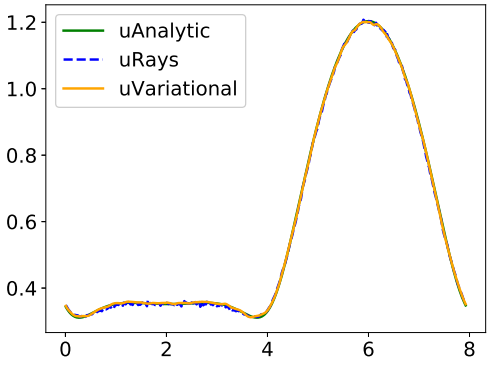
(a) Mesh $\mathcal{T}'$, $\omega = 1$



(b) Mesh $\mathcal{T}$, $\omega = 1$



(c) Mesh $\mathcal{T}$, $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$



(d) Mesh $\mathcal{T}$, $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$ (Fine mesh of Figure 4.5c)

Figure 4.12: Comparison between the results of a direct integration along rays and our variational method for the ellipse example of section 4.3.4.

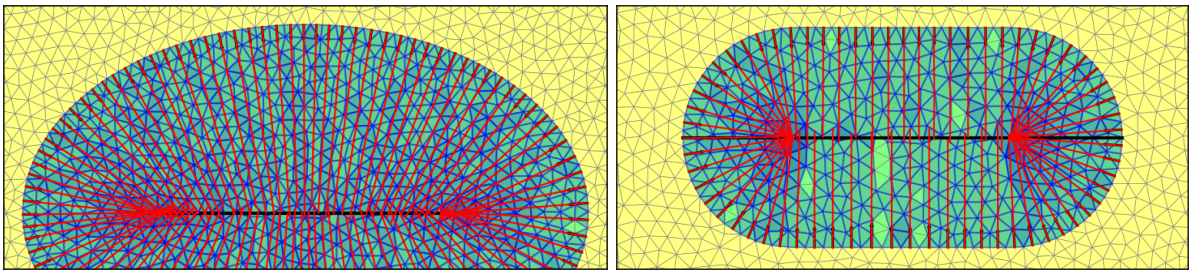
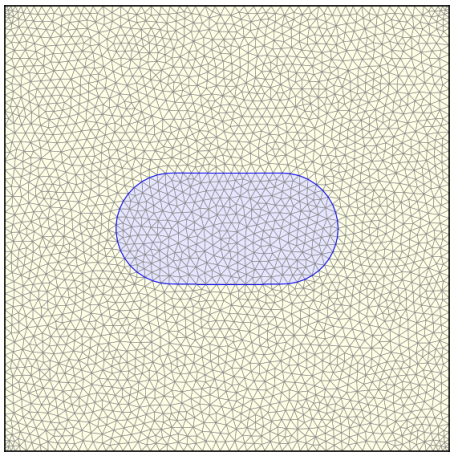
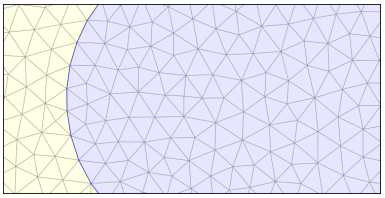


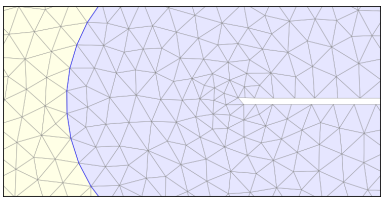
Figure 4.13: Numerical rays for the ellipse and stadium examples of section 4.3.4. The inaccuracy of the direct integration along rays is related to that of the skeleton detection within the last triangle. The triangle paths are colored in transparent blue (the darker, the more often visited); depicting that some mesh triangles might not be crossed by all the rays or some triangles may be visited more often than others.



(a) Mesh $\mathcal{T}$ of the bulk domain D with $\Omega \subset D$ (*in blue*) discretized as a submesh.



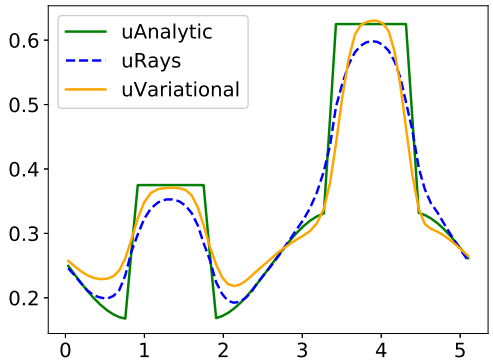
(b) Zoom on the mesh $\mathcal{T}$.



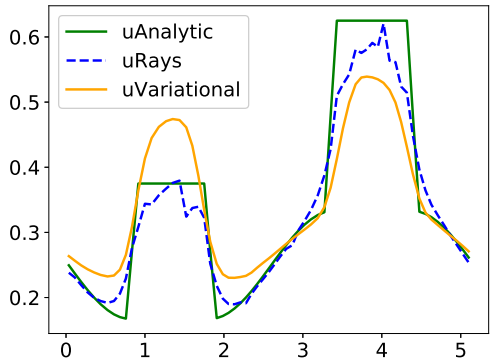
(c) Zoom on the mesh $\mathcal{T}'$ (skeleton Σ truncated)

Figure 4.14: Meshes used for the stadium example of section 4.3.4.

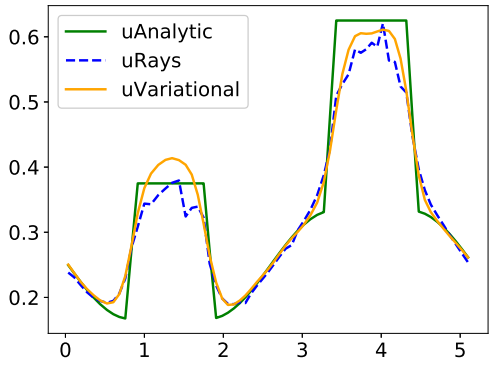
Both methods for the numerical evaluation of u (i.e. the direct integration along rays and our variational method) are applied, and the results are represented on Figure 4.15. Even though this example does not fit into the admissible setting of section 4.2, the results indicate that our numerical method still works, up to over-smoothing inaccuracies of the computed solution u near the discontinuities of the curvature κ.



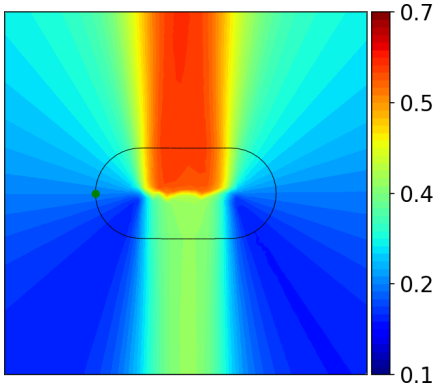
(a) Mesh $\mathcal{T}'$, $\omega = 1$



(b) Mesh $\mathcal{T}$, $\omega = 1$



(c) Mesh $\mathcal{T}$, $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$



(d) Variational solution u on the mesh $\mathcal{T}$, $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$ extended constantly along rays.

Figure 4.15: Comparison between the direct integration along rays and our variational method for the stadium example of section 4.3.4.

A test case involving a Lipschitz domain Ω with angular corners

In this last example, we consider the following situation:

$$D = \{(x_1, x_2) \in \mathbb{R}^2 \,|\, |x_1| \leq 2 \text{ and } |y| \leq 2\},\ \Omega = \{(x_1, x_2) \in D \,|\, \max(x_1^2/2 - x_2 - 1, x_2) \leq 0\}, \quad (4.3.30)$$

together with the datum function f:

$$f(x_1, x_2) = \cos(6x_1)^2,$$

as illustrated on Figure 4.16. Again, the theoretical framework of section 4.2 does not apply to the present situation because:

- The vector field ∇d_Ω is not of class $\mathcal{C}^1$ on U and is not U-filling: the reunion of all the rays emerging from points $y \in \partial\Omega$ does not cover the whole set $U = D\backslash\overline{\Sigma}$;
- The flow η of this vector field is not of class $\mathcal{C}^1$, because the curvature κ of the boundary (and even the normal vector) is discontinuous at the corners. Numerically, $|\Delta d_\Omega|$ blows up in the whole area filled by the skeleton and the set of points that are not covered by normal rays.

Still, the quantity u given by (4.2.6) can be calculated analytically wherever it makes sense. A few elementary, albeit technical computations yield:

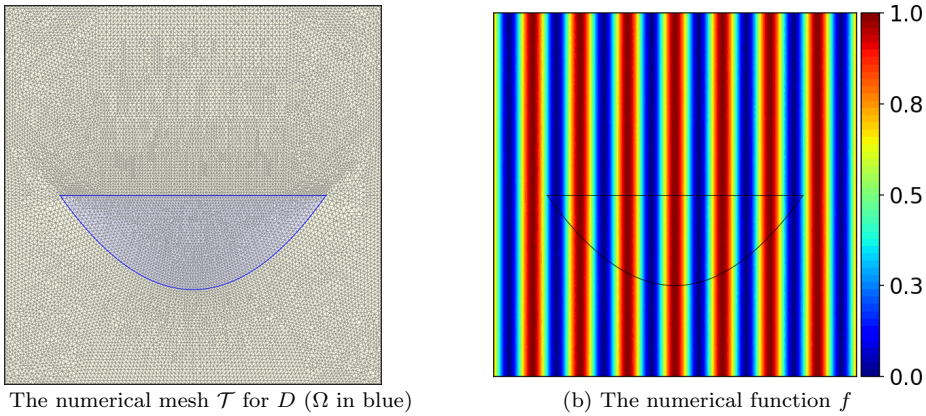
$$u(y_1, y_2) = \begin{cases} \cos(6y_1)^2(2 - \lambda_-(s^{-1}(y_1))) \text{ if } y_2 \geq 0, \\ \phi(\lambda_+(y_1)) - \phi(\lambda_-(\boldsymbol{y}_1)) \text{ if } y_2 \leq 0 \end{cases} \quad (4.3.31)$$

where

$$\lambda_-(t) = \frac{t^2/2 - 1}{1 + 1/\sqrt{1+t^2}},\ \lambda_+(t) = \min((2/|t| - 1), t^2/2)\sqrt{1+t^2}, \quad (4.3.32)$$

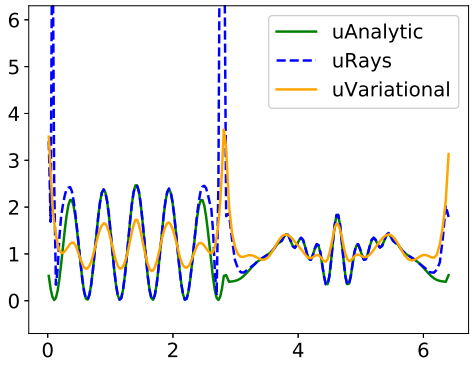
s^{-1} is the reciprocal function of $t \mapsto s(t) = \left(1 + \frac{\lambda_-(t)}{\sqrt{1+t^2}}\right) t$ and ϕ is a primitive function of $\lambda \mapsto \cos(6t(1 + \lambda/\sqrt{1+t^2}))^2(1 + \lambda/(1+t^2)^{3/2})$.

A numerical approximation to this function u is computed by using either a direct integration along rays or our variational method based on (4.2.5), on a single mesh $\mathcal{T}$ where the skeleton Σ is not removed, and using three possible choices of the weight ω. The results are represented on Figure 4.17. In all cases, the observed numerical inaccuracies, characterized by very high values, are concentrated near the angular corners of Ω, while the method yields satisfying accuracy on the remaining smooth parts of $\partial\Omega$. Again, we observe the poor accuracy associated to the choice of a constant weight $\omega = 1$. For this example, a uniformly small weight $\omega = 1e^{-10}$ seems to yield a better accuracy than the previous choice $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$.

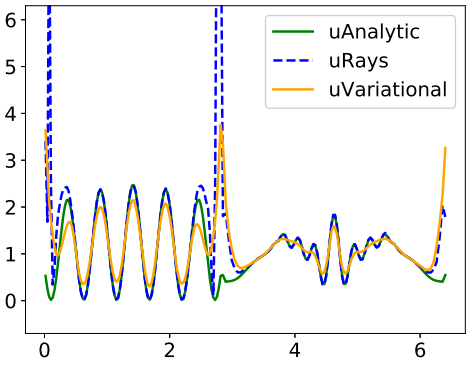


(a) The numerical mesh $\mathcal{T}$ for D (Ω in blue) (b) The numerical function f

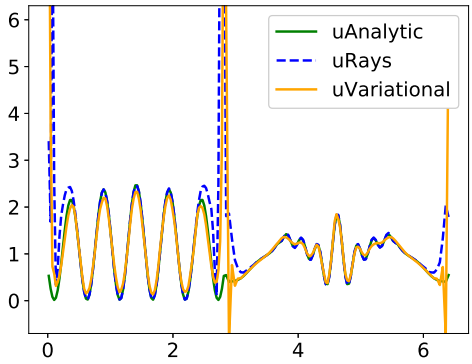
Figure 4.16: Setting for the non smooth example of section 4.3.4.



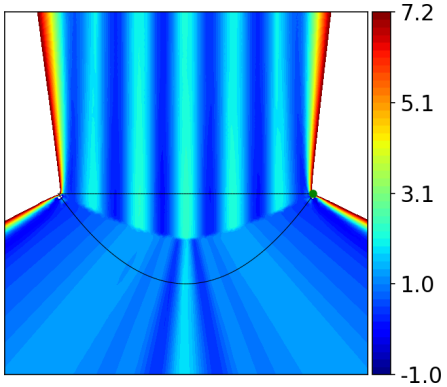
(a) $\omega = 1$



(b) $\omega = 2/(1 + |d_\Omega|^{3.5}|\Delta d_\Omega|^{3.5})$



(c) $\omega = 1e - 10$



(d) Numerical solution u obtained with $\omega = 1e - 10$ (extended along rays with (4.3.21))

Figure 4.17: Comparison between direct integration along normal rays and our variational method for the non smooth example of section 4.3.4.

4.4 APPLICATIONS TO MAXIMUM AND MINIMUM THICKNESS CONSTRAINTS IN SHAPE OPTIMIZATION

We now show how the general method proposed in section 4.2 and the inferred numerical methods in section 4.3 allow to efficiently implement geometric constraints, namely maximum and minimum thickness constraints in shape optimization, as the original motivation for our work. Here no comparison is made between our new variational approach and the previous method (using direct integration along rays) to evaluate the shape derivatives of the related shape functionals. The optimized shapes and topologies resulting from the variational method are very similar to those obtained in the previous works [30, 234]. There is no clear gain in computational time but there is a very substantial simplification of the implementation (which would be tremendous in 3-d).

4.4.1 Shape optimization setting for linearly elastic structures

In this section, we consider situations where only the mechanical displacement $\boldsymbol{u}$ of a structure comes into play. For compatibility with the notation of the present chapter, the domain to be optimized associated to the solid structure is denoted $\Omega \subset D$ (and not Ω_s as in chapter 2). Any such shape Ω is clamped on a part Γ_D of its boundary, and traction loads $\boldsymbol{g} \in L^2(\Gamma_N)$ are applied on a disjoint region of $\partial\Omega$; the complement $\Gamma := \partial\Omega \backslash (\overline{\Gamma_D} \cup \overline{\Gamma_N})$ is traction-free and body forces are omitted for simplicity; Γ is the only region of the boundary $\partial\Omega$ which is subject to optimization.

In this situation, the displacement $\boldsymbol{u}_\Omega$ of the shape is characterized as the unique solution in $H^1(\Omega, \mathbb{R}^d)$ to the linearized elasticity system:

$$\begin{cases} -\mathrm{div}(Ae(\boldsymbol{u}_\Omega)) = 0 \text{ in } \Omega, \\ Ae(\boldsymbol{u}_\Omega) \cdot \boldsymbol{n} = \boldsymbol{g} \text{ on } \Gamma_N, \\ Ae(\boldsymbol{u}_\Omega) \cdot \boldsymbol{n} = 0 \text{ on } \Gamma, \\ \boldsymbol{u}_\Omega = 0 \text{ on } \Gamma_D, \end{cases} \tag{4.4.1}$$

where $e(\boldsymbol{u}) := \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)$ is the strain tensor associated to the displacement $\boldsymbol{u}$ and A is the Hooke's law, defined for any symmetric $n \times n$ matrix by $Ae(\boldsymbol{u}) = 2\mu e(\boldsymbol{u}) + \lambda \mathrm{Tr}(e(\boldsymbol{u}))I$ involving the Lamé coefficients λ, μ which characterize the physical properties of the constituent material.

In this context, we consider structural optimization problems of the form

$$\min_{\Omega \subset D} J(\Omega), \quad \text{s.t. } P(\Omega) \leq 0, \tag{4.4.2}$$

where $J(\Omega)$ is a performance criterion, which will typically be the volume $\mathrm{Vol}(\Omega)$ or the compliance $C(\Omega)$ of shapes:

$$\mathrm{Vol}(\Omega) = \int_\Omega \mathrm{d}x, \quad C(\Omega) = \int_\Omega Ae(\boldsymbol{u}_\Omega) : e(\boldsymbol{u}_\Omega)\, \mathrm{d}x, \tag{4.4.3}$$

and $P(\Omega)$ is a geometric constraint

$$P(\Omega) = \int_D j(d_\Omega)\, \mathrm{d}x, \tag{4.4.4}$$

involving the signed distance function d_Ω to Ω and a given smooth function $j : \mathbb{R} \to \mathbb{R}$.

The shapes derivatives of the volume and compliance functions are classically given by (see chapter 2):

$$\mathrm{DVol}(\Omega)(\boldsymbol{\theta}) = \int_\Gamma \boldsymbol{\theta} \cdot \boldsymbol{n} \mathrm{d}y, \text{ and } \mathrm{D}C(\Omega)(\boldsymbol{\theta}) = -\int_\Gamma Ae(\boldsymbol{u}_\Omega) : e(\boldsymbol{u}_\Omega)(\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}y. \tag{4.4.5}$$

where we recall that $\mathrm{d}y$ stands for the surface measure on Γ. The shape derivative of the geometric constraint $P(\Omega)$ in (4.4.4) is given by (see the reminders of chapter 1, section 1.3.2)

$$\mathrm{D}P(\Omega)(\boldsymbol{\theta}) = \int_{D \backslash \overline{\Sigma}} j'(d_\Omega(x)) d'_\Omega(\boldsymbol{\theta})(x) \mathrm{d}x = \int_\Gamma u\, \boldsymbol{\theta} \cdot \boldsymbol{n} \mathrm{d}y. \tag{4.4.6}$$

The "Eulerian" derivative $d'_\Omega(\boldsymbol{\theta})$ of the signed distance function d_Ω is defined on $U = D \backslash \overline{\Sigma}$ by:

$$\forall x \in D \backslash \overline{\Sigma}, \quad d'_\Omega(\boldsymbol{\theta})(x) = -\boldsymbol{\theta}(p_{\partial\Omega}(x)) \cdot \boldsymbol{n}(p_{\partial\Omega}(x)), \tag{4.4.7}$$

and the scalar function $u : \Gamma \to \mathbb{R}$ reads:

$$u(y) = -\int_{s\in \mathrm{ray}(y)} j'(d_\Omega(s)) \prod_{1\le i\le n-1} (1+\kappa_i(y) d_\Omega(s)) \mathrm{d}s,$$

as follows from an application of the coarea formula with the help of the material recalled in chapter 1, section 1.3.2.

Using the conclusions of section 4.2, the function u in (4.4.6) may be conveniently evaluated by solving the variational problem:

$$\text{Find } u \in V_\omega \text{ such that } \forall v \in V_\omega, \int_{\partial\Omega} uv\mathrm{d}y + \int_{D\setminus\overline{\Sigma}} \omega(\nabla d_\Omega\cdot\nabla u)(\nabla d_\Omega\cdot\nabla v)\mathrm{d}x = -\int_{D\setminus\overline{\Sigma}} j'(d_\Omega(x))v\mathrm{d}x, \tag{4.4.8}$$

for a suitable weight ω satisfying (H1) to (H3), as discussed in section 4.3. Note how easily (4.4.6) is retrieved by taking the test function $v = -d'_\Omega(\boldsymbol{\theta})$ in (4.4.8). Since $\boldsymbol{\theta}\cdot\boldsymbol{n} \in L^\infty(\partial\Omega)$, the derivative given by (4.4.7) is indeed an admissible test function $d'_\Omega(\boldsymbol{\theta}) \in V_\omega$, for it belongs to $L^\infty(D\setminus\overline{\Sigma})$ and is constant along normal rays.

In our numerical implementation, we set $\omega = 1/(1+100|d_\Omega \Delta d_\Omega|^{3.5})$ in order to solve (4.4.8) with $\mathbb{P}_1$ finite elements, the constant 100 being selected to increase the slope of the weight near the skeleton.

4.4.2 Shape optimization under a maximum thickness constraint

We first consider structural optimization problems featuring a maximum thickness constraint. Following the work in [30, 234], a shape $\Omega \subset D$ is said to have maximum thickness lower than $d_{\max} > 0$ provided:

$$\forall x \in \Omega,\ d_\Omega(x) \ge -d_{\max}/2. \tag{4.4.9}$$

Loosely speaking, this amounts to saying that the skeleton Σ of Ω lies at a distance of at most $d_{\max}$ from the boundary $\partial\Omega$. Following closely [30, 234], the pointwise constraint (4.4.9) is relaxed into a single integral constraint formulated in terms of the following penalty functional $P_{\mathrm{MaxT}}(\Omega)$:

$$P_{\mathrm{MaxT}}(\Omega) \le \frac{d_{\max}}{2}, \text{ where } P_{\mathrm{MaxT}}(\Omega) := \left(\frac{\displaystyle\int_\Omega h(d_\Omega) d_\Omega^2 \mathrm{d}x}{\displaystyle\int_\Omega h(d_\Omega)\mathrm{d}x}\right)^{1/2}, \tag{4.4.10}$$

and h is a regularized Heaviside function centered at $d_{\max}/2$:

$$\forall x \in \mathbb{R},\ h(x) = \frac{1}{2}\left(1+\tanh\left(\frac{x-d_{\max}/2}{\alpha_f d_{\max}/2}\right)\right). \tag{4.4.11}$$

The parameter α_f in (4.4.11) tunes the level of regularization; in our context, it is set to $\alpha_f = \frac{4h_{\max}}{5d_{\max}}$, where $h_{\max}$ is the maximum edge length of the computational mesh of D.

A simple calculation yields the shape derivative of $P_{\mathrm{MaxT}}(\Omega)$:

$$\begin{aligned}\mathrm{D}P_{\mathrm{MaxT}}(\Omega)(\boldsymbol{\theta}) = &-\frac{1}{2P_{\mathrm{MaxT}}(\Omega)} \frac{\displaystyle\int_\Omega h(d_\Omega) d_\Omega^2 \mathrm{d}x}{\left(\displaystyle\int_\Omega h(d_\Omega)\mathrm{d}x\right)^2} h(0) \int_\Gamma \boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}y \\ &+ \frac{1}{2P_{\mathrm{MaxT}}(\Omega)} \int_\Omega \left[\frac{h'(d_\Omega)d_\Omega^2 + 2h(d_\Omega)d_\Omega}{\displaystyle\int_\Omega h(d_\Omega)\mathrm{d}z} - \frac{\displaystyle\int_\Omega h(d_\Omega)d_\Omega^2\mathrm{d}z}{\left(\displaystyle\int_\Omega h(d_\Omega)\mathrm{d}z\right)^2} h'(d_\Omega)\right] d'_\Omega(\boldsymbol{\theta})\mathrm{d}x,\end{aligned} \tag{4.4.12}$$

where the term involving $d'_\Omega(\boldsymbol{\theta})$ is then computed using the variational formulation (4.4.8).

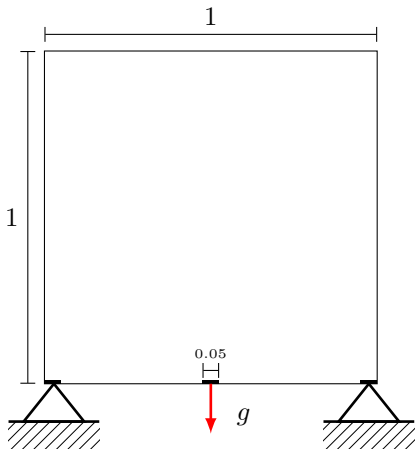
In the forthcoming examples of section 4.4.2, we solve the optimization problem of minimizing the volume $\mathrm{Vol}(\Omega)$ of the structure while imposing that the compliance $C(\Omega)$ do not exceed a given threshold $g_{\max}$, as well as the maximum thickness constraint (4.4.10), namely:

$$\begin{aligned}\min \quad & \mathrm{Vol}(\Omega) \\ \text{s.t.} \quad & \begin{cases} C(\Omega) \le g_{\max} \\ P_{\mathrm{MaxT}}(\Omega) \le \dfrac{d_{\max}}{2}. \end{cases}\end{aligned} \tag{4.4.13}$$

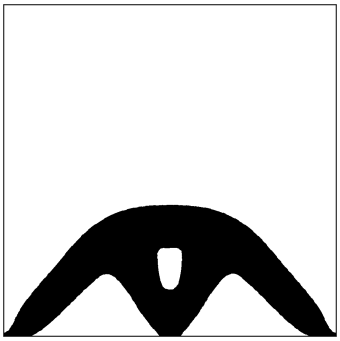
Optimization of the shape of a two-dimensional arch

Our first test-case reproduces that considered in §8.1.1 of [30], whose setting is displayed on Figure 4.18a: in a square-shaped domain D, the considered shapes are clamped on both their bottom left and bottom right corners, and a vertical load $\boldsymbol{g} = (0, -20)$ is applied at the middle of their bottom side.

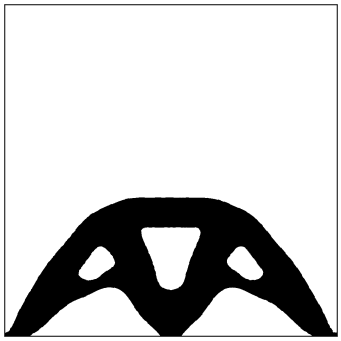
Starting from an initial shape arbitrarily perforated with several holes, the optimization problem (4.4.13) is solved with and without including the maximum thickness constraint, using the numerical values $g_{\max}$ =7.00 and $d_{\max}$ =0.12 when they are relevant.



(a) Setting of the optimization problem



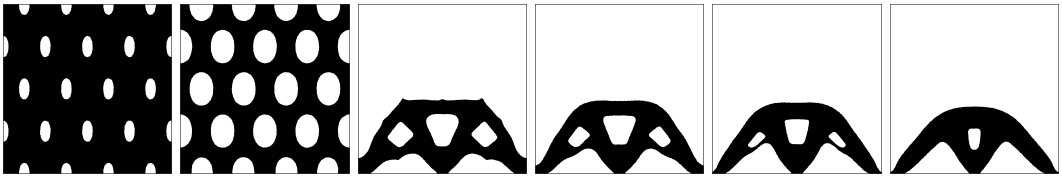
(b) Optimized shape without maximum thickness constraint ($\max|d_\Omega|$ =0.10).



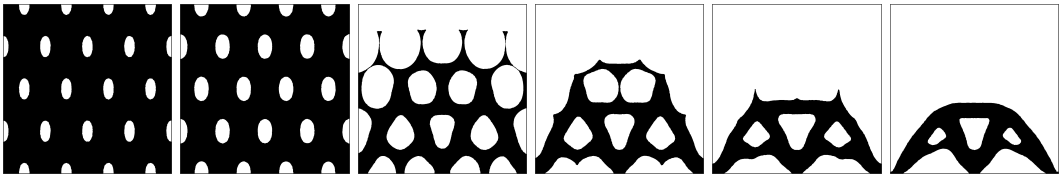
(c) Optimized shape with maximum thickness constraint ($\max|d_\Omega|$ =0.07).

Figure 4.18: Physical setting and optimized shapes obtained in the two-dimensional arch optimization test-case of section 4.4.2.

The resulting optimized shapes in both cases are displayed on Figs. 4.18b and 4.18c. Several intermediate shapes as well as the convergence histories of the computation are displayed on Figs. 4.19 to 4.21. The obtained shapes are quite analogous to those obtained by [30, 234] where the calculation of the shape derivative of $P_{\mathrm{MaxT}}(\Omega)$ relied on a direct numerical integration along rays.



(a) *Without* thickness constraint.



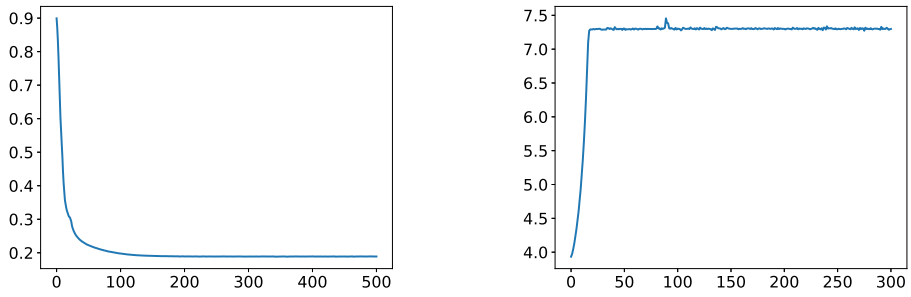
(b) *With* maximum thickness constraint.

Figure 4.19: Iterations 0, 5, 40, 65, 100 and 200 in the shape optimization example of a 2-d arch of section 4.4.2.

Optimization of a two-dimensional MBB-Beam

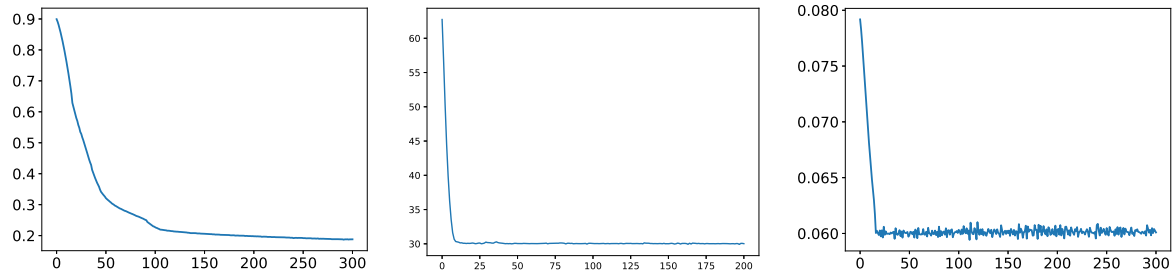
We consider now the classical MBB-Beam test-case depicted on Figure 4.22a: in a box D with dimensions 3×1, a material shape Ω is constrained to no horizontal motion on the left boundary, and to no vertical motion on the bottom right corner. A vertical load $\boldsymbol{g} = (0, -10)$ is applied on the top left corner, and the optimization problem (4.4.13) is considered again, with the numerical values for the thresholds $g_{\max}$ =30.00 and $d_{\max}$=0.16.

The resulting optimized shapes with and without including the maximum thickness constraint in (4.4.13) are represented on Figs. 4.22b and 4.22c and the convergence histories of the computation are



(a) Volume Vol(Ω) (b) Compliance $C(\Omega)$

Figure 4.20: Optimization histories for the optimization problem (4.4.13) *without* maximum thickness constraint. Final values: volume Vol(Ω) = 0.19, and compliance $C(\Omega)$ =7.00.



(a) Volume Vol(Ω) (b) Compliance $C(\Omega)$ (c) Thickness penalty $P_{\mathrm{MaxT}}(\Omega)$

Figure 4.21: Optimization histories for the optimization problem (4.4.13) *with* maximum thickness constraint. Final values: volume Vol(Ω) = 0.19, compliance $C(\Omega)$ =7.30, and penalty $P_{\mathrm{MaxT}}(\Omega)$ =0.06.

shown on Figs. 4.23 to 4.24.

4.4.3 Shape optimization examples under a minimum thickness constraint

We now turn to the implementation of a minimum thickness constraint thanks to our variational method. Following [30, 234], we say that a shape Ω has minimum thickness greater than $d_{\min}$ if

$$\forall y \in \partial\Omega,\ \zeta_-(y) < -d_{\min}/2. \tag{4.4.14}$$

In other words, the boundary $\partial\Omega$ is at a minimum distance $d_{\min}/2$ of the part $\Sigma \cap \Omega$ of the skeleton located inside the shape.

Enforcing a minimum thickness as a hard constraint (*i.e.* rather than a penalty term in the objective function) is by no means a straightforward task, because:

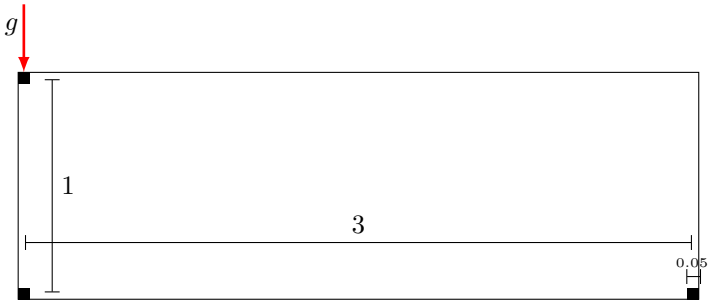
1. Our definition (4.4.14) of minimum thickness involve the distance to the skeleton ζ_-, which is not differentiable with respect to the shape,

2. It is not clear how to formulate (4.4.14) by mean of a penalty functional such as (4.4.10) to penalize localizations on the shape that do not meet the thickness requirement,

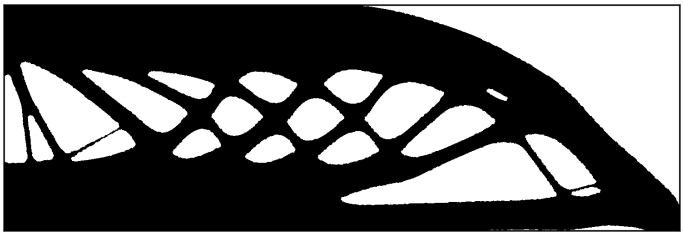
3. Even if we were able to enforce the constraint at each iteration, such would prevent topology changes to occur naturally, which would be an issue in numerical practice.

We propose in the following a more flexible setting to enforce such a requirement in a structural optimization problem. Elaborating on ideas proposed in [234, 88, 92, 222], the minimum thickness requirement is implemented in the objective function rather than in the constraints: we minimize a penalty functional $P_{\mathrm{MinT}}(\Omega)$ for the minimum thickness under constraints on the volume and compliance of shapes, i.e. we solve:

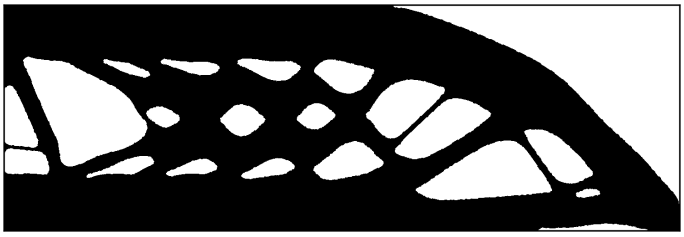
$$\begin{aligned} &\min \quad P_{\mathrm{MinT}}(\Omega) \\ &\text{s.t.} \begin{cases} C(\Omega) \le g_{\max} \\ \mathrm{Vol}(\Omega) \le V_{\max}. \end{cases} \end{aligned} \tag{4.4.15}$$



(a) Setting of the optimization problem. Small black areas correspond to non optimizable parts of the design domain.

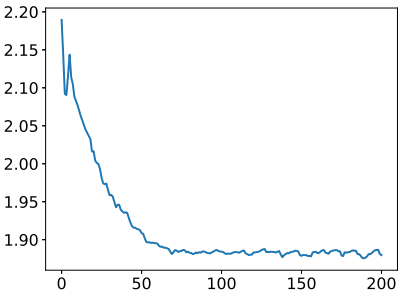


(b) Optimized shape without maximum thickness constraint ($\max d_\Omega$ =0.36).

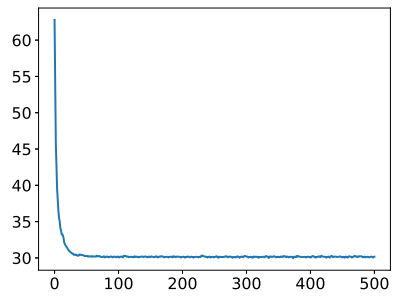


(c) Optimized shape with maximum thickness constraint.

Figure 4.22: Physical setting and obtained optimized shapes in the 2-d MBB beam optimization test-case of section 4.4.2 ($\max |d_\Omega|$ =0.26).

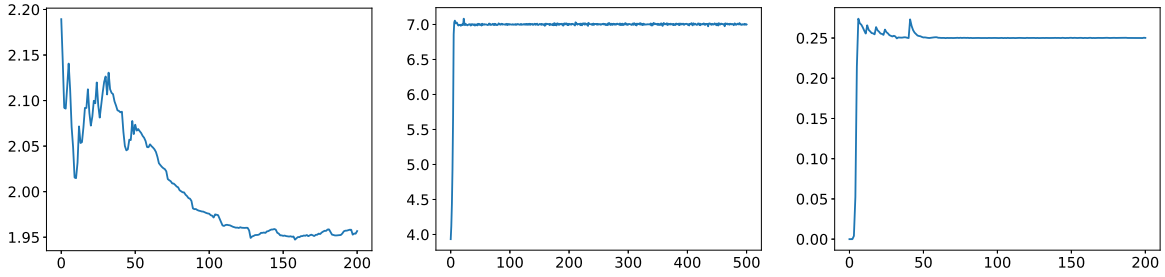


(a) Volume Vol(Ω)



(b) Compliance $C(\Omega)$

Figure 4.23: Optimization histories for the shape optimization problem (4.4.13) of the 2-d MBB-Beam of section 4.4.2 *without* including a maximum thickness constraint. Final values: volume Vol(Ω) =1.88, and compliance $C(\Omega)$ =29.96.



(a) Volume Vol(Ω) (b) Compliance $C(\Omega)$ (c) Thickness Penalty $P_{\mathrm{MaxT}}(\Omega)$

Figure 4.24: Optimization histories for the shape optimization problem (4.4.13) of the 2-d MBB-Beam of section 4.4.2 with maximum thickness constraint. Final values: Vol(Ω) =1.96, compliance: $C(\Omega)$ =30.03, and penalty $P_{\mathrm{MaxT}}(\Omega)$ =0.25.

This strategy is expected to work because our optimization algorithm is designed to satisfy violated constraints *first*, before then attempting to reduce the objective function while maintaining the constraints respected.

The penalty functional $P_{\mathrm{MinT}}(\Omega)$ we considered in our case when solving (4.4.15) is taken from [234] and is specially designed to have a zero derivative when the constraint (4.4.14) is satisfied:

$$P_{\mathrm{MinT}}(\Omega) = -\int_{\Omega} d_{\Omega}^2 \max(d_{\Omega} + d_{\min}/2, 0)^2 \mathrm{d}x. \tag{4.4.16}$$

The shape derivative of $P_{\mathrm{MinT}}(\Omega)$ is given by

$$\mathrm{D}P_{\mathrm{MinT}}(\Omega)(\boldsymbol{\theta}) = -\int_{\Omega} 2(d_{\Omega} \max(d_{\Omega} + d_{\min}/2, 0)^2 + d_{\Omega}^2 \max(d_{\Omega} + d_{\min}/2, 0)) d'_{\Omega}(\boldsymbol{\theta}) \mathrm{d}x. \tag{4.4.17}$$

Note that an increase in perimeter entails a decrease in the value of $P_{\mathrm{MinT}}(\Omega)$, but this behavior is tempered in our case by the volume constraint. Variants can be considered to address such an issue, and we refer to [234] for the details.

Optimization of the shape of a 2-d cantilever beam with minimum thickness constraint

We first consider the classical two-dimensional cantilever benchmark example, as depicted on Figure 4.25: in a box D with size 2×1, shapes Ω are clamped on their left-hand side, and a vertical load $\boldsymbol{g} = (0, -10)$ is applied on the middle of their right-hand side.

The optimization problem (4.4.15) is solved with the parameter values $g_{\max}$ =70.00 and $V_{\max}$ =0.80. The resulting optimized shapes are displayed on Figure 4.26 without including minimum thickness constraint (we minimize the volume Vol(Ω) subject to the compliance constraint as in (4.4.13)), and with the minimum thickness for two different values of $d_{\min}$. The corresponding convergence histories are shown on Figs. 4.27 to 4.29.

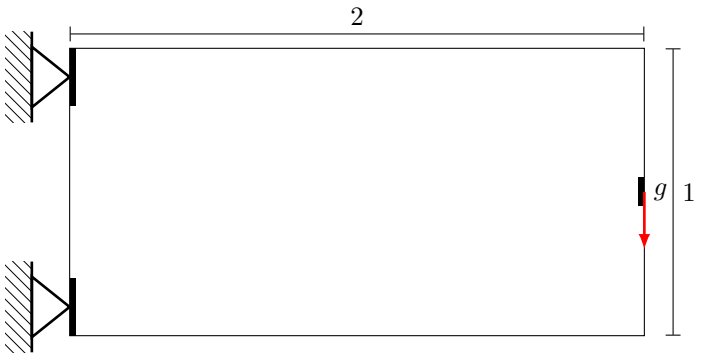
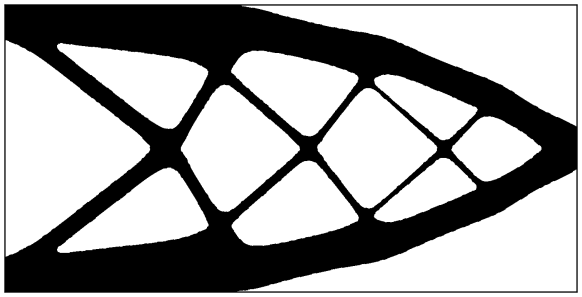
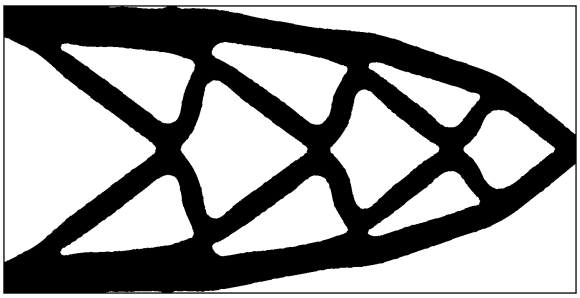


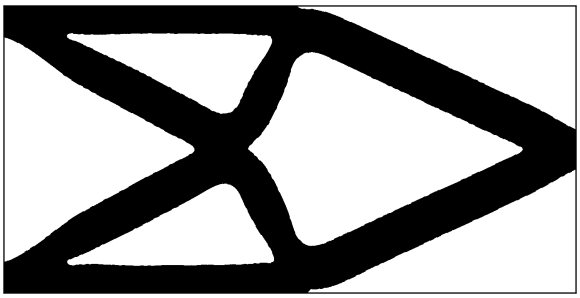
Figure 4.25: Setting of the cantilever test case of section 4.4.3.



(a) Optimized shape without including minimum thickness constraint.

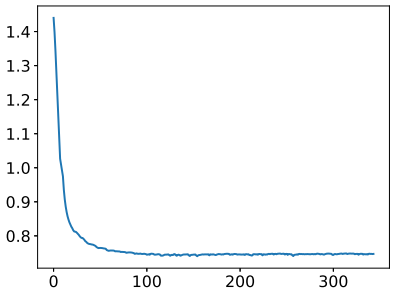


(b) Optimized shape with minimum thickness constraint ($d_{\min} = 0.1$).

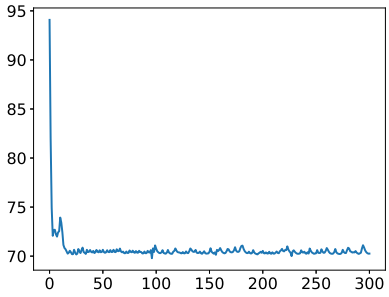


(c) Optimized shape with minimum thickness constraint ($d_{\min} = 0.2$).

Figure 4.26: Optimization of the shape of the 2-d cantilever of section 4.4.3 under minimum thickness constraint.

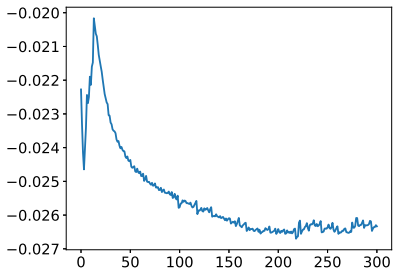


(a) Volume Vol(Ω)

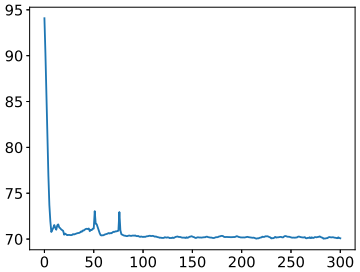


(b) Compliance $C(\Omega)$

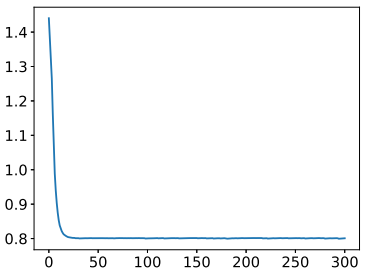
Figure 4.27: Optimization history for the 2-d cantilever optimization problem (4.4.13) of section 4.4.3 *without* minimum thickness constraint. Final values: volume Vol(Ω) =0.75, and compliance $C(\Omega)$ =70.29.



(a) Minimum thickness $P_{\mathrm{MinT}}(\Omega)$

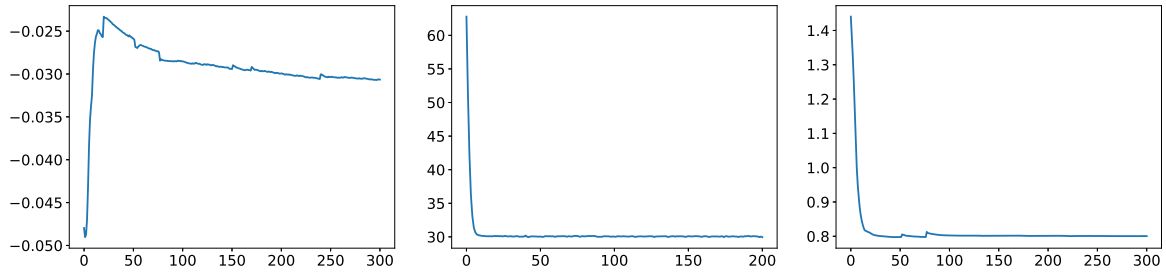


(b) Compliance $C(\Omega)$



(c) Volume Vol(Ω)

Figure 4.28: Optimization history for the 2-d cantilever optimization problem (4.4.13) of section 4.4.3 *with* minimum thickness constraint. ($d_{\min} = 0.1$) Final values : volume Vol(Ω) =0.80, compliance: $C(\Omega)$ =70.26.



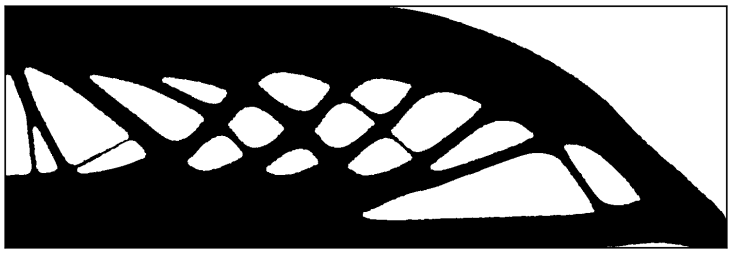
(a) Minimum thickness $P_{\mathrm{MinT}}(\Omega)$ (b) Compliance $C(\Omega)$ (c) Volume $\mathrm{Vol}(\Omega)$

Figure 4.29: Optimization history for the 2-d cantilever optimization problem (4.4.13) of section 4.4.3 *with* minimum thickness constraint. ($d_{\min} = 0.2$). Final values: volume $\mathrm{Vol}(\Omega)$ =0.80, and compliance: $C(\Omega)$ =70.09.

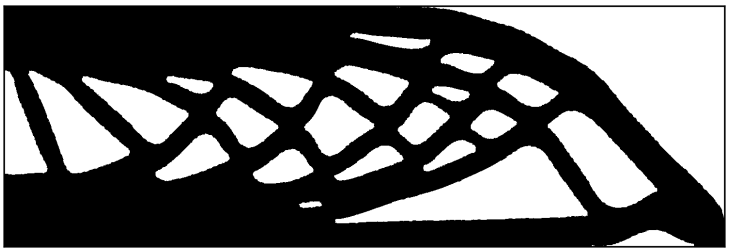
Shape optimization of a 2-d MBB Beam under a minimum thickness constraint

We now apply the same methodology on the MBB beam test-case of section 4.4.2. Optimized shapes are compared on Figure 4.30 without minimum thickness constraint (the result being that of Figure 4.22b), and for two values of $d_{\min}$. The corresponding convergence histories are shown on Figs. 4.23, 4.31 and 4.32. Finally, some intermediate shapes of the optimization process are reprinted on Figure 4.33.

One observes that for the last case of Figure 4.30c with $d_{\min} = 0.2$, the minimum thickness constraint is not satisfied everywhere but a substantial improvement is visible over the first design of Figure 4.30a. Notably, this approach is sufficiently flexible to guide the optimization path towards shapes with very different topologies.



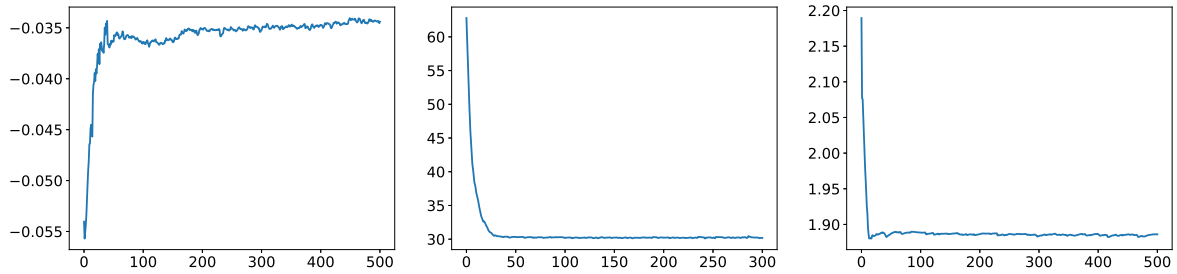
(a) Optimized shape without minimum thickness constraint.



(b) Optimized shape with minimum thickness constraint ($d_{\min} = 0.1$).

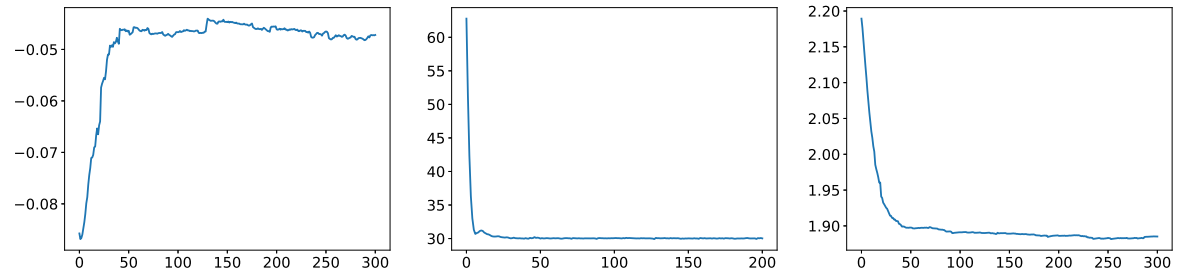
(c) Optimized shape with minimum thickness constraint ($d_{\min} = 0.2$).

Figure 4.30: Minimum thickness optimization for a cantilever test case of section 4.4.3.



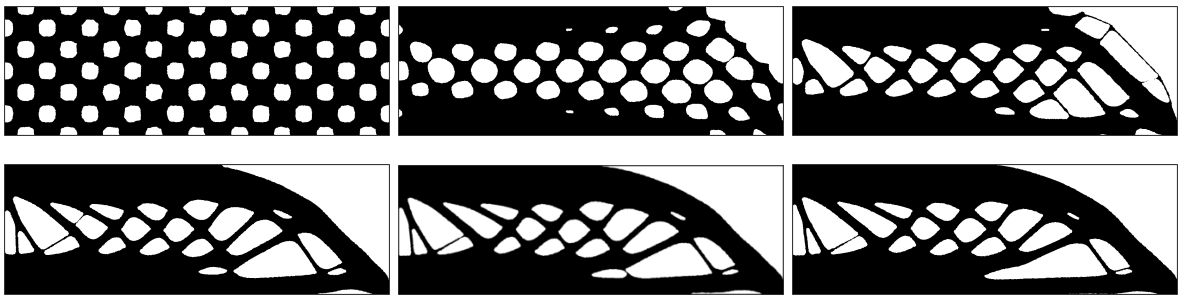
(a) Minimum thickness $P_{\mathrm{MinT}}(\Omega)$ (b) Compliance $C(\Omega)$ (c) Volume $\mathrm{Vol}(\Omega)$

Figure 4.31: Convergence histories for the MBB-Beam test-case of section 4.4.3 with minimum thickness constraint. ($d_{\min} = 0.1$). Final values: volume $\mathrm{Vol}(\Omega)$ =1.89, and compliance $C(\Omega)$ =30.17.

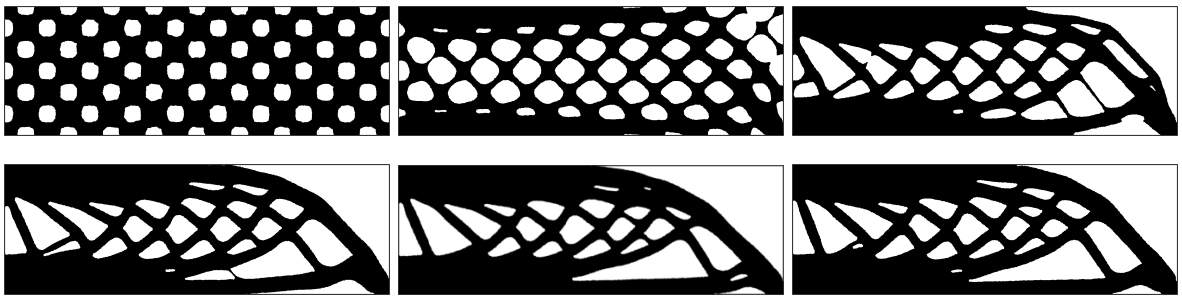


(a) Minimum thickness $P_{\mathrm{MinT}}(\Omega)$ (b) Compliance $C(\Omega)$ (c) Volume $\mathrm{Vol}(\Omega)$

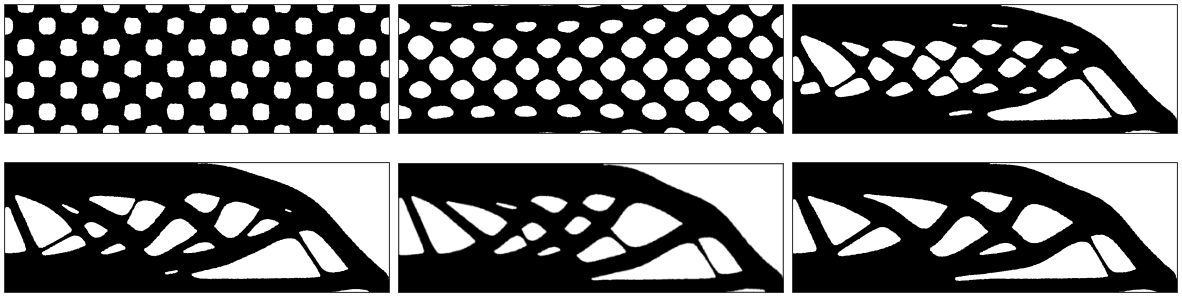
Figure 4.32: Convergence curves for the MBB-Beam test-case of section 4.4.3 with minimum thickness constraint ($d_{\min} = 0.2$). Final values: volume $\mathrm{Vol}(\Omega)$ =1.89, and compliance $C(\Omega)$ =30.18.



(a) No minimum thickness constraint (setting of Figure 4.22b).



(b) Minimum thickness constraint with $d_{\min} = 0.1$



(c) Minimum thickness constraint with $d_{\min} = 0.2$

Figure 4.33: Optimization histories for the MBB beam test-case with minimum thickness constraints of section 4.4.3: Iterations 0, 10, 40, 100, 150 and 200.

CHAPTER 5

TOPOLOGY OPTIMIZATION OF 2-D HEAT EXCHANGERS

Contents

Heat exchangers are a very topical issue in the topology optimization community with the publication of an increasing number of dedicated works [255, 258, 189, 271, 275, 205, 10, 104, 116, 225, 266, 303] This chapter reports on two independent case studies considering the shape optimization of two models of 2-d heat exchangers.

Section 5.1 considers a liquid-liquid heat exchanger featuring two fluid phases that must not interpenetrate. The cornerstone of the optimization problem is the treatment of the non-mixing condition: in our approach, it is imposed by using a minimum distance constraint between the two phases which lends itself to the variational method of chapter 4. This test case is inspired from a similar case study investigated by Papazoglou [255] with a very different (density based) method.

Section 5.2 then addresses the topology optimization of a different, air-oil heat exchanger. The aim is to determine the shape of the cross sections of oil pipes cooled down by an air flow. We rely on a different physical model, featuring a thermostatic boundary condition for the temperature on the optimized interface, which is very convenient to make the problem two dimensional. Since the optimization problem favors the apparition of very thin and elongated structures, a minimum thickness constraint is implemented to improve numerical convergence towards more manufacturable designs.

This part is an application of the material developed in the previous chapters: the formulas of chapter 2 for the shape derivatives of coupled thermal fluid problems, the null space algorithm of chapter 3 for optimization drives the resolution of the optimization problem, and we rely on the variational method of chapter 4 in order to enforce non penetration or minimum thickness constraints.

5.1 Design optimization of 2-d liquid-liquid heat exchangers with a non-mixing constraint

This section investigates the shape optimization of a 2-d heat exchanger featuring two liquid phases that must not interpenetrate. The physical setting considered is described in section 5.1.1. The mathematical formulation of the optimization problem, including the modeling of the non-penetration constraint is outlined in section 5.1.2. Details about the computation of associated shape derivatives are provided in section 5.1.3. Finally, numerical results are presented in section 5.1.4.

5.1.1 Physical setting

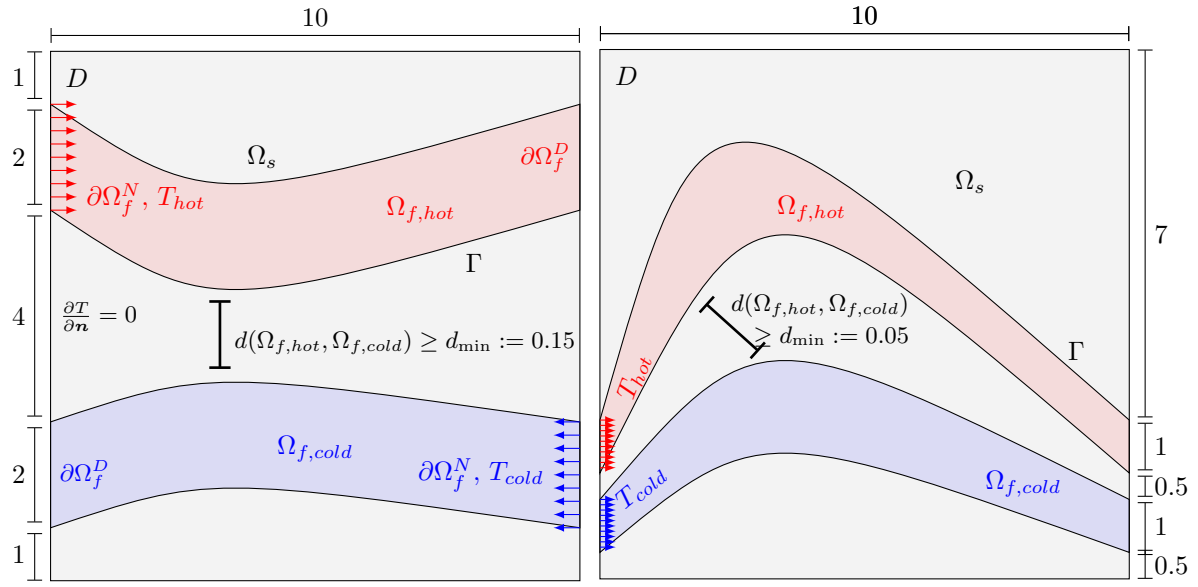
We consider a 'hold-all' domain $D = [0, 10] \times [0, 10] = \Omega_s \cup \Omega_f$ which is the disjoint union of a solid phase Ω_s and a fluid phase Ω_f. The fluid phase $\Omega_f = \Omega_{f,hot} \cup \Omega_{f,cold}$ is itself constituted of two distinct channels $\Omega_{f,hot}$ and $\Omega_{f,cold}$ whose shapes are to be optimized. The phases $\Omega_{f,hot}$ and $\Omega_{f,cold}$ enter the domain D with respective inlet temperatures $T_{hot} = 100$ and $T_{cold} = 0$. All other boundaries are adiabatic, corresponding to a zero Neumann condition $\partial T/\partial \boldsymbol{n} = 0$ for the temperature field T. Two

possible configurations are considered about the location of the inlets and described in the schematic of Figure 5.1 below:

- Test case 1 (Figure 5.1a): the two liquid phases enter D in opposite directions. The inlet and outlet cross sections share a common size $a = 2$.

- Test case 2 (Figure 5.1b): the two liquid phases enter from the same side of D with inlet and outlet cross sections having a smaller common size $a = 1$.

Following the notation convention of chapter 2, the reunion of the two inlets, of the two outlets, and of total fluid-solid interface to be optimized are denoted respectively by $\partial\Omega_f^D$, $\partial\Omega_f^N$ and Γ. These test cases are very similar to those considered by Papazoglou [255].



(a) Test case 1: A hot fluid phase $\Omega_{f,hot} \subset D$ is entering from the upper left side of D with a temperature T_{hot}, and a cold fluid phase $\Omega_{f,cold}$ is entering in the reverse direction from the lower right inlet.

(b) Test case 2: A hot fluid phase $\Omega_{f,hot} \subset D$ is entering from the upper left side of D with a temperature T_{hot}, and a cold fluid phase $\Omega_{f,cold}$ is entering in the same direction at the lower left inlet (boundary conditions not represented).

Figure 5.1: Settings of the two test cases considered in the optimal design of the heat exchangers of section 5.1 featuring the non-mixing condition $d(\Omega_{f,hot}, \Omega_{f,cold}) \geq d_{\min}$.

The physics involved in this problem are those described in chapter 2, section 3.6.3: the velocity of the fluid velocity and pressure $(\boldsymbol{v}, p)$ are characterized by the Navier-Stokes equations in the total fluid domain

$$\Omega_f := \Omega_{f,cold} \cup \Omega_{f,hot}$$

while the temperature field T in the whole domain D is determined by the equations of conduction-convection in both solid and liquid phases Ω_s and Ω_f:

$$\begin{cases} -\mathrm{div}(\sigma_f(\boldsymbol{v}, p)) + \rho \nabla \boldsymbol{v}\, \boldsymbol{v} = 0 & \text{in } \Omega_f \\ \mathrm{div}(\boldsymbol{v}) = 0 & \text{in } \Omega_f \\ \boldsymbol{v} = \boldsymbol{v}_0 & \text{on } \partial\Omega_f^D \\ \sigma_f(\boldsymbol{v}, p)\boldsymbol{n} = 0 & \text{on } \partial\Omega_f^N \\ \boldsymbol{v} = 0 & \text{on } \Gamma, \end{cases} \tag{5.1.1}$$

$$\begin{cases} -\mathrm{div}(k_f \nabla T_f) + \rho c_p \boldsymbol{v} \cdot \nabla T_f = 0 & \text{in } \Omega_f \\ -\mathrm{div}(k_s \nabla T_s) = 0 & \text{in } \Omega_s \\ T = 100 & \text{on } \partial\Omega_f^D \cap \partial\Omega_{f,hot} \\ T = 0 & \text{on } \partial\Omega_f^D \cap \partial\Omega_{f,cold} \\ -k_f \dfrac{\partial T_f}{\partial \boldsymbol{n}} = 0 & \text{on } \partial\Omega_T^N \cap \partial\Omega_f \\ -k_s \dfrac{\partial T_s}{\partial \boldsymbol{n}} = 0 & \text{on } \partial\Omega_T^N \cap \partial\Omega_s \\ T_f = T_s & \text{on } \Gamma \\ -k_f \dfrac{\partial T_f}{\partial \boldsymbol{n}} = -k_s \dfrac{\partial T_s}{\partial \boldsymbol{n}} & \text{on } \Gamma. \end{cases} \tag{5.1.2}$$

For each of the considered test cases, both fluid phases $\Omega_{f,cold}$ and $\Omega_{f,hot}$ are assumed to have the same density $\rho = 1$ and thermal conductivity $k_f = 10$. They enter the domain D from the inlet boundaries $\partial\Omega_f^D \subset \partial D$ with a parabolic velocity profile with maximum norm $||\boldsymbol{v}_0||_\infty = 1$, and they exit at outlet boundaries $\partial\Omega_f^N \subset \partial D$ with the vanishing normal stress condition $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = 0$, where we recall that the fluid stress tensor is defined by the Newton law

$$\sigma_f(\boldsymbol{v}, p) := 2\nu e(\boldsymbol{v}) - pI.$$

The viscosity ν is the same for both fluids; it is computed by the formula $\nu := \rho a ||\boldsymbol{v}_0||_\infty / \mathrm{Re}$ where $\mathrm{Re} = 60$ is the Reynolds number and a is the size of the inlet ($a = 2$ or $a = 1$ for the respective test cases 1 and 2). The capacity coefficient of the fluids is calculated by $c_p := k_f \mathrm{Pe}/(\nu \mathrm{Re})$ with the Péclet number given by $\mathrm{Pe} := 500$. The solid phase Ω_s is assumed to have a larger thermal conductivity coefficient $k_s = 110$ than the fluid.

5.1.2 Formulation of the shape optimization problem

The aim of the optimal design problem is to find the shape of both fluid phases $\Omega_f := \Omega_{f,cold} \cup \Omega_{f,hot}$ which maximizes the heat exchanged between the "hot" and "cold" phases under a maximal pressure drop constraint and a non-mixing condition. This exchanged heat is mathematically appraised by the opposite of an objective functional $J(\Omega_f, \boldsymbol{v}(\Omega_f), T(\Omega_f))$ which is to be minimized:

$$J(\Omega_f, \boldsymbol{v}(\Omega_f), T(\Omega_f)) := -\left(\int_{\Omega_{f,cold}} \rho c_p \boldsymbol{v} \cdot \nabla T \mathrm{d}x - \int_{\Omega_{f,hot}} \rho c_p \boldsymbol{v} \cdot \nabla T \mathrm{d}x \right).$$

This quantity can indeed be interpreted as the heat transferred because an integration by part (see also chapter 2, section 2.5.7) implies that

$$-J(\Omega_f, \boldsymbol{v}(\Omega_f), T(\Omega_f)) = \int_{\partial\Omega_{f,cold}} \rho c_p T\, \boldsymbol{v} \cdot \boldsymbol{n} \mathrm{d}y - \int_{\partial\Omega_{f,hot}} \rho c_p T\, \boldsymbol{v} \cdot \boldsymbol{n} \mathrm{d}y.$$

The above expression turns out to be exactly the heat exiting the cold phase minus the one leaving the hot phase (to be maximized), up to additional constant terms depending on inlet boundary values.

Following chapter 2, section 2.5.7, the pressure drop constraint reads

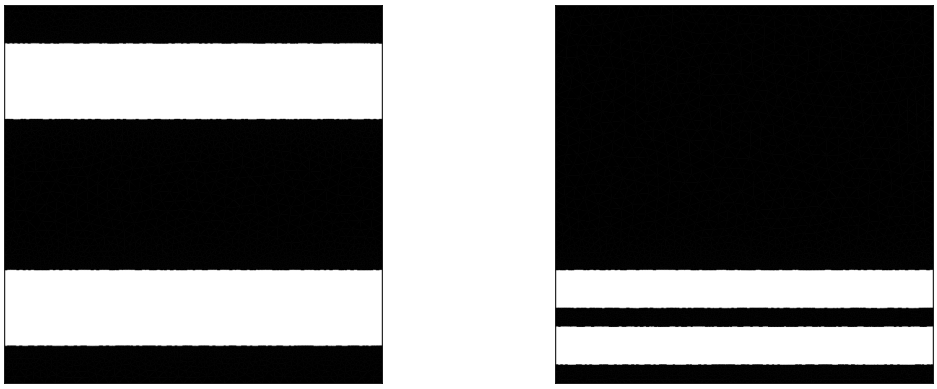
$$\mathtt{DP}(\Omega_f) := \int_{\partial\Omega_f^D} p \mathrm{d}y - \int_{\partial\Omega_f^N} p \mathrm{d}y \leq \mathtt{DP}_0$$

where $\mathtt{DP}_0$ is a given threshold value. In our implementation, this value is set according to the initial domain Ω_f^0:

$$\mathtt{DP}_0 := \begin{cases} 2\mathtt{DP}(\Omega_f^0) \text{ in test case 1,} \\ 5\mathtt{DP}(\Omega_f^0) \text{ in test case 2.} \end{cases}$$

The initial domain Ω_f^0 features two straight pipes which are displayed for each case 1 and 2 on Figure 5.2 below.

Let us now discuss the modeling of the non-mixing condition between both fluid phases. In his thesis, Papazoglou [255] proposed a multi-material model suited to the use of density based topology



(a) Test case 1 (b) Test case 2

Figure 5.2: Initial distribution of the fluid domain Ω_f^0 (in *white*).

optimization. In the context of the method of Hadamard for shape optimization, the "hot" and "cold" domains at stake are explicitly defined at every iteration which makes it possible to formulate the non-mixing constraint in a geometric fashion.

The key idea is to consider the *two* signed distance functions $d_{\Omega_{f,cold}}$ and $d_{\Omega_{f,hot}}$ associated with the two fluid phases $\Omega_{f,cold}$ and $\Omega_{f,hot}$. Following the introduction, this non-penetration condition is enforced by requiring the hot phase to remain at a minimum distance $d_{\min}$ from the cold phase, which can mathematically be formulated as:

$$\forall x \in \Omega_{f,cold},\ d_{\Omega_{f,hot}}(x) \geq d_{\min}, \tag{5.1.3}$$

or equivalently

$$\forall x \in \partial\Omega_{f,cold},\ d_{\Omega_{f,hot}}(x) \geq d_{\min}. \tag{5.1.4}$$

Note that the roles of $\Omega_{f,hot}$ and $\Omega_{f,cold}$ in (5.1.3) and (5.1.4) are interchangeable. For our applications, we set $d_{\min} = 0.15$ for the test case 1, and $d_{\min} = 0.05$ for the test case 2. Following the strategy of Allaire, Jouve and Michailidis in [30, 234], we consider averaged penalty functionals in order to approximate the pointwise constraint (5.1.4). Such can be obtained by reformulating (5.1.4) with the infinity norm:

$$\left|\left| \frac{1}{d_{\Omega_{f,hot}}} \right|\right|_{L^\infty(\partial\Omega_{f,cold})} \leq \frac{1}{d_{\min}}. \tag{5.1.5}$$

Classically, we consider an approximation of the infinity norm by a L^p norm, which yields an averaged penalty functional $P_{cold\to hot}(\Omega_f)$:

$$P_{cold\to hot}(\Omega_f) := \left(\int_{\partial\Omega_{f,cold}} \left(\frac{1}{d_{\Omega_{f,hot}}} \right)^p \mathrm{d}y \right)^{\frac{1}{p}} \simeq \left|\left| \frac{1}{d_{\Omega_{f,hot}}} \right|\right|_{L^\infty(\partial\Omega_{f,cold})}. \tag{5.1.6}$$

For our application, the parameter p involved is set to $p = 4$.

Remark 5.1. The formulation (5.1.4) involving $\partial\Omega_{f,hot}$ is preferred to that (5.1.3) set on the whole domain $\Omega_{f,hot}$, because we expect that averaging on a smaller set in (5.1.6) yields in some sense a more accurate approximation of the infinity norm.

Remark 5.2. This constraint on the distance between the two connected components $\Omega_{f,cold}$ and $\Omega_{f,hot}$ is substantially different to the requirement of a minimum thickness for the solid phase Ω_s. The use of the signed distance functions to $\Omega_{f,cold}$ (or to $\Omega_{f,hot}$) makes it possible to formulate an *actual* constraint, which is much more involved for a minimum thickness for Ω_s, in comparison with the approach proposed by [30] and in section 4.3.4.

In order to balance the effect of the constraint (5.1.6) over both fluid phases: we introduce the symmetrized version of (5.1.6):

$$P_{hot\to cold}(\Omega_f) := \left(\int_{\partial\Omega_{f,hot}} \left(\frac{1}{d_{\Omega_{f,cold}}} \right)^p \mathrm{d}y \right)^{\frac{1}{p}}. \tag{5.1.7}$$

It is then convenient to formulate the non-penetration constraint in terms of the harmonic mean of $P_{hot\to cold}$ and $P_{cold\to hot}$, which yields an averaged measure of the distance $d(\Omega_{f,cold},\Omega_{f,hot})$ between both phases; the non penetration constraint (5.1.5) is approximated by the following one:

$$Q_{hot\leftrightarrow cold}(\Omega_f) := \frac{2}{P_{hot\to cold}(\Omega_f)+P_{cold\to hot}(\Omega_f)} \geq d_{\min}. \tag{5.1.8}$$

All in all, the considered optimization problem reads:

$$\begin{aligned} \min_{\Omega_f\subset D} \quad & J(\Omega_f) = -\left(\int_{\Omega_{f,cold}} \rho c_p \boldsymbol{v}\cdot\nabla T \mathrm{d}x - \int_{\Omega_{f,hot}} \rho c_p \boldsymbol{v}\cdot\nabla T \mathrm{d}x\right) \\ s.t. \quad & \begin{cases} \mathtt{DP}(\Omega_f) = \displaystyle\int_{\partial\Omega_f^D} p\mathrm{d}s - \int_{\partial\Omega_f^N} p\mathrm{d}s \leq \mathtt{DP}_0 \\ Q_{hot\leftrightarrow cold}(\Omega_f) \geq d_{\min}. \end{cases} \end{aligned} \tag{5.1.9}$$

This optimization problem is solved with the null space gradient flow described in chapter 3 and the level set based mesh evolution algorithm of chapter 1, section 1.4.2. The shape derivatives of $J(\Omega_f)$ and $\mathrm{D}P(\Omega_f)$ are computed with the formulas given in volume form in chapter 2, proposition 2.3. The calculation of the shape derivative of the penalty functional $Q_{hot\leftrightarrow cold}(\Omega_f)$ is described in the next section.

5.1.3 Shape derivative of the non-mixing constraint $Q_{hot\leftrightarrow cold}(\Omega_f)$

The expression of the shape derivative of $Q_{hot\leftrightarrow cold}(\Omega_f)$ is easily obtained from those of $P_{cold\to hot}$ and $P_{hot\to cold}$. Since the latter have similar expressions, we content ourselves with an outline of the calculation of the shape derivative of $P_{cold\to hot}$. A straightforward computation yields

$$\begin{aligned} \mathrm{D}P_{cold\to hot}(\Omega_f)(\boldsymbol{\theta}) = \frac{1}{p}\left(\int_{\partial\Omega_{f,cold}} \frac{1}{|d_{\Omega_{f,hot}}|^p}\mathrm{d}y\right)^{\frac{1}{p}-1}\Bigg[\int_{\partial\Omega_{f,cold}}\left(\frac{\partial}{\partial\boldsymbol{n}}+\kappa\right)\left(\frac{1}{|d_{\Omega_{f,hot}}|^p}\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}y \\ - \int_{\partial\Omega_{f,cold}} \frac{p}{|d_{\Omega_{f,hot}}|^{p+1}} d'_{\Omega_{f,hot}}(\boldsymbol{\theta})\mathrm{d}y,\Bigg] \end{aligned} \tag{5.1.10}$$

where we recall that $d'_{\Omega_{f,hot}}(\boldsymbol{\theta})$ is the Eulerian derivative of the signed distance function $d_{\Omega_{f,hot}}$ to $\Omega_{f,hot}$ (see chapter 1, section 1.3.2). The first term of the above right-hand side has the usual structure of a boundary integral which can be easily numerically evaluated as a linear form in terms of $\boldsymbol{\theta}\cdot\boldsymbol{n}$. The last term is computed thanks to the variational method proposed in chapter 4. To be quite specific in this setting involving two distinct phases, we solve

Find $u_{hot}\in V_\omega$ such that $\forall v\in V_\omega$,

$$\int_{\partial\Omega_{f,hot}} u_{hot} v\mathrm{d}y + \int_{D\setminus\overline{\Sigma}_{hot}} \omega(\nabla u_{hot}\cdot\nabla d_{\Omega_{f,hot}})(\nabla v\cdot\nabla d_{\Omega_{f,hot}})\mathrm{d}x = \int_{\partial\Omega_{f,cold}} \frac{p}{|d_{\Omega_{f,hot}}|^{p+1}} v\mathrm{d}y, \tag{5.1.11}$$

where $\overline{\Sigma}_{hot}$ denotes the skeleton set of $\Omega_{f,hot}$ and the weight ω is given by

$$\omega = 1/(1+100|d_{\Omega_{f,hot}}\Delta d_{\Omega_{f,hot}}|^{3.5}.$$

The reader is referred to chapter 4, section 4.2 regarding the definition of the space V_ω, which is discretized with $\mathbb{P}_1$ finite elements. Finally, the last integral term in (5.1.10) is readily obtained from the formula

$$\int_{\partial\Omega_{f,cold}} \frac{p}{|d_{\Omega_{f,hot}}|^{p+1}} d'_{\Omega_{f,hot}}(\boldsymbol{\theta})\mathrm{d}y = -\int_{\partial\Omega_{f,hot}} u_{hot}\,\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}y,$$

where the right-hand side is a boundary integral involving $\boldsymbol{\theta}\cdot\boldsymbol{n}$ which can be easily discretized in our finite element setting.

5.1.4 Numerical results for two different configurations

Numerical results are depicted on Figs. 5.3 and 5.5 for the respective test cases 1 and 2. For each test case, we plot the initial and final designs, the temperature T and the velocity field $\boldsymbol{v}$, as well as some intermediate shapes obtained with our algorithm and the convergence histories for the objective and constraint functionals. We retrieve serpentine shapes as observed in [255].

This test case exhibits two striking features. At first, our approximation (5.1.6) of the L^∞ norm constraint by the averaged penalty functional $P_{cold\to hot}(\Omega_f)$ with a L^p norm ($p = 4$ in our case) works (surprisingly) very well: the 'ideal' pointwise constraint seems to be imposed at all intermediate iterations. On a different note, it is remarkable that our overall method is able to keep finding better shapes even after the saturation of the distant constraint, which happens very early in the optimization process. The reader will notice that some noise affect the convergence curves: we attribute it to inaccuracies related to the rather large Péclet and Reynolds number considered.

Finally, let us mention that we did not need a very fine mesh of the region of Ω_s in between the cold and the hot domains $\Omega_{f,cold}$ and $\Omega_{f,hot}$ to handle the distance constraint (5.1.8). We plot on Figs. 5.4 and 5.6 the final meshes of the optimized shape for the respective test cases 1 and 2. For the second test case, it is visible that a resolution of about ten mesh elements (which means a skeleton located at approximately *five* mesh elements only) in between the two pipes allows for a satisfying approximation of the shape derivative of the distance constraint from the resolution of the variational problem (5.1.11).

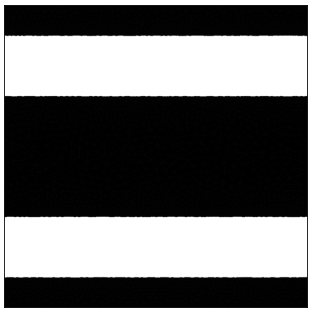
5.2 TOPOLOGY OPTIMIZATION OF A 2-D AIR-OIL HEAT EXCHANGER

We now report on a study issued from a collaboration with Safran Aero Boosters. The problem at hand is concerned with the optimization of the cross section of an air-oil heat exchanger. Although the original problem is three-dimensional, a few assumptions allow to reduce it to 2 dimensions. The approach followed in this part is different to that of the previous section 5.1: it is not necessary to prescribe a non-mixing constraint for the air and oil phases because we consider a different physical model in which there is no "solid" phase Ω_s; the air phase occupies a domain Ω_f to be optimized, and the oil phase is described by an isothermal boundary condition on $\partial\Omega_f$.

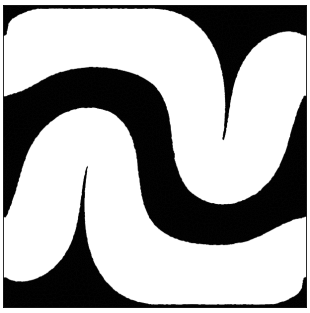
The general setting of the considered problem is described in section 5.2.1. Let us emphasize that the physical model considered in this setting is essentially the same as the fluid thermal model (5.1.1) and (5.1.2) of the previous section up to a change of boundary conditions on the optimized air-oil interface. The optimization problem is then mathematically formulated in section 5.2.2, where we define objective and constraint functions. A minimum thickness constraint is considered for the oil phase in order to improve the convergence towards manufacturable designs. Section 5.2.3 presents a variety of numerical results for different initializations and settings of the constraints. We mention the consideration of a different fluid model featuring a boundary condition on the stagnation pressure in section 5.2.4, which could be of direct interest for realistic industrial applications. The interesting point is the use of the rotational formulation of the Navier-Stokes equations [106]. However, our numerical results with this adaptation are not conclusive. We finally conclude in section 5.2.5 with a few general remarks about the extension of this approach to more realistic 3-d systems characterized by large Reynolds numbers.

5.2.1 Setting of the case study

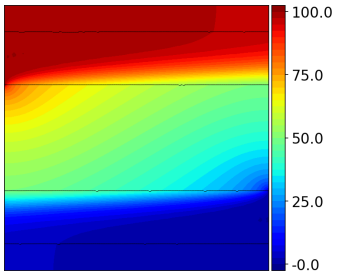
The objective of the study is the optimization of the shape and topology of a 2-d air-oil heat exchanger. The optimization domain D contains an air phase which occupies a subdomain $\Omega_f \subset D$ to be optimized. The complementary $D \setminus \Omega_f$ is filled up with oil. The setting of the problem is depicted on the schematic of Figure 5.7. After a 3-d extrusion in the z direction, the system is interpreted as a heat exchanger featuring air flowing in the x direction and cooling down oil channels flowing in the transverse z direction.



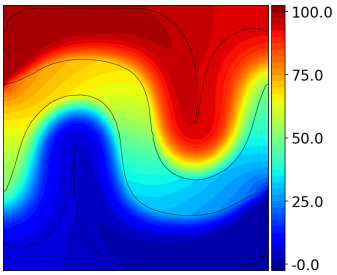
(a) Initial design



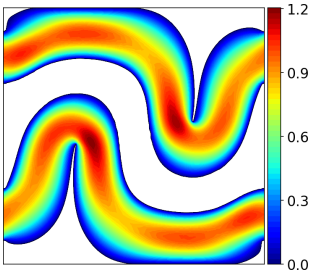
(b) Optimized design



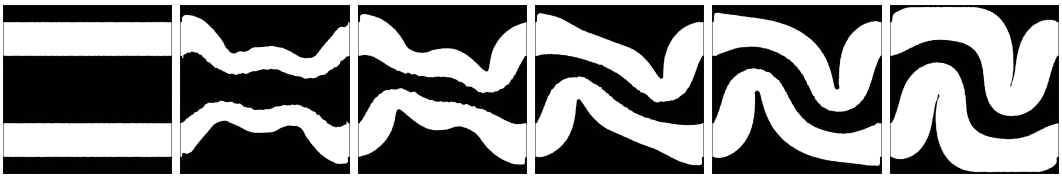
(c) Initial temperature field



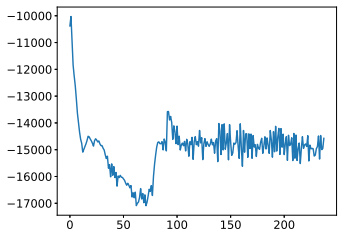
(d) Final temperature field.



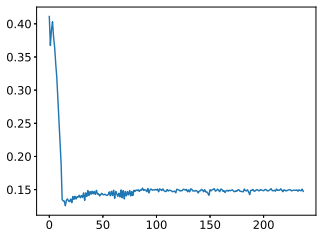
(e) Final norm fields for the fluid velocity.



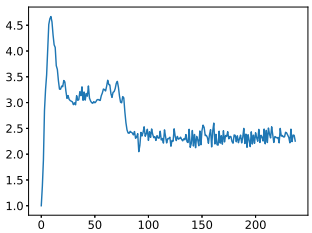
(f) Intermediate iterations 0, 10, 20, 30, 40 and 200



(g) Objective function $J(\Omega_f)$. Final value: $J(\Omega_f)$ =-1.5e+04.



(h) Penalty functional $Q_{cold\leftrightarrow hot}$ for the non-mixing constraint. Final value: $Q_{cold\leftrightarrow hot}(\Omega_f) = 0.15$.



(i) Pressure drop $\texttt{DP}(\Omega_f)/\texttt{DP}(\Omega_f^0)$. Final value: $\texttt{DP}(\Omega_f)/\texttt{DP}(\Omega_f^0) = 2.3$.

Figure 5.3: Optimization results for the test case 1 (cold and hot inlets entering D from opposite sides) of section 5.1.

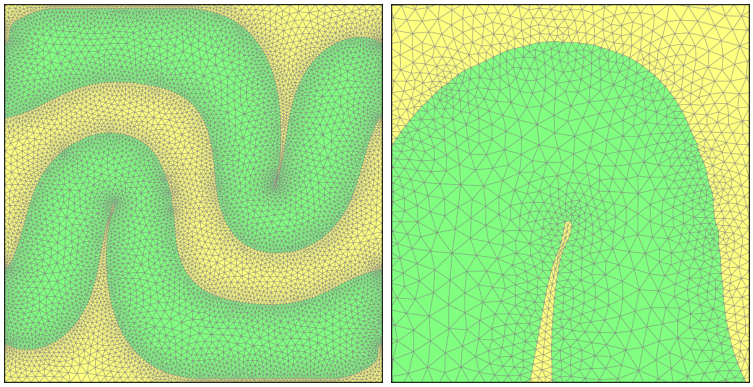
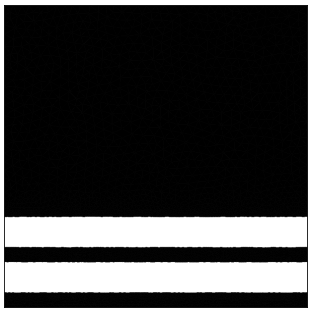
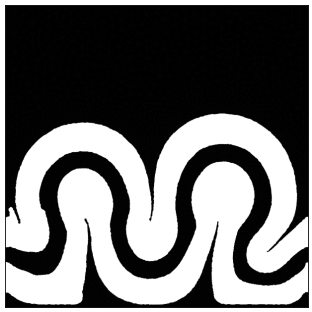


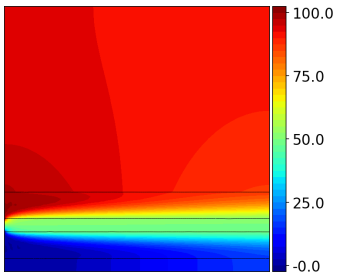
Figure 5.4: Zoom on a mesh of an intermediate optimization iteration.



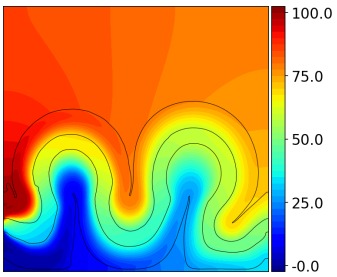
(a) Initial design



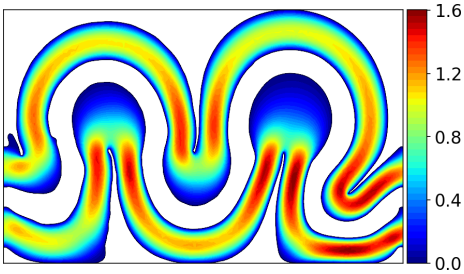
(b) Optimized design



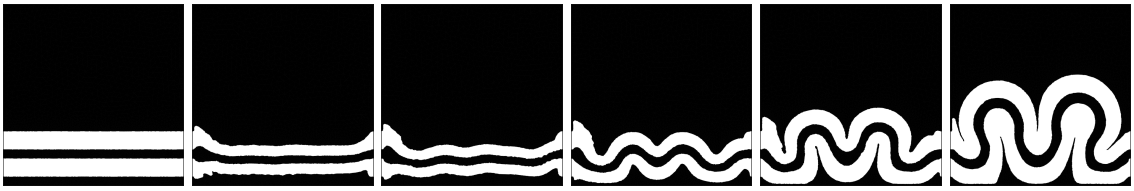
(c) Initial temperature field



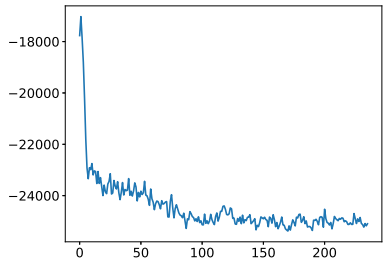
(d) Final temperature field.



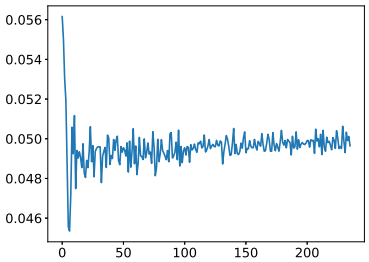
(e) Final norm fields for the fluid velocity.



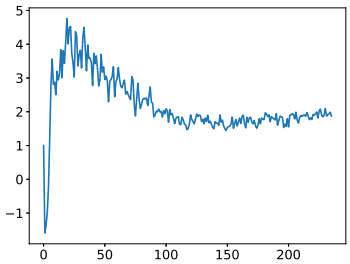
(f) Intermediate iterations 0, 8, 20, 50, 88 and 200



(g) Objective function $J(\Omega_f)$. Final value: $J(\Omega_f)$ =-2.5e+04.



(h) Penalty functional $Q_{cold \leftrightarrow hot}$ for the non-mixing constraint. Final value: $Q_{cold \leftrightarrow hot}(\Omega_f) = 0.05$.



(i) Pressure drop $\texttt{DP}(\Omega_f)/\texttt{DP}(\Omega_f^0)$. Final value: $\texttt{DP}(\Omega_f)/\texttt{DP}(\Omega_f^0) = 1.9$.

Figure 5.5: Optimization results for the test case 2 (cold and hot inlets entering D from the same side) of section 5.1.

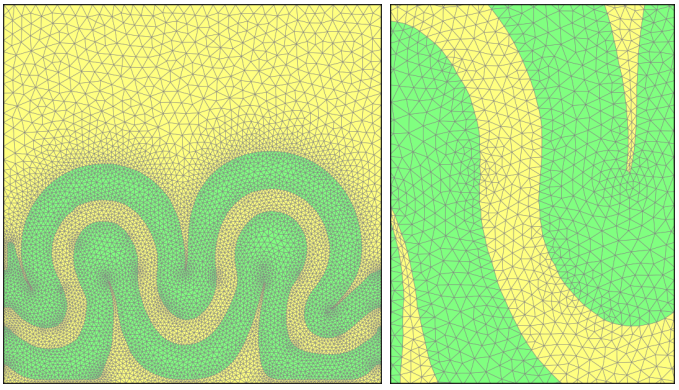


Figure 5.6: Zoom on a mesh of an intermediate optimization iteration.

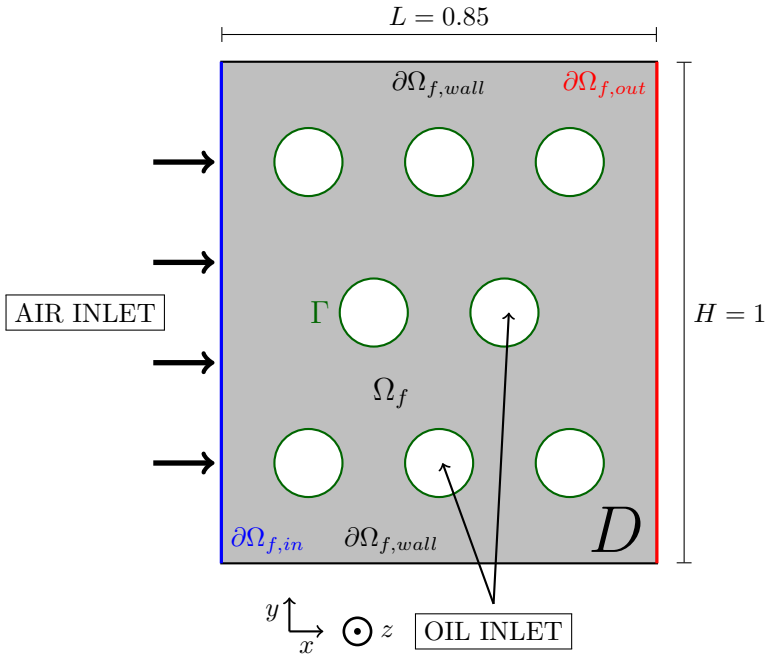


Figure 5.7: Setting of the air-oil heat exchanger of section 5.2.

The oil filling the phase $D\backslash\Omega_f$ has a much higher thermal conductivity than air so that its temperature T can be considered to be uniform: $T = T_{oil}$ in $D\backslash\Omega_f$. Assuming in addition the interface Γ to be made of a solid phase that is assumed to be a sufficiently thin and conductive material, it is sufficient to describe the system by the single physics of the air phase with a thermostatic temperature condition $T = T_{oil}$ on the interface Γ.

The objective of the study is to demonstrate the feasibility of topology optimization with the method of Hadamard for the maximization of the heat exchanged between the two phases, under a maximum constraint on the outlet pressure drop.

5.2.2 Physical modeling and formulation of optimization problems

The air phase is characterized by fluid velocity and pressure $(\boldsymbol{v}, p)$ solutions to the incompressible steady state Navier-Stokes equations:

$$\begin{cases} -\mathrm{div}(\sigma_f(\boldsymbol{v},p)) + \rho\nabla\boldsymbol{v}\,\boldsymbol{v} = 0 & \text{in } \Omega_f \\ \mathrm{div}(\boldsymbol{v}) = 0 & \text{in } \Omega_f \\ \boldsymbol{v} = \boldsymbol{v}_0 & \text{on } \partial\Omega_{f,in} \\ \sigma_f(\boldsymbol{v},p)\boldsymbol{n} = 0 & \text{on } \partial\Omega_{f,out} \\ \boldsymbol{v} = 0 & \text{on } \partial\Omega_{f,wall} \\ \boldsymbol{v} = 0 & \text{on } \Gamma, \end{cases} \tag{5.2.1}$$

where the fluid stress tensor is given by $\sigma_f(\boldsymbol{v},p) = 2\nu e(\boldsymbol{v}) - pI$ as in the previous section 5.1. The fluid velocity $\boldsymbol{v}$ then determines a temperature field T through the following convection diffusion equation:

$$\begin{cases} -\mathrm{div}(k_f\nabla T) + \rho c_p \boldsymbol{v}\cdot\nabla T = 0 & \text{in } \Omega_f \\ T = T_{in} & \text{on } \partial\Omega_{f,in} \\ -k_f\dfrac{\partial T}{\partial \boldsymbol{n}} = 0 & \text{on } \partial\Omega_f\backslash\partial\Omega_{f,in} \\ T = T_{oil} & \text{on } \Gamma. \end{cases} \tag{5.2.2}$$

This model differs from the one considered in the previous section (equations (5.1.1) and (5.1.2)) due to the fact that a thermostatic boundary condition $T = T_{oil}$ is imposed on the interface Γ rather than a diffusion model within the complementary subdomain $D\backslash\Omega_f$. Note that we assume, as in our previous studies, a zero normal temperature flux on the wall and outlet boundaries $\partial\Omega_{f,wall}$ and $\partial\Omega_{f,out}$. The

	Input velocity $\|\|\boldsymbol{v}_0\|\|$	Pressure loss threshold $\texttt{DP}_0$	Re	Pe
Configuration 1	10	1300	40	475
Configuration 2	25	1030	100	1200
Configuration 3	40	475	160	1900

Table 5.1: Configurations considered for the input velocity and pressure constraint values.

fluid is entering the left boundary $\partial\Omega_{f,in}$ with a parabolic velocity profile $\boldsymbol{v}_0$ with maximum norm value $||\boldsymbol{v}_0||_\infty$. The "cold" air flow Ω_f is entering with a temperature $T_{in} := 310$. The temperature of the "hot" oil phase $D\backslash\Omega_f$ flowing in the opposite direction is constant and it is given by $T_{oil} = 400$. The capacity and density coefficients of the fluid are set to $\rho = 1$, $c_p = 1$. Several values are considered for the viscosity ν, the conductivity coefficient k_f and for the intensity $||\boldsymbol{v}_0||_\infty$, which are discussed below.

Definition of objective and constraint functions The goal is to maximize the exchanged heat between the air phase and the oil phase while imposing an upper bound on the static pressure drop:

$$\begin{aligned} \min_{\Omega_f \subset D} \quad & J(\Omega_f) := -\int_{\Omega_f} \rho c_p \boldsymbol{v}\cdot\nabla T \mathrm{d}x \\ s.t. \quad & \texttt{DP}(\Omega_f) := \int_{\partial\Omega_{f,in}} p\mathrm{d}s - \int_{\partial\Omega_{f,out}} p\mathrm{d}s \leq \texttt{DP}_0. \end{aligned} \tag{5.2.3}$$

Upon integration by parts, the objective function $J(\Omega_f)$ rewrites as the opposite of the heat transported from the inlet boundary $\partial\Omega_{f,in}$ to the outlet boundary $\partial\Omega_{f,out}$:

$$J(\Omega_f) = -\left(\int_{\partial\Omega_{f,in}} \rho c_p T\boldsymbol{v}\cdot\boldsymbol{n}\mathrm{d}s + \int_{\partial\Omega_{f,out}} \rho c_p T\boldsymbol{v}\cdot\boldsymbol{n}\mathrm{d}s\right).$$

Although mathematically equivalent, we prefer to use the volume form in (5.2.3) which seems to us more accurate numerically (following the previous section and chapter 2, section 2.5.7).

The optimization problem (5.2.3) is solved for several values for the input velocity $||\boldsymbol{v}_0||_\infty$ and the pressure loss threshold $\texttt{DP}_0$. The details of the considered three situations are provided in Table 5.1 below.

Reynolds, Péclet numbers and numerical values of the physical parameters Since we do not rely on a turbulent model in the Navier-Stokes equations (5.2.1) for the determination of the fluid velocity and pressure $(\boldsymbol{v}, p)$, our study is restricted to moderate values of Reynolds and Péclet numbers. The viscosity and conductivity coefficients ν and k_f are computed so as to fix the Reynolds and Péclet number

$$\mathrm{Re} := \frac{\rho||\boldsymbol{v}_0||_\infty H}{\nu}, \qquad \mathrm{Pe} := \frac{\rho c_p||\boldsymbol{v}_0||_\infty H}{k_f}$$

to the values provided for each configuration in Table 5.1.

Minimum thickness constraint for the oil phase cross section The resolution of the optimization problem (5.2.3) tends to produce very thin and elongated shapes for the oil cross section. These are favorable for the optimization problem (5.2.3), however they are numerically unstable. Indeed, small components of the design thinner than the prescribed mesh size tend to disappear due to numerical diffusion during the optimization process. We therefore consider an alternative version of (5.2.3) which tends to impose a minimum thickness $d_{\min}$ of the oil channels. This constraint could also be of interest in the perspective of the manufacturing process. We follow the strategy proposed in chapter 4, section 4.4.3: the performance of the system $J(\Omega_f)$ is prescribed to be at least as good as a threshold value $\texttt{J}_0$, and

we minimize instead an energy $E(\Omega_f)$ which favors areas of $D\backslash\Omega_f$ thicker than $d_{\min}$:

$$\begin{aligned} \min_{\Omega_f \subset D} \quad & E(\Omega_f) := -\int_{D\backslash\Omega_f} d_{\Omega_f}^2 \max(-d_{\Omega_f} + d_{\min}/2, 0)^2 \mathrm{d}x \\ s.t. \quad & \begin{cases} \mathtt{DP}(\Omega_f) \leq \mathtt{DP}_0 \\ J(\Omega_f) \leq \mathtt{J}_0. \end{cases} \end{aligned} \tag{5.2.4}$$

The reader is referred to chapter 4 for the computation of the shape derivative of $E(\Omega_f)$. The threshold value $\mathtt{J}_0$ measuring the performance of the heat exchange is determined empirically from the values obtained with the resolution of the original problem (5.2.3) which does not feature the minimum thickness constraint. For our application, we took $\mathtt{J}_0 = 10000$.

Shape derivatives for the heat exchanger problem The physics considered for this test case is essentially the same as that of (5.1.1) and (5.1.2); the only difference is the Dirichlet boundary condition $T = T_{oil}$ on the interface Γ. It is easily verified that the shape derivative formulas provided in propositions 2.3 and 2.4 are still valid up to a change of boundary condition for the thermal adjoint equation: the adjoint problem (2.4.9) is solved with the boundary condition $S = 0$ on Γ. Then, it can be shown that the shape derivative of an arbitrary functional $J(\Omega_f, \boldsymbol{v}(\Omega_f), p(\Omega_f), T(\Omega_f))$ in volume and surface forms read respectively:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) &= \frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) + \int_{\Omega_f} \left[-(\sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{w} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{v})\mathrm{div}(\boldsymbol{\theta})\right] \mathrm{d}x \\ &+ \int_{\Omega_f} [\sigma_f(\boldsymbol{v}, p) : (\nabla \boldsymbol{w} \nabla \boldsymbol{\theta}) + \sigma_f(\boldsymbol{w}, q) : (\nabla \boldsymbol{v} \nabla \boldsymbol{\theta}) + \rho \boldsymbol{w} \cdot (\nabla \boldsymbol{v} \nabla \boldsymbol{\theta}) \cdot \boldsymbol{v}] \mathrm{d}x \\ &- \int_{\Omega_f} \mathrm{div}(\boldsymbol{\theta})(k_f \nabla T \cdot \nabla S + \rho c_p (\boldsymbol{v} \cdot \nabla T) S) \mathrm{d}x \\ &+ \int_{\Omega_f} \left[k_f (\nabla \boldsymbol{\theta} + \nabla \boldsymbol{\theta}^T) \nabla T \cdot \nabla S + \rho c_p \boldsymbol{v} \cdot (\nabla \boldsymbol{\theta}^T \nabla T) S\right] \mathrm{d}x, \end{aligned} \tag{5.2.5}$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big[J(\Gamma_{\boldsymbol{\theta}}, \boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}), p(\Gamma_{\boldsymbol{\theta}}), T(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) = \overline{\frac{\partial \mathfrak{J}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) + \int_{\Gamma} \left(2\nu e(\boldsymbol{w}) : e(\boldsymbol{v}) + k_f \frac{\partial T}{\partial \boldsymbol{n}} \frac{\partial S}{\partial \boldsymbol{n}}\right)(\boldsymbol{\theta} \cdot \boldsymbol{n}) \mathrm{d}s. \tag{5.2.6}$$

where the notation $\mathfrak{J}$ refers to the modified functional of (2.4.7) introduced in chapter 2.

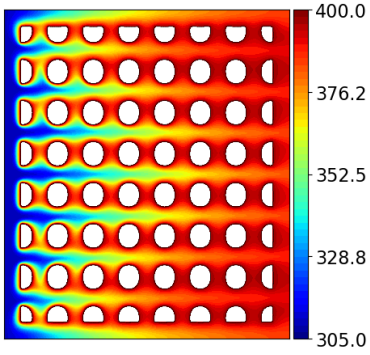
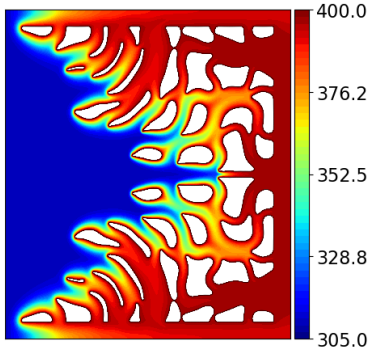
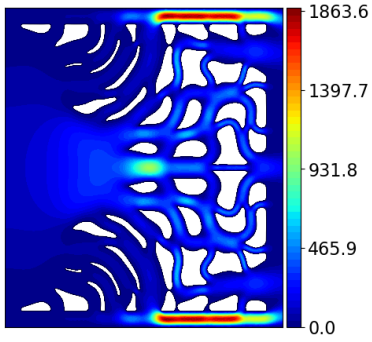
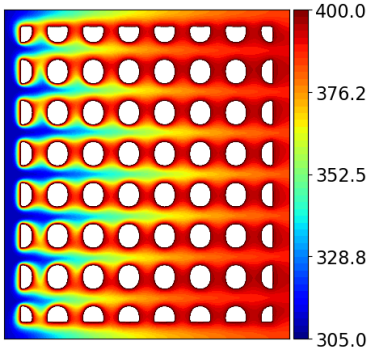
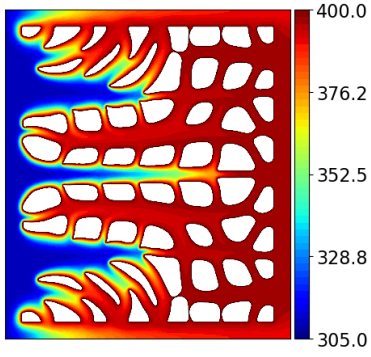
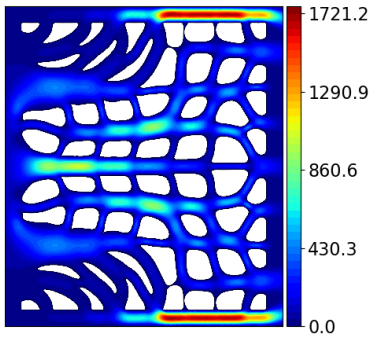
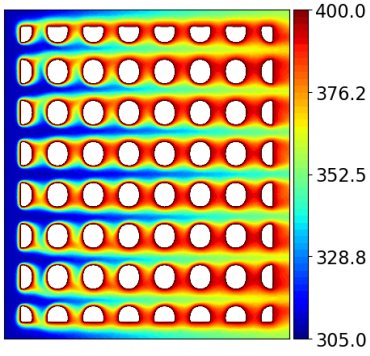
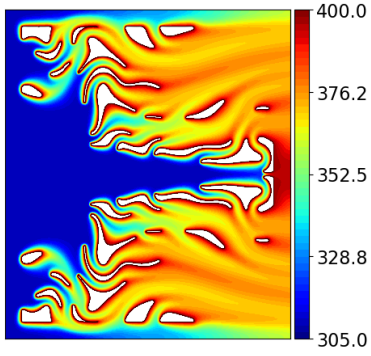
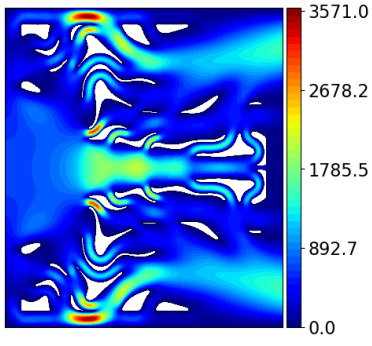
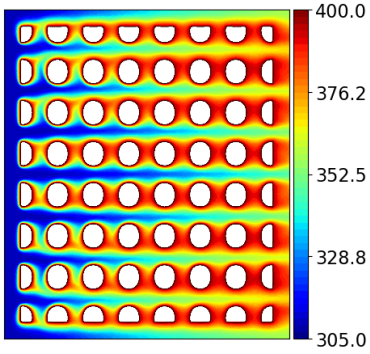
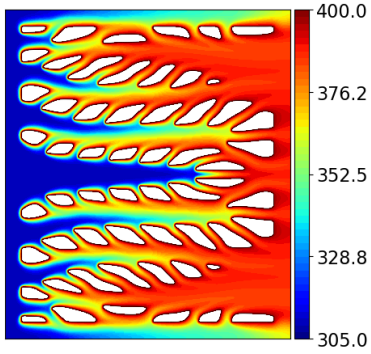
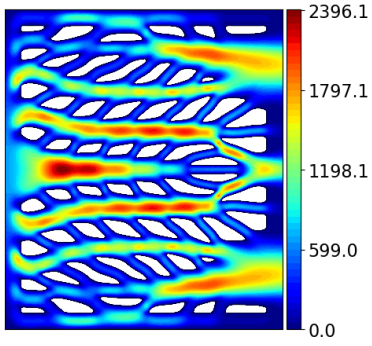
5.2.3 Numerical results

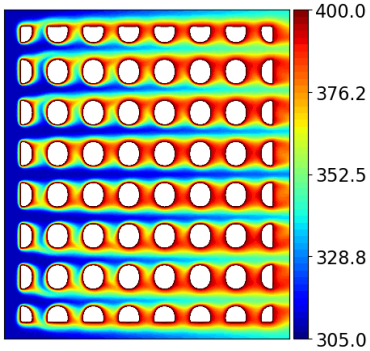
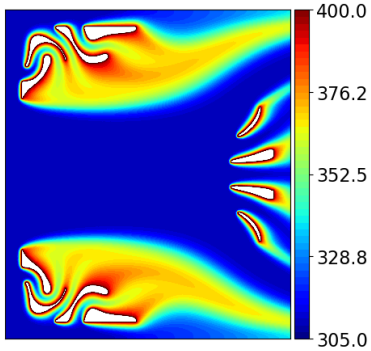
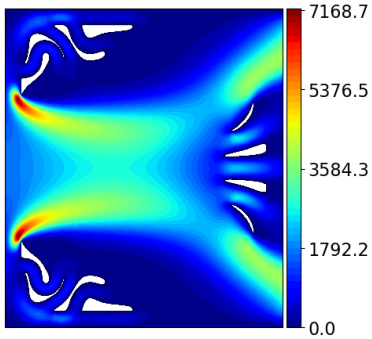
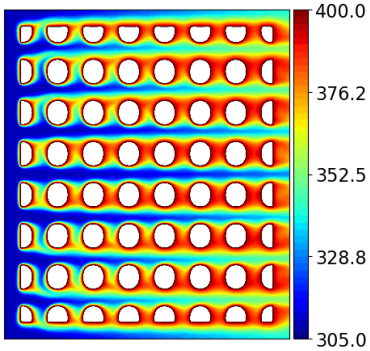
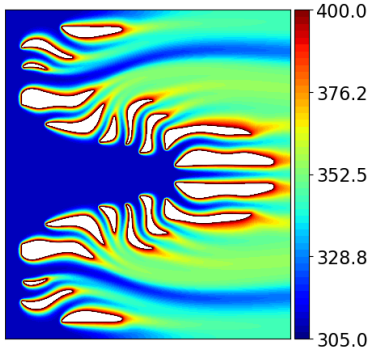
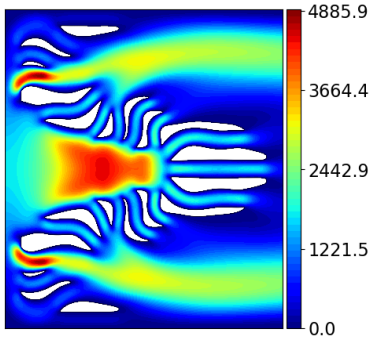
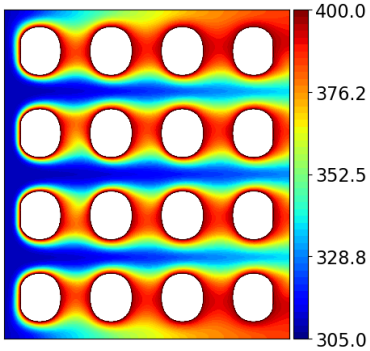
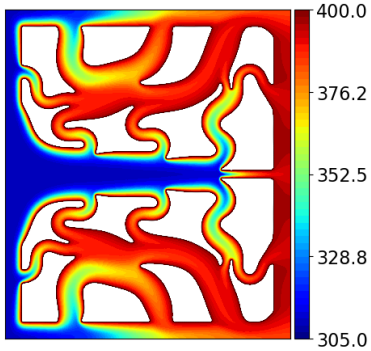
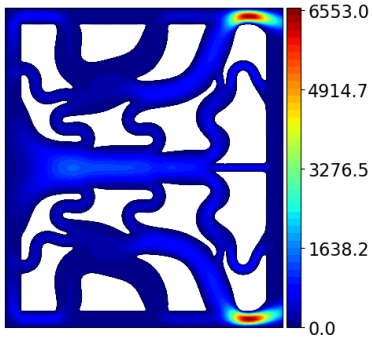
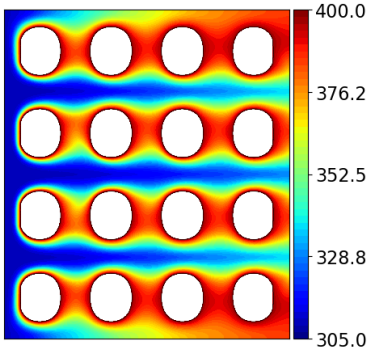
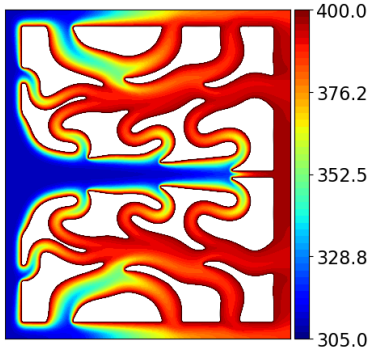
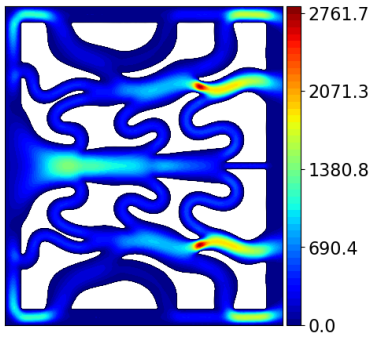
Both optimization problems (5.2.3) and (5.2.4) are solved for the three configurations considered in Table 5.1. Since there is no unique solution, we propose a set of 24 results obtained by varying the shape chosen for the initialization. These are reported in the Table 5.2 below where we display the final temperature and kinetic energy fields. We note that our optimization algorithm is able to (i) create recirculating fluid regions favorable to heat exchange and (ii) make the transverse oil channels assume an aerodynamic shape limiting the output pressure loss. We also observe that the effect of the minimum thickness constraint in (5.2.4) is most significant on configurations with maximal input velocity $||\boldsymbol{v}_0||_\infty = 40$ which feature the most elongated structures. A few intermediate shapes are reported in Figs. 5.8a to 5.8c.

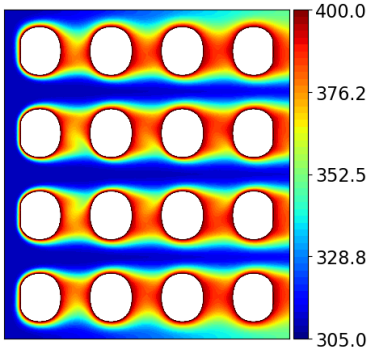
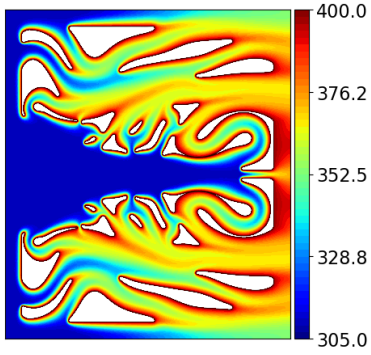
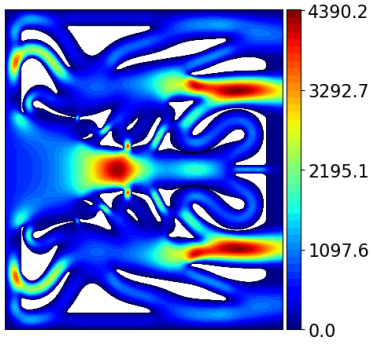
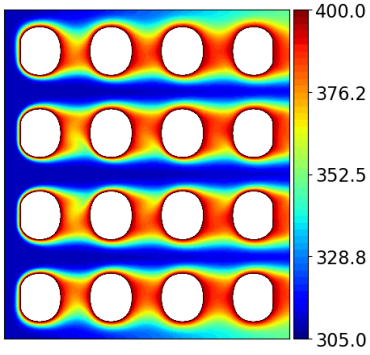
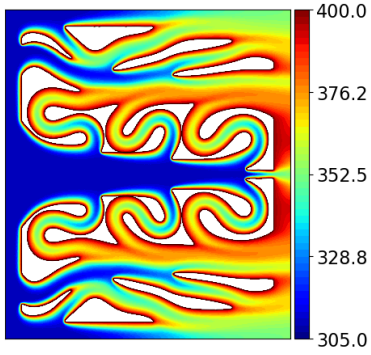
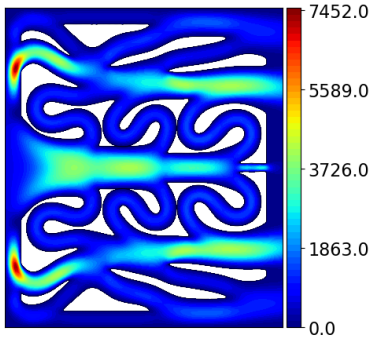
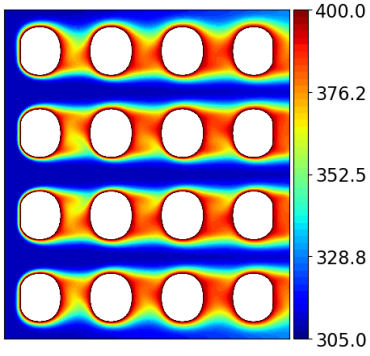
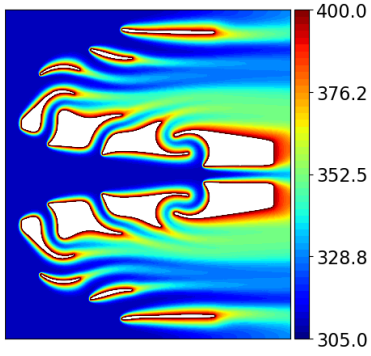
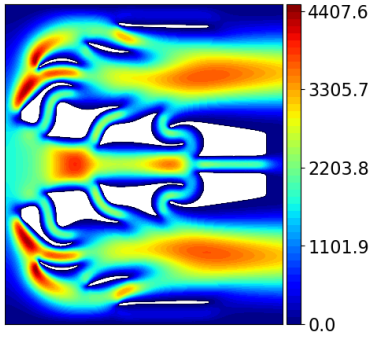
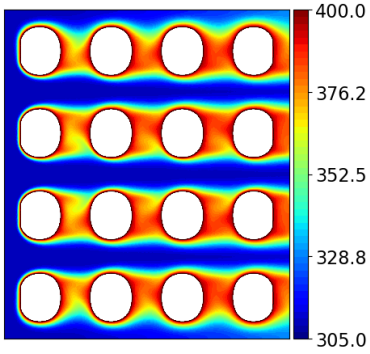
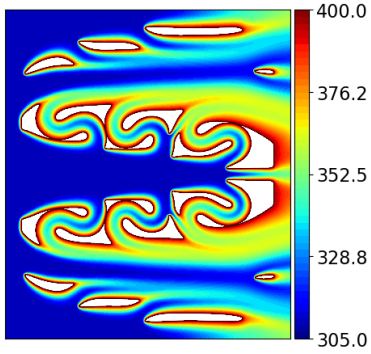
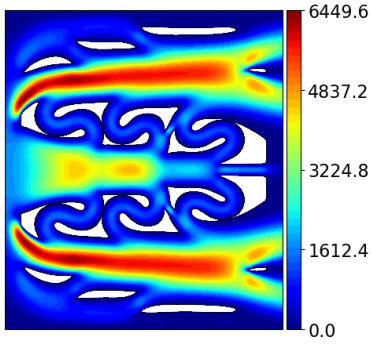
We relied on a rather fine mesh resolution (the minimum edge size was of the order of $h_{\min} = 0.003$), which is illustrated on Figure 5.9 where the mesh of the final shape is displayed for one of the test cases.

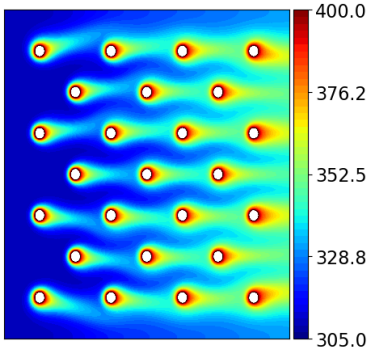
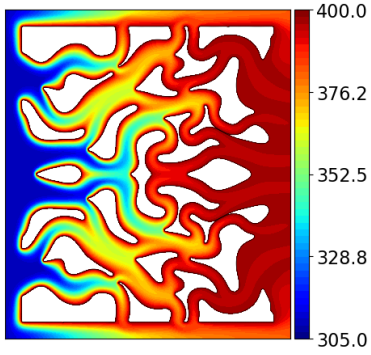
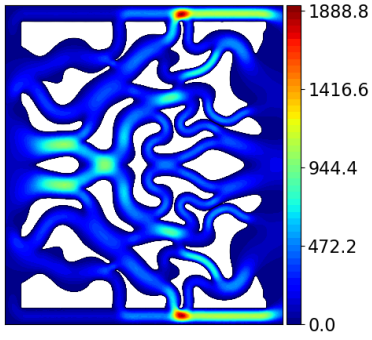
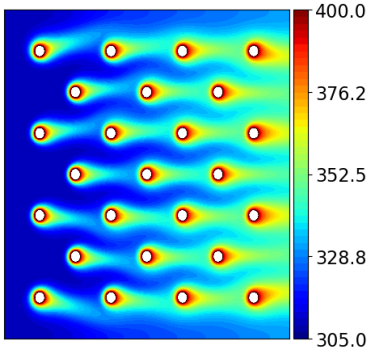
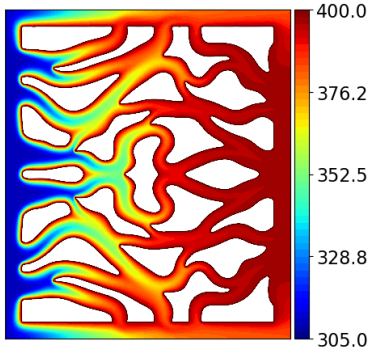
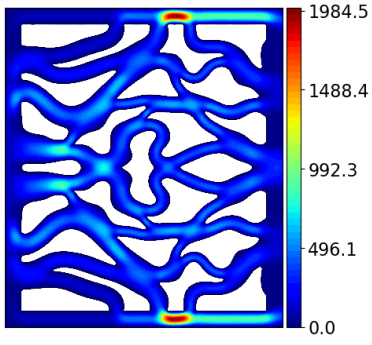
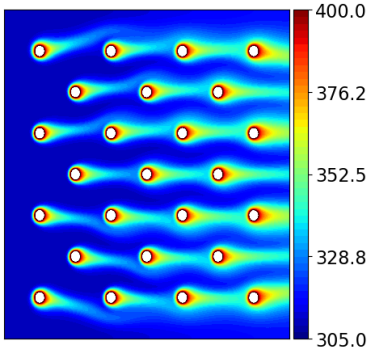
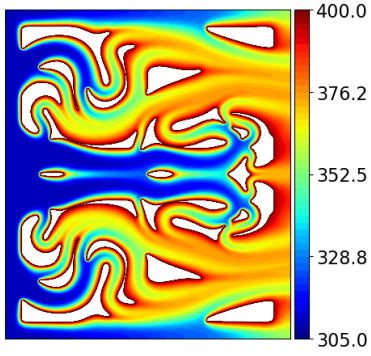
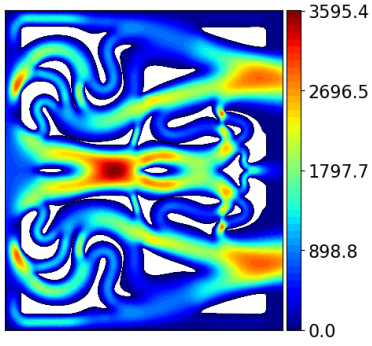
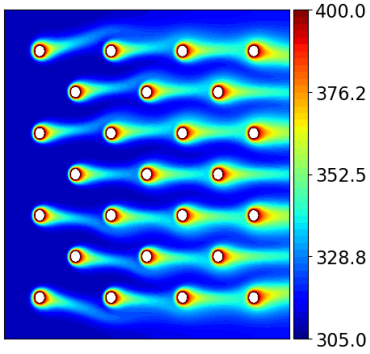
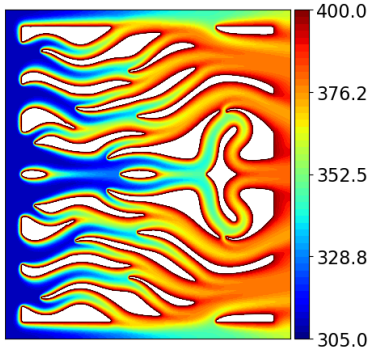
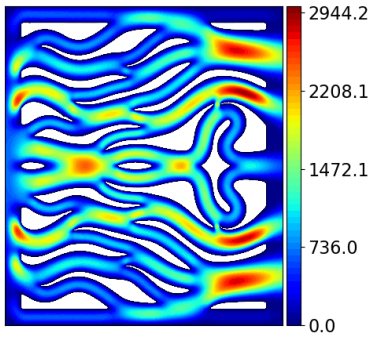
These results point out that a rather large variety of candidate shapes may be satisfactory solutions for the heat exchanger problem (5.2.3). Although the test case remains academic because we do not take turbulence into account, these rather unconventional designs further highlight longer term perspectives for topology optimization in the context of coupled fluid thermal industrial applications.

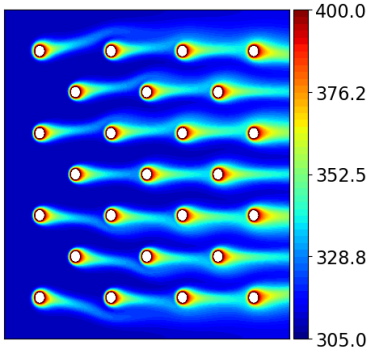
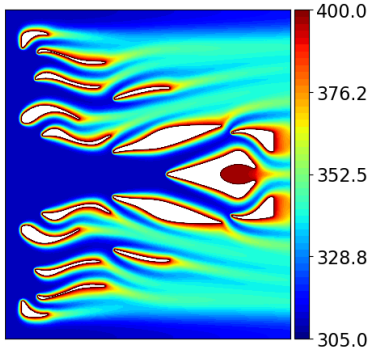
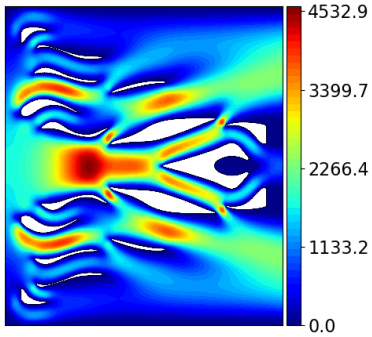
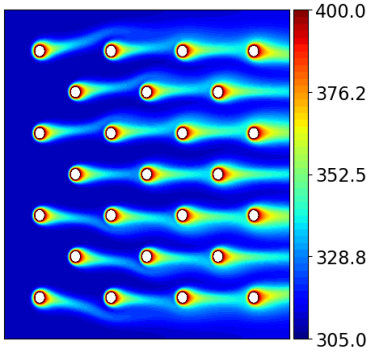
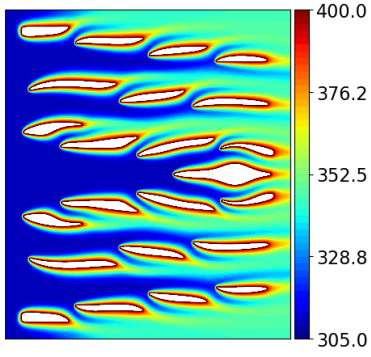
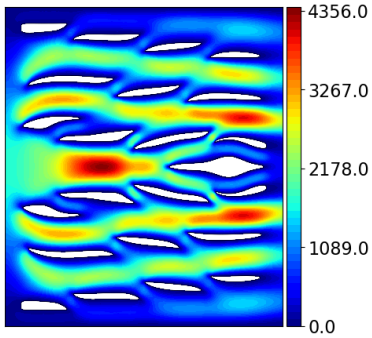
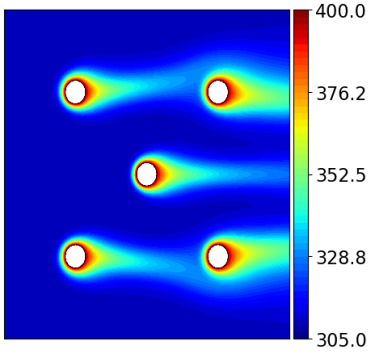
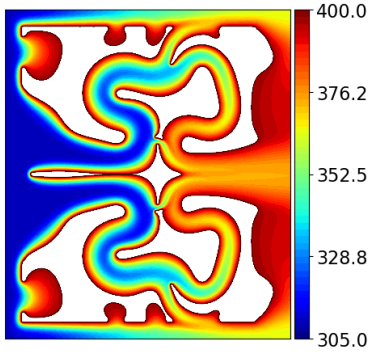
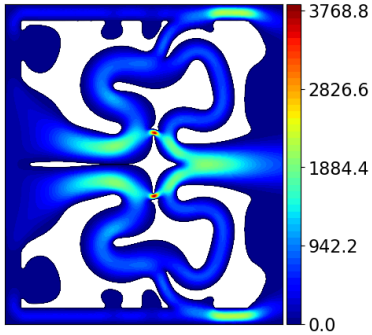
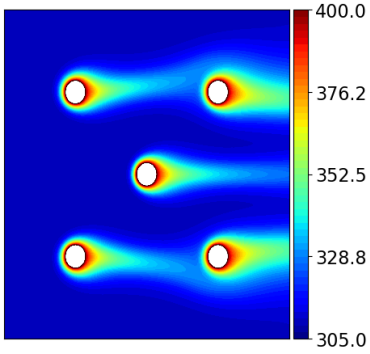
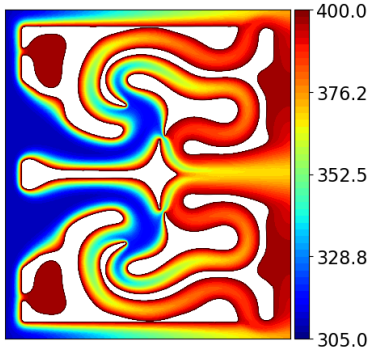
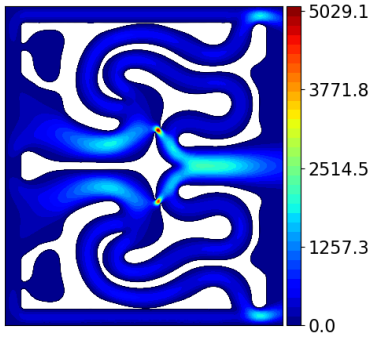
Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)

Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 1 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 DP_0=1300 J_{final} =4350 DP_{final} =1113 *Without* min. thickness constraint			
Test case 2 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 DP_0=1300 J_{final} =4346 DP_{final} =1217 *With* min. thickness constraint			
Test case 3 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 DP_0=1030 J_{final} =8089 DP_{final} =983 *Without* min. thickness constraint			
Test case 4 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 DP_0=1030 J_{final} =9742 DP_{final} =1030 *With* min. thickness constraint			

Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 5 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =3472 $\mathtt{DP}_{final}$ =392 *Without* min. thickness constraint			
Test case 6 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =7285 $\mathtt{DP}_{final}$ =520 *With* min. thickness constraint			
Test case 7 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 $\mathtt{DP}_0$=1300 J_{final} =4086 $\mathtt{DP}_{final}$ =1308 *Without* min. thickness constraint			
Test case 8 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 $\mathtt{DP}_0$=1300 J_{final} =4168 $\mathtt{DP}_{final}$ =1188 *With* min. thickness constraint			

Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 9 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =7667 $\mathtt{DP}_{final}$ =968 *Without* min. thickness constraint			
Test case 10 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =7508 $\mathtt{DP}_{final}$ =1112 *With* min. thickness constraint			
Test case 11 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =5731 $\mathtt{DP}_{final}$ =479 *Without* min. thickness constraint			
Test case 12 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =6847 $\mathtt{DP}_{final}$ =524 *With* min. thickness constraint			

Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 13 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 $\mathtt{DP}_0$=1300 J_{final} =4208 $\mathtt{DP}_{final}$ =1140 *Without* min. thickness constraint			
Test case 14 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 $\mathtt{DP}_0$=1300 J_{final} =4252 $\mathtt{DP}_{final}$ =1157 *With* min. thickness constraint			
Test case 15 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =7785 $\mathtt{DP}_{final}$ =1022 *Without* min. thickness constraint			
Test case 16 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =8711 $\mathtt{DP}_{final}$ =1106 *With* min. thickness constraint			

Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 17 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 DP_0=475 J_{final} =6236 DP_{final} =470 *Without* min. thickness constraint			
Test case 18 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 DP_0=475 J_{final} =7822 DP_{final} =498 *With* min. thickness constraint			
Test case 19 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 DP_0=1300 J_{final} =3361 DP_{final} =1149 *Without* min. thickness constraint			
Test case 20 $\|\|\boldsymbol{v}_0\|\|_\infty$=10 DP_0=1300 J_{final} =3582 DP_{final} =1064 *With* min. thickness constraint			

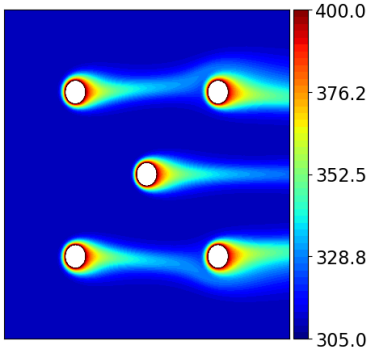
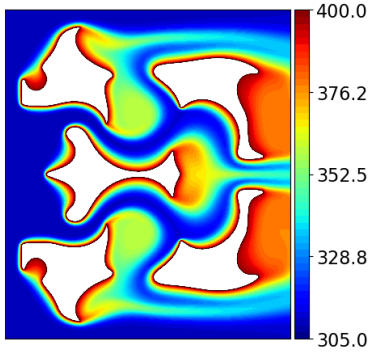
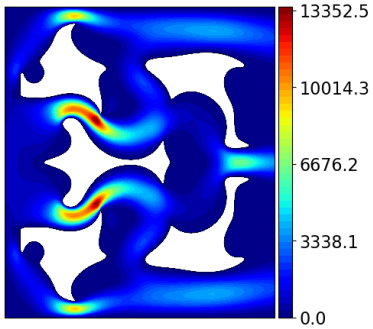
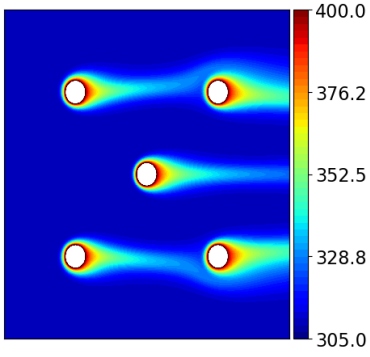
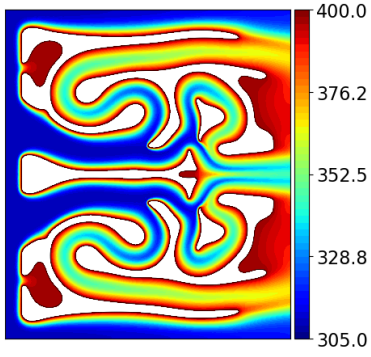
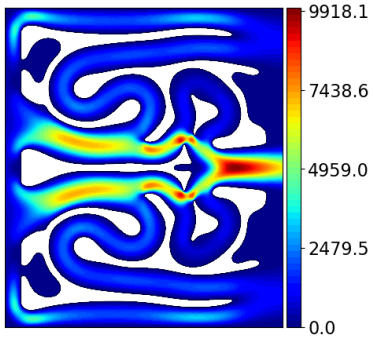
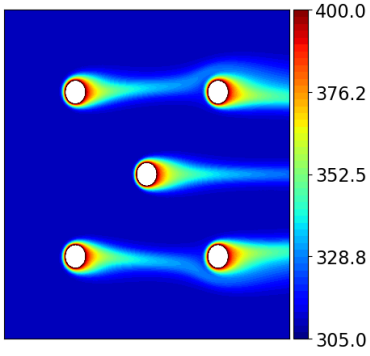
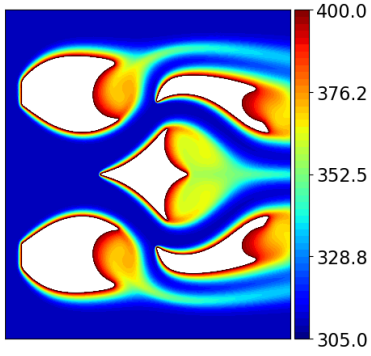
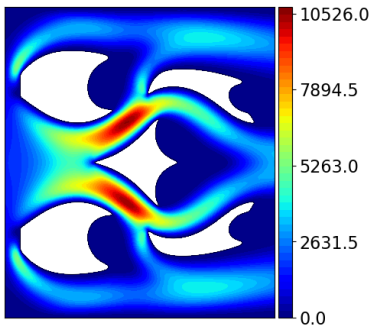
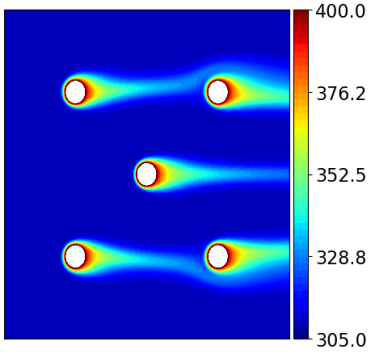
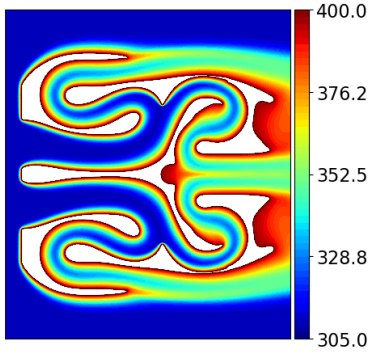
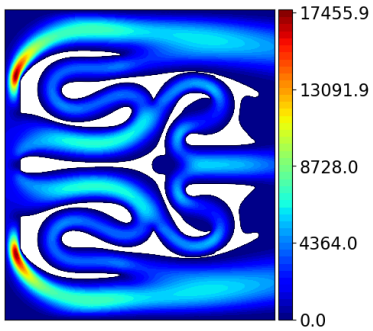
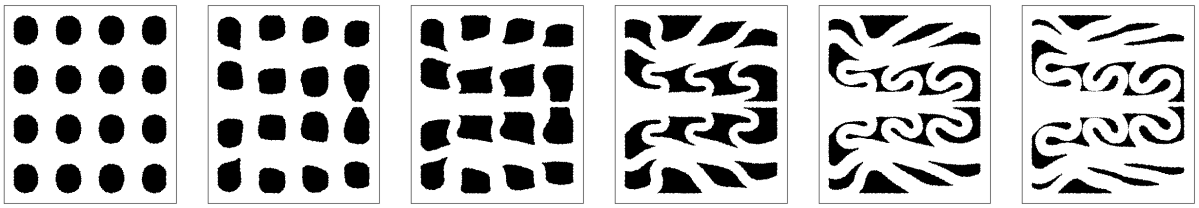
Test case	T (Initial design)	T (Optimized design)	$\|\|\boldsymbol{v}\|\|^2$ (Optimized design)
Test case 21 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =2972 $\mathtt{DP}_{final}$ =986 *Without* min. thickness constraint			
Test case 22 $\|\|\boldsymbol{v}_0\|\|_\infty$=25 $\mathtt{DP}_0$=1030 J_{final} =5330 $\mathtt{DP}_{final}$ =1589 *With* min. thickness constraint			
Test case 23 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =2847 $\mathtt{DP}_{final}$ =476 *Without* min. thickness constraint			
Test case 24 $\|\|\boldsymbol{v}_0\|\|_\infty$=40 $\mathtt{DP}_0$=475 J_{final} =4925 $\mathtt{DP}_{final}$ =1051 *With* min. thickness constraint			

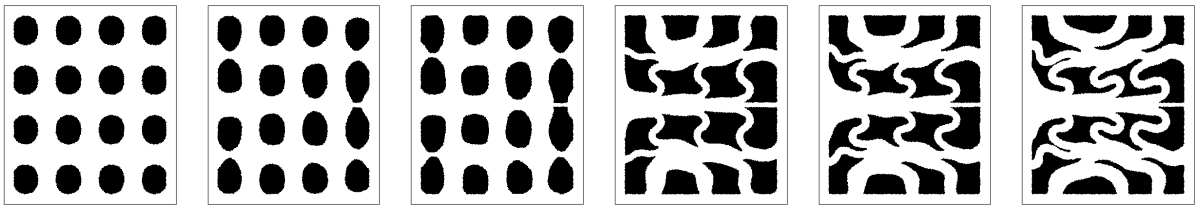
Table 5.2: Topology optimization results for the air-oil heat exchanger case study.

5.2.4 An alternative model featuring a stagnation pressure boundary condition

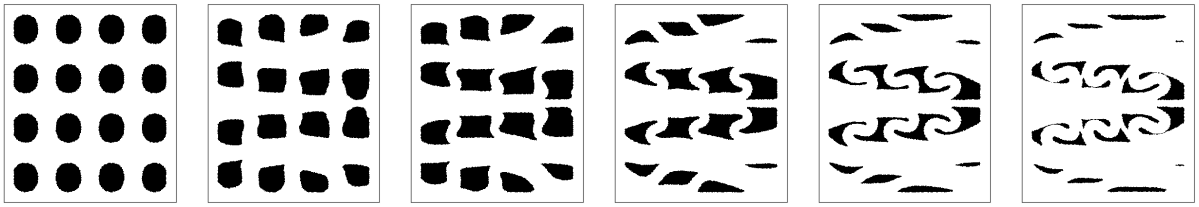
Industrial specifications of heat exchangers often impose prescribed input pressure values P_{in} and P_{out} at the inlet and the outlet of the system rather than the upper bound condition $\mathtt{DP}(\Omega_f) \leq \mathtt{DP}_0$ on the pressure loss. Furthermore, P_{in} and P_{out} correspond to the "stagnation" pressure $p + \rho v^2/2$ ($v = ||\boldsymbol{v}||^2$



(a) **Test case 8**, $||\boldsymbol{v}_0||_\infty$=25



(b) **Test case 10**, $||\boldsymbol{v}_0||_\infty$=10



(c) **Test case 12**, $||\boldsymbol{v}_0||_\infty$=40

Figure 5.8: Iterations 0, 10, 20, 100, 200, 400 for several test case configurations.

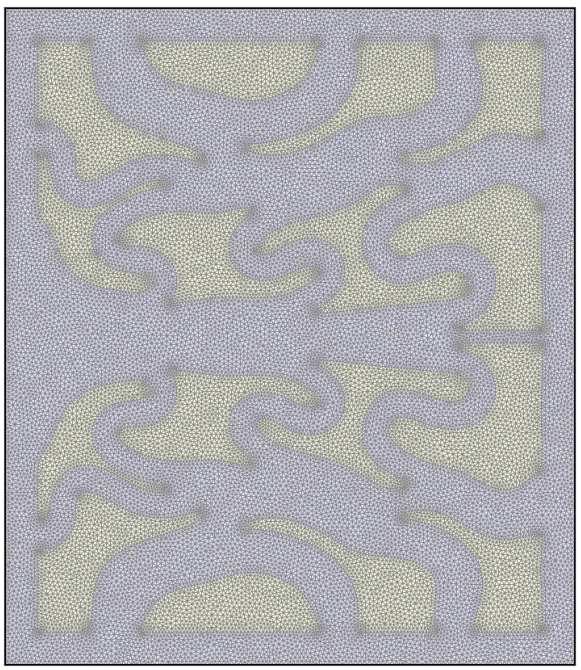


Figure 5.9: Mesh of the final shape of a test case featuring $||\boldsymbol{v}_0|| = 10$.

in our case) rather than the static pressure p (the "stagnation pressure" can be directly experimentally measured with pitot tubes). These considerations suggest to consider a variant of the optimization problem (5.2.3) where the pressure loss constraint would be enforced as a boundary condition involving the pressures P_{in} and P_{out} in the Navier-Stokes equations (5.1.1).

It turns out that it is indeed possible to impose the value of the stagnation pressure upon a judicious rewriting of the Navier-Stokes equations in their classical "rotational form".

Rewriting of the physical model and shape derivatives

The Navier-Stokes system in rotational form classically reads

$$\begin{cases} -\nu\Delta\boldsymbol{v} + \nabla\left(p + \frac{1}{2}||\boldsymbol{v}||^2\right) + \rho(\nabla\times\boldsymbol{v})\times\boldsymbol{v} = 0 & \text{in } \Omega_f \\ \operatorname{div}(\boldsymbol{v}) = 0 & \text{in } \Omega_f \\ p + \frac{1}{2}\rho||\boldsymbol{v}||^2 = P_{in} & \text{on } \partial\Omega_{f,in} \\ \boldsymbol{v}\times\boldsymbol{n} = 0 & \text{on } \partial\Omega_{f,in} \\ p + \frac{1}{2}\rho||\boldsymbol{v}||^2 = P_{out} & \text{on } \partial\Omega_{f,out} \\ \boldsymbol{v}\times\boldsymbol{n} = 0 & \text{on } \partial\Omega_{f,out} \\ \boldsymbol{v} = 0 & \text{on } \partial\Omega_{f,wall} \\ \boldsymbol{v} = 0 & \text{on } \Gamma. \end{cases} \tag{5.2.7}$$

The associated variational formulation then reads (see the works of Conca et. al. [105, 106]):

Find $(\boldsymbol{v}, q) \in V_{\boldsymbol{v},q}$ such that $\forall(\boldsymbol{w}, r) \in V_{\boldsymbol{v},q}$,

$$\int_{\Omega_f} [\nu(\nabla\times\boldsymbol{v})\cdot(\nabla\times\boldsymbol{w}) + \rho(\nabla\times\boldsymbol{v})\times\boldsymbol{v}\cdot\boldsymbol{w} - r\operatorname{div}(\boldsymbol{v}) - q\operatorname{div}(\boldsymbol{w})]\mathrm{d}x = -\int_{\partial\Omega_{f,in}} P_{in}\boldsymbol{w}\cdot\boldsymbol{n}\mathrm{d}s - \int_{\partial\Omega_{f,out}} P_{out}\boldsymbol{w}\cdot\boldsymbol{n}\mathrm{d}s \tag{5.2.8}$$

where $V_{\boldsymbol{v},q}$ is the functional space

$$V_{\boldsymbol{v},q} = \{(\boldsymbol{v}, q) \in H^1(\Omega, \mathbb{R}^d) \times L^2(\Omega)/\mathbb{R} \,|\, \boldsymbol{v} = 0 \text{ on } \Gamma \cup \partial\Omega_{f,wall} \text{ and } \boldsymbol{v}\times\boldsymbol{n} = 0 \text{ on } \partial\Omega_{f,in} \cup \partial\Omega_{f,out}\}.$$

We then consider the minimization problem (5.2.3) without the pressure loss constraint:

$$\min_{\Omega_f \subset D} \quad J(\Omega_f). \tag{5.2.9}$$

Since the variational formulation (5.2.8) is different to that associated with the more standard Navier-Stokes equations (5.2.1), the shape derivative of $J(\Omega_f)$ needs to be recomputed. Very briefly, the adjoint fluid variable, denoted $(\boldsymbol{w}, r)$, is obtained by solving the following variational problem:

Find $(\boldsymbol{w}, r) \in V_{\boldsymbol{v},q}$ such that $\forall(\boldsymbol{w}', r') \in V_{\boldsymbol{v},q}$,

$$\int_{\Omega_f} \nu(\nabla\times\boldsymbol{w})\cdot(\nabla\times\boldsymbol{w}') + \rho(\nabla\times\boldsymbol{w}')\times\boldsymbol{v}\cdot\boldsymbol{w} + \rho(\nabla\times\boldsymbol{v})\times\boldsymbol{w}'\cdot\boldsymbol{w} - r\operatorname{div}(\boldsymbol{w}') - r'\operatorname{div}(\boldsymbol{w})\mathrm{d}x = -\int_{\Omega_f} \rho c_p\boldsymbol{w}'\cdot\nabla T S\mathrm{d}x + \frac{\partial\mathfrak{J}}{\partial(\boldsymbol{v}, q)}\cdot(\boldsymbol{w}', r'). \tag{5.2.10}$$

The thermal adjoint variable S is computed by the previous equation of chapter 2, (2.4.9). It can then be shown that the shape derivative of an arbitrary functional $J(\Gamma, \boldsymbol{v}, q, T)$ reads in surface form:

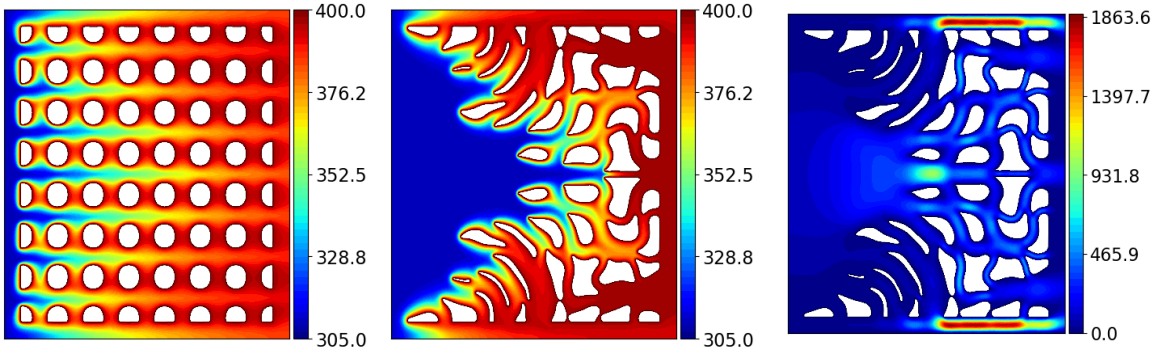
$$\left.\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=0} [J(\Omega_{\boldsymbol{\theta}}, \boldsymbol{v}_{\boldsymbol{\theta}}, q_{\boldsymbol{\theta}}, T_{\boldsymbol{\theta}})]\cdot\boldsymbol{\theta} = \overline{\frac{\partial\mathfrak{J}}{\partial\boldsymbol{\theta}}}\cdot\boldsymbol{\theta} + \int_{\partial\Omega_f}\left(\nu(\nabla\times\boldsymbol{w})\cdot(\nabla\times\boldsymbol{v}) + k_f\frac{\partial T}{\partial n}\frac{\partial S}{\partial n}\right)(\boldsymbol{\theta}\cdot\boldsymbol{n})\mathrm{d}s. \tag{5.2.11}$$

Remark 5.3. This rotational formulation does not allow to prescribe the stagnation pressure $P + \rho||\boldsymbol{v}||^2/2$ on some parts of the boundary $\partial\Omega_f$, and the normal stress flux $\sigma_f(\boldsymbol{v}, p)\cdot\boldsymbol{n}$ at some other parts, which could be also of industrial interest. However, this might be possible by resorting to a different variational treatment accounting for these boundary conditions, see [68].

Numerical results

Our numerical results do not seem conclusive; the main reason lies in that the boundary condition on the stagnation pressure does not allow to bound sufficiently the values of the velocity in the whole domain. Indeed, the prescribed input and output pressure tends to make the flow velocity very large between close obstacles: the Reynolds number is not well controlled, which makes the numerical treatment difficult.

However when the optimization succeeds, the optimized shapes look similar to those obtained previously with the pressure loss imposed as a constraint. An instance of such shapes is illustrated on Figure 5.10 below. For these reasons, it seems to us preferable to stick to the first approach (5.2.3) (or



(a) Initial shape (T) (b) Optimized shape (T) (c) Optimized shape ($||\boldsymbol{v}||$)

Figure 5.10: An optimization result for the shape optimization of the problem (5.2.9) featuring the Navier-Stokes equations (5.2.7) with a boundary condition on the stagnation pressure $p + \rho v^2/2$.

its reformulation (5.2.4) accounting for the minimum thickness constraint), which is also that followed by the majority of academics involving heat exchangers, e.g. [116, 255].

5.2.5 Conclusions

In this study, we have demonstrated the relevance of shape and topology optimization for generating unconventional heat exchanger designs. The method could be extended to other choices of physical models up to the use of different formulas for the shape derivatives: future work could take into account much larger, turbulent, Reynolds and Péclet numbers by considering turbulence models. Once again, we emphasize the ability of the method to take into account arbitrary and multiple optimization constraints. The extension to true 3-d test cases at low Reynolds number would not add additional theoretical difficulties in principle; however the implementation is delicate and still the object of ongoing work.

Let us mention, however, that several difficulties are still to be expected regarding the applications of such shape optimization methods to realistic industrial applications featuring large Reynolds and Péclet numbers (typically of the order of 30,000 or more).

First, when the Reynolds number is large, the assumption of stationary velocity and pressure fields does not make sense because it is well known that stationary solutions are unstable; the time-dependent solution varying rather on some attractor [304].

Second, even assuming the optimization to be numerically tractable, current heat exchanger designs suggest that optimal topologies would be characterized by a microstructure featuring very small details periodically repeated. Such designs would be very difficult to obtain with the method of Hadamard because very costly to mesh explicitly. It seems to us that a homogenization approach would be relevant for tackling such design problems. The recent works of [254, 166, 27] provide some perspectives regarding the automatic generation of optimized microstructures from homogenized models in the context of linear elasticity. In the final chapter 7, we provide a few preliminary theoretical contributions towards the application of such methods for fluid models.

CHAPTER 6

TOWARDS 3-D AND INDUSTRIAL APPLICATIONS: IMPLEMENTATION RECIPES FOR A VARIETY OF NUMERICAL TEST CASES

Contents

This chapter attempts to demonstrate the ability of multiphysics shape optimization by the method of Hadamard to deal with 3-d problems approaching industrial test cases.

From the numerical point of view, the extension of the shape optimization algorithm outlined in chapter 1, section 1.4 from 2-d to 3-d is delicate. Several additional ingredients are required in order achieve a satisfactory efficiency, or even to be able to achieve a single optimization run (for fluid applications). The first section 6.1 outlines the most important features of our implementation: we namely discuss the use of domain decomposition techniques and preconditioning for the resolution of the 3-d multiphysics state equations involved in our shape optimization test cases. We also provide beforehand a brief presentation of our `python`/`FreeFEM` implementation, and some details about various technical operations (rarely described in the litterature) applied in the course of optimization iterations in order to enforce non-optimizable regions, symmetry and non-degeneracy of discretized shapes.

In the next section 6.2, we present a variety of 3-d test cases solved thanks to our implementation. Four examples are considered: the first three of them are single physics applications in either linear elasticity, heat conduction, or fluid mechanics. The fourth one is a true multiphysics example; it is concerned with fluid-structure interaction. These test cases are described as *moderately* large-scale, in the sense that the problems considered make extensive use of parallel computing and preconditioning in order to be run in reasonable CPU time, however their size remains rather small (our largest test case features up to 1.7 millions degrees of freedom) when compared to that of industrial problems (reaching about the billion of degrees of freedom). Our results, however, are promising and are preliminary to more challenging applications.

6.1 Implementation recipes for 3-d constrained topology optimization of multiphysics system

In this section, we discuss several aspects of our implementation in `python` (version 3.6) and `FreeFEM` [183] for 3-d shape and topology optimization of coupled thermal-fluid elastic systems. The first two sections 6.1.1 and 6.1.2 provide information regarding several choices of programming paradigms. Section 6.1.3 describes several "hidden" operations applied to shapes in the course of optimization iterations, in order to account for non-optimizable regions, symmetry, or to avoid mesh degeneracy. Finally, our parallel implementation of finite element operations using domain decomposition and preconditioning is discussed in section 6.1.4.

6.1.1 Abstract programming paradigm for constrained shape optimization

A single implementation of the null space gradient flow algorithm as described in chapter 3, algorithm 3.1 is used for both 2-d and 3-d problems, which could in principle be used for any optimization problem set on a *manifold* as soon as a generic minimal set of ingredients (described below) is provided.

In a few words, the algorithm was implemented in `python` via a `nullspace` function whose prototype reads

```
def nullspace(problem: Optimizable, params=None, results=None):
    """
    Solve the optimization problem
        min      J(x)
        x in V
        under the constraints
        g_i(x)=0  for all i=1..p
        h_j(x)<=0 for all j=1..q

    problem: an instance of the class Optimizable
    params:  a dictionary of optimization parameters
    results: a previous output of the nullspace function
             (the algorithm will restart from the last iteration)
    """
```

The most important argument of the function `nullspace` is the variable `problem` which instantiates an abstract class `Optimizable`. An `Optimizable` object encodes a generic optimization problem

$$\min_{\mathtt{x}\in\mathcal{X}} \quad \mathtt{J}(\mathtt{x}) \qquad s.t. \begin{cases} \mathtt{g}_i(\mathtt{x}) = 0 \text{ for } 1 \le i \le \mathtt{p} \\ \mathtt{h}_j(\mathtt{x}) \le 0 \text{ for } 1 \le j \le \mathtt{q}. \end{cases} \tag{6.1.1}$$

The set $\mathcal{X}$ is assumed to be a manifold equipped with the structure outlined in chapter 3, section 3.6.1 whose representative illustration is reproduced on Figure 6.1 below with the current notation. In order

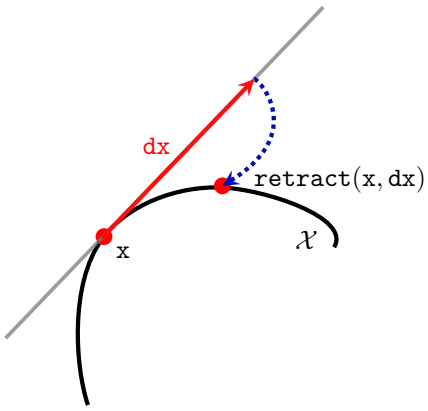


Figure 6.1: Tangential motion and retraction on an abstract manifold $\mathcal{X}$.

to solve (6.1.1), it is sufficient that the following information be provided (e.g. as object methods) by the supplied instance `problem` of an `Optimizable` object:

- $\mathtt{J}$, $\mathtt{g}_i$, $\mathtt{h}_j$: $\mathcal{X} \to \mathbb{R}$: objective functions and constraints;
- $\mathtt{DJ}$, $\mathtt{Dg}_i$, $\mathtt{Dh}_j$: $\mathcal{X} \to \mathbb{R}^n$: Fréchet derivatives of objective and constraints as functions. Here, n is thought of as the dimension of the tangent space of $\mathcal{X}$ to $\mathtt{x} \in \mathcal{X}$. For such a given $\mathtt{x}$, $\mathtt{DJ}(\mathtt{x})^T\mathtt{dx}$ is the variation of the objective function $\mathtt{J}$ at $\mathtt{x}$ along the tangent direction $\mathtt{dx} \in \mathbb{R}^n$;
- $\mathtt{A} : \mathcal{X} \to \mathbb{R}^{n\times n}$: local inner product needed for the computation of gradients

$$\nabla\mathtt{J}(\mathtt{x}) := \mathtt{A}(\mathtt{x})^{-1}\mathtt{DJ}(\mathtt{x}), \quad \nabla\mathtt{g}_i(\mathtt{x}) := \mathtt{A}(\mathtt{x})^{-1}\mathtt{Dg}_i(\mathtt{x}), \quad \nabla\mathtt{h}_j(\mathtt{x}) := \mathtt{A}(\mathtt{x})^{-1}\mathtt{Dh}_j(\mathtt{x}) \in \mathbb{R}^n.$$

Of course, $\texttt{A}(\texttt{x})$ must be a symmetric positive definite matrix.

- $\texttt{retract} : \mathcal{X} \times \mathbb{R}^n \to \mathcal{X}$: a *retraction* that convert a current point $\texttt{x}$ and a tangent vector $\texttt{dx} \in \mathbb{R}^n$ into a new point $\texttt{retract}(\texttt{x}, \texttt{dx}) \in \mathcal{X}$ on the manifold $\mathcal{X}$. This mapping must be compatible with the previous definitions of the derivatives $\texttt{DJ}, \texttt{Dg}_i$ and $\texttt{Dh}_j$ in the following sense:

$$\texttt{J}(\texttt{retract}(\texttt{x}, \Delta t \texttt{dx})) = \texttt{J}(\texttt{x}) + \Delta t\, \texttt{DJ}(\texttt{x})^T \texttt{dx} + o(\Delta t) \text{ as } \Delta t \to 0.$$

- `accept`: an optional function that is called by the optimization algorithm when the next point $\texttt{x} \in \mathcal{X}$ is accepted, which serves e.g. for saving current available information before proceeding to the next iteration.

The above ingredients constitute all the necessary information required by any first order optimization algorithm acting on a manifold (equipped with a retraction), and in particular by the null space algorithm outlined in chapter 3, algorithm 3.1. In our implementation for shape optimization, the current guess `x` contains the path to the current mesh file of the optimized domain, and `dx` is a finite element vector discretizing a deformation field. The function `A` returns the finite element matrix associated with the regularization, while `retract` encodes the advection and remeshing step. When passing from 2-d to 3-d, or if we were to rely on other numerical representations of shapes (e.g. using the level set method on a fixed), it is sufficient to update the above functions `A` and `retract` accordingly. This programming paradigm allows us to implement separately 2-d and 3-d test case (which required different solvers for the physical state equations), while calling the same null space optimization algorithm.

6.1.2 Interfacing `python` and `FreeFEM`

Our treatment of 2-d and 3-d shape optimization test cases with the method of Hadamard is implemented in both programming languages `python` and `FreeFEM` [183]:

1. on the one hand, the `nullspace` optimization routine of the previous section is conveniently implemented in `python`, which is a very user-friendly language and which allows for easy debugging. For instance, it is possible to pause a running instance of the code at any step of the optimization process (such as calling `mmg` for remeshing, computing the signed distance function with `mshdist`, etc...);

2. on the other hand, `FreeFEM` is used for all finite element related operations: the assembly of sparse matrices discretizing variational forms, integration on meshed subdomains or boundaries, the resolution of linear systems, the use of domain decomposition methods, mesh interpolation, etc... In particular, the language relies on a $C++$ kernel which allows to run these operations very efficiently. Furthermore, the syntax of the language is very close to mathematics and it allows to easily implement the expressions of the shape derivatives of chapter 2, propositions 2.3 and 2.4 (see also [34] for detailed examples of the use of `FreeFEM` in topology optimization).

The interface between both languages is realized thanks to a preprocessing meta-language: special instructions are added to `FreeFEM` (non executable) source files which are then parsed by a `python` routine and converted into a proper `.edp` file to be executed by `FreeFEM`. An example of such an augmented `.edp` code taken from our 2-d implementation is provided below:

```
SET (FLUID_DOMAIN,"3")
DEFAULT (THERMIC_ENABLED,"0")
IF THERMIC_ENABLED
    solve thermic(T,S)=
        int2d(Th,$FLUID_DOMAIN)(kf*grad(T)'*grad(S)
    IF FLUID_ENABLED
        \\+rho*cp*grad(T)'*[vx,vy]*S
    ENDIF
        \\)
        +int2d(Th,$SOLID_DOMAIN)(ks*grad(T)'*grad(S))
    IFDEF Qf
        -int2d(Thf)($Qf*S)
    ENDIF
```

```
    IFDEF Qs
        -int2d(Ths)($Qs*S)
    ENDIF
    IFDEF h
        +int1d(Th,$BCTN)($h*S)
    ENDIF
    IFDEF T0
        +on($BCTD,T=$T0)
    ENDIF
    \\;
ENDIF
```

This piece of code implements the variational form of chapter 2, (2.4.3):

$$\texttt{thermic(T,s)} = \int_{\Omega_s} k_s \nabla T \cdot \nabla S \mathrm{d}x + \int_{\Omega_f} (k_f \nabla T \cdot \nabla S + \rho c_p S \boldsymbol{v} \cdot \nabla T) \mathrm{d}x$$
$$= \int_{\Omega_s} Q_s S \mathrm{d}x + \int_{\Omega_f} Q_f S \mathrm{d}x + \int_{\partial\Omega_T^N} h S \mathrm{d}s.$$

The blue keywords `SET`, `DEFAULT`, `IF`, etc... correspond to preprocessing instructions parsed by `python`: they determine sections of the code to be included in the final executable depending on the values of the preprocessing variables prefixed by the dollar '`$`' symbol. For instance, the term $\rho c_p S \boldsymbol{v} \cdot \nabla T$ will not be included in the final code if `$FLUID_ENABLED` is not set to 1. The dollar prefixed variables can be easily accessed from the `python` implementation, or assigned directly in the FreeFem code with the instructions `SET` or `DEFAULT`. The double backslash symbols '\\' indicates that the previous line carriage return must be removed. For instance, if `$THERMIC_ENABLED`=1, `$FLUID_ENABLED`=1, `$T0`=100 and `$BCTD`=1, then the above source code is converted by our `python` parser as the following "standard" `FreeFEM` source code

```
    solve thermic(T,S)=
        int2d(Th,3)(kf*grad(T)'*grad(S)+rho*cp*grad(T)'*[vx,vy]*S)
        +int2d(Th,2)(ks*grad(T)'*grad(S))
        +on(1,T=100);
```

Such a preprocessing language proves to be very convenient in order (i) to use a single source code for all our test cases whatever the number of considered physical equations and (ii) to easily change test case parameter values from either `python` or `FreeFEM` thanks to the dollar prefixed variables and (iii) maintain a good readability of the implementation.

`FreeFEM` source files augmented with preprocessing instructions are called by the various functions `J`,`DJ`, `A`, `retract`, etc... declared in the implementation of an `Optimizable` object (see the previous section 6.1.1). Below is reported a typical output of a running instance of our 2-d code, which offers a clear picture of all the elementary steps of the shape optimization algorithm outlined in chapter 1, algorithm 1.1.

```
0. J=0.1079 G=[0.302] H=[]
FreeFem++ 01_cantilever/01_run/scripts/sensitivities.edp -nw (0.44)
FreeFem++ 01_cantilever/01_run/scripts/scalar_product.edp -nw (0.29)
advect -nocfl -dt 1.0 01_cantilever/01_run/meshes/Th_0000.mesh -c
    01_cantilever/01_run/scalars_dir/phi_0000.chi.sol
    -s 01_cantilever/01_run/tmp/theta.sol
    -o 01_cantilever/01_run/tmp/phi.o.sol (0.08)
FreeFem++ 01_cantilever/01_run/scripts/symmetrize.edp -nw (0.27)
mmg2d_O3 -nr -hmin 0.02 -hmax 0.1 -hgrad 1.3 -hausd 1e-4 -ls
    -sol 01_cantilever/01_run/tmp/phi_tmp.sol
    -out 01_cantilever/01_run/tmp/Th.o.mesh
    01_cantilever/01_run/meshes/Th_0000.mesh (0.16)
mv 01_cantilever/01_run/tmp/Th.o.mesh 01_cantilever/01_run/tmp/Th_tmp.mesh (0.00)
mshdist -v 0 -ncpu 1 01_cantilever/01_run/tmp/Th_tmp.mesh -dom -fmm (0.03)
```

```
FreeFem++ 01_cantilever/01_run/scripts/solve_state.edp -nw (0.43)
[...]
```

6.1.3 Processing operations on level set functions: generation of initial designs, taking into account non-optimizable regions, symmetrization and regularization

We now describe a few very classical operations that are implemented in most level-set based topology optimization codes but which are not often detailed in published works. The context and notation assumed are that of chapter 2: a hold-all domain $D \subset \mathbb{R}^d$ ($d = 2$ or 3) is considered, and the goal is to optimize the shape of the interface $\Gamma = \partial\Omega_s \cap \partial\Omega_f$ between a solid subdomain $\Omega_s \subset D$ and a fluid subdomain $\Omega_f = D\backslash\overline{\Omega_s}$.

Generation of initial meshes and initial designs

Our implementation relies on a fully explicit mesh discretization of the hold-all domain D featuring the design shape $\Omega_f \subset D$ as a submesh. This is very convenient for solving finite element problems on Ω_f or Ω_s, since boundary conditions can then be applied in a straightforward manner on the boundary $\partial\Omega_f$ or $\partial\Omega_s$.

It is worth mentioning that generating an initial mesh featuring a correct topology respecting user-defined labels on the boundary triangles is an art by itself: there exist several software programs in order to do this (such as `TetGen` [283] or `gmsh` [167]). For our purposes, we found very convenient to do it in a level-set fashion assisted with the help of the library `mmg` [108]:

1. a first mesh $\mathcal{T}_0$ for the computational domain D is generated. Usually, D is a box so that this step is very easily achieved, for instance using a `FreeFEM` command of the kind

   ```
   // Generate a mesh for the box domain [0,2]*[0,1]*[0,1]
   mesh3 Th = cube(60, 40, 40, [x*2,y,z]);
   ```

2. a level set function ϕ is generated on the domain D such that the negative subdomain of ϕ on ∂D,

   $$\{x \in \partial D \,|\, \phi_0(x) < 0\}$$

   delimits a set of connected components on the boundary of D associated with various boundary conditions. The level set of ϕ_0 is explicitly discretized within $\mathcal{T}_0$ and remeshed with `mmg` so that a new mesh $\mathcal{T}_1$ where these boundary patches are explicitly meshed is obtained;

3. the obtained boundary patches are tagged with specific labels according to the desired boundary conditions (see e.g. the red patches on Figure 2.3 below);

4. an initial design $D = \Omega_s \cup \Omega_f$ is proposed by the user under the form of a level-set function ϕ_1 (satisfying $\phi_1 \leq 0$ in Ω_f). An initial mesh $\mathcal{T}$ of D featuring the initially proposed design for Ω_f discretized as a submesh and the correct boundary labels is obtained by discretizing the negative subdomain of ϕ_1 in $\mathcal{T}_1$. Note that boundary labels do not need to be rewritten neither at this step nor at subsequent stages of the optimization because the remeshing library `mmg` preserves meshed interfaces between any two regions labeled with distinct tags.

Classically, complex initial domains can be generated by applying min-max operations to level-set functions associated with elementary shapes (such as sphere or half-spaces), since minimum and maximum of two level-set functions amount to considering to respectively the intersection and the reunion of the corresponding subdomains [247].

Non optimizable subdomains

It is very customary to impose that some non-optimizable subdomains ω belong either to the fluid or the solid part Ω_f or Ω_s. A particular attention must be given to the treatment of this constraint during the computation step of a descent direction $\boldsymbol{\theta} \in W^{1,\infty}(D, \mathbb{R}^d)$: this deformation $\boldsymbol{\theta}$ should satisfy

$$\boldsymbol{\theta} \cdot \boldsymbol{n}_\omega \geq 0 \text{ on } \partial\omega \tag{6.1.2}$$

where $\boldsymbol{n}_\omega$ denotes the outward normal to ω. The constraint (6.1.2) ensures that the boundary of the optimized shape does not penetrate the non optimizable region ω. In principle, this requirement could be imposed as an additional constraint in the mathematical program determining the current descent direction from the knowledge of the shape derivatives. For instance, in the context of the unconstrained minimization of an objective functional $J(\Gamma)$, a "best" descent direction $\boldsymbol{\theta}$ taking into account the non optimizable region ω could be obtained by solving the quadratic minimization problem

$$\min_{\boldsymbol{\theta}\in V} \quad \mathrm{D}J(\Gamma)(\boldsymbol{\theta}) \qquad s.t. \begin{cases} \boldsymbol{\theta}\cdot\boldsymbol{n}_\omega \geq 0 \\ ||\boldsymbol{\theta}||_V \leq 1, \end{cases} \tag{6.1.3}$$

where $||\cdot||_V$ is the regularizing norm considered for the identification of $\mathrm{D}J(\Gamma)$ with a gradient (see chapter 1, section 1.4.1).

In practice, it is easier to solve (6.1.3) by replacing the inequality constraint $\boldsymbol{\theta}\cdot\boldsymbol{n}_\omega \geq 0$ with an equality constraint $\boldsymbol{\theta}\cdot\boldsymbol{n}_\omega = 0$, or even by simply imposing that $\boldsymbol{\theta} = 0$ on ω. This is precisely what we do in all our numerical examples with satisfactory results. The descent direction $\boldsymbol{\theta}$ solving the optimization problem (6.1.3) is then used as an advection field driving the evolution of a level-set function ϕ for the optimized shape Ω_f (see the step 2 of the algorithm algorithm 1.2 recalled in chapter 1). In principle, imposing $\boldsymbol{\theta}\cdot\boldsymbol{n}_\omega \geq 0$ on $\partial\omega$ should guarantee at the continuous level that the interface of the optimized shape do not penetrate $\partial\omega$. Unfortunately, due to numerical inaccuracies, too large time-steps, or incorrect initial designs, a post-treatment of the advected function is used in order to correct small violations of the non-optimizable region. Let us introduce

- a level-set function Φ_ω associated with the non-optimizable region ω, i.e.

$$\forall x \in \omega,\ \Phi_\omega(x) < 0 \text{ and } \forall x \in D\backslash\overline{\omega},\ \Phi_\omega(x) \geq 0;$$

- a level-set function Ψ associated with the distribution $\mathcal{X} \subset \omega$ of material desired in the non-optimizable region of ω, i.e. it should always hold

$$\{x \in \omega \,|\, \Psi(x) < 0\} \subset \Omega_f \text{ and } \{x \in \omega \,|\, \Psi(x) > 0\} \subset \Omega_s.$$

Then denoting by ϕ the obtained level set function for the next iteration after the advection step, we perform the operation (see Figure 6.2 below)

$$\phi \leftarrow \min(\max(\phi, -\Phi_\omega), \max(\Psi, \Phi_\omega)), \tag{6.1.4}$$

which corresponds to the domain update

$$\Omega_f \leftarrow (\Omega_f \cap (D\backslash\overline{\omega})) \cup \mathcal{X}. \tag{6.1.5}$$

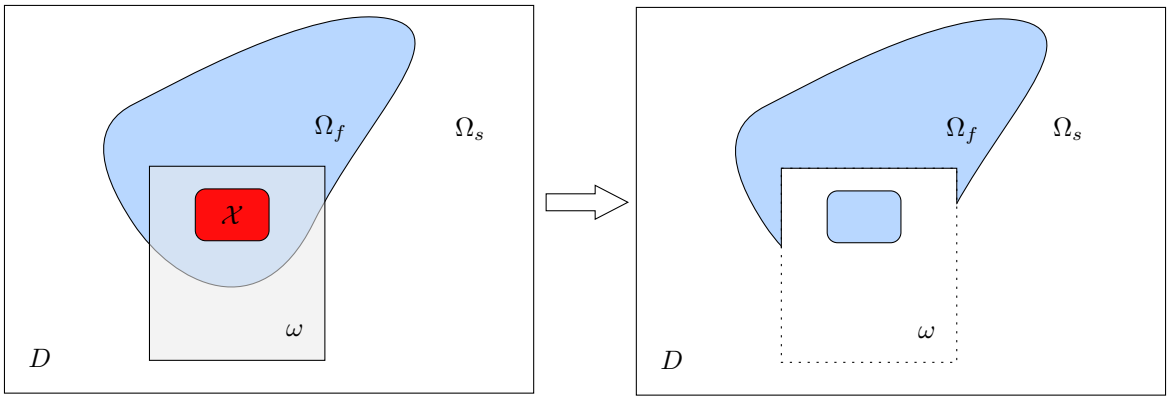


Figure 6.2: Illustration of the level-set operation (6.1.4) enforcing non-optimizable regions ω: the distribution of material Ω inside the domain ω should match exactly the red set $\mathcal{X}$. After the operation, Ω is the new blue color domain on the right.

Note that the post-treatment (6.1.4) is only a "projection" of Ω_f onto the set of shapes fulfilling the requirement that $\Omega_f \cap \omega = \mathcal{X}$. To be consistent with this operation, i.e. to guarantee that it does not affect the decreasing property (6.1.3) of $\boldsymbol{\theta}$, it is very important to impose the condition (6.1.2), or at least $\boldsymbol{\theta} \cdot \boldsymbol{n}_\omega = 0$ on $\partial\omega$.

Remark 6.1. In the context where a fixed computational domain D containing the optimized shape Ω is used, the boundaries of D should be set as non-optimizable as well, because deformations making Ω larger than D should not be allowed. A slight difference of implementation was considered regarding this point whether we use either the surface or the volume expression of the shape derivative:

- if the surface expression (2.4.14) of the shape derivative is used, then the considered space of deformation fields is
$$V = \{v\nabla d_{\Omega_f} \,|\, v \in H^1(D)\},$$
that is the set of vector fields aligned with the constant extension of the normal to $\partial\Omega_f$. Since shape derivatives are written as surface integrals over $\Gamma \subset \partial\Omega_f$, and we make sure numerically that Γ never becomes a subset of ∂D, no constraint is required on the value of v regarding the fact that ∂D is a non-optimizable boundary.

- if the volume expression (2.4.14) of the shape derivative is used, then
$$V = H^1(D, \mathbb{R}^d)$$
and it is necessary to enforce that $\boldsymbol{\theta} \cdot \boldsymbol{n}_D \geq 0$ on ∂D where $\boldsymbol{n}_D$ is the outward normal to D. In our implementation, we rather enforce $\boldsymbol{\theta} \cdot \boldsymbol{n}_D = 0$ in the identification problem (6.1.3).

Enforcing symmetries of the iterated domains

Many optimization test cases feature inherent domain symmetries; it is then expected that locally optimal designs for these problems should also be symmetric. In many contributions of the literature, symmetry is used in order to reduce the size of the computational domain: the overall optimization is performed on only one part of the domain D, and the final design on D as a whole is inferred by symmetry.

However, it is sometimes desirable to perform the computations involved involved in the resolution of the shape optimization problem on the whole domain D and to symmetrize shapes as a post-treatment. Such a post-treatment is needed because in general, numerical inaccuracies tend to quickly make optimized shapes nonsymmetric when the computational mesh is not symmetric. In our implementation where each intermediate design is obtained with a level set function ϕ, the symmetry of ϕ with respect to a given symmetry transformation S_1 (satisfying $S_1 \circ S_1 = I$) is enforced by the operation

$$\phi \leftarrow \frac{1}{2}(\phi + \phi \circ S_1) \qquad \text{(for one symmetry } S_1\text{)}.$$

Likewise, invariance of the design with respect to two symmetry transformations S_1, S_2 (satisfying $S_1 \circ S_1 = S_2 \circ S_2 = I$) is imposed via

$$\phi \leftarrow \frac{1}{5}(\phi + \phi \circ S_1 + \phi \circ S_2 + \phi \circ S_1 \circ S_2 + \phi \circ S_2 \circ S_1) \qquad \text{(for two symmetries } S_1, S_2\text{)}.$$

The compositions $\phi \circ S_1$, $\phi \circ S_2$ are computed by interpolation (which is an automatic feature in `FreeFEM`). These transformations ensure that the updated level set function satisfies $\phi \circ S_1 = \phi$ and $\phi \circ S_2 = \phi$.

Regularization of the level-set function to avoid mesh degeneracy

A very important ingredient for handling topological changes in the context of the mesh evolution method of [24] is a regularization of the level-set function. After the step 2 of algorithm 1.1, a level-set function ϕ is obtained at the vertices of a computational mesh of D whose negative subdomain corresponds to the new shape. A new mesh of D is then created in which this updated shape exists as a submesh (step 3 in algorithm 1.1). Before performing this remeshing step, ϕ is regularized into a new level set function $\widetilde{\phi}$ by solving the elliptic problem

$$\text{Find } \widetilde{\phi} \in H^1(D), \text{ such that } \forall \psi \in H^1(D), \int_D (\gamma^2 \nabla\widetilde{\phi} \cdot \nabla\psi + \widetilde{\phi}\psi)\mathrm{d}x = \int_D \phi\psi \mathrm{d}x. \tag{6.1.6}$$

The regularizing length scale is typically set to $\gamma = 0.01$`hmin` where `hmin` is the desired *minimum* mesh edge size. It may appear surprising that one regularizes ϕ at a scale lower than the mesh size. In fact, the minimum mesh size `hmin` provided to `mmg` might be violated if the input topology of the mesh prescribes it: for instance, `mmg` would never remove a bubble of arbitrary size in an input mesh because the topology of the computed output mesh would then change (see [108]). If the Hadamard's shape derivative dictates to continuously decrease the size of these bubbles at every optimization step, the resulting successive meshes quickly degenerate and the finite element resolution becomes impossible. Removing these bubbles (or any other topological features much smaller than `hmin`) by the pretreatment (6.1.6) allows to avoid these issues.

Handling connectivities of the meshes for finite element related operations.

A particular care must be paid in the implementation of the variational formulations (2.4.2) to (2.4.4) for the state equations, or in the calculation of the shape derivatives (2.4.13) and (2.4.14). Indeed, the state variables at play $(\boldsymbol{v}, p), T, \boldsymbol{u}$ (as well as the adjoint variables) are living on different meshes: the fluid variables $(\boldsymbol{v}, p)$ and the elastic variable $\boldsymbol{u}$ are for instance solved on submeshes of the global meshed domain D corresponding respectively to the fluid or to the solid subdomains Ω_f, Ω_s. As a result, interpolation matrices are extensively used in order to transfer nodal information from one mesh to another (e.g. for transferring the value of $\sigma_f(\boldsymbol{v}, p)$ on the *fluid* interface Γ seen as a boundary of the meshed domain Ω_f, to the *solid* interface Γ seen as a boundary of the meshed domain Ω_s). This is achieved in `FreeFEM` thanks to the instruction `interpolate`;

6.1.4 Domain decomposition and preconditioning for 3-d variational problems

Passing from the 2-d implementation to its 3-d counterpart is *theoretically* without difficulty, but it requires in practice a substantial amount of effort. We shall not discuss the (quite important) differences between 2-d and 3-d regarding *remeshing* issues (the reader is again referred to [108]); we shall focus instead on the difficulties related to the resolution of variational problems by the finite element method.

The cornerstone of the passage from 2-d to 3-d lies in the assembly and inversion of large sparse linear systems obtained from the discretization of the physical equations of chapter 2, (2.4.2) to (2.4.4). Generally, linear systems resulting from 2-d applications are sufficiently small so that a direct factorization based method can be used [305]. In 3-d, it is possible to use direct methods only for very low resolution problems: indeed, modern direct solvers based on LU factorizations such as `MUMPS` [38] have a complexity of order $O(N^2)$ (N is the number of degrees of freedom of the finite element approximation) which becomes quickly too expensive in terms of both CPU time and required memory. In this context, one classically resorts to *iterative* methods merely based on matrix-vector products, which are relatively inexpensive to compute due to the sparsity of the matrices involved in the context of finite element problems. The most popular iterative methods are the conjugate gradient method (CG) for symmetric positive definite problems, and GMRES for the general non symmetric case.

The treatment of large sparse systems as usually encountered in large scale 3-d applications involves two additional ingredients:

1. *preconditioning*: iterative methods may take many iterations to converge in reasonable CPU time for ill-conditioned linear systems [173]. It is often possible to accelerate the resolution of such linear systems

$$Ax = b, \tag{6.1.7}$$

by left or right multiplying the (large) square matrix A and the vector b with a *preconditioner* M:

$$MAx = Mb \text{ or } AMy = b. \tag{6.1.8}$$

A good preconditioner M is a square matrix approximating well the inverse of A. In that case, it is expected that iterative methods applied to (6.1.8) will converge in much less iterations than when applied to the original problem (6.1.7).

2. *domain decomposition*: the computational domain D is divided into a number m of subdomains: $D = D_1 \cup \cdots \cup D_m$. Very roughly, the inverses of the "restrictions" $A_1, \ldots, A_m$ of the finite element

matrix A to these subdomains are used to build a block preconditioner

$$M := \begin{bmatrix} A_1^{-1} & & \\ & \ddots & \\ & & A_m^{-1} \end{bmatrix}.$$

This allows to distribute the resolution of the linear system (6.1.7) on multiple cpus, because all operations involving the restriction inverses $A_1^{-1}, \ldots, A_m^{-1}$ (or their approximation using a limited number of GMRES iterations) can be performed in parallel.

For our applications, we rely on the PETSc library [54, 55, 56] and its interface in `FreeFEM` developed by Jolivet [199, 132] which allows to solve finite element problems with a large library of state-of-the-art preconditioned iterative methods. The domain decomposition step was achieved thanks to a macro `buildMinimalist` (see [198]):

```
// Partition the mesh Th for the finite element space Pk and compute
// associated partition of unity D and connectivity arrays arrayIntersection
// and restrictionIntersection
buildMinimalist(Th, D, arrayIntersection, restrictionIntersection, Pk);
//Now, Th is one of the submeshes of the domain decomposition [...]
```

Importantly, the use of these techniques requires a significant amount of effort, because the whole implementation needs to be thought parallel for scalability; this includes operations ranging from the finite element matrix assembly, the evaluation of volume or surface integrals, up to the numerical assembly and regularization of shape derivatives (2.4.13) and (2.4.14).

We now detail the choice (physics dependent) preconditioners for the weakly coupled fluid thermal mechanical system (2.2.1) to (2.2.3). For these matters, we have been very much assisted by Pierre Jolivet and his tutorial on `FreeFEM` [198].

Linear elasticity: the linear elasticity system (2.2.3) is an elliptic vector system. The Geometric Algebraic Multigrid preconditioner (GAMG) is known to be very efficient for solving linear elasticity problems [8] (among other smoothed aggregation methods). The idea of multigrid methods is to construct a preconditioner by using the (iterative) inverses of the elasticity operator restricted to coarser meshes; these are much cheaper to compute and allow to obtain good approximations of the low frequency component of the solution.

We used the GAMG preconditioner in our `FreeFEM` implementation by calling

```
// set GAMG preconditioner for elasticity matrix AElasticity
string petsc_options_elasticity = "-pc_type gamg -ksp_type gmres -ksp_max_it 200"
    +" -pc_gamg_threshold 0.01";
set(AElasticity, sparams = petsc_options_elasticity, nearnullspace = Rb);
```

The variable `Rb` contains the near null space of the elasticity operator (i.e. rigid body modes on the whole computational domain), it is required to help the preconditioner to achieve good performance.

Heat conduction-convection For the convective heat problem (2.2.2), we rely on the pointwise aggregation multigrid preconditioner (BoomerAMG) implemented in the `hypre` library [149] which is supposed to be slightly more efficient than `gamg` for scalar problems (in fact, for our moderately large applications, the difference was not obvious). This preconditioner was used in our implementation by calling

```
// set BoomerAMG preconditioner for the thermic system matrix AThermic
string petsc_options_thermic = "-pc_type hypre";
set(AThermic, sparams = petsc_options_thermic, nearnullspace = Rb);
```

This time, the near null space variable `Rb` corresponds to the constant functions. Note that for these first two considered physics, we observed that the use of the standard conjugate gradient method converges well without the need for a preconditioner (the latter allows nevertheless to significantly speed

up this convergence). This situation is very particular to the fact that the partial differential operators corresponding to these physics are elliptic (if no convection is involved).

Navier-Stokes equations Solving the 3-d steady state Navier-Stokes equations is much more challenging than the previous elasticity and heat conduction problems, because unpreconditioned iterative solvers fail to converge in a reasonable amount of time.

Our implementation of the resolution of the steady state incompressible Navier-Stokes equations relies on the Augmented Lagrangian Preconditionner recently described in [238, 279] and for which the source code is available in `FreeFEM`. The main ingredient of this method is the addition of a penalization term for the divergence constraint in the variational formulation (2.4.2) associated with the nonlinear Navier-Stokes problem (2.2.1):

Find $(\boldsymbol{v}, p) \in V_{\boldsymbol{v},p}(\Gamma)$ such that $\forall (\boldsymbol{w}, q) \in V_{\boldsymbol{v},p}(\Gamma)$,

$$\int_{\Omega_f} [\sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{w} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v} \cdot \boldsymbol{v} - q\mathrm{div}(\boldsymbol{v}) - p\mathrm{div}(\boldsymbol{w}) + \gamma \mathrm{div}(\boldsymbol{v})\mathrm{div}(\boldsymbol{w})]\,\mathrm{d}x = \int_{\Omega_f} \boldsymbol{f}_f \cdot \boldsymbol{w}\mathrm{d}x. \quad (6.1.9)$$

The coefficient $\gamma > 0$ penalizes the constraint $\mathrm{div}(\boldsymbol{v}) = 0$. The resolution of the nonlinear problem (6.1.9) is performed with the Newton method: at each step $k \geq 0$ of the process, an increment $(\delta \boldsymbol{v}_k, \delta p_k)$ is computed by solving the linearization of (6.1.9) around $(\boldsymbol{v}_k, p_k)$:

Find $(\delta \boldsymbol{v}_k, \delta p_k) \in V_{\boldsymbol{v},p}(\Gamma)$ such that $\forall (\boldsymbol{w}, q) \in V_{\boldsymbol{v},p}(\Gamma)$,

$$\int_{\Omega_f} [\sigma_f(\delta \boldsymbol{v}_k, \delta p_k) : \nabla \boldsymbol{w} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v}_k \cdot \delta \boldsymbol{v}_k + \rho \boldsymbol{w} \cdot \nabla(\delta \boldsymbol{v}_k) \cdot \boldsymbol{v}_k - q\mathrm{div}(\delta \boldsymbol{v}_k) - \delta p_k \mathrm{div}(\boldsymbol{w})]\,\mathrm{d}x$$
$$+ \int_{\Omega_f} \gamma \mathrm{div}(\delta \boldsymbol{v}_k)\mathrm{div}(\boldsymbol{w})\mathrm{d}x = \int_{\Omega_f} \boldsymbol{f}_f \cdot \boldsymbol{w}\mathrm{d}x. \quad (6.1.10)$$

The next iterate $(\boldsymbol{v}_{k+1}, p_{k+1})$ is then obtained by setting

$$\boldsymbol{v}_{k+1} := \boldsymbol{v}_k + \delta \boldsymbol{v}_k, \quad p_{k+1} := p_k + \delta p_k.$$

In our implementation, the initial guess $(\boldsymbol{v}_0, p_0)$ is the solution of the Stokes counterpart problem to (6.1.9) (obtained with $\rho = 0$). The difficult part of the method is the resolution of the so-called Oseen problem (6.1.10) which turns to be very poorly ill-conditioned. For our application we rely on the Augmented Lagrangian preconditioner of [238]. It relies on the use of a suitable preconditioner for the block matrix discrete operator associated to the problem (6.1.10), which is of the form:

$$\texttt{AOseen} = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \quad (6.1.11)$$

where A is the matrix discretizing the bilinear form

$$(\delta \boldsymbol{v}_k, \boldsymbol{w}) \mapsto \int_{\Omega_f} [\sigma_f(\delta \boldsymbol{v}_k, \delta p_k) : \nabla \boldsymbol{w} + \rho \boldsymbol{w} \cdot \nabla \boldsymbol{v}_k \cdot \delta \boldsymbol{v}_k + \rho \boldsymbol{w} \cdot \nabla(\delta \boldsymbol{v}_k) \cdot \boldsymbol{v}_k + \gamma \mathrm{div}(\delta \boldsymbol{v}_k)\mathrm{div}(\boldsymbol{w})]\,\mathrm{d}x$$

and B is the discretization of the divergence operator $(\boldsymbol{w}, q) \mapsto \int_{\Omega_f} q\mathrm{div}\boldsymbol{w}\mathrm{d}x$. The linear solver GMRES is very slow to converge if it is called directly on (6.1.11). A preconditioner is obtained by using the inverse of A (or an approximation of it); following [238], we use the (right) preconditioner `MOseen` given by

$$\texttt{MOseen} := \begin{bmatrix} A^{-1} & 0 \\ B & S, \end{bmatrix} \quad (6.1.12)$$

where the inverse A^{-1} is estimated thanks to an additive Schwarz method (ASM; see [313]). The matrix S is an approximation of the pressure Schur complement $S = -BA^{-1}B^T$, it is estimated from its inverse $S^{-1} = -(\gamma + 1/\mathrm{Re})M_p$ where M_p is the mass matrix associated with $(p, q) \mapsto \int_{\Omega_f} pq\mathrm{d}x$ and $\mathrm{Re} = \rho/\nu$ is the Reynolds number (see [238] for the details). Many variants of this technique using A^{-1} to obtain a preconditioner exist, which can be related to Uzawa-type methods [52]. Note that setting $\gamma > 0$ in (6.1.9) is required for the (flexible) GMRES solver to converge.

To be very specific, this preconditioner is implemented in `FreeFEM` thanks to the following lines of code:

```
// set Augmented Lagrangian  preconditioner for the Oseen matrix AOseen
string paramsOseen = " -ksp_type fgmres -pc_type fieldsplit "
    + " -pc_fieldsplit_type multiplicative -ksp_atol 1e-6";
    + " -fieldsplit_velocity_pc_type asm -fieldsplit_velocity_pc_asm_overlap 1 "
    + " -fieldsplit_velocity_sub_pc_type lu "
    + " -fieldsplit_velocity_sub_pc_factor_mat_solver_type mumps";
    + " -fieldsplit_velocity_ksp_type gmres -fieldsplit_velocity_ksp_rtol 1e-1 "
    + " -fieldsplit_velocity_ksp_pc_side right"
    + " -fieldsplit_velocity_ksp_restart 50";
    + " -fieldsplit_pressure_pc_type jacobi -fieldsplit_pressure_ksp_type cg "
    + " -fieldsplit_pressure_ksp_max_it 5";
set(AOseen, sparams=paramsOseen, fields=fields[], names=names,
        schurPreconditioner = S, schurList = listX[]);
```

The pressure block term S is inverted by a CG method. The parameter `listX` refers to the matrix blocks with the "`velocity`" or "`pressure`" labels.

Note that we also use this preconditioner for the resolution of the fluid adjoint problem (6.2.10) which involves the transpose of the Oseen matrix. Note that some adaptation of this method would be needed if the objective function were to depend on the pressure, because the divergence $\mathrm{div}(\boldsymbol{w})$ of the fluid adjoint variable $\boldsymbol{w}$ would not equal zero.

Comparative performance of various stages of the parallel finite element implementation We report in Table 6.1 various running times corresponding to the main operations performed during one iteration of our optimization algorithm. For simplicity, we considered a situation where only the fluid physics is involved and modeled by a Stokes system. More precisely, the following operations of our overall implementation are accelerated by parallel computing thanks to the use of domain decomposition:

1. all steps required for solving the state equations (2.2.1) to (2.2.3);
2. the computation of the values of the objective and constraint functions;
3. all steps required for computing the shape derivative of these functionals, including the resolution of adjoint systems;
4. the resolution of identification problems of the form (1.4.4) which feature the linear inversion of an elliptic problem.

We note, in Table 6.1, that the most computationally expensive tasks consisted in the resolution of the Stokes problem and its adjoint, as well as the various finite element matrix or vector assembly steps. All things considered, the scaling obtained is quite satisfactory and allowed us to run fluid shape optimization test cases in reasonable CPU time.

6.2 A FEW (MODERATELY) LARGE-SCALE THREE DIMENSIONAL MULTIPHYSICS APPLICATIONS

In this whole section, we go back to the three physics setting of chapter 2: a computational domain $D = \Omega_s \cup \Omega_f \subset \mathbb{R}^3$ is given, which is the disjoint union of solid and fluid phases Ω_s and Ω_f. The behavior of the fluid-solid system is described by the weakly coupled system of partial differential equations (2.2.1) to (2.2.3), which determines the fluid velocity and pressure fields $(\boldsymbol{v}, p)$ in Ω_f, the:temperature field T in D and the elastic displacement $\boldsymbol{u}$ for Ω_s. The ultimate goal is to optimize the shape of the interface $\Gamma = \Omega_s \cap \Omega_f$ (a three dimensional surface) in order to solve constrained optimization problems of the form

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \\ s.t. \quad & \begin{cases} g_i(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) = 0, & 1 \le i \le p, \\ h_j(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma), T(\Gamma), \boldsymbol{u}(\Gamma)) \le 0, & 1 \le j \le q, \end{cases} \end{aligned} \tag{6.2.1}$$

where the notation convention is that introduced in chapter 2. In what follows, we treat four instances of the program (6.2.1). Our first three sets of examples involve only one physics at a time. In section 6.2.1, we examine very classical problems in linear elasticity, by considering the optimization of mechanical

Action	1 cpu	4 cpus	8 cpus
Loaded meshes	0.48	0.48	0.49
Partitioning	0.05	7.0	6.5
Saved partitioning data	0.82	0.3	0.21
Build Stokes Matrix and Rhs	10.73	2.7	1.38
Solved Stokes problem	56.42	18.72	10.98
Built interpolate Fhf1vp matrix	0.38	0.16	0.09
Computed objective J	6.96	2.12	1.17
Read partitioning data	1.53	0.71	0.51
Adjoint fluid matrix assembly	18.8	5.51	2.59
Connectivity operation for non optimizable subdomains	3.97	1.27	0.69
Riesz matrix built	4.14	1.31	0.71
Assemble adjoint fluid RHS	6.25	1.8	0.89
Adjoint fluid problem resolution	103.84	38.04	23.24
Assemble volumetric shape derivative	72.68	20.9	10.92
Identification of the shape derivative of J to a gradient	1.67	0.66	0.5
Total `FreeFEM` running CPU time	308.14	112.53	71.55

Table 6.1: Running cpu times (2.60 GHz) for all finite element operations processed at every optimization iteration. The test case considered is the drag minimization around an obstacle for a 3-d Stokes problem featuring approx 22,000 vertices.

structures subject to either traction or torsion loads. The originality of our work is the treatment of these problems with the mesh evolution technique of [24] at much higher resolutions than in the seminal paper. We also discuss qualitative differences between the use of the surface expression (2.4.14) of the shape derivative and that of the volume expression (2.4.13).

The next section 6.2.2 considers an optimal design problem in pure heat conduction, which is a 3-d extension of the 2-d case treated in section 2.5.7. Although the physics at play is the least complicated to solve among the four considered situations (it involves only a scalar elliptic problem), the obtained final design is very intricated and illustrates well the efficiency of our mesh evolution method and the remeshing library `mmg`.

The third context of interest is that of shape and topology optimization in fluid mechanics; we tackle in section 6.2.3 the very classical problem of finding optimal aerodynamic designs with respect to the induced lift and drag forces. This problem has been the object of much effort in the literature, see e.g. [191, 260]; however, these contributions most often consider industrial contexts featuring very high Reynolds numbers and where the shape to optimize is parameterized by a small number of parameters (which make sense because very small design update can lead to a substantial increase of performance). Often, automatic differentiation is used rather than Hadamard's shape derivatives in order to obtain the sensitivity to these parameters in the context of the resolution of the physics with industrial codes. The novelty of our work is the application of our *topology optimization* method, relying on analytic Hadamard's shape derivatives, which allows to compute optimal aerodynamic designs (at low Reynolds number) without resorting to any parameterization of the shape.

Finally, our fourth and last test case of section 6.2.4 features two weakly coupled physics: a vertical plate is pushed down by a fluid; the problem at hand is to find a distribution of solid material around the plate in order to make the whole structure the least compliant as possible. This is the most computationally involved test case considered in this thesis: discretization meshes at play contained up to 250,000 vertices in the fluid domain (and the same in the solid domain), which required the resolution of linear systems featuring more than 2×10^6 degrees of freedom.

All these results can still be considered as preliminary: future work will seek to find optimal designs for convective heat transfer problems where both fluid mechanics and thermal conduction interact. Many

improvements could be considered in order to reach much larger, industrial size problems: one of the limiting factor which prevented us to obtain such results lies in the fact that many important steps of the optimization algorithm are still *sequential*, including for instance the remeshing step or the computation of the signed distance function.

6.2.1 Cantilever beam subject to traction or torsion loads

We start by reproducing the classical benchmark test case of a 3-d cantilever beam subject to either a flexural or a torsion load. The computational domain D is a box of dimensions $2 \times 1 \times 1$. The solid structure is fixed at four squares of size 0.3×0.3 located on the left-hand side of the boundary ∂D as depicted on Figure 6.3. A force $\boldsymbol{g}$ is applied on a disk-shaped region at the center of the right-hand side of ∂D; two cases are considered:

- *traction load*: we set $\boldsymbol{g} := -\boldsymbol{e}_y$ where $(\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z)$ denotes the canonical basis of $\mathbb{R}^3$, i.e. the load $\boldsymbol{g}$ is vertical, pointing downward;
- *torsion load*: we set

$$\boldsymbol{g} = \begin{pmatrix} 0 \\ \frac{(z-0.5)}{\sqrt{(y-0.5)^2+(z-0.5)^2}} \\ -\frac{(y-0.5)}{\sqrt{(y-0.5)^2+(z-0.5)^2}} \end{pmatrix}$$

  which corresponds to a torsion force field.

The goal is to minimize the compliance of the structure Ω_s under a volume constraint:

$$\begin{aligned} \min \quad & J(\Omega_s, \boldsymbol{u}(\Omega_s)) := \int_{\Omega_s} Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_s) := \int_{\Omega_s} \mathrm{d}x = V_{target}. \end{aligned} \tag{6.2.2}$$

where V_{target} is a target volume, set to 0.15 in this example.

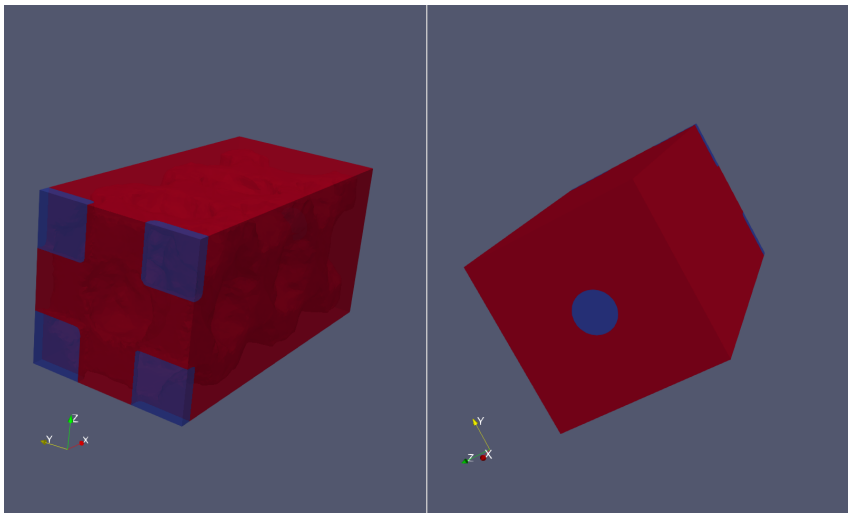


Figure 6.3: Computational domain for the 3-d cantilever test case of section 6.2.1 subjected to a bending load. The blue regions correspond to four square-shaped Dirichlet boundaries to which the whole structure is fixed, and a disk-shaped Neumann boundary to which either a traction or a torsion force field is applied.

The former situation, involving traction load has already been treated with the level set mesh evolution algorithm in [24], with a much smaller resolution however (18,081 vertices in [24] vs. 108,605 in our case for the first mesh).

Note that for the four numerical examples presented below, convergence was not fully attained as the objective function was still decreasing at the end of the performed iterations. However it is expected that more iterations would not substantially change the physical outline of the shape.

Traction test case Two numerical results are shown for the traction test case on Figs. 6.4 and 6.5 corresponding to the use of respectively the surface and volume expressions of the shape derivative. The optimization paths are quite different. Qualitatively, the use of the surface expression of the shape derivative (identified with a gradient via the inner product of chapter 1, (1.4.8)) seems to favor the creation of holes while less topological changes occur when the volume expression is used (identified with a gradient via the inner product of chapter 1, (1.4.7)) which favors the occurrence of walls rather than bars on this example. These plots further illustrate the use of remeshing: the level set mode of `mmg` (option `-ls`) makes it possible to obtain high quality meshes of the shape at every optimization iteration.

The convergence histories for the objective and constraint functions in both cases are provided in Figure 6.6. The final designs obtained by using either the surface or volume expression of the shape derivative show very similar performance.

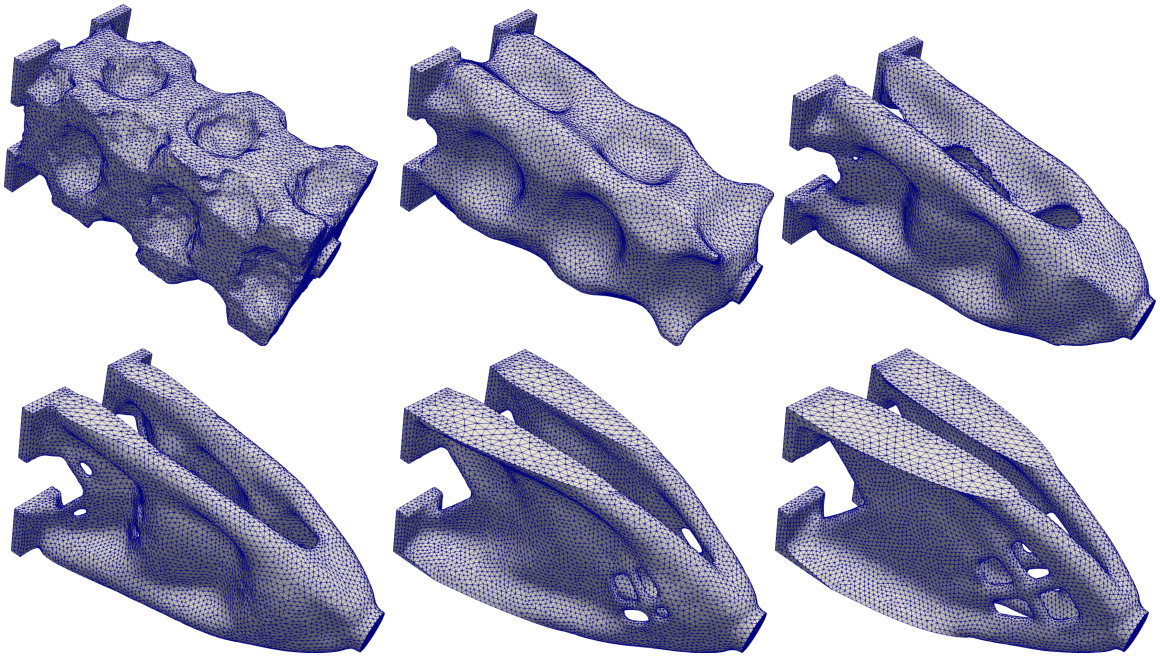


Figure 6.4: From top left to bottom right: iterations 0, 10, 20, 40, 100 and 400 of the optimization of a 3-d cantilever under *traction* load (section 6.2.1) using the *surface* expression of the shape derivative.

Torsion test case Optimization results for the *torsion* test case are shown on Figs. 6.7 and 6.8 corresponding to the use of either the surface or volume expressions of the shape derivative. The convergence histories for the objective and constraint functions are plotted on Figure 6.9. For this test case, a clear difference is observed between both computations the use of the volume expression favors a shell-shaped structure without bars at the upper end at the end of the iterations, which seems to be more efficient at least for the first iterations (note again that convergence is not fully attained).

6.2.2 Optimal design for pure thermal heat conduction

We now focus on the optimization of a pure heat conduction test case (only the thermal equation of chapter 2, (2.2.2) is solved). The problem considered is the direct extension to 3-d of the test case previously treated in chapter 2, section 2.5.3. The setting is represented on Figure 6.10: the hold-all domain D is a box with size $1 \times 1 \times 1$. It is divided between two phases Ω_s with (low) conductivity $k_s = 1$ and Ω_f with (high) conductivity $k_f = 100$. A Dirichlet boundary condition is imposed on a small square of size 0.4×0.4 at the bottom face of ∂D where the temperature is prescribed to $T = 0$. All other external boundaries of the cube D are adiabatic ($\partial T/\partial \boldsymbol{n} = 0$). The whole domain is heated with a source $Q_s = Q_f = 10^4$ and the goal is to find the shape of the interface $\Gamma = \partial\Omega_s \cap \partial\Omega_f$ of the two materials which minimizes the average temperature over D subject to a volume constraint:

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, T(\Gamma)) = \int_D T \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_f) = V_{target}. \end{aligned} \tag{6.2.3}$$

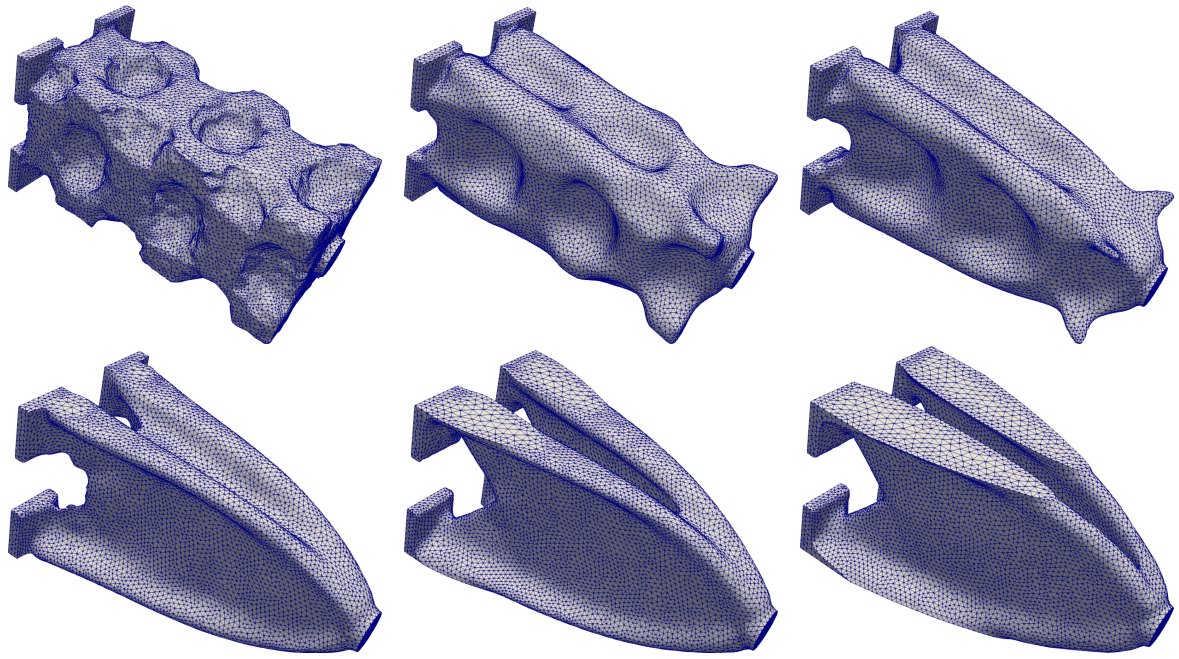
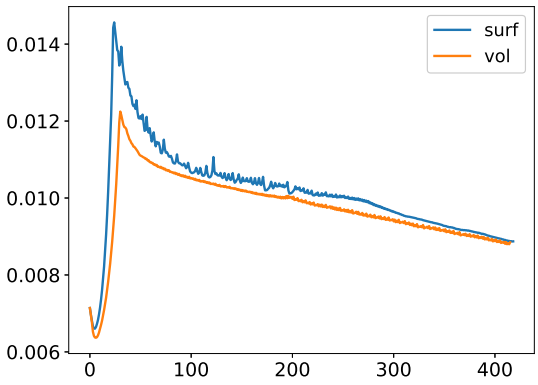
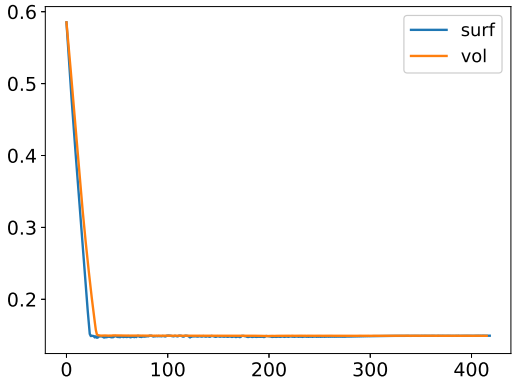


Figure 6.5: From top left to bottom right: iterations 0, 10, 20, 40, 100 and 280 of the optimization of a 3-d cantilever subjected to a *traction* load (section 6.2.1) using the *volume* expression of the shape derivative. Note the angle view is not the same as in Figure 6.4 for better visualization.



(a) Objective function $J(\Omega_s)$.



(b) Volume fraction $\mathrm{Vol}(\Omega_s)/\mathrm{Vol}(D)$.

Figure 6.6: Convergence histories for the 3-d cantilever test cases subjected to a *traction* load of section 6.2.1 using either the *surface* or *volume* expression of the shape derivative.

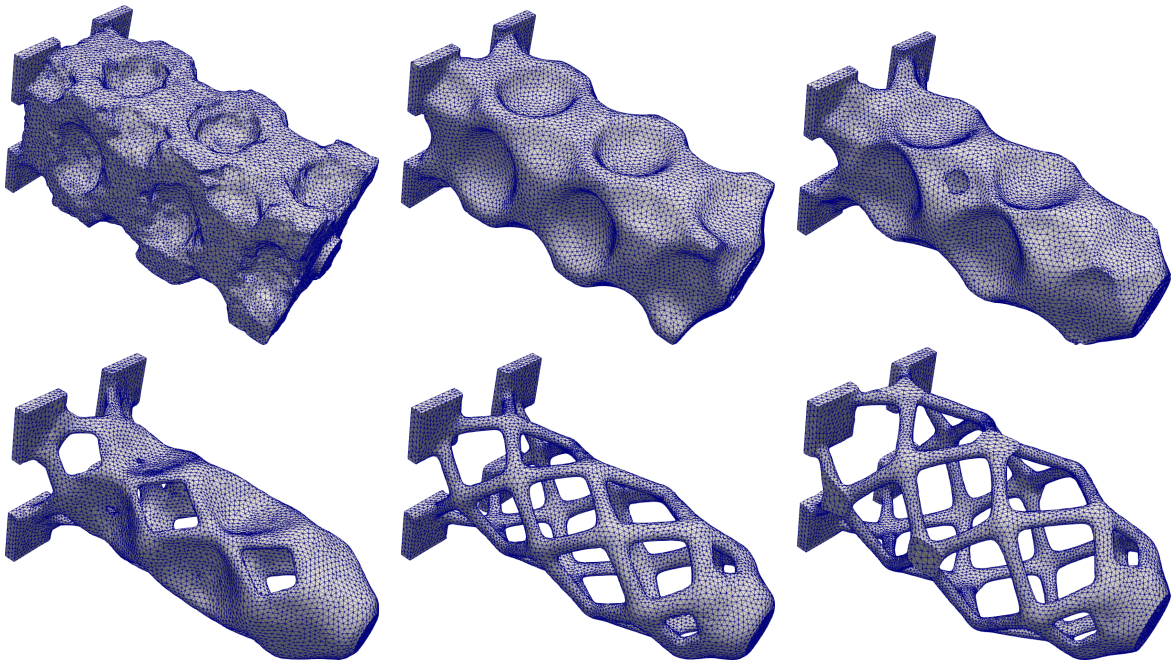


Figure 6.7: From top left to bottom right: iterations 0, 10, 20, 40, 100 and 400 of the optimization of a 3-d cantilever subjected to a *torsion* load (section 6.2.1) using the *surface* expression of the shape derivative.

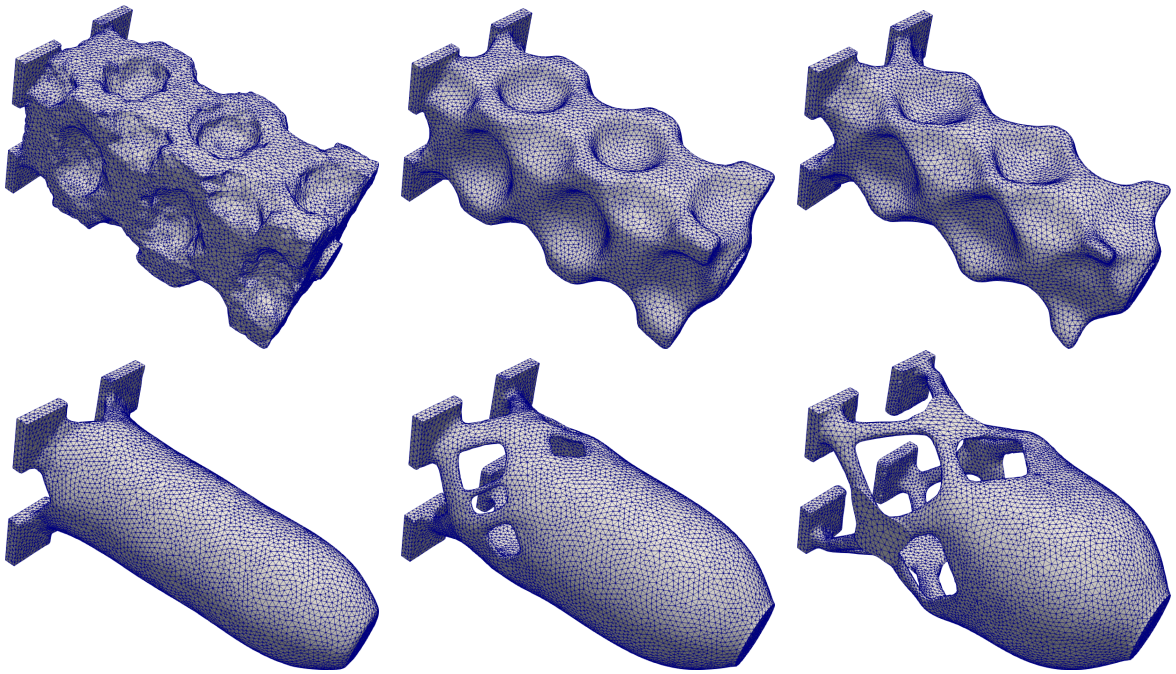
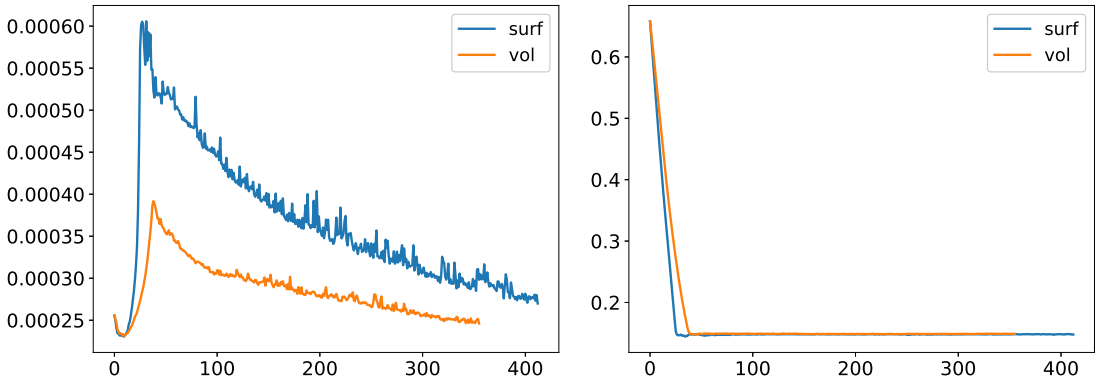


Figure 6.8: From top left to bottom right: iterations 0, 10, 20, 40, 100 and 350 of the optimization of a 3-d cantilever subjected to a *torsion* load (section 6.2.1) using the *surface* expression of the shape derivative.



(a) Objective function. (b) Constraint G0.

Figure 6.9: Convergence histories for the 3-d cantilever test case (subjected to a *torsion* load) of section 6.2.1 using either the *surface* or *volume* expression of the shape derivative.

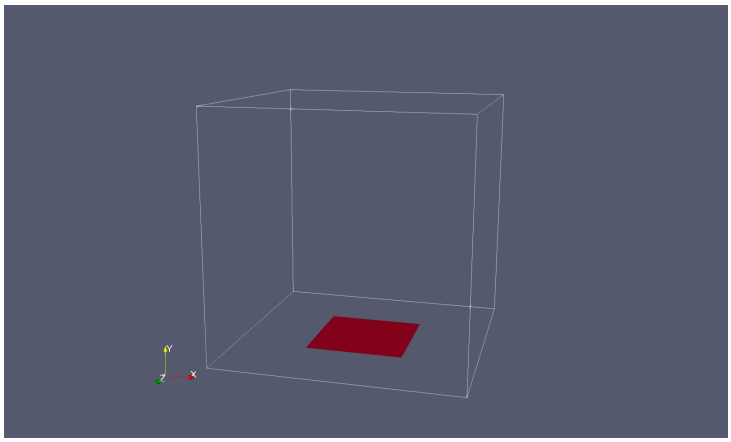


Figure 6.10: Setting for the pure conduction test case.

In this three-dimensional context, the volume constraint is set to $V_{target} := 0.05$.

Results are reported on Figs. 6.11 and 6.12, relying on the *volume* expression for the shape derivative. The resolution of the mesh varies from 63,761 vertices for the initial design to 206,464 for the final design. Remarkably, the remeshing software `mmg` and our optimization method are able to capture sheet like structures not thicker than one or two mesh element size. For this numerical example, the difference with the use of the surface expression of the shape derivative is very striking (we report a design obtained with the surface expression on Figure 6.13): the optimized shape presents much less details in the latter case and achieves a worse performance, as illustrated on Figure 6.14. Naturally, a lot of other factors come into play which could also account for the differences (such as the level of regularization of the shape derivative), for which we do not have a definitive explanation.

6.2.3 Lift–Drag topology optimization for aerodynamic design

This section investigates 3-d shape and topology optimization for lift-drag problems in aerodynamic design. The lift functional is the vertical force generated by a flow around an obstacle; it is commonly defined as a *surface integral* involving the normal stress tensor. The drag is the energy dissipated by the fluid around the obstacle. Lift-Drag shape optimization is a very classical problem which has been the object of a very large amount of contributions, see e.g. [259, 236, 191, 192, 142, 211, 175]. However, these references have considered situations very close to realistic applications where

(i) the physics is more challenging than in our case, featuring for instance compressible fluids or characterized by much larger Reynolds numbers;

(ii) the shape design is usually described by means of CAD parameters to optimize;

(iii) very small updates of the shape may lead to substantial gains of efficiency. In this context, it makes sense to seek for improved geometries by means of very small deformations of the proposed CAD design.

Few works have actually tried to apply shape and topology optimization techniques, were the design shape is allowed to deform freely, to lift-drag problems: we are essentially aware of [100, 205, 164, 307]. In what follows, we treat a lift-drag optimization problem with the method of Hadamard and our topology optimization framework on 2-d and 3-d examples featuring a very small Reynolds number (Re = 200). We are not aware, to the best of our knowledge, of analogous results in the 3-d setting.

A first part of this section is devoted to the computation of the shape derivative of this functional: although it has already been considered in several works of the literature, the calculation and the numerical implementation of the resulting formulas do not seem completely standard to us; we propose a special treatment based on a classical idea of [63, 137].

In a second part, we present 2-d and 3-d numerical designs for an instance of the lift-drag optimization problem.

Shape derivatives of the lift functional

Let $D = \mathbb{R}^d$ be a computational domain featuring a liquid phase Ω_f flowing around a solid obstacle $\Omega_s \subset\subset D$. The notation convention is again that assumed in chapter 2: the flow is entering the domain from a Dirichlet boundary $\partial\Omega_f^D$ with a given velocity $\boldsymbol{v} = \boldsymbol{v}_0$ and exits the domain with a zero normal stress boundary condition $\sigma_f(\boldsymbol{v}, p)\cdot\boldsymbol{n} = 0$ on $\partial\Omega_f^N$. The remaining part of the fluid boundary is the solid interface $\Gamma = \partial\Omega_f \backslash (\partial\Omega_f^D \cup \partial\Omega_f^N)$ which is to be optimized.

The *lift* generated by the obstacle is the total force exerted by the fluid on the solid interface $\Gamma = \Omega_f \cap \Omega_s = \partial\Omega_s$ in the vertical, y-direction:

$$\mathrm{Lift}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma)) := -\int_\Gamma \boldsymbol{e}_y \cdot \sigma_f(\boldsymbol{v}, p)\cdot\boldsymbol{n}\mathrm{d}s, \tag{6.2.4}$$

where we recall that the fluid stress tensor is given by $\sigma_f(\boldsymbol{v}, p) = 2\nu e(\boldsymbol{v}) - pI$, and the notation $\boldsymbol{a}\cdot M\cdot\boldsymbol{b} := \boldsymbol{a}^T M\boldsymbol{b}$ for any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ and matrix $M \in \mathbb{R}^{d\times d}$. Note that the minus sign accounts for our convention of chapter 2 whereby the normal $\boldsymbol{n}$ is pointing outward the fluid domain. Several authors have considered the optimization of the lift functional, e.g. in [292], [214] based on surrogate models, [206] based on control points, [205] with the SIMP method and a different objective functional, and [262, 246] in a time varying setting, [219]. Several difficulties are involved in the evaluation of the shape derivative of the lift functional:

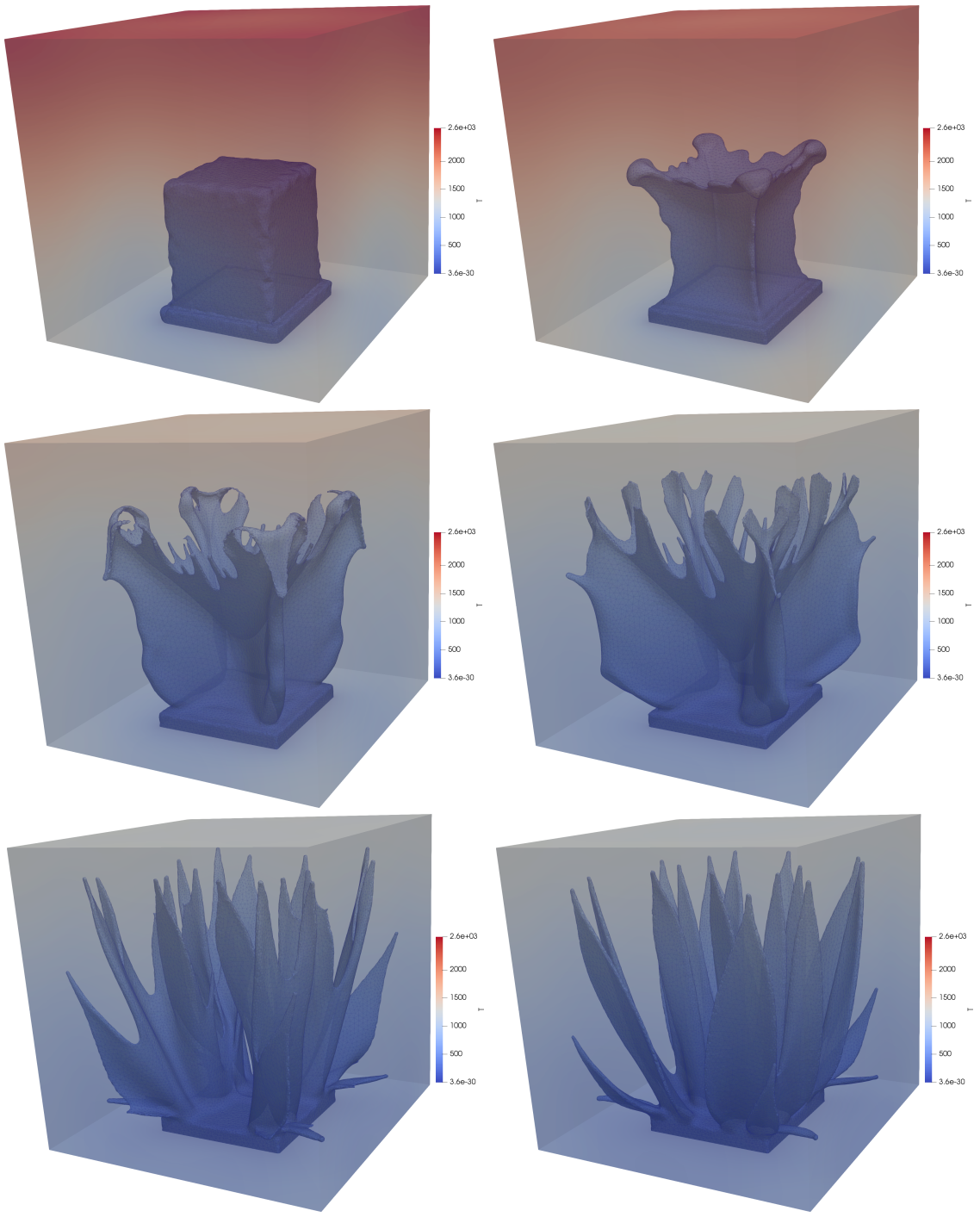


Figure 6.11: From top left to bottom right: iterations 0, 5, 15, 30, 100 and 258 of the optimization of a 3-d design for heat conduction using the *volume* expression of the shape derivative (test case of section 6.2.2).

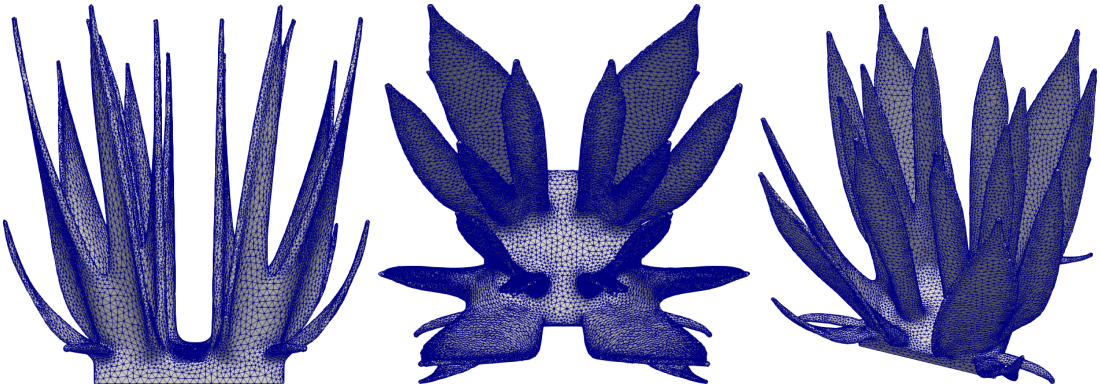


Figure 6.12: Different 3-d views of the optimized design for the heat conduction test case of section 6.2.2.

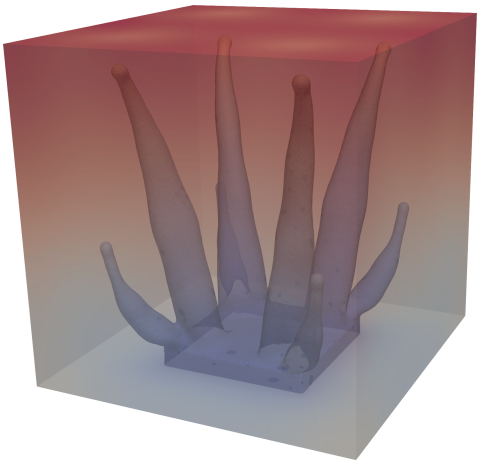
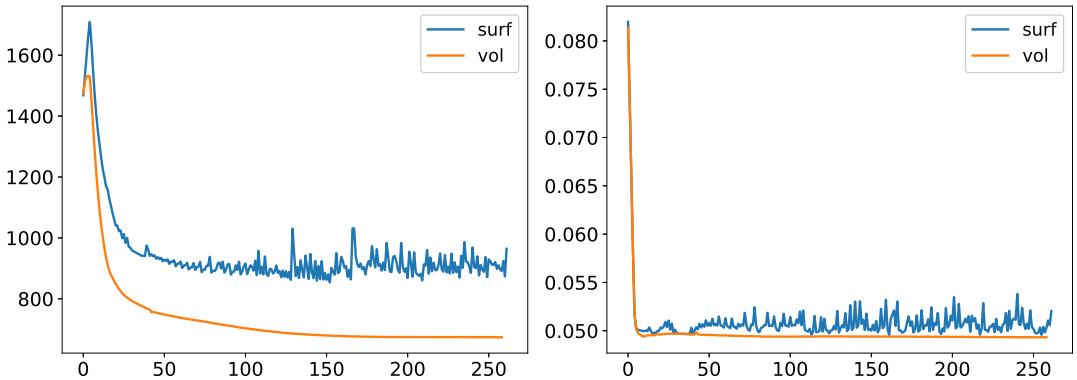


Figure 6.13: Optimized 3-d design for heat conduction obtained with the *surface* expression of the shape derivative (test case of section 6.2.2).



(a) Objective function $J(\Gamma)$.

(b) Volume constraint $\mathrm{Vol}(\Omega_f)$.

Figure 6.14: Convergence histories for the 3-d heat conduction test case of section 6.2.2 using either the volume or the surface expression of the shape derivative.

1. from the theoretical point of view, the partial derivative $(\boldsymbol{v}, p) \mapsto \partial \mathrm{Lift}/\partial(\boldsymbol{v}, p)$ is not a continuous linear form of $H^1(\Omega_f) \times L^2(\Omega_f)$: the fact that the normal derivative $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n}$ has a trace on Γ (as an element of $H^{-1/2}(\Omega_f, \mathbb{R}^d)$) is related to the fact that $(\boldsymbol{v}, p)$ solves the Navier-Stokes system (see chapter 2, section 2.4) but it may not be true for arbitrary variations $(\boldsymbol{w}, q) \in H^1(\Omega_f) \times L^2(\Omega_f)$;

2. even assuming a different variational setting where these partial derivatives would make sense (see [296] for an example of such setting), the functional is posed on the boundary. Introducing as in chapter 2 the transported function

$$\mathfrak{Lift}(\boldsymbol{\theta}, \widehat{\boldsymbol{v}}, \widehat{p}) := \mathrm{Lift}(\Gamma_{\boldsymbol{\theta}}, \widehat{\boldsymbol{v}} \circ (I + \boldsymbol{\theta})^{-1}, \widehat{p} \circ (I + \boldsymbol{\theta})^{-1}), \tag{6.2.5}$$

the computation of the partial derivative

$$\overline{\frac{\partial \mathfrak{Lift}}{\partial \boldsymbol{\theta}}}$$

is not straightforward since (i) it involves the gradient of v, and (ii) the variations of the normal $\boldsymbol{n}$ come into play. This approach has been followed by [292] for the compressible Navier-Stokes equations.

Here, we propose an alternative method, which relies on a reformulation of the lift functional (6.2.4) as a volume integral: let $\mathcal{X} \in H^1(\Omega_f)$ be any extension of the constant function 1 on Γ (i.e. $\mathcal{X} = 1$ on Γ) and vanishing on the complementary boundary (i.e. $\mathcal{X} = 0$ on $\partial\Omega_f \backslash \Gamma$). In our particular implementation, $\mathcal{X}$ is obtained as the solution to the following Poisson problem:

$$\begin{cases} -\Delta \mathcal{X} = 0 \text{ in } \Omega_f \\ \quad \mathcal{X} = 1 \text{ on } \Gamma \\ \quad \mathcal{X} = 0 \text{ on } \partial\Omega_f \backslash \Gamma \end{cases} \tag{6.2.6}$$

The function $\mathcal{X}\boldsymbol{e}_y$ yields then an extension of the vector field $\boldsymbol{e}_y$ vanishing on $\partial\Omega_f \backslash \Gamma$, which enables to rewrite $\mathrm{Lift}(\Gamma)$ as a volume integral:

$$\begin{aligned} \mathrm{Lift}(\Gamma) &= -\int_\Gamma \mathcal{X}\boldsymbol{e}_y \cdot \sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} \mathrm{d}s \\ &= -\int_{\Omega_f} \mathrm{div}(\mathcal{X}\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{e}_y) \mathrm{d}x \\ &= -\int_{\Omega_f} (\nabla \mathcal{X} \cdot \sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{e}_y + \mathcal{X}\mathrm{div}(\sigma_f(\boldsymbol{v}, p)) \cdot \boldsymbol{e}_y) \mathrm{d}x \\ &= \int_{\Omega_f} (\mathcal{X}\boldsymbol{f}_f \cdot \boldsymbol{e}_y - \rho\mathcal{X}\boldsymbol{e}_y \cdot \nabla\boldsymbol{v} \cdot \boldsymbol{v} - \nabla\mathcal{X} \cdot \sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{e}_y) \mathrm{d}x \end{aligned} \tag{6.2.7}$$

where we have used the state equation $-\mathrm{div}(\sigma_f(\boldsymbol{v}, p)) + \rho\nabla\boldsymbol{v}\boldsymbol{v} = \boldsymbol{f}_f$ to obtain the last line. This rewriting is rather classical and is often used in numerical applications, since it is known to yield a more accurate evaluation of the lift functional [63, 137].

Remark 6.2. The last equality of (6.2.7) can be considered as a *definition* for the meaning of the normal derivative $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n}$ in (6.2.4), since it does not depend on the chosen extension $\mathcal{X} \in H^1(\Omega_f)$ satisfying $\mathcal{X} = 1$ on Γ and $\mathcal{X} = 0$ on $\partial\Omega_f \backslash \Gamma$ (from the variational formulation (2.4.2)).

Equation (6.2.7) can now be easily differentiated with respect to the shape. Let us introduce the functional spaces $V_{\boldsymbol{v},p}(\Gamma)$ and $V_{\mathcal{X}}(\Gamma)$ associated with the solutions $(\boldsymbol{v}, p)$ and $\mathcal{X}$:

$$V_{\boldsymbol{v},p}(\Gamma) := \{(\boldsymbol{w}, q) \in H^1(\Omega_f, \mathbb{R}^d) \times L^2(\Omega_f)\mathbb{R} \,|\, \boldsymbol{w} = 0 \text{ on } \partial\Omega_f\},$$

$$V_{\mathcal{X}}(\Gamma) := \{\Psi \in H^1(\Omega_f) \,|\, \Psi = 0 \text{ on } \partial\Omega_f.\}$$

We also denote by $\boldsymbol{v}_0 + V_{\boldsymbol{v},p}(\Gamma)$ and $1 + V_{\mathcal{X}}(\Gamma)$ the affine spaces

$$\boldsymbol{v}_0 + V_{\boldsymbol{v},p}(\Gamma) := \{(\boldsymbol{v}, p) \in H^1(\Omega_f, \mathbb{R}^d) \times L^2(\Omega_f)\mathbb{R} \,|\, \boldsymbol{v} = 0 \text{ on } \partial\Omega_f \backslash \partial\Omega_f^D \text{ and } \boldsymbol{v} = \boldsymbol{v}_0 \text{ on } \partial\Omega_f^D\},$$

$$1 + V_{\mathcal{X}}(\Gamma) := \{\mathcal{X} \in H^1(\Omega_f) \,|\, \mathcal{X} = 0 \text{ on } \partial\Omega_f \backslash \Gamma \text{ and } \mathcal{X} = 1 \text{ on } \Gamma.\}$$

In what follows, we assume that the extension $\mathcal{X}$ of the constant function 1 on Γ is given by the solution to the Poisson problem (6.2.6). We denote, for any $\boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d)$, $(\widehat{\boldsymbol{v}},\widehat{p}) \in \boldsymbol{v}_0 + V_{\boldsymbol{v},p}(\Gamma)$ and $\widehat{\mathcal{X}} \in 1 + V_{\mathcal{X}}$,

$$\mathfrak{Lift}(\boldsymbol{\theta},\widehat{\boldsymbol{v}},\widehat{p},\widehat{\mathcal{X}}) := \mathrm{Lift}(\Gamma_{\boldsymbol{\theta}},\widehat{\boldsymbol{v}}\circ(I+\boldsymbol{\theta})^{-1},\widehat{p}\circ(I+\boldsymbol{\theta})^{-1},\widehat{\mathcal{X}}\circ(I+\boldsymbol{\theta})^{-1}) \tag{6.2.8}$$

the transported functional on the reference situation. Note that since $\mathfrak{Lift}$ does not depend on the choice of the extension $\mathcal{X}$ satisfying $\mathcal{X}=1$ on Γ, (6.2.8) and (6.2.5) coincide for such $\mathcal{X}$. Then, with the notation of chapter 2:

$$\begin{aligned}
\forall \boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d),\ \frac{\partial \mathfrak{Lift}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) &= \int_{\Omega_f} (\mathcal{X}\boldsymbol{f}_f\cdot\boldsymbol{e}_y - \rho\mathcal{X}\boldsymbol{e}_y\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v} - \nabla\mathcal{X}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{e}_y)\mathrm{div}(\boldsymbol{\theta})\mathrm{d}x \\
&+ \int_{\Omega_f} (\rho\mathcal{X}\boldsymbol{e}_y\cdot\nabla\boldsymbol{v}\nabla\boldsymbol{\theta}\cdot\boldsymbol{v} + (\nabla\boldsymbol{\theta}^T\nabla\mathcal{X})\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{e}_y + \nu\nabla\mathcal{X}\cdot(\nabla\boldsymbol{v}\nabla\boldsymbol{\theta}+\nabla\boldsymbol{\theta}^T\nabla\boldsymbol{v}^T)\cdot\boldsymbol{e}_y)\mathrm{d}x, \\
\forall \boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d),\ \overline{\frac{\partial \mathfrak{Lift}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) &= \int_{\Gamma} (\mathcal{X}\boldsymbol{f}_f\cdot\boldsymbol{e}_y - \rho\mathcal{X}\boldsymbol{e}_y\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v} - \nabla\mathcal{X}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{e}_y)\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s \\
&+ \int_{\Gamma} ((\nabla\mathcal{X}\cdot\boldsymbol{n})(\boldsymbol{n}\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{e}_y) + \nu(\nabla\mathcal{X}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{n})(\boldsymbol{e}_y\cdot\boldsymbol{n}) + \nu(\boldsymbol{e}_y\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{n})(\nabla\mathcal{X}\cdot\boldsymbol{n}))\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s \\
&= \int_{\Gamma} [\boldsymbol{f}_f\cdot\boldsymbol{e}_y + 2\nu e(\boldsymbol{v}) : (\nabla\mathcal{X}\otimes\boldsymbol{e}_y)]\boldsymbol{\theta}\cdot\boldsymbol{n}\mathrm{d}s, \\
\forall (\boldsymbol{w}',q') \in V_{\boldsymbol{v},p}(\Gamma)\ \frac{\partial \mathfrak{Lift}}{\partial(\boldsymbol{v},p)}(\boldsymbol{w}',q') &= -\int_{\Omega_f} (\rho\mathcal{X}\boldsymbol{e}_y\cdot\nabla\boldsymbol{w}'\cdot\boldsymbol{v} + \rho\mathcal{X}\boldsymbol{e}_y\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{w}' + \nabla\mathcal{X}\cdot\sigma_f(\boldsymbol{w}',q')\cdot\boldsymbol{e}_y)\mathrm{d}x,
\end{aligned} \tag{6.2.9}$$

The simplifications in the second equality of (6.2.9) are classical consequences of the tangential gradients $\nabla_\Gamma \boldsymbol{v} = 0$ and $\nabla_\Gamma \mathcal{X}$ being 0 on Γ.

Finally, the partial derivative with respect to $\mathcal{X}$ is zero, because an easy integration by part implies that

$$\forall \Psi' \in V_{\mathcal{X}}(\Gamma),\ \frac{\partial \mathfrak{Lift}}{\partial \mathcal{X}}(\Psi') = -\int_\Gamma \Psi'\boldsymbol{e}_y\cdot\sigma_f(\boldsymbol{v},p)\cdot\boldsymbol{n}\mathrm{d}s = 0.$$

Therefore, there is no contribution of the variations of the extension $\mathcal{X}$ defined in (6.2.6) to the shape derivative of $\mathfrak{Lift}$.

All in all, applying the result of chapter 2, propositions 2.3 and 2.4 yields expressions for the shape derivative of Lift (eqn. (6.2.4)) in surface and volume forms:

Proposition 6.1. *Let $(\boldsymbol{w},q) \in V_{\boldsymbol{v},p}(\Gamma)$ the adjoint fluid variables solutions to*

$$\forall (\boldsymbol{w}',q') \in V_{\boldsymbol{v},p}(\Gamma),\quad \int_{\Omega_f} \Big(\sigma_f(\boldsymbol{w},q) : \nabla\boldsymbol{w}' + \rho\boldsymbol{w}\cdot\nabla\boldsymbol{w}'\cdot\boldsymbol{v} + \rho\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{w}' - q'\mathrm{div}(\boldsymbol{w})\Big)\mathrm{d}x = \frac{\partial \mathfrak{Lift}}{\partial(\boldsymbol{v},p)}(\boldsymbol{w}',q'). \tag{6.2.10}$$

The lift functional $\Gamma \mapsto \mathrm{Lift}(\Gamma,\boldsymbol{v}(\Gamma),p(\Gamma))$ is differentiable with respect to the shape and the shape derivative reads (in volumetric form):

$$\begin{aligned}
\forall \boldsymbol{\theta} \in W^{1,\infty}(D,\mathbb{R}^d),\ &\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=0} \Big[\mathrm{Lift}(\Gamma_{\boldsymbol{\theta}},\boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}),p(\Gamma_{\boldsymbol{\theta}}),T(\Gamma_{\boldsymbol{\theta}}),\boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) \\
&= \frac{\partial \mathfrak{Lift}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}) + \int_{\Omega_f} [\boldsymbol{w}\cdot\mathrm{div}(\boldsymbol{f}_f\otimes\boldsymbol{\theta}) - (\sigma_f(\boldsymbol{v},p) : \nabla\boldsymbol{w} + \rho\boldsymbol{w}\cdot\nabla\boldsymbol{v}\cdot\boldsymbol{v})\mathrm{div}(\boldsymbol{\theta})]\,\mathrm{d}x \\
&\quad + \int_{\Omega_f} [\sigma_f(\boldsymbol{v},p) : (\nabla\boldsymbol{w}\nabla\boldsymbol{\theta}) + \sigma_f(\boldsymbol{w},q) : (\nabla\boldsymbol{v}\nabla\boldsymbol{\theta}) + \rho\boldsymbol{w}\cdot(\nabla\boldsymbol{v}\nabla\boldsymbol{\theta})\cdot\boldsymbol{v}]\mathrm{d}x
\end{aligned} \tag{6.2.11}$$

or (in surface form):

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=0} \Big[\mathrm{Lift}(\Gamma_{\boldsymbol{\theta}},\boldsymbol{v}(\Gamma_{\boldsymbol{\theta}}),p(\Gamma_{\boldsymbol{\theta}}),T(\Gamma_{\boldsymbol{\theta}}),\boldsymbol{u}(\Gamma_{\boldsymbol{\theta}}))\Big](\boldsymbol{\theta}) = \overline{\frac{\partial \mathfrak{Lift}}{\partial \boldsymbol{\theta}}}(\boldsymbol{\theta}) + \int_\Gamma (\boldsymbol{f}_f\cdot\boldsymbol{w} + \sigma_f(\boldsymbol{w},q) : \nabla\boldsymbol{v})\mathrm{d}s. \tag{6.2.12}$$

The above formulas are easily implemented in our shape and topology optimization framework; they are at the basis of the resolution of the test cases presented in the next paragraph.

Numerical test cases

We consider the problem of maximizing the Lift generated by a flow obstacle $\Omega_s \subset\subset D$ subject to an upper bound constraint on the drag (so that the obstacle remains aerodynamic). In addition, the volume occupied by the obstacle and the location of its center of mass are prescribed:

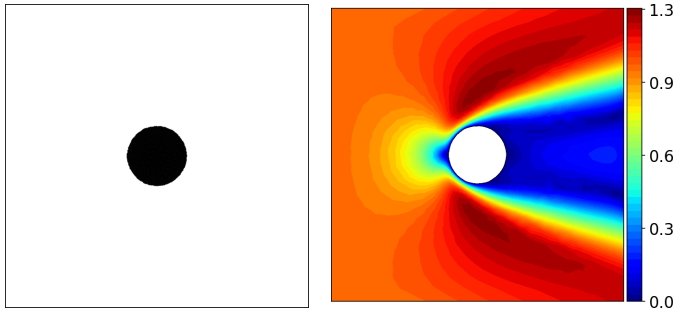
$$\begin{aligned} \min \quad & -\text{Lift}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma)) \\ s.t. \quad & \begin{cases} \text{Drag}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma)) \leq \texttt{DRAG}_0 \\ \text{Vol}(\Omega_f) = V_0 \\ \boldsymbol{X}(\Omega_s) := \dfrac{1}{|\Omega_s|} \displaystyle\int_{\Omega_s} \boldsymbol{x} \mathrm{d}x = \boldsymbol{x}_0. \end{cases} \end{aligned} \tag{6.2.13}$$

The drag functional has been considered in chapter 2, section 2.5.4; it is defined by

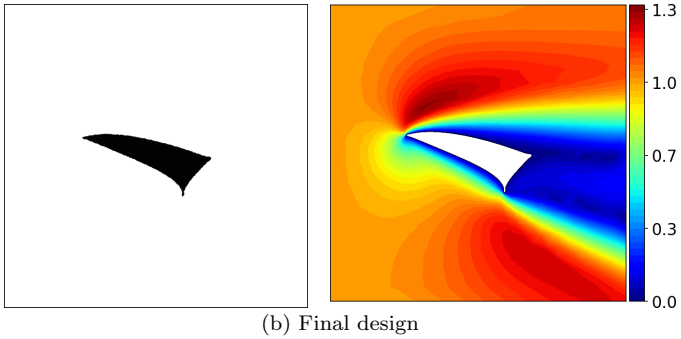
$$\text{Drag}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma)) := \int_{\Omega_f} \sigma_f(\boldsymbol{v}, p) : \nabla \boldsymbol{v} \mathrm{d}x = \int_{\Omega_f} 2\nu e(\boldsymbol{v}) : e(\boldsymbol{v}) \mathrm{d}x.$$

The constant $\texttt{DRAG}_0$ is set to $\alpha \texttt{DRAG}^*$ ($\alpha = 1.1$ in 2-d and $\alpha = 1.5$ in 3-d) where $\texttt{DRAG}^*$ is the value of the minimum drag problem subject to the same volume and position constraints.

We start by solving the problem in 2-d; the setting is identical to that of section 2.5.4: $D = [0,1]\times[0,1]$, $V_{target} = 0.03$, $x_0 = (0.5, 0.5)$, $||\boldsymbol{v}_0|| = 1$, $\rho = 1$, $\nu = 1/200$ and $\text{Re} = 200$. The flow is entering the left-hand boundary with velocity $\boldsymbol{v}_0 = \boldsymbol{e}_x$ and it exits the right-hand boundary with zero normal stress ($\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = 0$). A slip boundary condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ is prescribed at other boundaries. The minimum drag value computed for this example is $\texttt{DRAG}^* \simeq 0.11$, so that the upper bound constraint is given by $\texttt{DRAG}_0 = \texttt{DRAG}^* \times 1.1 = 0.121$. The initial and optimized shapes with the corresponding velocity fields are shown on respectively Figure 6.15. A few intermediate iterations are shown on Figure 6.16, and the optimization histories are reported on Figure 6.18. Finally, the mesh discretizing the final shape is plotted on Figure 6.17; both the boundary of the obstacle and the outlet were refined in order to numerically capture boundary layers and vortex patterns.



(a) Initial design



(b) Final design

Figure 6.15: Optimization results for the 2-d lift-drag optimization problem of section 6.2.3. The norm of the velocity field $\boldsymbol{v}$ is plotted on the right.

Our design is slightly similar to those obtained in the work [205] with a completely different method: a density based approach was used and a different expression was used to estimate the lift functional.

We then solve the same problem in 3-d: the hold-all domain is the box $D = [0,1] \times [0,1] \times [0,1]$. A flow is entering with a velocity $\boldsymbol{v}_0 = \boldsymbol{e}_x$ on the left-hand side of the domain. A slip boundary condition

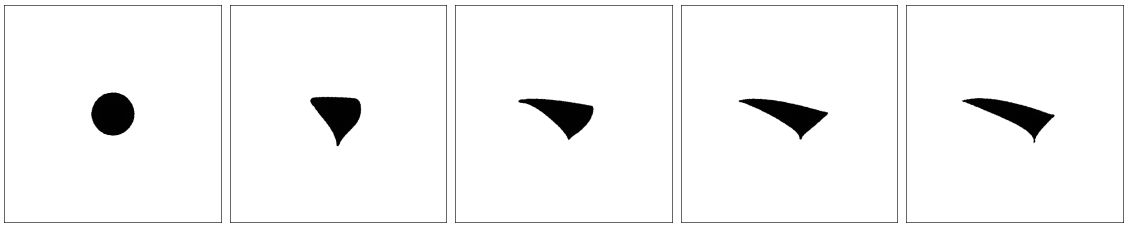


Figure 6.16: Intermediate iterations 0, 8, 30, 80 and 200 for the 2-d lift-drag optimization problem of section 6.2.3

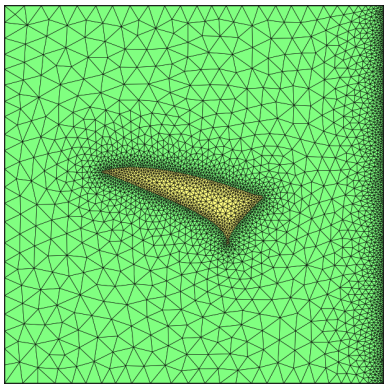
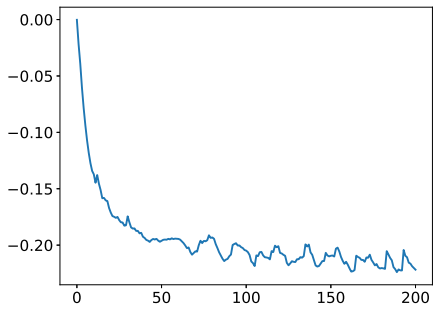
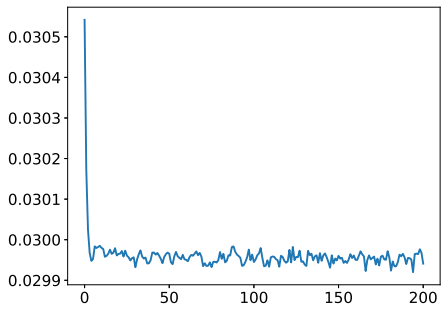


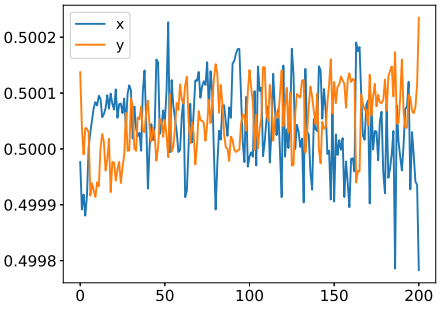
Figure 6.17: Mesh of the final shape for the 2-d lift-drag optimization problem of section 6.2.3.



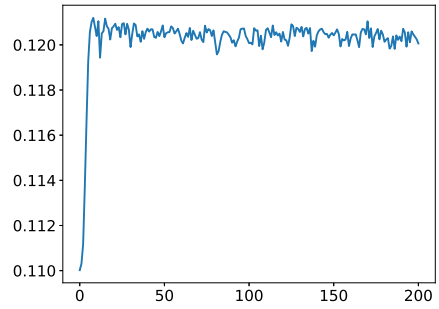
(a) Objective function $\text{Lift}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma))$.



(b) Volume constraint $\text{Vol}(\Omega_f, \boldsymbol{v}(\Gamma), p(\Gamma))$.



(c) Center of mass constraint $\mathbf{X}(\Omega_s) = (x, y)$.

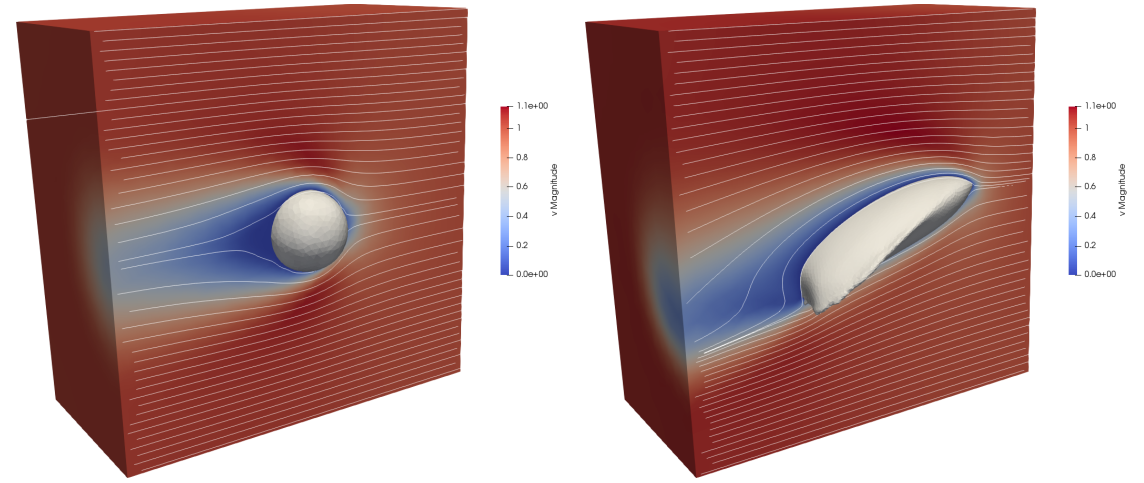


(d) Drag constraint $\text{Drag}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma))$.

Figure 6.18: Convergence histories for the 2-d lift-drag optimization test case of section 6.2.3.

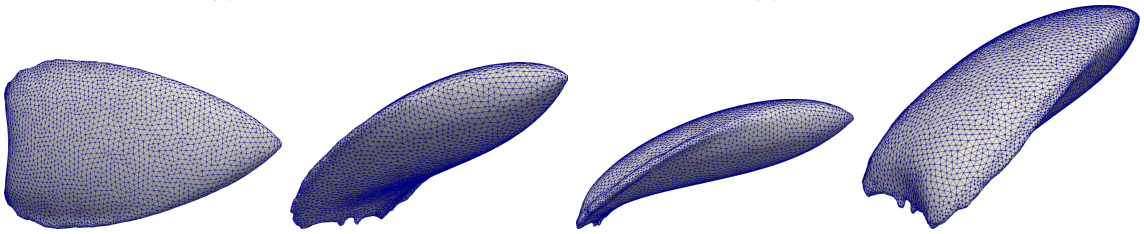
$\boldsymbol{v} \cdot \boldsymbol{n} = 0$ is imposed at other boundaries of the cube. The physical parameters are still set to $\nu = 1/200$ and $\rho = 1$ corresponding to $\mathrm{Re} = \frac{\rho h v_{\max}}{\nu} = 200$ (the characteristic length is $h = 1$). The volume fraction target is set to $V_{target} = 0.01$. For this configuration, the computed value for the minimum drag problem is $\mathtt{DRAG}^* = 0.0304$ which yields an upper bound $\mathtt{DRAG}_0 = 1.5 \times \mathtt{DRAG}^* = 0.0456$.

The optimized shape with associated velocity profile is shown on Figure 6.19. The final fluid mesh features 54,299 vertices. The finite element problems associated with the Navier-Stokes and adjoint equations are solved by using the domain decomposition technique described in section 6.1.4 with 12 cpus (2.60GHz). Convergence histories are shown on Figure 6.18. A single resolution of the state equations including domain decomposition and the Newton loop with the Newton method takes approximately 2 minutes. The computation of the shape derivative, including adjoing system resolutions takes approximately the same time. Every remeshing step (not performed in parallel) is achieved within approximately one minute.



(a) Initial shape

(b) Optimized design



(c) Optimized design (other 3-d views)

Figure 6.19: Optimized shape for the 3-d lift-drag maximization problem of section 6.2.3.

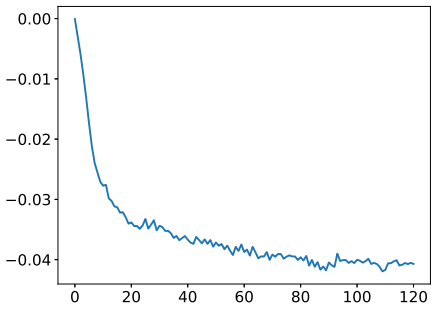
6.2.4 A 3-d fluid-structure interaction test case

Our last test case is concerned with a rather large-scale fluid-structure interaction problem. A fluid is entering the box $D = [0, 4] \times [0, 1] \times [0, 1]$ with an input velocity $\boldsymbol{v}_0 = y\boldsymbol{e}_x$ on the left-hand boundary. A no-slip boundary condition $\boldsymbol{v} = 0$ is prescribed on the bottom face of the domain. The top and side faces assume a slip boundary condition $\boldsymbol{v} \cdot \boldsymbol{n} = 0$. The flow exits the domain with a zero normal stress boundary condition $\sigma_f(\boldsymbol{v}, p) \cdot \boldsymbol{n} = 0$. A mechanical structure $\Omega_s \subset D$ is fixed ($\boldsymbol{u} = 0$) on a square patch of the bottom face and is subjected to the stress induced by the fluid (namely, $\boldsymbol{u}$ is the solution to (2.2.3)).

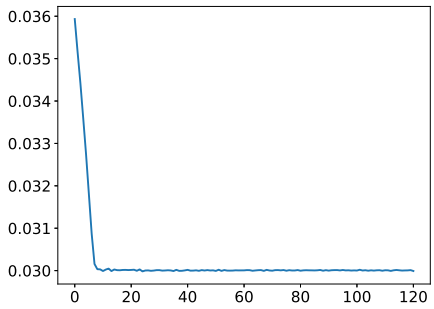
A vertical plate is set as a non optimizable part of the mechanical structure (as well as a small layer above the bottom Dirichlet boundary): the setting is made visible on Figure 6.21. The goal is to find how to distribute additional material in order to make the structure Ω_s as rigid as possible. The problems features of course a volume constraint on the mechanical structure, so that it reads

$$\begin{aligned} \min_{\Gamma} \quad & J(\Gamma, \boldsymbol{u}(\Gamma)) = \int_{\Omega_s} Ae(\boldsymbol{u}) : e(\boldsymbol{u}) \mathrm{d}x \\ s.t. \quad & \mathrm{Vol}(\Omega_s) = V_{target}. \end{aligned} \tag{6.2.14}$$

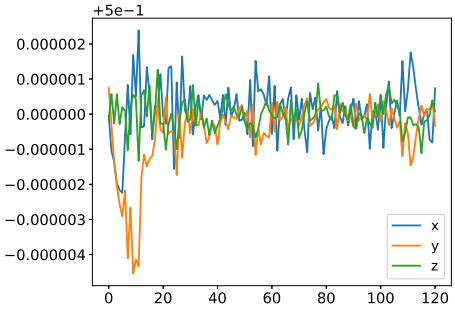
The Reynolds number, fluid density and viscosity are respectively set to $\mathrm{Re} = 60$, $\rho = 1$ and $\nu = \rho h ||\boldsymbol{v}_0||_\infty / \mathrm{Re} = 0.012$ (the characteristic length is $h = 0.7$ corresponding to the height of the non-



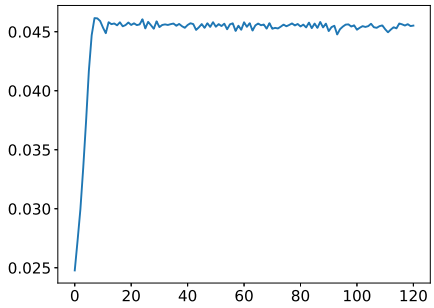
(a) Objective function $\text{Lift}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma))$.



(b) Volume constraint $\text{Vol}(\Omega_f, \boldsymbol{v}(\Gamma), p(\Gamma))$.



(c) Center of mass constraint $\mathbf{X}(\Omega_s) = (x, y, z)$.



(d) Drag constraint $\text{Drag}(\Gamma, \boldsymbol{v}(\Gamma), p(\Gamma))$.

Figure 6.20: Convergence histories for the 3-d lift-drag optimization test case of section 6.2.3.

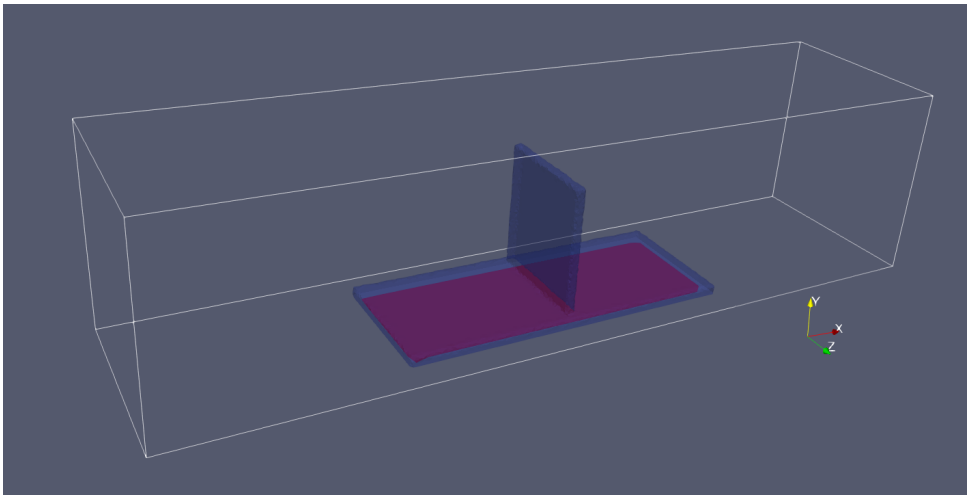


Figure 6.21: Setting of the fluid-structure interaction test case. A flow is entering from the left-hand side; a no-slip boundary condition is imposed at the bottom wall, and a slip boundary condition on the other side walls. The flow is pushing against a non optimizable vertical mechanical plate tightened to the bottom wall with a zero Dirichlet boundary condition (on the red surface).

Solving state equations	24 cpus	15’
Computing shape derivatives and gradients	24 cpus	2’
Advection of the level-set function	sequential	24”
Symmetrization and regularization of the level-set function	24 cpus	13”
Remeshing	sequential	2’
Computation of the signed distance function	sequential	3’

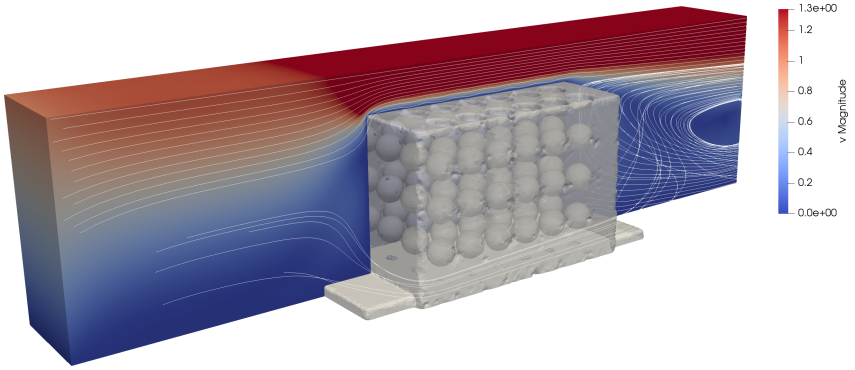
Table 6.2: Running times for the first iteration (220,283 mesh nodes) of the fluid-structure interaction test case of section 6.2.4.

optimizable plate and $||\boldsymbol{v}_0||_\infty = 1$). The Lamé coefficients of the mechanical structure are $\lambda = 0.00529$ and $\mu = 0.0476$.

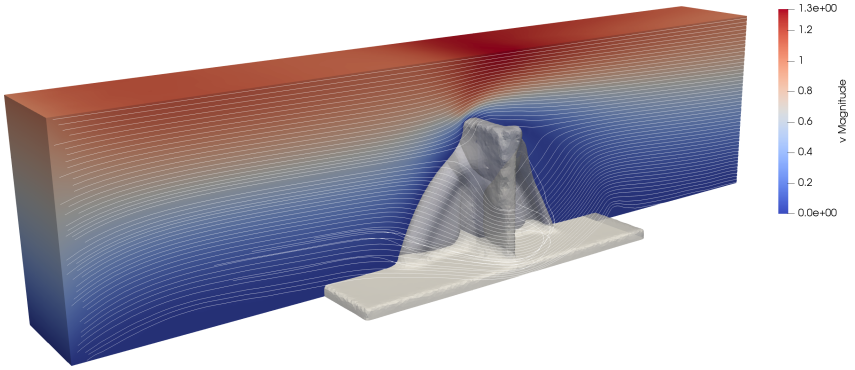
The optimized shape obtained with the *volume* expression of the shape derivative is plotted on Figure 6.22. Note that for this example, we did not see a clear difference in performance between the results arising from the use of the volume and the surface expressions of the shape derivative. Not surprisingly, the final design has an aerodynamic profile in order to reduce the stress applied by the fluid flow.

This example is our most computationally involved test case. Finite element computations were run in parallel on 24 cpus (2.60GHz) . The number of mesh nodes varies from 220,283 for the first iteration (including 132,775 nodes in the fluid domain, which means approx. 1,7 millions degrees of freedom for the linearized fluid system) to 82,454 (including 66,021 nodes in the fluid domain) at the last iteration. Running cpu times for the first mesh (220,283 mesh nodes) are listed in Table 6.2. The most intensive task is the resolution of the state equations (the running time mentioned includes all domain decomposition steps, finite element matrix assembly, the Newton loop for the Navier-Stokes system, etc. . .).

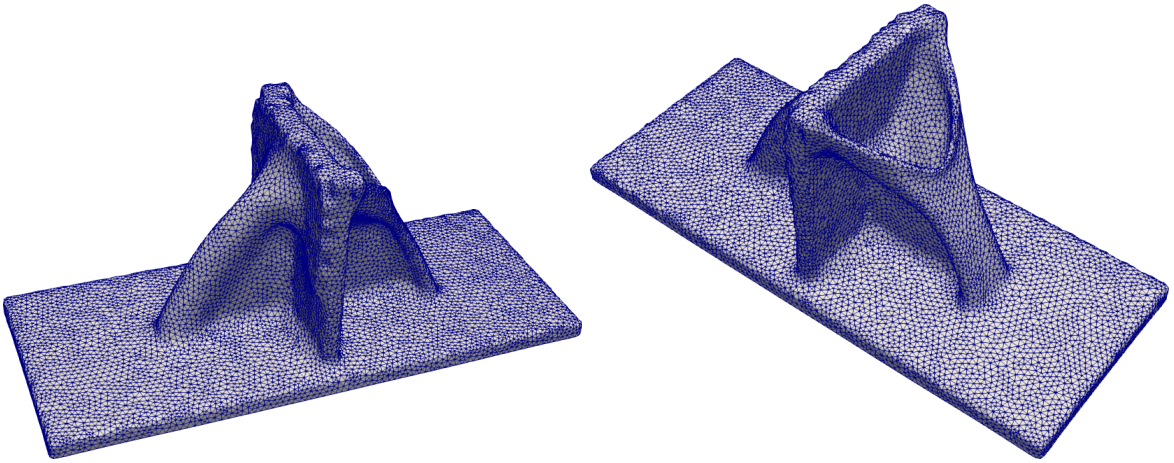
Although these results are promising, we are still far from realistic industrial system sizes (featuring the order of the billion of elements). We had difficulties in going to mesh sizes reaching the million of vertices because the cost of remeshing becomes prohibitive. So far, this step is performed sequentially; future works could benefit from making this step run in parallel as well.



(a) Initial shape



(b) Final design



(c) Final design (front and back views respectively on the left and on the right).

Figure 6.22: Optimized design for the 3-d fluid structure interaction test case of section 6.2.4.

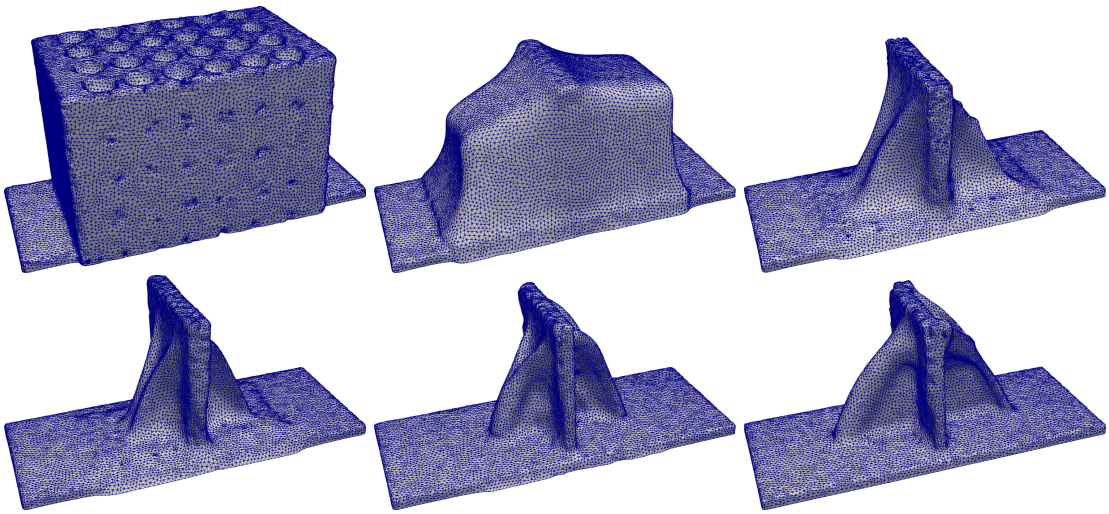


Figure 6.23: Intermediate iterations 0, 40, 100, 125, 175 and 300 for the fluid structure test case of (6.2.14).

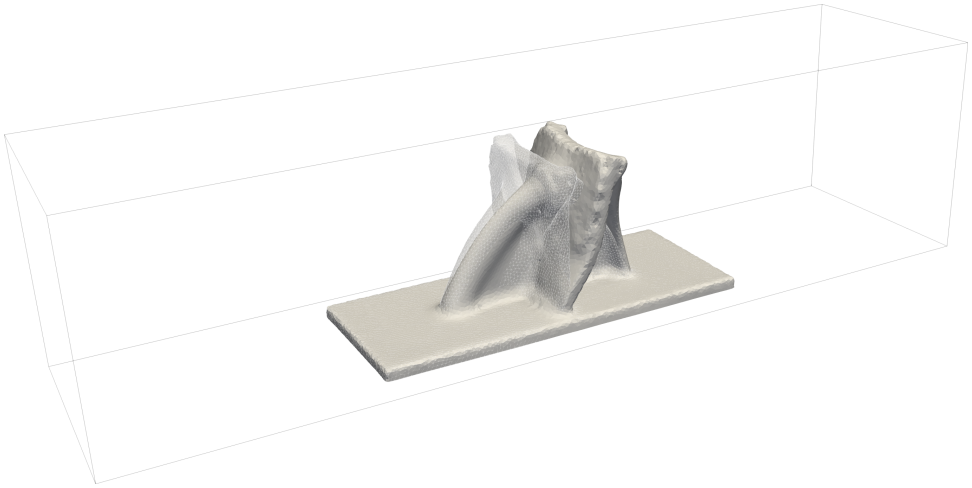
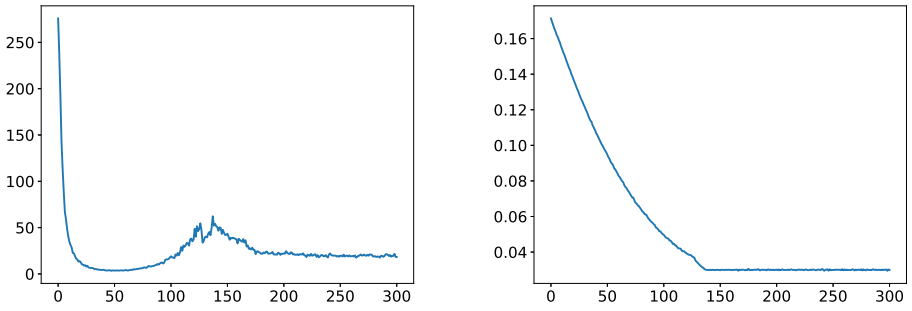


Figure 6.24: Linear elastic deformation of the solid structure under the load force applied by the fluid for the test case of (6.2.14).



(a) Objective function $J(\Omega_s, \boldsymbol{u}(\Omega_s))$. (b) Volume fraction $\mathrm{Vol}(\Omega_s)/\mathrm{Vol}(D)$.

Figure 6.25: Convergence history for the fluid-structure interaction optimization test case of section 6.2.4.

CHAPTER 7

HIGH ORDER HOMOGENIZED EQUATIONS FOR PERFORATED PROBLEMS: TOWARDS FLUID TOPOLOGY OPTIMIZATION BY THE HOMOGENIZATION METHOD

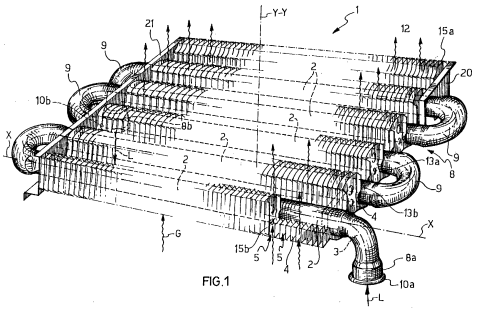
Contents

7.1 Introduction

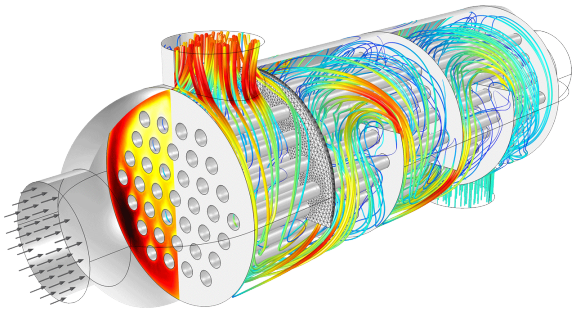
This final chapter is an opening towards the use of topology optimization for the design of fluid systems by the homogenization method. Our motivation originates from the observation that many industrial applications in the aeronautic industry involve multi-scale designs. For instance, heat exchangers feature periodic patterns visible at a microscopic scale which are geometrically modulated over larger scales (Figure 7.1). They are integrated into a suitable macroscopic structure so as to maximize the exchange surface between hot and cold phases, while limiting the output pressure loss.

When it comes to automatically generate such multi-scale structures, the topology optimization techniques based on the method of Hadamard described in the previous chapters reach their limits. Indeed, the numerical optimization of the shapes of highly resolved composite structures would require the use of very fine meshes so as to capture the most microscopic details of the fluid flowing in the whole structure and to update the tiniest details of its boundaries; the numerical computation would be very costly and probably very slow to converge.

The goal of this chapter is to lay down theoretical material that would allow, ultimately, to design fluid systems by homogenization methods similar to those available in the context of topology optimization of mechanical structures [65, 58, 18, 254, 27]. In the latter contexts, they specifically allow to generate



(a) Industrial gas-liquid heat exchanger design featuring blade patterns (Figure from [169]).



(b) Industrial air-water heat exchanger design featuring tube patterns (Figure from COMSOL multiphysics [239]).

Figure 7.1: Two examples of heat exchanger designs featuring periodically repeated patterns.

multi-scale designs by optimizing simultaneously both the microstructure of a mechanical system and its macroscopic shape. Unfortunately, these methods do not extend in a straightforward manner to fluid systems, because the homogenization theory which accounts for the effective physics of the fluid in a porous medium is different. Several homogenized models exist depending on various scaling regimes assumed by the microstructure pattern (namely, the Darcy, Brinkman, or Stokes regimes [272, 14]) which makes it unclear which effective model should be used to describe a context featuring all possible regimes simultaneously at different locations in the domain.

In this chapter, we derive high order homogenized models for the Stokes system (in periodic domains) which unify the three classical regimes. The mathematical methodology is inspired from the works of Bakhvalov and Panasenko [53], Smyshlyaev and Cherednichenko [287], and Allaire et. al. [33]; it is based on (non standard) two-scale asymptotic expansions and formal operations on related power series which give rise to several families of tensors and homogenized equations at any order. These formal expansions are then justified thanks to rigorous error estimates.

The obtained higher order homogenized models are higher order corrections of the Darcy model: in the low-volume fraction limit where the obstacle size assume one of the three classical regimes of the literature, they specialize to either the Stokes, the Brinkman, or the Darcy equation. Since these higher order formulations unify the three regimes, we expect these could open the way to the development of homogenization methods for the topology optimization of fluid systems. Furthermore, a by-product of the use of higher order models is to yield a more accurate description of the effective physics characterizing porous media in contexts where the size of the microstructure (a parameter ε which is assumed to be close to zero in the homogenized model) is not so small.

The remainder of the chapter is organized as follows. Section 7.2 outlines the motivations for the use of higher order homogenized model for the Stokes system and provides a summary of our main mathematical results. The notation and mathematical setting used in the remainder of the chapter are also introduced.

In order to highlight the main features of our derivation, and for pedagogical purposes, the next sections investigate higher order homogenized equations for several elliptic problems with an increasing order of complexity. We start with a study of the perforated Poisson problem in section 7.3, which is a simplified, scalar version of the Stokes system. In Section 7.4 we consider the elasticity system with Dirichlet boundary conditions on the holes (rather than Neumann ones as is more customarily considered). It is the direct vectorial analogue of the Poisson system which is characterized by an important difference: high order homogenized models of the scalar problem feature only differential operators of *even* order, while differential operators of *odd* orders are present in the vectorial context. However these vanish in case of symmetries. Finally, the Stokes system in a porous medium (7.2.13) is treated in section 7.5, which requires additional work due to the pressure variable and the associated divergence constraint.

7.2 MOTIVATIONS FROM SHAPE OPTIMIZATION AND SUMMARY OF RESULTS

This section motivates and present our main results developed in thorough details in the next parts. We start by providing in section 7.2.1 a brief account of the homogenization method for topology optimization of mechanical structures. The difficulties that arise when considering the application of these methods

for the design of fluid systems, related to the emergence of three different homogenized regimes, are stressed in section 7.2.2. We then provide a summary of our main results in section 7.2.3 regarding the high order homogenization of Stokes flow in porous media which could address some of these difficulties in future works. Finally, the general mathematical setting and the notation conventions used in the remainder of the work are introduced in section 7.2.4.

7.2.1 Shape optimization by the homogenization method in linear elasticity

The homogenization method is a powerful technique for the topology optimization of mechanical structures: it allows to capture multi-scale designs characterized by geometrically modulated micro-structures [65, 18, 254, 177, 27]. Let $D = [-L, L]^d$ be a given computational rectangular domain and consider the problem of finding the shape assumed by a distribution of two materials characterized by constant elastic Hooke's tensors A_0 and A_1 in the respective domains $\Omega \subset D$ and $D \backslash \Omega$:

$$\min_{\Omega \subset D} \quad J(\Omega, \boldsymbol{u}(\Omega)) \qquad s.t. \begin{cases} -\mathrm{div}(A(\mathbf{1}_\Omega)\nabla \boldsymbol{u}) = \boldsymbol{f} \text{ in } D \\ \boldsymbol{u} = 0 \text{ on } \partial D, \end{cases} \tag{7.2.1}$$

The elastic tensor in (7.2.1) is given by $A(\mathbf{1}_\Omega) := 1_\Omega A_0 + (1 - \mathbf{1}_\Omega)A_1$ where $\mathbf{1}_\Omega$ denotes the characteristic function of Ω and $\boldsymbol{f}$ is a source term. The main point of the homogenization method is to replace the design problem (7.2.1) of finding *shapes* Ω by a relaxed one expressed in terms of composite *microstructures* which are not shapes but rather limits of minimizing sequence of shapes. In theory, one can formulate a relaxed version of (7.2.1) posed over the whole set of such possible limits (the G-closure, see [18]). However, for practical applications, it may be sufficient to consider periodic microstructures with a geometrically modulated pattern: such microstructures can be parameterized by varying parameters $(a_1(x), \dots, a_m(x))$, as illustrated on Figure 7.2. This allows to replace (7.2.1) with a "homogenized"

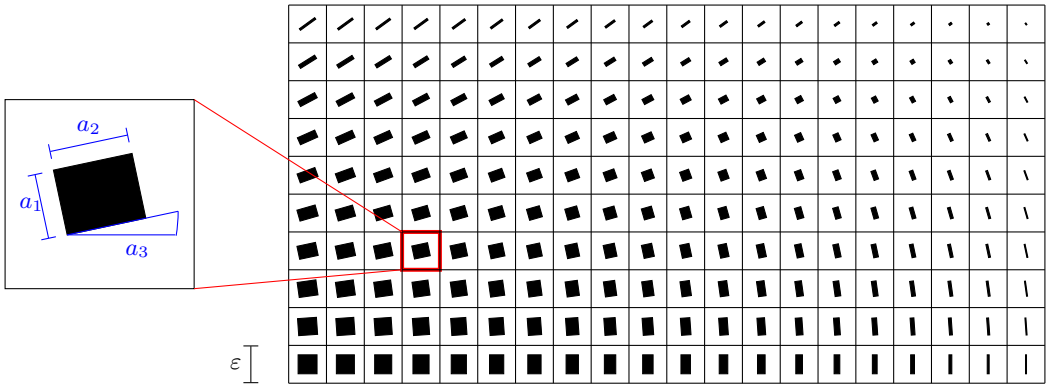


Figure 7.2: A periodic medium with varying microstructure parameterized by the size (a_1, a_2) and orientation a_3 of a rectangular inclusion. Note that in [27], the parameter a_3 rather parameterizes the orientation of the periodic cells themselves than only their inner rectangular pattern.

relaxed problem

$$\min_{(a_1,\dots,a_m) \in L^\infty(D, \mathbb{R}^d)} \quad J^*(a_1, \dots, a_m, \boldsymbol{u}(a_1, \dots, a_m)) \qquad s.t. \begin{cases} -\mathrm{div}(A^*(a_1, \dots, a_m)\nabla \boldsymbol{u}) = \boldsymbol{f} \text{ in } D \\ \boldsymbol{u} = 0 \text{ on } \partial D, \end{cases} \tag{7.2.2}$$

where $A^*(a_1, \dots, a_m)$ is an effective tensor characterizing the stiffness of the composite material obtained by mixing A_0 and A_1 according to the microstructure pattern determined by $a_1, \dots a_m$.

The problem (7.2.2) is easier to solve than (7.2.1): it reduces to rather standard parametric optimization with respect to the $(a_1, \dots, a_m)$. Furthermore, the state variable $\boldsymbol{u}$ can be evaluated in the "homogeneous" domain D, without the need for discretizing all the geometric details of the microstructure. The difficulty of the method lies the post-treatment needed to actually reconstruct a sequence of shapes approaching the optimal parameterized microstructure; see the contributions [254, 35, 166] for a description of some (delicate) algorithms that allow to perform this task in both 2-d and 3-d.

7.2.2 Three homogenized regimes for a Stokes flow in a porous medium

The ultimate motivation would be to apply similar methods for the topology optimization of fluid systems. The optimization problem at play would be formulated in terms of the Stokes (or the Navier-Stokes) equation (instead of (7.2.1)):

$$\min_{\Omega\subset D} \quad J(\Omega, \boldsymbol{u}(\Omega), p(\Omega)),$$
$$s.t. \begin{cases} -\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } \Omega \\ \operatorname{div}(\boldsymbol{u}) = 0 \text{ in } \Omega, \\ \boldsymbol{u} = 0 \text{ on } \partial\Omega, \end{cases} \tag{7.2.3}$$

where $(\boldsymbol{u}, p)$ denotes the fluid velocity and pressure couple (extended by zero outside the optimization domain Ω). However, it is not straightforward to derive a relaxed formulation such as (7.2.3) because the available homogenization theory for the Stokes system (7.2.3) is substantially different to the one for the elasticity system (7.2.1). This difference comes from the fact that the shape Ω characterizing the mixture (of material A_0 and A_1 for the elasticity system, or of fluid and solid in the solid system) arises in the elasticity problem (7.2.1) through the piecewise constant Hooke's tensor $A(\mathbf{1}_\Omega)$, whereas it occurs in the Stokes system (7.2.3) in terms of a *zero Dirichlet* (no-slip) boundary condition $\boldsymbol{u} = 0$ on the boundary of the holes $\partial\Omega$. In contrast with (7.2.1), there is no known theory that would characterize the G-closure of the Stokes system featured in (7.2.3), i.e. the set of all possible limit models obtained by considering minimizing sequences of domains. Instead, the literature [272, 12, 14] describes the occurrence of three possible homogenized regimes (described below) depending on the scaling ratio between the size of the periodic inclusions and their relative distance.

To be more precise, let us place ourselves in the classical context of periodic homogenization for (7.2.3), which as motivated above, may be sufficient for shape optimization purposes. We assume $D := [-L, L]^d$ to be a d-dimensional box filled with periodic obstacles $\omega_\varepsilon := \varepsilon(\mathbb{Z}^d + \eta T) \cap D$. The fluid domain Y is assumed to be connected, see the introduction of section 7.5 for precise assumptions. we denote by $P = (0,1)^d$ the unit cell, and by $Y := P \backslash \eta T$ the unit perforated cell. The boundary of the obstacle T is assumed to be smooth. The setting is illustrated on Figure 7.3. The parameter ε denotes the size of the periodic cell and is assumed to be small. The parameter η is another rescaling of the obstacle T within each cell of size ε (which will allow us later to consider the so-called low volume fraction limit case where η is small comparatively to ε). The working, porous domain is denoted by $D_\varepsilon := D \backslash \omega_\varepsilon$, and we consider

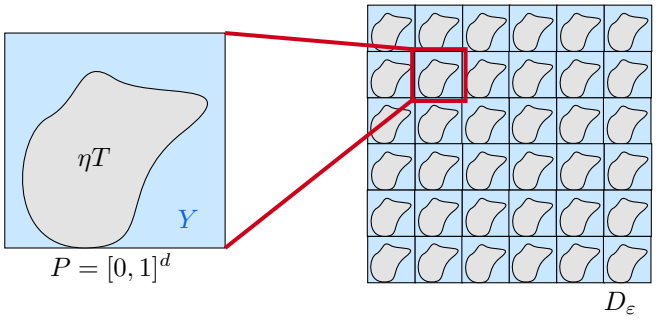


Figure 7.3: The perforated domain D_ε and the unit cell $Y = P \backslash (\eta T)$.

the solution $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ of the following Stokes system:

$$\begin{cases} -\Delta \boldsymbol{u}_\varepsilon + \nabla p_\varepsilon = \boldsymbol{f} \text{ in } D_\varepsilon \\ \operatorname{div}(\boldsymbol{u}_\varepsilon) = 0 \\ \boldsymbol{u}_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon \\ \boldsymbol{u}_\varepsilon \text{ is } D\text{–periodic}, \end{cases} \tag{7.2.4}$$

where $\boldsymbol{f}$ is now required to be a smooth, D–periodic right hand-side. The periodicity assumption for $\boldsymbol{u}_\varepsilon$ is classical in homogenization and is used to avoid difficulties related to the arising of boundary layers (see [216, 326, 19]). The literature accounts for several homogenized equations depending on how the size $a_\varepsilon = \eta\varepsilon$ of the holes compares to the critical size $\sigma_\varepsilon := \varepsilon^{d/(d-2)}$ in dimension $d \geq 3$ or $\sigma_\varepsilon := \exp(-1/\varepsilon^2)$ for $d = 2$ [242, 105, 11, 15]:

- if $a_\varepsilon = o(\sigma_\varepsilon)$, then the holes are "too small" and $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ converges as $\varepsilon \to 0$ to the solution $(\boldsymbol{u}, p)$ of the Stokes equation in the homogeneous domain D:

$$\begin{cases} -\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } D \\ \quad \operatorname{div}(\boldsymbol{u}) = 0 \\ \quad\quad \boldsymbol{u} \text{ is } D\text{–periodic.} \end{cases} \tag{7.2.5}$$

- if $a_\varepsilon = \sigma_\varepsilon$, then $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ converges as $\varepsilon \to 0$ to the solution $(\boldsymbol{u}, p)$ of the Brinkman equation

$$\begin{cases} -\Delta \boldsymbol{u} + \Psi^* \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } D \\ \quad\quad \operatorname{div}(\boldsymbol{u}) = 0 \\ \quad\quad\quad \boldsymbol{u} \text{ is } D\text{–periodic,} \end{cases} \tag{7.2.6}$$

  where the so-called *strange term* $\Psi^* \boldsymbol{u}$ involves a symmetric positive definite $d \times d$ matrix Ψ^* that can be computed by means of an exterior problem in $\mathbb{R}^d \backslash T$ when $d \geq 3$, and which is equal to πI if $d = 2$ (see [11]).

- if $\sigma_\varepsilon = o(a_\varepsilon)$ and $a_\varepsilon = \eta\varepsilon$ with $\eta \to 0$ as $\varepsilon \to 0$, then the holes are "large" and $(a_\varepsilon^{d-2} \varepsilon^{-d} \boldsymbol{u}_\varepsilon, p_\varepsilon)$ converges to the solution $(\boldsymbol{u}, p)$ of the Darcy problem

$$\begin{cases} \Psi^* \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } D \\ \quad \operatorname{div}(\boldsymbol{u}) = 0 \text{ in } D \\ \quad\quad \boldsymbol{u} \text{ is } D\text{–periodic,} \end{cases} \tag{7.2.7}$$

  where Ψ^* is the same symmetric positive definite $d \times d$ matrix as in (7.2.6).

- if $a_\varepsilon = \eta\varepsilon$ with the ratio η *fixed*, then $(\varepsilon^{-2} \boldsymbol{u}_\varepsilon, p_\varepsilon)$ converges to the solution $(\boldsymbol{u}, p)$ of the Darcy problem

$$\begin{cases} M^0 \boldsymbol{u} + \nabla p = \boldsymbol{f} \text{ in } D \\ \quad \operatorname{div}(\boldsymbol{u}) = 0 \text{ in } D \\ \quad\quad \boldsymbol{u} \text{ is } D\text{–periodic,} \end{cases} \tag{7.2.8}$$

  where M^0 is another positive symmetric $d \times d$ matrix (which depends on η). Furthermore it can be shown that $M^0 / |\log(\eta)| \to \Psi^*$ if $d = 2$, and $M^0 / \eta^{d-2} \to \Psi^*$ (if $d \geq 3$) when $\eta \to 0$, so that there is a continuous transition from (7.2.8) to (7.2.7), see [13].

These three different regimes, namely Stokes, Brinkman, and Darcy, raise practical difficulties in view of applying the homogenization method for shape optimization. Indeed, it is not clear which regime to use in order to write down a relaxed problem (7.2.2) in a context where the shape of the holes $\eta T \equiv \eta(x) T(x)$ would be allowed to be modulated along the domain D: in regions featuring very tiny obstacles, one should use the Stokes or Brinkman equation (7.2.5) and (7.2.6), however one should use the Darcy model (7.2.8) when the obstacles become large enough.

In fact, there is a continuous transition between all regimes which can be captured by higher order homogenized equations, and which is the purpose of the present chapter. Corrective terms scaled by higher powers of ε can be added to the Darcy equation (7.2.8), which reduce to one of (7.2.5) to (7.2.7) in the other scaling regimes considered.

7.2.3 High order homogenized equations for the Stokes problem: summary of results

The main result of this chapter is the derivation of higher order homogenized equations for the Stokes system (7.2.4). For a desired order $K \in \mathbb{N}$, there exists a homogenized model of order $2K+2$ which can be written

$$\begin{cases} \displaystyle\sum_{k=0}^{2K+2} \varepsilon^{k-2} \mathbb{D}_K^k \cdot \nabla^k \boldsymbol{v}_K^* + \nabla q_K^* = \boldsymbol{f}, \\ \quad\quad \operatorname{div}(\boldsymbol{v}_K^*) = 0, \\ \quad\quad\quad \boldsymbol{v}_K^* \text{ is } D\text{–periodic,} \end{cases} \tag{7.2.9}$$

and which determines a so-called higher order homogenized solution $(\boldsymbol{v}_K^*, q_K^*)$ (definition 7.8). Note that $(\boldsymbol{v}_K^*, q_K^*)$ depends on ε, but this dependence is implicitly understood for simplicity of notation. The notation

$$\mathbb{D}_K^k := (\mathbb{D}_{K,i_1\dots i_k,lm}^k)_{1\leq i_1\dots i_k\leq d, 1\leq l,m\leq d}$$

designates a k-th order *matrix valued* tensor involving k indices $i_1 \dots i_k$ associated with partial derivatives and two indices l, m associated with spatial coordinates; it can be computed thanks to a procedure involving the resolution of cell problems. The k-th order differential operator $\mathbb{D}_K^k \cdot \nabla^k$ is obtained by contracting the k indices $i_1 \dots i_k$ with k partial derivatives and performing a matrix product with the other two indices l, m:

$$(\mathbb{D}_K^k \cdot \nabla^k \boldsymbol{v}_K^*)_l := \mathbb{D}_{K,i_1\dots i_k,lm}^k \partial_{i_1\dots i_k}^k v_{K,m}^*,$$

where $(v_{K,m}^*)_{1\leq m\leq d}$ denote the components of $\boldsymbol{v}_K^*$ and an implicit summation over the repeated indices $i_1, \dots i_k$ and m is assumed.

Equation (7.2.9) is said to be of "higher order" (namely of order $2K+2$ with $K \in \mathbb{N}$) because $\boldsymbol{v}_K^*$ (respectively q_K^*) yields an approximation of $\boldsymbol{u}_\varepsilon$ (respectively p_ε)) of order ε^{K+3} (respectively ε^{K+1}) in the $L^2(D)$ norm: in proposition 7.39 below, we prove the following error bounds with a constant $C_K(\boldsymbol{f})$ that depends only on $\boldsymbol{f}$, K (and *a priori* on the shape of the obstacle ηT):

$$\left\| \boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}_K^* \right\|_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+2}$$

$$\left\| p_\varepsilon - \left(q_K^* + \sum_{k=0}^{K-1} \varepsilon^{k-1} \boldsymbol{\beta}^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}_K^* \right) \right\|_{L^2(D)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+1}.$$

The variables $N^k(\cdot/\varepsilon)$ and $\boldsymbol{\beta}^k(\cdot/\varepsilon)$ are respectively matrix and vector valued corrector tensors which are periodic in the cell P (Figure 7.3) and which do not depend on $\boldsymbol{f}, \boldsymbol{v}_K^*, p_K^*$. Furthermore, N^k vanishes on the boundary of the obstacle $\partial(\eta T)$ so that a function multiplied by $N^k(\cdot/\varepsilon)$ vanishes on the boundary of the holes $\partial\omega_\varepsilon$.

Our methodology is based on the existence of "criminal" ansatz for the velocity and pressure solution $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$. The "classical" ansatz reads formally

$$\boldsymbol{u}_\varepsilon = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \boldsymbol{u}_i(x, x/\varepsilon), \quad p_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i (p_i^*(x) + \varepsilon p_i(x, x/\varepsilon)), \quad x \in D_\varepsilon. \tag{7.2.10}$$

where the functions $\boldsymbol{u}_i(x,y)$ and $p_i(x,y)$ are periodic in the variable $y \in P$ (see eqn. (7.5.2) below). To our knowledge, very few works have investigated higher order homogenized models for the Stokes equation. Most of the available works have considered situations with low regularity for $\boldsymbol{f}$ and D (see [272, 12]), where the homogenization process can be justified only for the first terms of the ansatz (7.2.10). Error bounds for the higher order terms (namely the result of proposition 7.37) have been obtained in [228, 74]. A few additional works have sought corrector terms from physical modelling considerations [159, 47, 46].

By introducing a suitable family of k-th order tensors $(\mathcal{X}^k, \boldsymbol{\alpha}^k)$ obtained as the solutions to cell problems, we obtain that (7.5.2) rewrites more explicitly (see proposition 7.29)

$$\begin{cases} \boldsymbol{u}_\varepsilon(x) = \displaystyle\sum_{i=0}^{+\infty} \varepsilon^i \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i (\boldsymbol{f}(x) - \nabla p_\varepsilon^*(x)), \\ \\ p_\varepsilon(x) = p_\varepsilon^*(x) + \displaystyle\sum_{i=0}^{+\infty} \varepsilon^{i+1} \boldsymbol{\alpha}^i(x/\varepsilon) \cdot \nabla^i (\boldsymbol{f}(x) - \nabla p_\varepsilon^*(x)), \end{cases} \quad x \in D_\varepsilon. \tag{7.2.11}$$

The ansatz for $\boldsymbol{u}_\varepsilon$ is a first instance of what [33] called an "asymptotic crime", because the function p_ε^* it features is a homogenized pressure which is itself a formal power series in ε (see (7.5.3)).

Despite being explicit, the ansatz (7.2.11) is not fully satisfactory because it requires the knowledge of the partial derivatives of f at any order, which may be difficult to obtain numerically. One of our

main results is proposition 7.33, where we show that $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ can be formally decomposed as

$$\begin{cases} \boldsymbol{u}_\varepsilon(x) = \displaystyle\sum_{i=0}^{+\infty} \varepsilon^i N^i(x/\varepsilon) \cdot \nabla^i \boldsymbol{u}_\varepsilon^*(x), \\ \\ p_\varepsilon(x) = p_\varepsilon^*(x) + \displaystyle\sum_{i=0}^{+\infty} \varepsilon^{i-1} \boldsymbol{\beta}^i(x/\varepsilon) \cdot \nabla^i \boldsymbol{u}_\varepsilon^*(x) \end{cases} \qquad \forall x \in D_\varepsilon. \tag{7.2.12}$$

The ansatz (7.2.12) is not standard and is different from (7.2.11); it is "even more" criminal because it expresses *both* oscillating solution $\boldsymbol{u}_\varepsilon$, p_ε in terms of their formal, non-oscillating averaged $\boldsymbol{u}_\varepsilon^*$, p_ε^* (which both depend themselves on ε, see (7.5.3)). Since the average of the tensor N^0 over the unit cell P is the 2×2 identity tensor while the average of the other tensors $(N^k)_{k\geq 1}$ and $(\boldsymbol{\beta}^k)_{k\geq 0}$ is zero (see proposition 7.34), formally averaging (7.2.12) with respect to the fast variable x/ε shows that the variable $(\boldsymbol{u}_\varepsilon^*, p_\varepsilon^*)$ can be indeed interpreted as a formal homogenized average of $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ of "infinite" order.

Criminal ansatz have been first derived in Bakhvalov and Panasenko (1989) [53] concerned with the conductivity equation with heterogeneous diffusivity tensor (and no holes). They have then been used in Smyshlyaev and Cherednichenko (2000) [287] to obtain higher order homogenized equation for a scalar elasticity model with discontinuous Hooke's tensor, and in [274, 33] in the context of the wave equation.

In proposition 7.32, we obtain that $(\boldsymbol{u}_\varepsilon^*, p_\varepsilon^*)$ solves the following formal "infinite-order" homogenized equation,

$$\begin{cases} \displaystyle\sum_{k=0}^{+\infty} \varepsilon^{k-2} M^k \cdot \nabla^k \boldsymbol{u}_\varepsilon^* + \nabla p_\varepsilon^* = \boldsymbol{f}, \\ \\ \operatorname{div}(\boldsymbol{u}^*) = 0, \\ \\ \boldsymbol{u}_\varepsilon^* \text{ is } D\text{–periodic.} \end{cases} \tag{7.2.13}$$

which involves a different family of constant tensors $(M^k)_{k\in\mathbb{N}}$. Our higher order homogenized models (7.2.9) of order $2K+2$ turn out to be truncations of (7.2.13) in the sense that the first $K+1$ coefficients coincide: $\mathbb{D}_K^k = M^k$ for $0 \leq k \leq K$. Truncating directly (7.2.13) yields, in general, an ill-posed model [20]. This fact is classical and several methods have been proposed to address this issue and obtain nonetheless well-posed higher order models see e.g. [35, 22, 3, 4, 33]. In our case, we adapt an idea from [287], whereby the coefficients $\mathbb{D}_K^k$ for $K+1 \leq k \leq 2K+2$ are obtained thanks to a minimization principle which makes indeed (7.2.9) well-posed. Let us stress, however, that this choice is not unique.

A rather surprising feature lies in the fact that (7.2.9) and (7.2.13) feature operators of *odd* order terms. This fact is to relate the vectorial context: the tensors $\mathbb{D}_K^k$ and M^k are symmetric and anti-symmetric valued matrices for respectively even and odd k (corollary 7.9). This property ensures that eventually, the operators $\mathbb{D}_K^k \cdot \nabla^k$ and $M^k \cdot \nabla^k$ are symmetric operators (see remark 7.20). Tensors $\mathbb{D}_K^k$ (with $k = 2p+1$, $p \in \mathbb{N}$) are *a priori* not zero (this assertion may be difficult to prove, however numerical results supporting this assertion are provided in section 7.4.5 below); we shall see nevertheless that these tensors vanish in the case where the obstacle ηT is symmetric with respect to the cell axes.

A legitimate question is to ask whether (7.2.9) or (7.2.13) reduces indeed to either of the equations (7.2.5) to (7.2.8) under the various possible scaling regimes for the sizes of the periodic obstacles. The answer is affirmative let us first mention that keeping the terms of lowest powers in (7.2.13) yields the Darcy model (7.2.8) (the tensor M^0 is the same). Furthermore, we prove the following asymptotics in the low volume fraction limit $\eta \to 0$ (in corollary 7.12) and assuming $d \geq 3$:

$$\varepsilon^{-2} M^0 \sim \eta^{d-2}/\varepsilon^2 \Psi^*, \tag{7.2.14}$$

$$\varepsilon^{-1} M^1 = o\left(\varepsilon(\eta^{d-2}/\varepsilon^2)\right), \tag{7.2.15}$$

$$\varepsilon^0 M^2 \to -I, \tag{7.2.16}$$

$$\forall k \geq 1,\ \varepsilon^{2k-2} M^{2k} = o\left(\left(\frac{\varepsilon^2}{\eta^{(d-2)}}\right)^{k-1}\right), \tag{7.2.17}$$

$$\forall k \geq 1,\ \varepsilon^{2k-1} M^{2k+1} = o\left(\varepsilon\left(\frac{\varepsilon^2}{\eta^{(d-2)}}\right)^{k-1}\right). \tag{7.2.18}$$

where I is the d-by-d identity matrix. The first convergence result (7.2.14) has been obtained in [13], it expresses the continuous transition of the Darcy tensor M^0 in (7.2.7) towards the Brinkman tensor Ψ^* in

(7.2.6). These asymptotics involve the ratio η^{d-2}/ε^2, which naturally brings into play the classical critical scaling $\eta \sim \varepsilon^{2/(d-2)}$ (i.e. $\sigma_\varepsilon = \varepsilon\varepsilon^{2/(d-2)} = \varepsilon^{d/(d-2)}$ for the holes). Therefore, the Stokes, Brinkman, or Darcy regimes (7.2.5) to (7.2.7) are retrieved in sense that the coefficients M^k of (7.2.13) converge to those of these three equations as $\eta, \varepsilon \to 0$.

To conclude, let us outline how such higher order homogenized models could be used for topology optimization of fluid flows. For $K = 0$, (7.2.9) rewrites in the form of a second order elliptic problem:

$$\begin{cases} \mathbb{D}_0^2 \cdot \nabla^2 \boldsymbol{v}_0^* + \varepsilon^{-1}\mathbb{D}_0^1 \cdot \nabla \boldsymbol{v}_0^* + \varepsilon^{-2}\mathbb{D}_0^0 \boldsymbol{v}_0^* + \nabla q_K^* = \boldsymbol{f}, \\ \qquad\qquad \operatorname{div}(\boldsymbol{v}_0^*) = 0, \\ \qquad\qquad \boldsymbol{v}_0^* \text{ is } D\text{–periodic.} \end{cases} \tag{7.2.19}$$

We shall see that (7.2.19) yields the same order of approximation $O(\varepsilon^3)$ of $\boldsymbol{u}_\varepsilon$ as the more classical Darcy equation (7.2.8). However, (7.2.19) makes sense for all three regimes simultaneously: it provides missing corrector terms that are not visible, at first order, in the Darcy regime (7.2.7). Further, it can be verified that the tensors $\mathbb{D}_0^2$ satisfy the same asymptotics as the tensors M^k for $0 \leq k \leq 2$. The formulation (7.2.19) is expected to be amenable for the development of a numerical topology optimization method: upon a suitable choice of parameterization $(a_1, \dots, a_m)$ of the microstructure, a relaxed formulation of (7.2.3) could be

$$\min_{(a_1,\dots,a_m)\in L^\infty(D,\mathbb{R}^d)} \quad J^*(a_1, \dots, a_m, \boldsymbol{v}_K^*(a_1, \dots, a_m), q_K^*(a_1, \dots, a_m))$$
$$s.t. \begin{cases} \mathbb{D}_0^2(a_1, \dots, a_m) \cdot \nabla^2 \boldsymbol{v}_0^* + \varepsilon^{-1}\mathbb{D}_0^1(a_1, \dots, a_m) \cdot \nabla \boldsymbol{v}_0^* + \varepsilon^{-2}\mathbb{D}_0^0(a_1, \dots, a_m)\boldsymbol{v}_0^* + \nabla q_K^* = \boldsymbol{f}, \\ \boldsymbol{v}_K^* \text{ is } D\text{–periodic,} \end{cases} \tag{7.2.20}$$

where J^* would be a consistent relaxation of J and the functions $(a_1, \dots, a_m) \mapsto \mathbb{D}_0^k(a_1, \dots, a_m)$ could be computed in an off-line step (following [27]). This model is second order for any type of microstructure and contains all three regimes simultaneously, which solves our initial issue. Higher order models could be used similarly to capture contexts where the size ε is only moderately small. Our work opens interesting perspectives for the topology optimization of liquid-liquid heat exchangers because our model can be extended in a straightforward manner to multicomponent fluid domains, see the appendix of section 7.5.6) for a discussion.

Finally let us notice that (7.2.20) is a generalization of the model of Borrvall and Petersson [71, 72] which is commonly used in density based methods for fluid topology optimization: it is a version of (7.2.6) where the Brinkman tensor Ψ^* is assumed to be a scalar and is used as a penalization to enforce $\boldsymbol{v} = 0$ in "solid" regions modeled by large values of Ψ^*. Some variants, however, have been proposed such as the Darcy model [324].

7.2.4 Setting and notation conventions related to tensors

In all what follows, the setting considered is that of Figure 7.3. $D := [-L, L]^d$ is a d-dimensional box filled with periodic obstacles $\omega_\varepsilon := \varepsilon(\mathbb{Z}^d + \eta T) \cap D$. $P = (0, 1)^d$ is the unit cell, and $Y := P \backslash \eta T$ the unit perforated cell. The boundary of the obstacle T is assumed to be smooth. The perforated domain is $D_\varepsilon := D \backslash \omega_\varepsilon$.

Below and further on, we consider scalar and vectorial functions such as

$$u : D \times P \to \mathbb{R}, \qquad \boldsymbol{u} : D \times P \to \mathbb{R}^d$$

which are both D and P–periodic with respect to respectively the first and the second variable. Their respective values are denoted by $u(x, y)$ and $\boldsymbol{u}(x, y)$ where the arguments x and y are called respectively the "slow" and the "fast" variable. The partial derivative with respect to the variable y_j (respectively x_j) is written simply ∂_j instead of ∂_{y_i} (respectively ∂_{x_j}) where the context is clear, *i.e.* when the function to which it is applied depends only on y (respectively only on x).

The star–"$*$"– symbol is used to denote the average of such functions with respect to the y variable:

$$\boldsymbol{u}^*(x) := \int_P \boldsymbol{u}(x, y)\mathrm{d}y, \qquad u^*(x) := \int_P u(x, y)\mathrm{d}y.$$

For functions depending only on the y variable, it will be sometimes more convenient to write this average with the usual angle bracket symbols. For instance, if $\boldsymbol{\mathcal{X}} : P \to \mathbb{R}^d$ is a vector field in the cell P, then its average is defined by

$$< \boldsymbol{\mathcal{X}} >:= \int_P \boldsymbol{\mathcal{X}}(y)\mathrm{d}y.$$

In all what follows, unless otherwise specified, the Einstein summation convention over repeated *subscript* indices is assumed (but never on *superscript* indices). Vectors of $\mathbb{R}^d$ are written in bold face notation.

The notation conventions used for tensor related operations are summarized in the nomenclature below.

Scalar, vector, and matrix valued tensors and their coordinates

$\boldsymbol{b}$	Vector of $\mathbb{R}^d$
$(b_j)_{1\leq j\leq d}$	Coordinates of the vector $\boldsymbol{b}$.
b^k	Scalar valued tensor of order k ($b^k_{i_1\cdots_k} \in \mathbb{R}$ for $1 \leq i_1, \ldots, i_k \leq d$)
$\boldsymbol{b}^k$	Vector valued tensor of order k ($\boldsymbol{b}^k_{i_1\ldots i_k} \in \mathbb{R}^d$ for $1 \leq i_1, \ldots, i_k \leq d$)
B^k	Matrix valued tensor of order k ($B^k_{i_1\ldots i_k} \in \mathbb{R}^{d\times d}$ for $1 \leq i_1, \ldots, i_k \leq d$)
$(b^k_j)_{1\leq j\leq d}$	Coordinates of the vector valued tensor $\boldsymbol{b}^k$ seen as *scalar* tensors of order k.
$(B^k_{lm})_{1\leq l,m\leq d}$	Coefficients of the matrix valued tensor B^k seen as *scalar* tensors of order k.
$b^k_{i_1\ldots i_k,j}$	Coefficient of the vector valued tensor $\boldsymbol{b}^k$ ($1 \leq i_1, \ldots i_k \leq d$ and $1 \leq j \leq d$).
$B^k_{i_1\ldots i_k,lm}$	Coefficient of the matrix valued tensor B^k ($1 \leq i_1, \ldots i_k \leq d$ and $1 \leq l, m \leq d$).

Tensor products

$b^p \otimes c^{k-p}$ Tensor product of scalar tensors of order p and k:

$$(b^p \otimes c^{k-p})_{i_1\ldots i_p} := b^p_{i_1\ldots i_p} c^{k-p}_{i_{p+1}\ldots i_k}. \tag{7.2.21}$$

$a^p \otimes \boldsymbol{b}^{k-p}$ Tensor product of a p-th order scalar tensor a^p and a $k-p$-th order vector valued tensor $\boldsymbol{b}^{k-p}$:

$$(a^p \otimes \boldsymbol{b}^{k-p})_{i_1\ldots i_k} := a^p_{i_1\ldots i_p} \boldsymbol{b}^{k-p}_{i_{p+1}\ldots i_k}. \tag{7.2.22}$$

$B^p \otimes C^{k-p}$ Tensor product of the p-th order and a $k-p$-th order $d \times d$ matrix valued tensors $B^p = (B^p_{lm})_{1\leq l,m\leq d}$ and $C^{k-p} = (C^{k-p}_{lm})_{1\leq l,m\leq d}$:

$$(B^p \otimes C^{k-p})_{i_1\ldots i_k,lm} := B^p_{i_1\ldots i_p,lj} C^{k-p}_{i_{p+1}\ldots i_k,jm}, \tag{7.2.23}$$

Hence a matrix product is implicitly assumed in the notation $B^p \otimes C^{k-p}$.

$B^p : C^{k-p}$ Tensor product of two p-th order and $k-p$-th order matrix valued tensors:

$$(B^p : C^{k-p})_{i_1\ldots i_k} := B^p_{i_1\ldots i_p,lm} C^{k-p}_{i_{p+1}\ldots i_k,lm}. \tag{7.2.24}$$

$\boldsymbol{b}^p \cdot \boldsymbol{c}^{k-p}$ Tensor product of two vector valued tensors $\boldsymbol{b}^p$ and $\boldsymbol{c}^{k-p}$:

$$(\boldsymbol{b}^p \cdot \boldsymbol{c}^{k-p})_{i_1\ldots i_k} := b^p_{i_1\ldots i_k,m} c^{k-p}_{i_{p+1}\ldots i_k,m}. \tag{7.2.25}$$

$B^p \cdot \boldsymbol{c}^{k-p}$ Tensor product of a matrix valued tensor B^p and a vector valued tensors $\boldsymbol{c}^{k-p}$:

$$(B^p \cdot \boldsymbol{c}^{k-p})_{i_1\ldots i_k,l} := B^p_{i_1\ldots i_k,lm} c^{k-p}_{i_{p+1}\ldots i_k,m}. \tag{7.2.26}$$

Contraction with partial derivatives

$b^k \cdot \nabla^k$ Differential operator of order k associated with a scalar tensor b^k: for any smooth scalar field v,

$$b^k \cdot \nabla^k := b^k_{i_1 \dots i_k} \partial^k_{i_1 \dots i_k}. \tag{7.2.27}$$

$\boldsymbol{b}^k \cdot \nabla^k$ Differential operator of order k associated with a vector tensor $\boldsymbol{b}^k$: for any smooth vector field $\boldsymbol{v}$,

$$\boldsymbol{b}^k \cdot \nabla^k \boldsymbol{v} \ = b^k_{i_1 \dots i_k, l} \partial^k_{i_1 \dots i_k} v_l. \tag{7.2.28}$$

$B^k \cdot \nabla^k$ Differential operator of order k associated with a matrix valued tensor B^k: for any smooth vector field $\boldsymbol{v}$,

$$(B^k \cdot \nabla^k \boldsymbol{v})_l \ = B^k_{i_1 \dots i_k, lm} \partial^k_{i_1 \dots i_k} v_m. \tag{7.2.29}$$

Special tensors

δ_{ij} Kronecker symbol: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

I Identity tensor of order 2:

$$I_{i_1 i_2} = \delta_{i_1 i_2} = e_j \otimes e_j.$$

The identity tensor is another notation for the Kronecker tensor.

I^{2k} Identity tensor of order $2k$:

$$I^{2k} := \overbrace{I \otimes I \otimes \cdots \otimes I}^{k \text{ times}}.$$

$(\boldsymbol{e}_j)_{1 \leq j \leq d}$ Vectors of the canonical basis of $\mathbb{R}^d$.

e_j Scalar valued tensor of order 1 given by $e_{j,i_1} := \delta_{i_1 j}$ (with $1 \leq j \leq d$).

$\mathbb{B}^{k,l}_K$ Bilinear tensor of order $k + l$ which may be either scalar or matrix valued:

- if $\mathbb{B}^{k,l}_K$ is a *scalar* tensor, then for any smooth scalar fields v and w:

$$\mathbb{B}^{k,l}_K \nabla^k v \nabla^l w := \mathbb{B}^{k,l}_{K, i_1 \dots i_k, j_1 \dots j_l} \partial^k_{i_1 \dots i_k} v \, \partial^l_{j_1 \dots j_l} w.$$

- if $\mathbb{B}^{k,l}_K$ is a *matrix* valued tensor, then for any smooth vector fields $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{C}^\infty(\mathbb{R}^d)$:

$$\mathbb{B}^{k,l}_K \nabla^k v \nabla^l w := \mathbb{B}^{p,k}_{K, i_1 \dots i_p, j_1 \dots j_k, lm} \partial^p_{i_1 \dots i_p} v_l \, \partial^k_{j_1 \dots j_k} w_m.$$

With a small abuse of notation, we consider zeroth order tensors b^0 to be constants (i.e. $b^0 \in \mathbb{R}$) and we still denote by $b^0 \otimes c^k := b^0 c^k$ the tensor product with a k-th order tensor c^k. The same convention also applies to vector valued and matrix valued tensors.

Since a k-th order tensor (scalar, vector, or matrix valued) B^k makes sense when contracted with k partial derivatives as in (7.2.27) to (7.2.29), the order in which the indices $i_1, \dots, i_k$ are written in $B^k_{i_1, \dots i_k}$ does not matter in general. However the ordering of spatial indices l, m of a matrix valued tensor $B^k = (B^k_{i_1 \dots i_k, lm})$ *does* matter.

Finally, in the whole work, we write C, C_K or $C_K(\boldsymbol{f})$ to denote universal constants that do not depend on ε but whose values may change from lines to lines (and which may depend on η).

Remark 7.1. In a limited number of places, the superscript or subscript indices p and q are used. Naturally, these are not to be confused with the pressure variables p_ε or q_ε introduced in section 7.5.

7.3 High order homogenization for the perforated Poisson problem

Our first study is concerned with the high order homogenization of the perforated Poisson problem:

$$\begin{cases} -\Delta u_\varepsilon = f \text{ in } D_\varepsilon \\ \quad u_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon \\ \quad u_\varepsilon \text{ is } D\text{–periodic.} \end{cases} \tag{7.3.1}$$

where $f \in \mathcal{C}^\infty(D)$ is a smooth D–periodic right-hand side. The system (7.3.1) can be considered as a simplified scalar version of the Stokes problem (7.2.4). Our objective is to derive well-posed high order homogenized equations for (7.3.1) of arbitrary order, which reduce to the three classical regimes (see [103]) analogous to (7.2.5) to (7.2.7) in the low volume fraction limit $\eta \to 0$.

Remark 7.2. The D–periodicity assumption for f and u_ε is classical in homogenization; it is used in order to avoid difficulties related to boundary layers. There exist other settings in which we expect our analysis would also work, for instance if the domain D is smooth, if $f \in \mathcal{C}^\infty_c(D)$ is compactly supported in D, and if (7.3.1) is supplemented with a Dirichlet boundary condition on ∂D (see [217]).

Let us summarize the main steps of our analysis. We start by writing the traditional (see e.g. [217]) two-scale ansatz for the solution of (7.3.1):

$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} u_i(x, x/\varepsilon), \tag{7.3.2}$$

where $u_i(x,y)$ is a P–periodic function in y satisfying $u_i(x,y) = 0$ for $y \in \partial(\eta T)$. Here and in all what follows, equalities involving infinite power series such as (7.3.2) are formal and without a precise meaning of convergence (which shall rather be justified in the section 7.3.4 dedicated to error estimates). We search for higher order "homogenized" equations for the "homogenized" macroscopic approximation

$$u^*_\varepsilon(x) := \sum_{i=0}^{+\infty} \varepsilon^{i+2} \int_Y u_i(x,y)\mathrm{d}y. \tag{7.3.3}$$

Our method can then be decomposed into the following steps:

1. introducing a family of k-th order scalar tensors $(\mathcal{X}^k(y))_{k\in\mathbb{N}}$ (vanishing on ηT and obtained as the solutions of cell problems (see definition 7.1)), we show in proposition 7.1 that (7.3.2) reads explicitly:
$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i f(x). \tag{7.3.4}$$
Introducing the averaged tensors of order i, $\mathcal{X}^{i*} := \int_Y \mathcal{X}^i(y)\mathrm{d}y$, (7.3.3) reads similarly
$$u^*_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i f(x). \tag{7.3.5}$$

2. We construct constant tensors M^i by inversion of the formal equality
$$\left(\sum_{i=0}^{+\infty} \varepsilon^{i-2} M^i \cdot \nabla^i\right)\left(\sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i\right) = I, \tag{7.3.6}$$
which yields (by left multiplication in (7.3.5)) the following formal, infinite order homogenized equation for $u^*_\varepsilon(x)$:
$$\sum_{i=0}^{+\infty} \varepsilon^{i-2} M^i \cdot \nabla^i u^*_\varepsilon(x) = f(x). \tag{7.3.7}$$

3. We substitute the expression for $f(x)$ given by (7.3.7) into the ansatz (7.3.4) so as to recognize a formal double series product
$$u_\varepsilon(x) = \left(\sum_{i=0}^{+\infty} \varepsilon^i \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i\right)\left(\sum_{i=0}^{+\infty} \varepsilon^i M^i \cdot \nabla^i\right) u^*_\varepsilon(x). \tag{7.3.8}$$

Introducing a new family of tensors $N^k(y)$ defined by the Cauchy product

$$N^k(y) := \sum_{p=0}^{k} \mathcal{X}^p(y) \otimes M^{k-p}, \qquad y \in Y,$$

we are able to express the oscillating solution u_ε in terms of its formal average u_ε^* as follows:

$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i N^i(x/\varepsilon) \cdot \nabla^i u_\varepsilon^*(x). \tag{7.3.9}$$

4. In order to obtain well-posed homogenized equations of finite order, we consider truncated versions of functions of the form of (7.3.9):

$$w_\varepsilon(v)(x) := \sum_{i=0}^{K} \varepsilon^i N^i(x/\varepsilon) \cdot \nabla^i v(x).$$

We then formulate a minimization problem for v by restriction of the one associated with u_ε:

$$\begin{aligned} \min_{v \in H^{K+1}(D)} \quad & \int_D \left(\frac{1}{2} \left| \nabla w_\varepsilon(v) \right|^2 - f(x) w_\varepsilon(v)(x) \right) \mathrm{d}x \\ & s.t. v \text{ is } D\text{–periodic.} \end{aligned} \tag{7.3.10}$$

Averaging over the fast variable x/ε (by using lemma 7.3), we obtain a limit minimization problem involving an approximate energy J_K^* (see section 7.3.3) which does not depend on x/ε,

$$\begin{aligned} \min_{v \in H^{K+1}(D)} \quad & J_K^*(v, f, \varepsilon) \\ s.t. \quad & v \text{ is } D\text{–periodic.} \end{aligned} \tag{7.3.11}$$

Its Euler-Lagrange equation (see definition 7.5) yields a well-posed homogenized equation of order $2K+2$:

$$\sum_{k=0}^{K+1} \varepsilon^{2k-2} \mathbb{D}_K^{2k} \cdot \nabla^{2k} v_K^* = f, \tag{7.3.12}$$

where the tensors $\mathbb{D}_K^{2k}$ are inferred from J_K^*. We then prove in proposition 7.13 that its solution $v_K^* \in H^{K+1}(D)$ yields a high order approximation of u_ε thanks to the following error estimate:

$$\left\| u_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k v_K^* \right\|_{L^2(D)} \leq C_K(f) \varepsilon^{K+3}, \tag{7.3.13}$$

for a constant $C_K(f)$ which depends only on K and f.

The most essential step of the methodology is the derivation of the non-classical ansatz (7.3.9). We shall see (proposition 7.9) that the tensor N^0 is of average $\int_Y N^0(y)\mathrm{d}y = 1$ and N^k is of average $\int_Y N^k(y)\mathrm{d}y = 0$ for $k \geq 1$: averaging (7.3.9) with respect to the fast variable x/ε yields consistently the formal "homogenized average" u_ε^*.

Let us stress that in the available works of the literature concerned with high order homogenization of scalar conductivity equations and its variants [53, 287, 33] (where the inhomogeneity arises through the discontinuity of the conductivity coefficient), the criminal ansatz (analogous to (7.3.9)) is readily obtained from the classical one (analogous to (7.3.2)) because the tensors N^k and $\mathcal{X}^k$ coincide in these contexts (check for instance [53, 20]). This does not occur in our case because of the Dirichlet boundary condition on $\partial(\eta T)$.

The remainder of this part is organized as follows: in section 7.3.1, we examine the definition of the family of tensors $\mathcal{X}^k$ yielding the classical ansatz (7.3.3). We state several properties of the tensor $\mathcal{X}^k$ and review classical error estimate results for the traditional ansatz. In section 7.3.2, we detail the procedure which allows to obtain the family of constant tensors M^k and the criminal ansatz (7.3.9) involving the new tensors $N^k(y)$. This allows us to derive in section 7.3.3 the higher order "homogenized energy" J_K^* and its associated high order homogenized equation (7.3.12). We prove the well-posedness of this

equation and that its first $K+1$ coefficients coincide with those of (7.3.7). Section 7.3.4 is dedicated to the proof of the error estimate (7.3.13) justifying the higher order homogenization process. We show in section 7.3.5 that the coefficients of the infinite order homogenized equation (7.3.7) converge to those of the original Poisson equation (7.3.1) in the low-volume fraction limit when the obstacle's size $\eta \to 0$, which allows us to retrieve the three classical regimes and the arising of the celebrated "strange term" (see [103]) for the critical size $\eta \sim \varepsilon^{2/(d-2)}$. Lastly, we establish in section 7.3.6 symmetry properties for the homogenized tensors $\mathcal{X}^{k*}$ and M^k when the obstacle ηT is invariant with respect to cell symmetries.

7.3.1 Formal infinite order two-scale expansions and tensors $\mathcal{X}^k$

The first step of our methodology is to insert formally the ansatz (7.3.2) into the Poisson system (7.3.1). Because it will help highlight the occurrence of double series structures, we also assume (although it is not fully necessary) that the right-hand side f depends on ε and admits the following formal expansion:

$$f(x) = \sum_{i=0}^{+\infty} \varepsilon^i f_i(x).$$

Evaluating the Laplace operator against (7.3.2) yields then formally

$$-\Delta u_\varepsilon = \sum_{i=-2}^{+\infty} \varepsilon^{i+2}(-\Delta_{yy} u_{i+2} - \Delta_{xy} u_{i+1} - \Delta_{xx} u_i)$$

where we use the convention $u_{-2}(x,y) = u_{-1}(x,y) = 0$, and where $-\Delta_{yy}$, $-\Delta_{xy}$, $-\Delta_{xx}$ are the operators

$$-\Delta_{xx} := -\mathrm{div}_x(\nabla_x \cdot),\ -\Delta_{xy} := -\mathrm{div}_x(\nabla_y \cdot) - \mathrm{div}_y(\nabla_x \cdot),\ -\Delta_{yy} := -\mathrm{div}_y(\nabla_y \cdot).$$

Identifying all powers in ε yields then the traditional cascade of equations (obtained e.g. in [217]):

$$\begin{cases} -\Delta_{yy} u_{i+2} = f_{i+2} + \Delta_{xy} u_{i+1} + \Delta_{xx} u_i & \text{for all } i \geq -2 \\ u_{-2}(x,y) = u_{-1}(x,y) = 0. \end{cases} \tag{7.3.14}$$

Definition 7.1. We define the family of tensors $(\mathcal{X}^k_{i_1 \dots i_k}(y))_{k \in \mathbb{N}}$ of order k by recurrence as follows:

$$\begin{cases} -\Delta_{yy} \mathcal{X}^0 = 1 \text{ in } Y \\ -\Delta_{yy} \mathcal{X}^1 = 2\partial_j \mathcal{X}^0 \otimes e_j \text{ in } Y \\ -\Delta_{yy} \mathcal{X}^{k+2} = 2\partial_j \mathcal{X}^{k+1} \otimes e_j + \mathcal{X}^k \otimes I \text{ in } Y, \quad \text{for all } k \geq 0 \\ \mathcal{X}^k = 0 \text{ on } \partial(\eta T) \\ \mathcal{X}^k \text{ is } P\text{–periodic.} \end{cases} \tag{7.3.15}$$

Lemma 7.1. *Assume that the boundary of the perforated cell Y is smooth. Then, for any $k \in \mathbb{N}$, the tensor $\mathcal{X}^k$ is well-defined and is smooth, namely it holds $\mathcal{X}^k \in \mathcal{C}^\infty(\overline{Y})$. In particular, $\mathcal{X}^k \in L^\infty(Y) \cap H^1(Y)$.*

Proof. Since the constant function 1 is smooth, standard regularity theory for the Laplace operator $-\Delta_{yy}$ (see [176, 76, 146]) implies $\mathcal{X}^0 \in \mathcal{C}^\infty(\overline{Y})$. The result follows by induction by repeating this argument to $\mathcal{X}^1$ and $\mathcal{X}^{k+2}$ for any $k \geq 0$. □

Proposition 7.1. *The solutions $(u_i(x,y))_{i \geq 0}$ to the cascade of equations (7.3.14) are given by*

$$\forall i \geq 0,\ u_i(x,y) = \sum_{k=0}^{i} \mathcal{X}^k(y) \cdot \nabla^k f_{i-k}(x), \qquad x \in D,\ y \in Y. \tag{7.3.16}$$

Recognizing a Cauchy product, the ansatz (7.3.2) can be formally written as the following infinite power series product:

$$u_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i f(x) = \varepsilon^2 \left(\sum_{i=0}^{+\infty} \varepsilon^i \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i \right) \left(\sum_{i=0}^{+\infty} \varepsilon^i f_i(x) \right). \tag{7.3.17}$$

Proof. We proceed by induction. The cases $i = 0$ and $i = 1$ are simple consequences of (7.3.14). If the result holds until rank $i+1$, then

$$\begin{aligned}
-\Delta_{yy}u_{i+2}(x,y) &= f_{i+2}(x) + \sum_{k=0}^{i+1} 2\partial_j \mathcal{X}^k(y)\cdot\nabla^k(\partial_j f_{i+1-k}(x)) + \sum_{k=0}^{i}\mathcal{X}^k\cdot\nabla^k(\Delta f_{i-k}(x)) \\
&= f_{i+2}(x) + \sum_{k=-1}^{i} 2(\partial_j\mathcal{X}^{k+1}(y)\otimes e_j)\cdot\nabla^{k+2} f_{i-k}(x) + \sum_{k=0}^{i}(\mathcal{X}^k(y)\otimes I)\cdot\nabla^{k+2} f_{i-k}(x) \\
&= f_{i+2}(x) + 2\partial_j\mathcal{X}^0(y)\partial_j f_{i+1}(x) + \sum_{k=0}^{i}(-\Delta_{yy}\mathcal{X}^{k+2}(y))\cdot\nabla^{k+2} f_{i-k}(x).
\end{aligned}$$

Whence

$$u_{i+2}(x,y) = \mathcal{X}^0(y)f_{i+2}(x) + \mathcal{X}^1_j(y)\partial_j f_{i+1}(x) + \sum_{k=0}^{i}\mathcal{X}^{k+2}(y)\cdot\nabla^{k+2}f_{i-k}(x) = \sum_{k=0}^{i+2}\mathcal{X}^k(y)\cdot\nabla^k f_{i+2-k}(x).$$

□

In what follows, we are going to derive high order PDEs for the "infinite order" homogenized solution (7.3.3). Following the notation conventions introduced in section 7.2.4, we find convenient to introduce averaged functions u_i^* and tensors $\mathcal{X}^{i*}$ labelled with the star "*" notation:

Definition 7.2. For any $i \in \mathbb{N}$, the averaged functions and tensors with respect to the fast variable are denoted by:

$$u_i^*(x) := \int_Y u_i(x,y)\mathrm{d}y. \tag{7.3.18}$$

$$\mathcal{X}^{i*} := \int_Y \mathcal{X}^i(y)\mathrm{d}y. \tag{7.3.19}$$

Remark 7.3. These are average on the unit cell $P = [0,1]^d$, although the integral is written on the domain Y, because $u_i(x,\cdot)$ and $\mathcal{X}^i$ vanish on $P\backslash Y = \overline{\eta T}$ (see Figure 7.3).

In the next proposition we show that $\mathcal{X}^{2p*}$ depends only on the lower order tensors $\mathcal{X}^p$ and $\mathcal{X}^{p-1}$ and that $\mathcal{X}^{2p+1*}$ is null for any $p \in \mathbb{N}$. Similar formulas have been obtained for the wave equation in heterogeneous media, see e.g. Theorem 3.5 in [4], and also [3, 264].

Proposition 7.2. *For any $0 \le p \le k$, the following identity holds for the tensor $\mathcal{X}^{k*}$:*

$$\mathcal{X}^{k*} = \int_Y \mathcal{X}^k\mathrm{d}y = (-1)^p\int_Y(\mathcal{X}^{k-p}\otimes(-\Delta_{yy}\mathcal{X}^p) - \mathcal{X}^{k-p-1}\otimes\mathcal{X}^{p-1}\otimes I)\mathrm{d}y, \tag{7.3.20}$$

with the convention that $\mathcal{X}^{-1} = 0$. In particular, for any $p \in \mathbb{N}$:

- $\mathcal{X}^{2p+1*} = 0$
- *$\mathcal{X}^{2p*}$ depends only on the tensors $\mathcal{X}^p$ and $\mathcal{X}^{p-1}$:*

$$\mathcal{X}^{2p*} = (-1)^p\int_Y(\partial_j\mathcal{X}^p\otimes\partial_j\mathcal{X}^p - \mathcal{X}^{p-1}\otimes\mathcal{X}^{p-1}\otimes I)\mathrm{d}y. \tag{7.3.21}$$

Proof. We proceed again by induction. Formula (7.3.20) holds true for $p = 0$ by using the convention $\mathcal{X}^{-1} = 0$ and $-\Delta_{yy}\mathcal{X}^0 = 1$. Assuming now the result true for $p < k$, we may perform the following integration by parts thanks to the boundary conditions satisfied by the tensors $\mathcal{X}^k$:

$$\begin{aligned}
\mathcal{X}^{k*} =& (-1)^p\int_Y(\mathcal{X}^{k-p}\otimes(-\Delta_{yy}\mathcal{X}^p) - \mathcal{X}^{k-p-1}\otimes\mathcal{X}^{p-1}\otimes I)\mathrm{d}y \\
=& (-1)^p\int_Y(-\Delta_{yy}\mathcal{X}^{k-p}\otimes\mathcal{X}^p - \mathcal{X}^{k-p-1}\otimes\mathcal{X}^{p-1}\otimes I)\mathrm{d}y \\
=& (-1)^p\int_Y((2\partial_j\mathcal{X}^{k-p-1}\otimes e_j + \mathcal{X}^{k-p-2}\otimes I)\otimes\mathcal{X}^p - \mathcal{X}^{k-p-1}\otimes\mathcal{X}^{p-1}\otimes I)\mathrm{d}y \\
=& (-1)^p\int_Y(-2\partial_j\mathcal{X}^p\otimes e_j - \mathcal{X}^{p-1}\otimes I)\otimes\mathcal{X}^{k-p-1} + \mathcal{X}^{k-p-2}\otimes\mathcal{X}^p\otimes I)\mathrm{d}y \\
=& (-1)^{p+1}\int_Y((-\Delta_{yy}\mathcal{X}^{p+1})\otimes\mathcal{X}^{k-p-1} - \mathcal{X}^{k-p-2}\otimes\mathcal{X}^p\otimes I)\mathrm{d}y.
\end{aligned}$$

Hence the formula is proved at order $p+1$.
Now, the formula at order $p=k$ reads

$$\mathcal{X}^{k*} = -\int_Y \mathcal{X}^0(-\Delta_{yy}\mathcal{X}^k)\mathrm{d}y = -\int_Y \mathcal{X}^k(-\Delta_{yy}\mathcal{X}^0)\mathrm{d}y = -\mathcal{X}^{k*},$$

which implies $\mathcal{X}^{k*}=0$ if k is odd. Formula (7.3.21) follows easily from (7.3.20) with $k=2p$. □

The following result demonstrates that $\mathcal{X}^k(y)$ is not identically equal to zero (although some components $\mathcal{X}^k_{i_1\dots i_k}(y)$ could vanish for some set of indices $i_1,\dots,i_k$, e.g. in case of symmetries of the obstacle ηT).

Proposition 7.3. *The following identity holds:*

$$-\Delta_{yy}(\partial^k_{i_1\dots i_k}\mathcal{X}^k_{i_1\dots i_k}) = (-1)^k(k+1). \tag{7.3.22}$$

Proof. The results naturally holds true for $k=0$. For $k=1$, it holds

$$-\Delta_{yy}\partial_i\mathcal{X}^1_i = \partial_i(2\partial_i\mathcal{X}^0) = 2\Delta\mathcal{X}^0 = -2.$$

Assuming the result till rank $k-1$, the formula still holds at rank k because

$$\begin{aligned}
-\Delta_{yy}\partial^k_{i_1\dots i_k}\mathcal{X}^k_{i_1\dots i_k} &= \partial^k_{i_1\dots i_k}(2\partial_{i_k}\mathcal{X}^{k-1}_{i_1\dots i_{k-1}} + \mathcal{X}^{k-2}_{i_1\dots i_{k-2}}\delta_{i_{k-1}i_k}) \\
&= 2\Delta_{yy}(\partial^{k-1}_{i_1\dots i_{k-1}}\mathcal{X}^{k-1}_{i_1\dots i_{k-1}}) + \Delta_{yy}(\partial^{k-2}_{i_1\dots i_{k-2}}\mathcal{X}^{k-2}_{i_1\dots i_{k-2}}) \\
&= -2(-1)^{k-1}k - (-1)^{k-2}(k-1) \\
&= (-1)^k(k+1).
\end{aligned}$$

□

The next result can be found in classical textbooks, see e.g. [217]. It states error estimates for the truncated classical ansatz (7.5.2):

Proposition 7.4. *Denote $u_{\varepsilon,K}$ the truncated ansatz of (7.3.17) at order K:*

$$\forall x\in D,\ u_{\varepsilon,K}(x) := \sum_{i=0}^{K}\varepsilon^{i+2}\mathcal{X}^i(x/\varepsilon)\cdot\nabla^i f(x). \tag{7.3.23}$$

Then assuming f is D–periodic, the following bound holds:

$$||u_\varepsilon - u_{\varepsilon,K}||_{H^1(D_\varepsilon)} \le C_K\varepsilon^{K+2}||f||_{H^{K+2}(D)} \tag{7.3.24}$$

for a constant C_K independent of f and ε (but depending on K).

Proof. It is sufficient to observe that the remainder $r_\varepsilon(x) = u_\varepsilon(x) - u_{\varepsilon,K}(x)$ satisfies

$$-\Delta r_\varepsilon = (\varepsilon^{K+1}(\Delta_{xy}u_K + \Delta_{xx}u_{K-1}) + \varepsilon^{K+2}\Delta_{xx}u_K))(x,x/\varepsilon).$$

and the terms $(\Delta_{xy}u_K)(x,x/\varepsilon)$, $(\Delta_{xx}u_{K-1})(x,x/\varepsilon)$ and $(\Delta_{xx}u_K)(x,x/\varepsilon)$ can be bounded in the $L^2(D)$ norm by $C_K||f||_{H^{K+2}}$. The result follows from

$$||\nabla r_\varepsilon||^2_{L^2(D_\varepsilon,\mathbb{R}^d)} \le C_K\varepsilon^{K+1}||f||_{H^{K+2}(D)}||r_\varepsilon||_{L^2(D_\varepsilon)} \le C_K\varepsilon^{K+2}||f||_{H^{K+2}(D)}||\nabla r_\varepsilon||_{L^2(D_\varepsilon,\mathbb{R}^d)},$$

where we have used the classical Poincaré inequality of lemma 7.2 below. □

Lemma 7.2 (see e.g. Lions 1981 [217])**.** *There exists a constant C independent of ε such that for any $\phi\in H^1(D_\varepsilon)$ satisfying $\phi=0$ on the boundary $\partial\omega_\varepsilon$ of the holes, the following Poincaré inequality holds:*

$$||\phi||_{L^2(D_\varepsilon)} \le C\varepsilon||\nabla\phi||_{L^2(D_\varepsilon,\mathbb{R}^d)}.$$

In the final part of this section, we provide a few insights highlighting in which sense the homogenized average u^*_ε of (7.3.3) is an approximation of u_ε. Let us recall the following important lemma (see e.g. Appendix C of [287] or [16, 138]):

Lemma 7.3. *Let ϕ a $P=[0,1]^d$–periodic function and $f\in\mathcal{C}^\infty(D^d)$ a smooth D–periodic function. Then for any $k\in\mathbb{N}$ arbitrarily high, the following inequality holds:*

$$\left|\int_D f(x)\phi(x/\varepsilon)\mathrm{d}x-\int_D\int_P f(x)\phi(y)\mathrm{d}y\mathrm{d}x\right|\leq\frac{(2L)^{d/2}}{|2\pi|^k}\left|\left|\phi-\int_P\phi\mathrm{d}y\right|\right|_{L^2(P)}||f||_{H^k(D)}\varepsilon^k. \tag{7.3.25}$$

Proof. The key idea is to perform integration by parts by differentiating f and integrating ϕ. Without loss of generality, we may assume ϕ to be of average zero, i.e. $\int_P\phi\mathrm{d}y=0$. We show by induction that for any $k\geq 0$, there exists a function $\phi^k_{i_1\dots i_k}$ with $1\leq i_1\dots i_k\leq d$ satisfying

$$\sum_{1\leq i_1,\dots i_k\leq d}||\phi^k_{i_1\dots i_k}||^2_{L^2(P)}\leq\frac{1}{|4\pi^2|^k}||\phi||^2_{L^2(P)}, \tag{7.3.26}$$

$$\int_P\phi^k_{i_1\dots i_k}\mathrm{d}y=0, \tag{7.3.27}$$

and such that

$$\int_D f(x)\phi(x/\varepsilon)\mathrm{d}y=(-1)^k\int_D\varepsilon^k\partial^k_{i_1\dots i_k}f\phi^k_{i_1\dots i_k}(x/\varepsilon)\mathrm{d}x. \tag{7.3.28}$$

Obviously, the above identities hold true for $k=0$. Assuming now that the results holds till rank $k\geq 0$, we introduce $\psi^k_{i_1\dots i_k}\in H^1(P)$ the unique solution to the following Laplace problem:

$$\begin{cases}-\Delta\psi^k_{i_1\dots i_k}=\phi^k_{i_1\dots i_k}\text{ in }P\\ \psi^k_{i_1\dots i_k}\text{ is }P\text{–periodic}\\ \displaystyle\int_P\psi^k_{i_1\dots i_k}\mathrm{d}y=0.\end{cases} \tag{7.3.29}$$

Such function actually exists because of the compatibility condition (7.3.27); this allows us to define

$$\phi^{k+1}_{i_1\dots i_{k+1}}:=\partial_{i_{k+1}}\psi^k_{i_1\dots i_k}.$$

It is clear from the periodicity of $\psi^k_{i_1\dots i_k}$ that (7.3.27) holds at rank $k+1$, and it holds from the variational formulation of (7.3.29)

$$\sum_{i_{k+1}=1}^d||\phi^{k+1}_{i_1\dots i_{k+1}}||^2_{L^2(P)}=||\nabla\psi^k_{i_1\dots i_k}||^2_{L^2(P,\mathbb{R}^d)}\leq\frac{1}{4\pi^2}||\phi^k_{i_1\dots i_k}||^2_{L^2(P)},$$

because $1/(4\pi^2)$ is the Poincaré constant of $[0,1]^d$. Summing over the indices $i_1\dots i_k$ and applying hypothesis (7.3.26) at rank k yields the same estimate (7.3.26) at rank $k+1$. We may subsequently write

$$\begin{aligned}\int_D\partial^k_{i_1\dots i_k}f\phi^k_{i_1\dots i_k}(x/\varepsilon)\mathrm{d}x&=\int_D\varepsilon\partial^k_{i_1\dots i_k}f\mathrm{div}(\phi^{k+1}_{i_1\dots i_{k+1}}(x/\varepsilon)\boldsymbol{e}_{i_{k+1}})\mathrm{d}x\\&=-\int_D\varepsilon\partial^{k+1}_{i_1\dots i_{k+1}}f(x)\phi^{k+1}_{i_1\dots i_{k+1}}(x/\varepsilon)\mathrm{d}x,\end{aligned}$$

from where (7.3.28) follows at rank $k+1$.

Let us finally prove (7.3.25). From (7.3.28) and the Cauchy-Schwartz inequality, we obtain

$$\left|\int_D f(x)\phi(x/\varepsilon)\mathrm{d}x\right|^2\leq\varepsilon^{2k}||f||^2_{H^k(D)}\sum_{1\leq i_1\dots i_k}\int_D|\phi^k_{i_1\dots i_k}(x/\varepsilon)|^2\mathrm{d}x\leq\varepsilon^{2k}||f||^2_{H^k(D)}\frac{(2L)^d}{|4\pi^2|^k}||\phi||^2_{L^2(P)},$$

whence the result. □

Remark 7.4. The constant 2π in (7.3.25) is related to the fact that the unit cell is of size 1 (it would be one for a cell of length 2π), it is therefore no surprise that it scales with the same power k than ε.

This lemma implies that u^*_ε is an approximation of u_ε in a distributional sense:

Proposition 7.5. *Denote by $u^*_{\varepsilon,K}$ the average of the truncated ansatz $u_{\varepsilon,K}$ (eqn. (7.3.23)) with respect to the fast variable, namely*

$$u^*_{\varepsilon,K}(x) := \sum_{i=0}^{K} \varepsilon^{i+2} u^*_i(x) = \sum_{i=0}^{K} \varepsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i f(x). \tag{7.3.30}$$

Then $u^*_{\varepsilon,K}$ is a $(K+3)-th$ order approximation of u_ε in the following weak sense:

$$\forall \psi \in \mathcal{C}^\infty_c(D), \left| \int_D u_\varepsilon(x)\psi(x)\mathrm{d}x - \int_D u^*_{\varepsilon,K}(x)\psi(x)\mathrm{d}x \right| \leq C_K(\psi)||f||_{H^{K+2}(D)}\varepsilon^{K+3},$$

for a constant $C_K(\psi)$ depending only on K and ψ.

Proof. We write

$$\begin{aligned} &\left| \int_D u_\varepsilon(x)\psi(x)\mathrm{d}x - \int_D u^*_{\varepsilon,K}(x)\psi(x)\mathrm{d}x \right| \\ &\qquad \leq \int_D |(u_\varepsilon(x) - u_{\varepsilon,K}(x))\psi(x)|\mathrm{d}x + \left| \int_D (u_{\varepsilon,K}(x) - u^*_{\varepsilon,K}(x))\psi(x)\mathrm{d}x \right| \end{aligned} \tag{7.3.31}$$

Let $M_K(\psi) > 0$ a constant such that $||\nabla^k \psi||_{L^\infty(D,\mathbb{R}^d)} \leq M_K(\psi)$ for any $0 \leq k \leq K+1$. The first term is bounded by $||u_\varepsilon - u_{\varepsilon,K}||_{L^2(D)}||\psi||_{L^2(D)} \leq C_K M_K(\psi)\varepsilon^{K+3}||f||_{H^{K+2}(D)}$ according to proposition 7.4. The second term is bounded by $C_K M_K(\psi)||f||_{H^{K+1}(D)}\varepsilon^{K+3}$ (up to the use of a larger constant C_K) as a consequence of the important Lemma 7.3 stated above. □

Remark 7.5. More elaborate arguments can be proposed for providing $u^*_\varepsilon(x)$ with a physical interpretation, e.g. by considering shifted cell averages (see [287, 102]).

7.3.2 Formal infinite order homogenized equation and criminal ansatz: tensors M^k and N^k

We now detail the steps (2) and (3) of the procedure outlined in the introduction of this section. Let us recall that the first tensor $\mathcal{X}^{0*}$ is a strictly positive number, since (7.3.21) implies $\mathcal{X}^{0*} = \int_Y |\nabla \mathcal{X}^0|^2 \mathrm{d}y$.

Proposition 7.6. *Let $(M^k)_{i\in\mathbb{N}}$ be the family of k-th order tensors defined by induction as follows:*

$$\begin{cases} M^0 = (\mathcal{X}^{0*})^{-1}, \\ M^k = -(\mathcal{X}^{0*})^{-1} \displaystyle\sum_{p=0}^{k-1} \mathcal{X}^{k-p*} \otimes M^p. \end{cases} \tag{7.3.32}$$

Then it holds, given the definitions (7.3.14) and (7.3.18) for u^*_i:

$$\forall i \geq 0,\ f_i(x) = \sum_{k=0}^{i} M^k \cdot \nabla^k u^*_{i-k}(x). \tag{7.3.33}$$

Recognizing a Cauchy product, this can be rewritten formally in terms of the following "infinite order" homogenized equation for u^*_ε (defined in (7.3.5)):

$$\sum_{i=0}^{+\infty} \varepsilon^{i-2} M^i \cdot \nabla^i u^*_\varepsilon = f. \tag{7.3.34}$$

Proof. We proceed by induction. The case $i = 0$ results from the identity $u^*_0(x) = \mathcal{X}^{0*} f_0(x)$ which yields $f_0(x) = (\mathcal{X}^{0*})^{-1} u^*_0(x)$. Then, assuming the result (7.3.33) holds till rank $i-1$, we average equation (7.3.16) with respect to the y variable to obtain

$$u^*_i = \sum_{p=0}^{i} \mathcal{X}^{p*} \cdot \nabla^p f_{i-p} = \mathcal{X}^{0*} f_i + \sum_{p=1}^{i} \mathcal{X}^{p*} \cdot \nabla^p f_{i-p}.$$

By inversion of the nonzero coefficient $\mathcal{X}^{0*}$, we obtain the following expression for f_i:

$$\begin{aligned}
f_i &= (\mathcal{X}^{0*})^{-1}\left(u_i^* - \sum_{p=1}^{i}\sum_{q=0}^{i-p}(\mathcal{X}^{p*}\otimes M^q)\cdot\nabla^{p+q}u_{i-p-q}^*\right)\\
&= (\mathcal{X}^{0*})^{-1}\left(u_i^* - \sum_{p=1}^{i}\sum_{k=p}^{i}(\mathcal{X}^{p*}\otimes M^{k-p})\cdot\nabla^{k}u_{i-k}^*\right) \qquad \text{(change of indices } k=p+q\text{)}\\
&= (\mathcal{X}^{0*})^{-1}\left(u_i^* - \sum_{k=1}^{i}\sum_{p=1}^{k}(\mathcal{X}^{p*}\otimes M^{k-p})\cdot\nabla^{k}u_{i-k}^*\right) \qquad \text{(inversion of summation)}\\
&= (\mathcal{X}^{0*})^{-1}\left(u_i^* - \sum_{k=1}^{i}\left(\sum_{p=0}^{k-1}\mathcal{X}^{k-p*}\otimes M^p\right)\cdot\nabla^{k}u_{i-k}^*\right) \qquad \text{(change of index } p\leftrightarrow k-p\text{)}\\
&= M^0u_i^* + \sum_{k=1}^{i}M^k\cdot\nabla^k u_{i-k}^*,
\end{aligned}$$

which concludes the proof. □

Corollary 7.1. *$M^k = 0$ for any odd value of k.*

Proof. This follows from proposition 7.2 and the recurrence formula (7.3.32). □

It is possible to write a more explicit formula for the tensors M^k:

Proposition 7.7. *The tensors M^k are explicitly given by $M^0 = (\mathcal{X}^{0*})^{-1}$ and:*

$$\forall k\geq 1,\ M^k = \sum_{p=1}^{k}\frac{(-1)^p}{(\mathcal{X}^{0*})^{p+1}}\sum_{\substack{i_1+\cdots+i_p=k\\ 1\leq i_1\ldots i_p\leq k}}\mathcal{X}^{i_1*}\otimes\cdots\otimes\mathcal{X}^{i_p*}. \tag{7.3.35}$$

Proof. We prove it by induction. For $k=1$, the result is true because

$$M^1 = -(\mathcal{X}^{0*})^{-1}M^0\mathcal{X}^{1*} = -(\mathcal{X}^{0*})^{-2}\mathcal{X}^{1*}$$

which is exactly (7.3.35). Assuming (7.3.35) holds till rank $k\geq 1$, we now compute

$$\begin{aligned}
M^{k+1} &= -(\mathcal{X}^{0*})^{-1}\sum_{p=0}^{k}\mathcal{X}^{k+1-p*}\otimes M^p\\
&= -(\mathcal{X}^{0*})^{-1}M^0\mathcal{X}^{k+1*} - (\mathcal{X}^{0*})^{-1}\sum_{p=1}^{k}\sum_{q=1}^{p}\frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}}\mathcal{X}^{k+1-p*}\otimes\sum_{\substack{i_1+\cdots+i_q=p\\ 1\leq i_1\ldots i_q\leq p}}\mathcal{X}^{i_1*}\otimes\cdots\otimes\mathcal{X}^{i_q*}\\
&= -(\mathcal{X}^{0*})^{-2}\mathcal{X}^{k+1*} - (\mathcal{X}^{0*})^{-1}\sum_{q=1}^{k}\frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}}\sum_{p=q}^{k}\sum_{\substack{i_1+\cdots+i_q=p\\ 1\leq i_1\ldots i_q\leq p}}\mathcal{X}^{k+1-p*}\otimes\mathcal{X}^{i_1*}\otimes\cdots\otimes\mathcal{X}^{i_q*}\\
&= -(\mathcal{X}^{0*})^{-2}\mathcal{X}^{k+1*} - (\mathcal{X}^{0*})^{-1}\sum_{q=1}^{k}\frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}}\sum_{\substack{i_1+\cdots+i_{q+1}=k+1\\ 1\leq i_1\ldots i_{q+1}\leq k+1}}\mathcal{X}^{i_{q+1}*}\otimes\mathcal{X}^{i_1*}\otimes\cdots\otimes\mathcal{X}^{i_q*}\\
&= -(\mathcal{X}^{0*})^{-2}\mathcal{X}^{k+1*} + \sum_{q=2}^{k+1}\frac{(-1)^q}{(\mathcal{X}^{0*})^{q+1}}\sum_{\substack{i_1+\cdots+i_q=k+1\\ 1\leq i_1\ldots i_q\leq k+1}}\mathcal{X}^{i_1*}\otimes\cdots\otimes\mathcal{X}^{i_q*},
\end{aligned}$$

from where the result follows. □

Remark 7.6. This result essentially states that $\sum_{k=0}^{+\infty}\varepsilon^k M^k\cdot\nabla^k$ is the formal series expansion of

$$\left(\sum_{k=0}^{+\infty}\varepsilon^k\mathcal{X}^k\cdot\nabla^k\right)^{-1}.$$

Indeed, it is elementary to show the following identity for the inverse of a power series $\sum_{k=0}^{+\infty} a_k z^k$ with $(a_k) \in \mathbb{C}^N$, $z \in \mathbb{C}$ and radius of convergence $R > 0$:

$$\left(\sum_{k=0}^{+\infty} a_k z^k\right)^{-1} = a_0^{-1} + \sum_{k=1}^{+\infty} \left(\sum_{p=1}^{k} \frac{(-1)^p}{a_0^{p+1}} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\le i_1 \dots i_p \le k}} a_{i_1} a_{i_2} \dots a_{i_p} \right) z^k. \tag{7.3.36}$$

The infinite order homogenized equation (7.3.34) is not fully satisfactory because it is not immediately clear how to truncate the operator $\sum_{i=0}^{+\infty} \varepsilon^i M^i \cdot \nabla^i$ so as to obtain a well-posed problem. Several techniques have been proposed in the literature to obtain well posed equations in the context of the homogenization of the conductivity or wave equation; we can mention among them a Boussinesq trick [35, 22, 3, 4] or a filtering method [33]. Here, we propose to follow a "variational" method inspired by [287], which is based on the existence of a "criminal" ansatz of the form (7.3.9). This ansatz is obtained by writing the oscillatory part $u_i(x, y)$ in terms of the non oscillatory part $u_i^*(x)$, which is obtained from the formal equality (7.3.8):

Proposition 7.8. *Given the previous definitions of $\mathcal{X}^i, M^i, u_i^*$, the following equality holds:*

$$\forall i \ge 0,\ u_i(x, y) = \sum_{k=0}^{i} \left(\sum_{p=0}^{k} M^p \otimes \mathcal{X}^{k-p}(y) \right) \cdot \nabla^k u_{i-k}^*(x). \tag{7.3.37}$$

Proof. It suffices to substitute (7.3.33) into (7.3.16), which yields

$$\begin{aligned} u_i(x, y) &= \sum_{p=0}^{i} \sum_{q=0}^{i-p} (\mathcal{X}^p(y) \otimes M^q) \cdot \nabla^{p+q} u_{i-p-q}^*(x) \\ &= \sum_{p=0}^{i} \sum_{k=p}^{i} (\mathcal{X}^p(y) \otimes M^{p-k}) \cdot \nabla^k u_{i-k}^*(x) \quad \text{(change of indices } k = p+q) \\ &= \sum_{k=0}^{i} \sum_{p=0}^{k} (\mathcal{X}^p(y) \otimes M^{p-k}) \cdot \nabla^k u_{i-k}^*(x) \quad \text{(interversion of summation)} \end{aligned} \tag{7.3.38}$$

The result follows by performing a last change of indices $p \to k - p$. □

We are now able to introduce the tensors N^i of (7.3.9) and their related differential operators:

Definition 7.3. For any $k \ge 0$, we denote by $N^k(y)$ the k-th order tensor

$$N^k(y) := \sum_{p=0}^{k} M^p \otimes \mathcal{X}^{k-p}(y). \tag{7.3.39}$$

Recognizing a Cauchy product, the identity (7.3.37) can be formally written in terms of the following "criminal" ansatz which expresses the oscillating solution u_ε in terms of its formal homogenized averaged u_ε^* (defined in (7.3.5)):

$$u_\varepsilon(x) = \sum_{i=0}^{\infty} \varepsilon^i N^i(x/\varepsilon) \cdot \nabla^i u_\varepsilon^*(x), \quad x \in D_\varepsilon. \tag{7.3.40}$$

The last proposition of this section gathers several important properties for the tensors N^k that are dual to those of the tensors $\mathcal{X}^k$.

Proposition 7.9. *The tensor $N^k(y)$ satisfies:*

1. *$\int_Y N^0(y)\mathrm{d}y = 1$ and $\int_Y N^k(y)\mathrm{d}y = 0$ for any $k \ge 1$,*
2. *For any $k \ge 0$, using the convention $N^{-1} = N^{-2} = 0$:*

$$\begin{cases} -\Delta_{yy} N^0 = M^0 \\ -\Delta_{yy} N^1 = 2\partial_j N^0 \otimes e_j + M^1 \\ -\Delta_{yy} N^{k+2} = 2\partial_j N^{k+1} \otimes e_j + N^k \otimes I + M^{k+2}. \end{cases} \tag{7.3.41}$$

3. For any $k \geq 0$,

$$-\Delta_{yy}(\partial^k_{i_1\dots i_k} N^k_{i_1\dots i_k}) = (-1)^k(k+1)M^0. \tag{7.3.42}$$

4. For any $1 \leq p \leq k-1$,

$$M^k = (-1)^{p+1}\int_Y (N^{k-p}\otimes(-\Delta_{yy}N^p) - N^{k-p-1}\otimes N^{p-1}\otimes I)\mathrm{d}y, \tag{7.3.43}$$

In particular, M^{2p} depends only on the tensors N^p and N^{p-1}, which depend themselves only on the first $p+1$ tensors $\mathcal{X}^0 \dots \mathcal{X}^p$.

Proof. 1. is a consequence of the definition (7.3.32) for the tensor M^k which can be rewritten as

$$\forall k \geq 1,\ \int_Y N^k(y)\mathrm{d}y = \sum_{p=0}^{k} \mathcal{X}^{k-p*}\otimes M^p = 0,$$

and for $k=0$, it holds $\int_Y N^0(y)\mathrm{d}y = M^0\mathcal{X}^{0*} = 1$.

2. The first two equalities of (7.3.41) are easily verified. The third line is obtained by writing

$$\begin{aligned}
-\Delta_{yy}N^k &= \sum_{p=0}^{k} M^{k-p}\otimes(-\Delta_{yy}\mathcal{X}^p)\\
&= \sum_{p=2}^{k} M^{k-p}\otimes(2\partial_j\mathcal{X}^{p-1}\otimes e_j + \mathcal{X}^{p-2}\otimes I) + M^{k-1}\otimes(2\partial_j\mathcal{X}^0\otimes e_j) + M^k\\
&= 2\partial_j\left(\sum_{p=1}^{k} M^{k-p}\otimes\mathcal{X}^{p-1}\right)\otimes e_j + \left(\sum_{p=0}^{k-2} M^p\otimes\mathcal{X}^{k-p-2}\right)\otimes I + M^k,
\end{aligned}$$

from where the result follows.

3. The proof of (7.3.42) is identical to that of proposition 7.3.

4. We start by proving the result for $p=1$ with $k>1$: using the point 1., we may write

$$\begin{aligned}
M^k &= \int_Y N^0\otimes M^k\mathrm{d}y = \int_Y N^0\otimes(-\Delta N^k - 2\partial_j N^{k-1}\otimes e_j - N^{k-2}\otimes I)\mathrm{d}y\\
&= \int_Y M^0\otimes N^k\mathrm{d}y + \int_Y (2\partial_j N^0\otimes e_j\otimes N^{k-1} - N^0\otimes N^{k-2}\otimes I)\mathrm{d}y\\
&= \int_Y ((-\Delta_{yy}N^1)\otimes N^{k-1} - N^0\otimes N^{k-2}\otimes I)\mathrm{d}y.
\end{aligned}$$

If the result now holds until rank p with $1 \leq p \leq k-2$, we obtain the result at rank $p+1$ with analogous computations:

$$\begin{aligned}
M^k &= (-1)^{p+1}\int_Y (M^{k-p} + 2\partial_j N^{k-p-1}\otimes e_j + N^{k-p-2}\otimes I)\otimes N^p - N^{k-p-1}\otimes N^{p-1}\otimes I)\mathrm{d}y\\
&= (-1)^{p+1}\int_Y ((-2\partial_j N^p\otimes e_j - N^{p-1}\otimes I)\otimes N^{k-p-1} + N^{k-p-2}\otimes N^p\otimes I)\mathrm{d}y\\
&= (-1)^{p+1}\int_Y ((\Delta_{yy}N^{p+1} + M^{p+1})\otimes N^{k-p-1} + N^{k-p-2}\otimes N^p\otimes I)\mathrm{d}y.
\end{aligned}$$

□

7.3.3 Homogenized equations of order $2K+2$: tensors $\mathbb{B}_K$ and $\mathbb{D}_K$

We now detail the step (4) of the introduction of this section which allows to to obtain a well-posed variational problem of order $2K+2$ by truncating the infinite homogenized equation (7.3.7). Error estimates are then proved in section 7.3.4.

For any $u \in H^1(D)$ and $f \in L^2(D)$, we denote by $J(u,f)$ the energy

$$J(u,f) := \int_D \left(\frac{1}{2}|\nabla u|^2 - fu\right) \mathrm{d}x.$$

It is well known that u_ε (identified with its extension by 0 in $D\backslash D_\varepsilon$) is the solution to the following minimization problem:

$$u_\varepsilon = \arg \min_{w \in H^1(D_\varepsilon)} \quad J(w,f)$$
$$s.t. \quad \begin{cases} w = 0 \text{ on } \partial\omega_\varepsilon \\ w \text{ is } D\text{–periodic.} \end{cases}$$

In this part, we follow the method of [53, 287] in order to obtain a minimization problem for an approximation of the formal homogenized average u_ε^*. The main idea to obtain well-posed homogenized equations of finite order is to consider truncations $w_{\varepsilon,K}$ of the "criminal" ansatz (7.3.9) of the form:

$$w_{\varepsilon,K}(v)(x) := \sum_{k=0}^{K} \varepsilon^k N^k(x/\varepsilon) \cdot \nabla^k v(x), \quad x \in D_\varepsilon, \tag{7.3.44}$$

where the function v does not depend on the fast variable x/ε and is sought to approximate the formal homogenized average u_ε^*. Note that $N^k(y)$ is extended by 0 in $D\backslash D_\varepsilon$ for (7.3.44) to make sense. We consider the following approximate minimization problem for the function v:

$$\min_{v \in H^{K+1}(D)} \quad J(w_\varepsilon(v), f)$$
$$s.t. \quad v \text{ is } D\text{–periodic.} \tag{7.3.45}$$

The next step is to eliminate the fast variable x/ε in (7.3.45) (by using lemma 7.3) and to read a higher order homogenized equation from the first order optimality condition. In order to do so, we introduce several additional tensors that arise in the averaging process.

Definition 7.4 (Tensors $\widetilde{N}_j^k$ and $\mathbb{B}_K^{l,m}$)**.** For any $K \in \mathbb{N}$, $1 \le j \le d$ and $0 \le k \le K+1$, let $\widetilde{N}_j^k(y)$ (with implicit dependence with respect to K) be the k-th order tensor defined by

$$\widetilde{N}_j^k(y) = \begin{cases} \partial_j N^0(y) \text{ if } k = 0 \\ \partial_j N^k(y) + N^{k-1}(y) \otimes e_j \text{ if } 1 \le k \le K \\ N^K(y) \otimes e_j \text{ if } k = K+1. \end{cases} \tag{7.3.46}$$

We define a family of constant bilinear tensors $\mathbb{B}_K^{l,m}$ of order $l+m$ by the formula

$$\mathbb{B}_K^{l,m} := \int_Y \widetilde{N}_j^l(y) \otimes \widetilde{N}_j^m(y) \mathrm{d}y, \text{ for any } 0 \le l,m \le K+1, \tag{7.3.47}$$

where the Einstein summation convention still assumed over the repeated subscript index $1 \le j \le d$.

Remember that, following the convention of section 7.2.4, we denote by $\mathbb{B}_K^{l,m} \nabla^l v \nabla^m v$ the bilinear contraction

$$\mathbb{B}_K^{l,m} \nabla^l v \nabla^m v := (\mathbb{B}_K^{l,m})_{i_1 \dots i_l j_1 \dots j_m} (\partial^l_{i_1 \dots i_l} v)(\partial^m_{j_1 \dots j_m} v).$$

The bilinear tensors allow us to formulate an approximate energy $J_K^*(v,f,\varepsilon)$ defined for any periodic function $v \in H^{K+1}(D)$ by:

$$J_K^*(v,f,\varepsilon) := \int_D \left(\frac{1}{2} \sum_{l,m=0}^{K+1} \varepsilon^{l+m-2} \mathbb{B}_K^{l,m} \nabla^l v \nabla^m v - fv\right) \mathrm{d}x. \tag{7.3.48}$$

The definition of the energy $J_K^*(v,f,\varepsilon)$ is motivated by the following asymptotic provided by lemma 7.3, which holds with any $m \ge 0$ arbitrarily high:

$$J(w_{\varepsilon,K}(v), f) = J_K^*(v,f,\varepsilon) + o(\varepsilon^m).$$

More precisely, the following result holds:

Proposition 7.10. *Assume $f \in \mathcal{C}^\infty(D)$ and D–periodic. Let $v \in \mathcal{C}^\infty(D)$ be a smooth D–periodic function and $w_{\varepsilon,K}(v) \in \mathcal{C}^\infty(D_\varepsilon)$ be the truncated ansatz of the form of (7.3.44). The following asymptotic estimate holds true with $m \geq 0$ arbitrarily high:*

$$|J(w_{\varepsilon,K}(v), f) - J^*_K(v, f, \varepsilon)| \leq C_{K,m}(||\nabla^{m+2} v||^2_{L^\infty(D,\mathbb{R}^d)} + ||f||^2_{H^m(D)})\varepsilon^m.$$

for a constant $C_{K,m}$ depending only on m and K.

Proof. For any $1 \leq j \leq d$, the partial derivative $\partial_{x_j} w_{\varepsilon,K}(v)$ reads

$$\begin{aligned}\partial_{x_j} w_{\varepsilon,K}(v) &= \sum_{k=0}^{K} \left(\varepsilon^{k-1} \partial_{y_j} N^k(\cdot/\varepsilon) \cdot \nabla^k v + \varepsilon^k N^k(\cdot/\varepsilon) \otimes e_j \cdot \nabla^{k+1} v\right) \\ &= \sum_{i=0}^{K+1} \varepsilon^{k-1} \widetilde{N}^i_j(\cdot/\varepsilon) \cdot \nabla^k v,\end{aligned}$$

by definition (7.3.46) of the tensors $\widetilde{N}^k_j$. The computation of the energy $J(w_{\varepsilon,K}(v), f)$ yields then

$$\begin{aligned}&J(w_{\varepsilon,N}(v), f) \\ &\quad = \int_D \left(\frac{1}{2} \sum_{l,m=0}^{K+1} \varepsilon^{l+m-2} (\widetilde{N}^l_j(x/\varepsilon) \cdot \nabla^l v)(\widetilde{N}^m_j(x/\varepsilon) \cdot \nabla^m v) - \sum_{l=0}^{K} \varepsilon^l (N^l(x/\varepsilon) \cdot \nabla^l v(x)) f(x) \right) \mathrm{d}x.\end{aligned}$$

The result follows from the application of lemma 7.3 and by using that $N^k(x/\varepsilon)$ is of average 1 if $k = 0$ and 0 otherwise (proposition 7.9). □

The approximate energy (7.3.48) is used (instead of $J(w_\varepsilon(v), f)$ in (7.3.45)) in order to build a (well-posed) higher order homogenized equation:

Definition 7.5. For any $K \in \mathbb{N}$, we call homogenized equation of order $2K + 2$ the Euler-Lagrange equation associated with the minimization problem

$$\min_{\substack{v \in H^{K+1}(D), \\ D\text{–periodic}}} J^*_K(v, f, \varepsilon). \tag{7.3.49}$$

This equation reads explicitly in terms of a higher order homogenized solution $v^*_K \in H^{K+1}(D)$ as

$$\begin{cases} \displaystyle\sum_{k=0}^{K+1} \varepsilon^{2k-2} \mathbb{D}^{2k}_K \cdot \nabla^{2k} v^*_K = f \\ v^*_K \text{ is } D\text{–periodic,} \end{cases} \tag{7.3.50}$$

where the constant tensors $\mathbb{D}^{2k}_K$ are defined by the formula for any $0 \leq k \leq K + 1$:

$$\mathbb{D}^{2k}_K := \sum_{l=0}^{2k} (-1)^l \mathbb{B}^{l,2k-l}_K, \tag{7.3.51}$$

assuming the convention $\mathbb{B}^{l,m}_K = 0$ for any $l > K + 1$ or $m > K + 1$.

Proof. Let us detail slightly the derivation of (7.3.50). The Euler-Lagrange Equation of (7.3.49) reads, after a suitable integration by parts:

$$\sum_{l,m=0}^{l+1} \varepsilon^{l+m-2} \frac{1}{2} ((-1)^l + (-1)^m) \mathbb{B}^{l,m}_K \nabla^{l+m} v^*_K = f.$$

Since $(-1)^k + (-1)^l$ vanishes when k and l are not of the same parity, only terms such that $k + l$ is even are not zero in the above equation. Hence, it rewrites as (7.3.50) with

$$\mathbb{D}^{2k}_K = \sum_{l+m=2k} \frac{1}{2} ((-1)^l + (-1)^m) \mathbb{B}^{l,m}_K = \sum_{l=0}^{2k} \frac{1}{2} ((-1)^l + (-1)^{2k-l}) \mathbb{B}^{l,2k-l}_K,$$

which eventually yields the desired expression (7.3.51). □

Remark 7.7. Naturally, the higher order homogenized solution v_K^* of (7.3.50) depends on ε, but we omit this dependence for the sake of notational simplicity.

Remark 7.8. Let us examine equation (7.3.50) at order $K=0$. It reads

$$\mathbb{D}_0^2 \cdot \nabla^2 v_0^* + \varepsilon^{-2}\mathbb{D}_0^0 v_0^* = f. \tag{7.3.52}$$

The constant coefficient $\mathbb{D}_0^0$ is given by

$$\mathbb{D}_0^0 = \mathbb{B}_0^{0,0} = \int_Y \widetilde{N}_{0,j}^0 \widetilde{N}_{0,j}^0 \mathrm{d}y = \int_Y |\nabla N^0|^2 \mathrm{d}y = (M^0)^2 \int_Y |\nabla \mathcal{X}^0|^2 \mathrm{d}y = (\mathcal{X}^{0*})^{-1}.$$

The tensor $\mathbb{D}_0^{2*}$ reads

$$\begin{aligned}
\mathbb{D}_0^2 = -\mathbb{B}_0^{1,1} &= -\int_Y \widetilde{N}_{0,j}^1 \otimes \widetilde{N}_{0,j}^1 \mathrm{d}y = -\int_Y |N^0(y)|^2 e_j \otimes e_j \mathrm{d}y \\
&= -\left(\int_Y |N^0(y)|^2 \mathrm{d}y\right) I = -(M^0)^2 \left(\int_Y |\mathcal{X}^0(y)|^2 \mathrm{d}y\right) \\
&= -\left(\frac{1}{(\mathcal{X}^{0*})^2}\int_Y |\mathcal{X}^0(y)|^2 \mathrm{d}y\right) I.
\end{aligned}$$

Importantly, (7.3.52) does not coincide with the two term truncation of the infinite order homogenized equation (7.3.34), which would read instead:

$$M^2 \cdot \nabla^2 v_0^* + \varepsilon^{-2} M^0 v_0^* = f.$$

Indeed, formulas (7.3.21) and (7.3.35) imply

$$M^2 = \frac{1}{(\mathcal{X}^{0*})^2}\int_Y (\partial_j \mathcal{X}^1 \otimes \partial_j \mathcal{X}^1 - |\mathcal{X}^0|^2 I)\mathrm{d}y \neq \mathbb{D}_0^2.$$

However, the coefficients M^0 and $\mathbb{D}_0^0$ do coincide,

Before stating the next proposition establishing the existence and uniqueness of the high order homogenized solution v_K^*, let us underline the following result which is an obvious, but somewhat important consequence of the definition (7.3.47):

Lemma 7.4. *The dominant tensor $\mathbb{B}_K^{K+1,K+1}$ is symmetric and non-negative.*

Proposition 7.11. *Assume further that the dominant tensor $\mathbb{B}_K^{K+1,K+1}$ is non-degenerate, that is there exists a constant $\nu > 0$ such that*

$$\forall \boldsymbol{\xi} = \boldsymbol{\xi}_{i_1 \dots i_{K+1}} \in \mathbb{R}^{d^{K+1}},\ \mathbb{B}_K^{K+1,K+1} \cdot \boldsymbol{\xi}\boldsymbol{\xi} \geq \nu |\boldsymbol{\xi}|^2. \tag{7.3.53}$$

Then there exists a unique periodic solution $v_K^* \in H^{K+1}(D)$ to the homogenized equation (7.3.50) of order $2K+2$.

Proof. Let us consider the space $V_K := \{v \in H^{K+1}(D) \,|\, v \text{ is } D\text{–periodic }\}$ and introduce $a : V_K \times V_K \to \mathbb{R}$ and $b : V_K \to \mathbb{R}$ the respective bilinear and linear forms defined for any $v \in V_K$ by

$$a(v,v) = \int_D \sum_{k,l=0}^{K+1} \varepsilon^{k+l-2} \mathbb{B}_K^{k,l} \nabla^k v \nabla^l v \mathrm{d}x, \tag{7.3.54}$$

$$b(v) = \int_D f v \mathrm{d}x. \tag{7.3.55}$$

The homogenized equation (7.3.50) reduces then to the following variational problem:

$$\text{Find } v_K^* \in V_K \text{ such that } \forall v \in V_K,\ a(v_K^*, v) = b(v). \tag{7.3.56}$$

From there, one could directly rely on the theory of Fredholm operators [231] to conclude. However we shall show with elementary arguments that a is coercive, which will allow us to apply Lax-Milgram theorem.

Under the non-degeneracy assumption (7.3.53), it is readily obtained that there exists a constant C_ε (depending on ε) such that

$$\forall v \in V_K(D),\ a(v,v) \geq (\varepsilon^{2K}\nu)||\nabla^{K+1}v||^2_{L^2(D,\mathbb{R}^d)} + \varepsilon^{-2}M^0||v||^2_{L^2(D)} - C_\varepsilon||v||_{H^{K+1}(D)}||v||_{H^K(D)}.$$

Remembering $M^0 > 0$ and applying the following Young's inequality

$$\forall x, y \in \mathbb{R},\ |xy| \leq \frac{x^2}{2\gamma} + \frac{\gamma y^2}{2}$$

for a sufficiently small $\gamma > 0$, we obtain the existence of two constants $\alpha_{\varepsilon,K} > 0$ and $\beta_{\varepsilon,K} > 0$ (that depend on ε and K) such that

$$\forall v \in V_K(D),\ a(v,v) \geq \alpha_{\varepsilon,K}||v||^2_{H^{K+1}(D)} - \beta_{\varepsilon,K}||v||^2_{H^K(D)}. \tag{7.3.57}$$

Furthermore, (7.3.47) together with the proof of proposition 7.10 allow to rewrite $a(v,v)$ as

$$a(v,v) = \int_D \int_Y \left|\left|(\varepsilon^{-1}\nabla_y + \nabla_x)\left(\sum_{k=0}^{K}\varepsilon^k N^k(y)\cdot\nabla^k v(x)\right)\right|\right|^2 \mathrm{d}x.$$

Then, $\int_D \int_Y u(x,y)^2 \mathrm{d}y\mathrm{d}x \geq \int_D \left|\int_Y u(x,y)\mathrm{d}y\right|^2 \mathrm{d}x$ implies the following inequality:

$$\forall v \in V_K,\ a(v,v) \geq ||\nabla v||^2_{L^2(D,\mathbb{R}^d)}. \tag{7.3.58}$$

We shall now prove that (7.3.57) and (7.3.58) together imply the coercivity of a on the space V_K, that is we claim there exists a constant $c_{\varepsilon,K} > 0$ such that

$$\forall v \in V_K,\ a(v,v) \geq c_{\varepsilon,K}||v||^2_{H^{K+1}(D)}. \tag{7.3.59}$$

Assume the contrary is true, then one can find a sequence (v_n) of functions satisfying $||v_n||_{H^{K+1}(D)} = 1$ and such that $a(v_n, v_n) \to 0$. Up to extracting a relevant subsequence, we may assume that $v_n \rightharpoonup v$ weakly in $H^{K+1}(D)$ and $v_n \to v$ strongly in $H^K(D)$. Then the polarization identity together with (7.3.57) and the positivity of a allow to show that (v_n) is a Cauchy sequence in V_K:

$$\begin{aligned}\forall l, m \in \mathbb{N},\ \alpha_{\varepsilon,K}||v_l - v_m||^2_{H^{K+1}(D)} &\leq a(v_l - v_m, v_l - v_m) + \beta_{\varepsilon,K}||v_l - v_m||^2_{H^K(D)} \\ &= 2a(v_l, v_l) + 2a(v_m, v_m) - a(v_l + v_m, v_l + v_m) + \beta_{\varepsilon,K}||v_l - v_m||^2_{H^K(D)} \\ &\leq 2a(v_l, v_l) + 2a(v_m, v_m) + \beta_{\varepsilon,K}||v_l - v_m||^2_{H^K(D)} \xrightarrow{l,m\to\infty} 0.\end{aligned}$$

Therefore $v_n \to v$ strongly in V_K. The continuity of a implies then $a(v,v) = \lim_{n\to+\infty} a(v_n, v_n) = 0$. The property (7.3.58) yields then that v is a constant. Therefore, $0 = a(v,v) = \varepsilon^{-2}M^0||v||^2_{L^2(D)}$, which implies $v = 0$. This is in contradiction with the fact that $||v_n||_{H^{K+1}(D)} = 1$ for any $n \geq 0$ and the strong convergence of (v_n). Finally, the coercivity (7.3.59) and the continuity of a and b over V_K ensure that all the assumptions of the Lax-Milgram theorem are fulfilled, which yields existence and uniqueness to the problem (7.3.56). □

Remark 7.9. The non degeneracy assumption (7.3.53) is automatically fulfilled for any shape of obstacle ηT when $K = 0$. It could fail to be satisfied for particular obstacle shapes for $K \geq 1$ (e.g. under strong symmetries with respect to the cell axes).

Before going to the proof of error estimates for the higher homogenized solution v^*_K, we provide a last result which shows that the high order homogenized equation (7.3.50) is in some sense a truncation of the formal infinite order homogenized equation (7.3.34). This fact was also observed by [287] for a (scalar) antiplane elasticity model.

Proposition 7.12. *The first $K+1$ homogenized coefficients of the homogenized equation (7.3.50) of order $2K+2$ coincide with those of the formal equation (7.3.34):*

$$\forall k \leq \lfloor K/2 \rfloor, \quad \mathbb{D}^{2k}_K = M^{2k}.$$

Proof. We show the following, slightly more general, result:

$$\forall k \leq K,\ M^k = \sum_{l=0}^{k} (-1)^l \mathbb{B}_K^{l,k-l}, \tag{7.3.60}$$

which is sufficient for our purpose because of (7.3.51). For $k, l \leq K$, the coefficient $\mathbb{B}_K^{l,k-l}$ is given by (from (7.3.46))

$$\mathbb{B}_K^{l,k-l} = \int_Y (\partial_j N^l + N^{l-1} \otimes e_j) \otimes (\partial_j N^{k-l} + N^{k-l-1} \otimes e_j) \mathrm{d}y,$$

where we use the convention $N^{-1} = N^{-2} = 0$. After an integration by parts, we rewrite $\mathbb{B}_K^{l,k-l}$ as follows:

$$\begin{aligned}
\mathbb{B}_K^{l,k-l} &= \int_Y (-\Delta N^l - 2\partial_j N^{l-1} \otimes e_j - N^{l-2} \otimes I) \otimes N^{k-l} \mathrm{d}y \\
&+ \int_Y (\partial_j N^l \otimes N^{k-l-1} \otimes e_j + N^{l-1} \otimes N^{k-l-1} \otimes I + \partial_j N^{l-1} \otimes N^{k-l} \otimes e_j + N^{l-2} \otimes N^{k-l} \otimes I) \mathrm{d}y \\
&= \int_Y (M^l \otimes N^{k-l}) \mathrm{d}y + B^{k,l} + B^{k,l-1}
\end{aligned}$$

where $B^{k,l}$ is the k-th order tensor defined by

$$B^{k,l} := \int_Y (\partial_j N^l \otimes N^{k-l-1} \otimes e_j + N^{l-1} \otimes N^{k-l-1} \otimes I) \mathrm{d}y.$$

Using now the point 1. of proposition 7.9 and recognizing a telescopic series, we finally obtain

$$\begin{aligned}
\sum_{l=0}^{k} (-1)^l \mathbb{B}_K^{l,k-l} &= (-1)^k M^k + \sum_{l=0}^{k} ((-1)^l B^{k,l} - (-1)^{l-1} B^{k,l-1}) \\
&= (-1)^k M^k + (-1)^k B^{k,k} - (-1)^{-1} B^{k,-1}.
\end{aligned}$$

The result follows from the fact that $M^k = 0$ when k is odd and $B^{k,k} = B^{k,-1} = 0$ with our convention $N^{-1}0$. □

7.3.4 Error estimates and justification of the higher order homogenization process

The main result of this subsection is proposition 7.13 below: we show that the truncated ansatz $w_{\varepsilon,K}(v_K^*)$ (equation (7.3.44)) built from the solution v_K^* of the homogenized equation (7.3.50) yields an approximation of the original solution u_ε of order $K+2$ in the $H^1(D)$ norm, and of order $K+3$ in the $L^2(D)$ norm.

The first step of the proof consists in showing that under periodicity assumptions, u_ε can be effectively approximated by truncated ansatz of the form of (7.3.44):

Lemma 7.5. *Let $u_{\varepsilon,K}^*$ be the average of the truncated expansion $u_{\varepsilon,K}$ (7.3.23)):*

$$\forall x \in D,\ u_{\varepsilon,K}^*(x) := \sum_{k=0}^{K} \varepsilon^{k+2} u_k^*(x).$$

There exists a constant $C_K(f)$ independent of ε such that

$$\left\| u_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k u_{\varepsilon,K}^* \right\|_{H^1(D)} \leq C_K(f) \varepsilon^{K+2}.$$

Proof. We use proposition 7.8 to rewrite

$$\begin{aligned}
u_{\varepsilon,K}(x) &= \sum_{i=0}^{K} \sum_{k=0}^{i} \varepsilon^{2+i} N^k(x/\varepsilon) \cdot \nabla^k u_{i-k}^*(x) = \sum_{k=0}^{K} \sum_{i=k}^{K} \varepsilon^k \varepsilon^{i-k+2} N^k(x/\varepsilon) \cdot \nabla^k u_{i-k}^*(x) \\
&= \sum_{k=0}^{K} \varepsilon^k N^k(x/\varepsilon) \cdot \nabla^k \left(\sum_{i=0}^{K-k} \varepsilon^{i+2} u_i^*(x) \right) \\
&= \sum_{k=0}^{K} \varepsilon^k N^k(x/\varepsilon) \cdot \nabla^k u_{\varepsilon,K}^*(x) - \sum_{k=0}^{K} \sum_{i=K-k+1}^{K} \varepsilon^{k+i+2} N^k(x/\varepsilon) \cdot \nabla^k u_i^*(x).
\end{aligned}$$

Hence using the result of proposition 7.4,

$$\begin{aligned}\left|\left|u_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot\ \ /\varepsilon)\cdot\nabla^k u^*_{\varepsilon,K}\right|\right|_{H^1(D)} &\leq ||u_\varepsilon - u_{\varepsilon,K}||_{H^1(D)} + \left|\left|u_{\varepsilon,K} - \sum_{k=0}^{K}\varepsilon^k N^k(x/\varepsilon)\cdot\nabla^k u^*_{\varepsilon,K}(x)\right|\right|_{H^1(D)}\\ &\leq C_K||f||_{H^{K+2}(D)}\varepsilon^{K+2} + \varepsilon^{K+3}\sum_{k=0}^{K}\sum_{i=K-k+1}^{K}||N^k(\cdot/\varepsilon)\cdot\nabla^k u^*_i||_{H^1(D)}\\ &\leq C_K||f||_{H^{K+2}(D)}\varepsilon^{K+2} + C_K\varepsilon^{K+2}||f||_{H^{2K+1}(D)}\end{aligned}$$

whence the result. □

In order to state our final result, we need the following lemma providing uniform estimates of the partial derivatives of v^*_K:

Lemma 7.6. *The solution v^*_K of (7.3.50) belongs to $\mathcal{C}^\infty(D)$ and for any $m\in\mathbb{N}$, there exists a constant C_m that does not depend on ε such that*

$$||v^*_K||_{H^{m+2}(D)} \leq C_m||f||_{H^m(D)}\varepsilon^2. \tag{7.3.61}$$

Proof. The result is obvious by solving (7.3.50) explicitly with Fourier expansions. □

Using the previous results, ideas that resemble those of Cea's Lemma allow finally to derive the following error estimate for the higher order ansatz $w_{\varepsilon,K}(v^*_K)$ (eqn. (7.3.44)):

Proposition 7.13. *Let v^*_K be the solution to the homogenized equation (7.3.50) of order $2K+2$. There exists a constant $C_K(f)$ independent of ε (but depending on K and f) such that the following error estimates hold:*

$$\left|\left|u_\varepsilon - \sum_{k=0}^{K}\varepsilon^k N^k(\cdot/\varepsilon)\cdot\nabla^k v^*_K\right|\right|_{H^1(D)} \leq C_K(f)\varepsilon^{K+2}. \tag{7.3.62}$$

$$\left|\left|u_\varepsilon - \sum_{k=0}^{K}\varepsilon^k N^k(\cdot/\varepsilon)\cdot\nabla^k v^*_K\right|\right|_{L^2(D)} \leq C_K(f)\varepsilon^{K+3}. \tag{7.3.63}$$

Proof. From, lemma 7.3 and by the definition (7.3.49) of $w_{\varepsilon,K}(v^*_K)$, itnfer holds for any smooth periodic function $\phi\in\mathcal{C}^\infty(D)$ and with any $m\geq 0$ arbitrarily high:

$$\begin{aligned}\left|\int_D \nabla v_{\varepsilon,K}\cdot\nabla\left(\sum_{k=0}^{K}\varepsilon^k N^k(x/\varepsilon)\cdot\nabla^k\phi(x)\right)\mathrm{d}x - \int_D f\phi\mathrm{d}x\right| &\leq C_m\varepsilon^m||\phi v^*_K||_{H^{m+2}(D)}\\ &\leq C'_m\varepsilon^m||\phi||_{H^{m+2}(D)}||f||_{H^m(D)}.\end{aligned}$$

Furthermore by definition of u_ε it also holds:

$$\begin{aligned}&\left|\int_D \nabla u_\varepsilon\cdot\nabla\left(\sum_{k=0}^{K}\varepsilon^k N^k(x/\varepsilon)\cdot\nabla^k\phi(x)\right)\mathrm{d}x - \int_D f\phi\mathrm{d}x\right|\\ &\qquad\leq C_m\varepsilon^m||f\phi||_{H^m(D)} \leq C'_m\varepsilon^m||f||_{H^m(D)}||\phi||_{H^{m+2}(D)}.\end{aligned}$$

Therefore $u_\varepsilon - v_{\varepsilon,K}$ is nearly orthogonal in the H^1 scalar product (up to some terms of order $O(\varepsilon^m)\leq C'_m\varepsilon^m(||\phi||_{H^{m+2}(D)}||f||_{H^m(D)})$ to the test fields of the form $w_{\varepsilon,K}(\phi) := \sum_{k=0}^{K}\varepsilon^k N^k(\cdot/\varepsilon)\cdot\nabla^k\phi$. Hence with successively $\phi = v^*_K$ and $\phi = u^*_{\varepsilon,K}$:

$$\begin{aligned}||\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))||^2_{L^2(D,\mathbb{R}^d)} &= \int_D|\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))|^2\mathrm{d}x\\ &\leq \int_D\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))\cdot\nabla u_\varepsilon\mathrm{d}x + C'_m\varepsilon^m||f||_{H^m(D)}||v^*_K||_{H^{m+2}(D)}\\ &\leq \int_D\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))\cdot\nabla\left(u_\varepsilon - w_{\varepsilon,K}(u^*_{\varepsilon,K})\right)\mathrm{d}x\\ &\quad + C'_m\varepsilon^m||f||_{H^m(D)}||v^*_K||_{H^{m+2}(D)} + C''_m\varepsilon^m||f||_{H^m(D)}||u^*_{\varepsilon,K}||_{H^{m+2}(D)}.\end{aligned}$$

From the definition (7.3.30) and the regularity estimate lemma 7.6, we infer that

$$||u^*_{\varepsilon,K}||_{H^{m+1}(D)} \leq C_m||f||_{H^{K+m+1}} \text{ and } ||v^*_K||_{H^{m+1}(D)} \leq C_m||f||_{H^{m+3}(D)}.$$

Hence we finally obtain

$$\begin{aligned}||\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))||^2_{L^2(D,\mathbb{R}^d)} &\leq \frac{1}{2}||\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))||^2_{L^2(D,\mathbb{R}^d)} + \frac{1}{2}\left|\left|\nabla\left(u_\varepsilon - w_{\varepsilon,K}(u^*_{\varepsilon,K})\right)\right|\right|^2_{L^2(D,\mathbb{R}^d)} \\ &\quad + C_m\varepsilon^m||f||_{H^{K+m+1}(D)} \\ &\leq \frac{1}{2}||\nabla(u_\varepsilon - w_{\varepsilon,K}(v^*_K))||^2_{L^2(D,\mathbb{R}^d)} + C_K\varepsilon^{2K+4}||f||_{H^{K+2}(D)} + C_m\varepsilon^m||f||_{H^{K+m+1}(D)}\end{aligned}$$

where we have applied Young's inequality $|ab| \leq \frac{1}{2}(a^2+b^2)$ and the error bound (7.3.24) in the last two lines. Setting $m = 2K+4$ implies the first error estimate (7.3.62). Then (7.3.63) follows from the Poincaré inequality of lemma 7.2. □

Remark 7.10. We knew already from proposition 7.4 with $K = 0$ that $||u_\varepsilon - N^0(x/\varepsilon)\varepsilon^2\mathcal{X}^{0*}f||_{H^1(D)} = O(\varepsilon^2)$. Therefore, the solution $w_{\varepsilon,K}(v^*_K)$ obtained from the homogenized equation (7.3.50) of order 2 with $K = 0$ does not provide a better order of magnitude. However, we may argue that the approximation $N^0(x/\varepsilon)v^*_0$ is better in the energy norm, since from the minimization principle (7.3.49), it must hold

$$||u_\varepsilon - N^0(\cdot/\varepsilon)v^*_0||_{H^1(D)} \leq ||u_\varepsilon - N^0(x/\varepsilon)\varepsilon^2\mathcal{X}^{0*}f||_{H^1(D)} + O(\varepsilon^m)$$

for any arbitrarily large value of m. The next section will provide additional supporting arguments that (7.3.50) provides more robust homogenized approximations since this equation does not degenerate when the size η of the obstacle ηT vanishes (it converges to the Poisson problem without holes).

Remark 7.11. We could have hoped, at first glance, to obtain similar bounds for (7.3.1) in case where the periodicity condition on ∂D is replaced by a Dirichlet boundary conditions $v^*_K = 0$ on ∂D. Indeed, the truncated ansatz $w_{\varepsilon,K}(v^*_K)$ would then satisfy simultaneously the boundary conditions $v_{\varepsilon,K} = 0$ on ∂D and $\partial\omega_\varepsilon$. However our proposed error analysis does not extend in a straightforward manner: one of the difficulty lies in that the result of lemma 7.6 does not hold because of the arising of of boundary layers of $v_{\varepsilon,K}$ and u_ε near ∂D. See [216, 19] for some examples of treatments of related problems.

Remark 7.12. There are other ways to build well-posed higher order homogenized equations, for instance by filtering the right-hand side f [33], or by resorting to a Boussinesq trick [20, 3, 264]. These approaches yield order K estimates by solving order K equations only (instead of $2K+2$ as in (7.3.50)), however they resort to the evaluation of some partial derivatives of f to the right-hand side (which may be a numerically delicate task).

Remark 7.13. The proof of proposition 7.13 is inspired from [287], however it does not extend straightforwardly to the context of the Stokes system (7.2.3) because of the divergence constraint. A different proof is proposed in the dedicated section 7.5.3.

7.3.5 Low volume fraction limits when the size of the obstacle tends to zero

The purpose of this section is to establish asymptotics for the tensors $\mathcal{X}^{k*}$ and M^k in the low volume fraction limit, namely when the size η of the obstacle vanishes to zero. Our main result is stated in corollary 7.2, which implies that the infinite order homogenized equation (7.3.34) converges formally to either of the three classical regimes of the literature (namely, to the original Laplace equation (7.3.1), or to the analogue of the Brinkman or Darcy equation (7.2.6) and (7.2.7)).

In this whole subsection, it is assumed, for simplicity, that the space dimension is greater than 3:

$$d \geq 3. \tag{7.3.64}$$

The case $d = 2$ requires a specific treatment (see e.g. [13]), although very similar results could be stated.

The hole ηT is assumed to be strictly included in the unit cell for any $\eta \leq 1$ (it does not touch the boundary): $\eta T \subset\subset P$. For a given function $\widetilde{v} \in L^2(\eta^{-1}P)$, we denote by $<\widetilde{v}>$ the average $<\widetilde{v}> := \eta^d\int_{\eta^{-1}P}\widetilde{v}(y)\mathrm{d}y$. Following the methodology of [13] (and also [196]), we shall use several times the following useful lemma:

Lemma 7.7. *Assume $d \geq 3$. There exists a constant $C > 0$ independent of $\eta > 0$ such that for any $\widetilde{v} \in H^1(\eta^{-1}P\backslash T)$ which vanishes on the hole ∂T and which is $\eta^{-1}P$ periodic, the following inequalities hold:*

$$||\widetilde{v}||_{L^2(\eta^{-1}P\backslash T)} \leq C\eta^{-d/2}||\nabla\widetilde{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}, \tag{7.3.65}$$

$$|<\widetilde{v}>| \leq C||\nabla\widetilde{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}, \tag{7.3.66}$$

$$||\widetilde{v}-<\widetilde{v}>||_{L^2(\eta^{-1}P\backslash T)} \leq C\eta^{-1}||\nabla\widetilde{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}, \tag{7.3.67}$$

$$||\widetilde{v}-<\widetilde{v}>||_{L^{2d/(d-2)}(\eta^{-1}P\backslash T)} \leq C||\nabla\widetilde{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}. \tag{7.3.68}$$

Proof. See [13, 210]. □

Let us also recall that for any $v \in H^1(P\backslash(\eta T))$, if $\widetilde{v}$ is taken to be the rescaled function $\widetilde{v}(y) \equiv v(\eta y)$ in the rescaled cell $\eta^{-1}P\backslash T$, then the L^2 norms of the functions and their gradients are related by the following identities:

$$||v||_{L^2(P\backslash(\eta T))} = \eta^{d/2}||\widetilde{v}||_{L^2(\eta^{-1}P\backslash T)} \text{ and } ||\nabla v||_{L^2(P\backslash(\eta T),\mathbb{R}^d)} = \eta^{d/2-1}||\nabla\widetilde{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}.$$

We also need to consider the so-called Deny-Lions space $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$ for which we recall the definition (the reader is referred to [13, 11, 15] and also [243], p.59. for more details).

Definition 7.6 (Deny-Lions space)**.** The Deny-Lions space $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$ is the completion of the space of smooth functions by the L^2 norm of their gradients:

$$\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T) := \overline{\mathcal{D}(\mathbb{R}^d\backslash T)}^{||\nabla\cdot||_{L^2(\mathbb{R}^d\backslash T,\mathbb{R}^d)}}.$$

When $d \geq 3$, it is admits the following characterization:

$$\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T) = \{\phi \text{ measurable } | \, ||\phi||_{L^{2d/(d-2)}(\mathbb{R}^d\backslash T)} < +\infty \text{ and } ||\nabla\phi||_{L^2(\mathbb{R}^d\backslash T,\mathbb{R}^d)} < +\infty\}.$$

We introduce Ψ the unique solution to the exterior problem

$$\begin{cases} -\Delta\Psi = 0 \text{ in } \mathbb{R}^d - T \\ \Psi = 0 \text{ on } \partial T \\ \Psi \to 1 \text{ at } \infty, \end{cases} \tag{7.3.69}$$

and we denote by Ψ^* the normal flux

$$\Psi^* := \int_{\mathbb{R}^d\backslash T} |\nabla\Psi|^2 \mathrm{d}x = -\int_{\partial T} \nabla\Psi \cdot \boldsymbol{n} \mathrm{d}s,$$

where $\boldsymbol{n}$ is the normal pointing *inward* T. The condition $\Psi \to 1$ at ∞ is to be understood in the sense that $\Psi - 1 \in \mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$.

The following result provides asymptotics for the tensors $\mathcal{X}^k$ and their averages $\mathcal{X}^{k*}$. The special case $k = 0$ is a reformulation of Theorem 3.1 of [13] (see also [196]) to the case of the Poisson system.

Proposition 7.14. *Assume $d \geq 3$. For any $k \geq 0$, denote by $\widetilde{\mathcal{X}}^{2k}$ and $\widetilde{\mathcal{X}}^{2k+1}$ the rescaled tensors in $\eta^{-1}P\backslash T$ defined by:*

$$\forall x \in \eta^{-1}P\backslash T,\ \widetilde{\mathcal{X}}^{2k}(x) := \eta^{(d-2)(k+1)}\mathcal{X}^{2k}(\eta x) \text{ and } \widetilde{\mathcal{X}}^{2k+1}(x) := \eta^{(d-2)(k+1)}\mathcal{X}^{2k+1}(\eta x).$$

Then:

1. *there exists a constant $C > 0$ independent of $\eta > 0$ such that:*

$$\forall \eta > 0,\ ||\nabla\widetilde{\mathcal{X}}^{2k}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} \leq C \text{ and } ||\nabla\widetilde{\mathcal{X}}^{2k+1}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} \leq C; \tag{7.3.70}$$

2. *the following asymptotic convergences hold:*

$$\widetilde{\mathcal{X}}^{2k} \rightharpoonup \frac{\Psi}{|\Psi^*|^{k+1}} I^{2k}, \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T) \tag{7.3.71}$$

$$\widetilde{\mathcal{X}}^{2k+1} \rightharpoonup 0 \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T) \tag{7.3.72}$$

$$\mathcal{X}^{2k*} \sim \frac{1}{\eta^{(d-2)(k+1)}|\Psi^*|^{k+1}} I^{2k}, \tag{7.3.73}$$

where we recall the definition $I^{2k} := \overbrace{I \otimes I \otimes \cdots \otimes I}^{k \text{ times}}$ *of section 7.2.4.*

Remark 7.14. Let us recall that we already know $\mathcal{X}^{2k+1*} = 0$ for any $k \in \mathbb{N}$ (proposition 7.2). However this is not the case in the vectorial context; in that case we shall adapt the arguments in order to obtain asymptotics also for $\mathcal{X}^{2k+1*}$ (proposition 7.27 below).

Proof. The result is proved by induction.

1. *Case* $2k$ *for* $k = 0$. The tensors $\widetilde{\mathcal{X}}^0$ satisfies $-\Delta\widetilde{\mathcal{X}}^0 = \eta^d$ in $\eta^{-1}P\backslash T$, hence for any $\Phi \in H^1(\eta^{-1}P\backslash T)$ which is $\eta^{-1}P$–periodic, it holds

$$\int_{\eta^{-1}P\backslash T} \nabla\widetilde{\mathcal{X}}^0 \cdot \nabla\Phi \mathrm{d}y = \eta^d \int_{\eta^{-1}P\backslash T} \Phi \mathrm{d}y + \int_{\partial T} (\nabla\widetilde{\mathcal{X}}^0 \cdot \boldsymbol{n})\Phi \mathrm{d}s. \tag{7.3.74}$$

Setting $\Phi = \widetilde{\mathcal{X}}^0$ in (7.3.74) and using successively the Cauchy-Schwartz inequality and lemma 7.7 yields

$$||\nabla\widetilde{\mathcal{X}}^0||^2_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} = \eta^d \int_{\eta^{-1}P\backslash T} \widetilde{\mathcal{X}}^0 \mathrm{d}y \leq \eta^d \eta^{-d/2} ||\widetilde{\mathcal{X}}^0||_{L^2(\eta^{-1}P\backslash T)} \leq C||\nabla\widetilde{\mathcal{X}}^0||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}.$$

This implies the first equality of (7.3.70) for $k = 0$. Furthermore, (7.3.66) also shows that $< \widetilde{\mathcal{X}}^0 >$ is bounded by a constant independent of η. Hence, up to extracting a subsequence, there exists a constant $c^0 \in \mathbb{R}$ and a function $\widehat{\Psi}^0$ such that

$$< \widetilde{\mathcal{X}}^0 > \to c^0 \text{ and } \widetilde{\mathcal{X}}^0 \rightharpoonup \widehat{\Psi}^0 \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T) \text{ when } \eta \to 0.$$

Furthermore, the lower semi-continuity of the $H^1_{loc}(\mathbb{R}^d\backslash T)$ norm and (7.3.68) imply that $\widehat{\Psi}^0 - c^0$ belongs to $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$ (see [13] for a detailed justification). Setting now Φ with compact support in $\mathbb{R}^d\backslash\overline{T}$ in (7.3.74) and passing to the limit when $\eta \to 0$ yields that $\widehat{\Psi}^0$ is solution to the exterior problem

$$\begin{cases} -\Delta\widehat{\Psi}^0 = 0 \text{ in } \mathbb{R}^d\backslash T \\ \widehat{\Psi}^0 = 0 \text{ on } \partial T \\ \widehat{\Psi}^0 \to c^0 \text{ at } \infty. \end{cases} \tag{7.3.75}$$

Obviously, the unique solution to this problem is given by $\widehat{\Psi} = c^0\Psi$. Finally, the constant c^0 can be identified by setting $\Phi = 1$ in (7.3.74), which yields

$$-\int_{\partial T} \nabla\widehat{\Psi}^0 \cdot \boldsymbol{n} \mathrm{d}s = \lim_{\eta\to 0} -\int_{\partial T} \nabla\widetilde{\mathcal{X}}^0 \cdot \boldsymbol{n} \mathrm{d}s = 1.$$

Therefore, $c_0 = 1/\Psi^*$ from where (7.3.71) follows for $k = 0$. Since the obtained limit is unique, the convergence holds for the whole sequence. Then (7.3.73) follows from a simple change of variable.

2. *Case* $2k + 1$ *for* $k = 0$. It holds $-\Delta\widetilde{\mathcal{X}}^1 = 2\eta\partial_j\widetilde{\mathcal{X}}^0 \otimes e_j$, hence for any $\Phi \in H^1(\eta^{-1}P\backslash T)$ which is $\eta^{-1}P$–periodic:

$$\int_{\eta^{-1}P\backslash T} \nabla\widetilde{\mathcal{X}}^1 \cdot \nabla\Phi \mathrm{d}y = \int_{\eta^{-1}P\backslash T} 2\eta(\partial_j\widetilde{\mathcal{X}}^0 \otimes e_j)\Phi \mathrm{d}y + \int_{\partial T} (\nabla\widetilde{\mathcal{X}}^1 \cdot \boldsymbol{n})\Phi \mathrm{d}s. \tag{7.3.76}$$

Taking $\Phi = \widetilde{\mathcal{X}}^1_j$ for a given fixed j yields

$$\begin{aligned} ||\nabla\widetilde{\mathcal{X}}^1_j||^2_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} &= \int_{\eta^{-1}P\backslash T} 2\eta\partial_j\widetilde{\mathcal{X}}^0(\widetilde{\mathcal{X}}^1_j - < \widetilde{\mathcal{X}}^1_j >)\mathrm{d}y \\ &\leq 2\eta||\partial_j\widetilde{\mathcal{X}}^0||_{L^2(\eta^{-1}P\backslash T)}||\widetilde{\mathcal{X}}^1_j - < \widetilde{\mathcal{X}}^1_j > ||_{L^2(\eta^{-1}P\backslash T)} \\ &\leq C\eta\eta^{-1}||\nabla\widetilde{\mathcal{X}}^1_j||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}. \end{aligned} \tag{7.3.77}$$

This implies the second part of (7.3.70) for $k=0$. As before, this means that up to extracting a subsequence, we may assume the existence of a constant vector $c^1 \in \mathbb{R}^d$ and a first order tensor $\widehat{\Psi}^1$ such that $<\widetilde{\mathcal{X}}^1> \to c^1$ and $\widetilde{\mathcal{X}}^1 \rightharpoonup \Psi^1$ in $H^1_{loc}(\mathbb{R}^d\backslash T)$ (note that we already know that $<\widetilde{\mathcal{X}}^1>=0$ and $c^1=0$ but the coming ideas will extend to the vectorial case) with $\widetilde{\mathcal{X}}^1 - c^1$ belonging to $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$. Setting Φ with compact support in $\mathbb{R}^d\backslash\overline{T}$ in (7.3.76) yields that $\widehat{\Psi}^1$ is solution to the exterior problem

$$\begin{cases} -\Delta\widehat{\Psi}^1 = 0 \text{ in } \mathbb{R}^d\backslash T \\ \widehat{\Psi}^1 = 0 \text{ on } \partial T \\ \widehat{\Psi}^1 \to c^1 \text{ at } \infty. \end{cases} \tag{7.3.78}$$

Therefore, $\widehat{\Psi}^1 = c^1\Psi$ and the constant c^1 is identified by taking $\Phi = 1$ in (7.3.76), which states that:

$$-\int_{\partial T} \nabla\widehat{\Psi}^1 \cdot \boldsymbol{n} \mathrm{d}s = \lim_{\eta\to 0} -\int_{\partial T} \nabla\widetilde{\mathcal{X}}^1 \cdot \boldsymbol{n}\mathrm{d}s = 0.$$

This implies $c^1=0$ and (7.3.72) is obtained for $k=0$.

3. *General case.* We now complete the proof by induction on k. Assuming the result holds till rank k, we compute

$$-\Delta\widetilde{\mathcal{X}}^{2k+2} = 2\eta^{d-1}\partial_j\widetilde{\mathcal{X}}^{2k+1}\otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k}\otimes I,$$
$$-\Delta\widetilde{\mathcal{X}}^{2k+3} = 2\eta\partial_j\widetilde{\mathcal{X}}^{2k+2}\otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k+1}\otimes I.$$

Hence for any $\Phi \in H^1(\eta^{-1}P\backslash T)$ which is $\eta^{-1}P$–periodic:

$$\int_{\eta^{-1}P\backslash T} \nabla\widetilde{\mathcal{X}}^{2k+2}\cdot\nabla\Phi\mathrm{d}y = \int_{\eta^{-1}P\backslash T}(2\eta^{d-1}\partial_j\widetilde{\mathcal{X}}^{2k+1}\otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k}\otimes I)\Phi\mathrm{d}y + \int_{\partial T}(\nabla\widetilde{\mathcal{X}}^{2k+2}\cdot\boldsymbol{n})\Phi\mathrm{d}s, \tag{7.3.79}$$

$$\int_{\eta^{-1}P\backslash T} \nabla\widetilde{\mathcal{X}}^{2k+3}\cdot\nabla\Phi\mathrm{d}y = \int_{\eta^{-1}P\backslash T}(2\eta\partial_j\widetilde{\mathcal{X}}^{2k+2}\otimes e_j + \eta^d\widetilde{\mathcal{X}}^{2k+1}\otimes I)\Phi\mathrm{d}y + \int_{\partial T}(\nabla\widetilde{\mathcal{X}}^{2k+3}\cdot\boldsymbol{n})\Phi\mathrm{d}s. \tag{7.3.80}$$

Setting $\Phi = \widetilde{\mathcal{X}}^{2k+2}$ and $\Phi = \widetilde{\mathcal{X}}^{2k+3}$ respectively in the above variational formulations (for a fixed set of indices), applying the Cauchy-Schwartz inequality and using lemma 7.7 yield

$$\begin{aligned} ||\nabla\widetilde{\mathcal{X}}^{2k+2}||^2_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} &\le C\eta^{d-1}||\nabla\widetilde{\mathcal{X}}^{2k+1}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}||\widetilde{\mathcal{X}}^{2k+2} - <\widetilde{\mathcal{X}}^{2k+2}>||_{L^2(\eta^{-1}P\backslash T)} \\ &\quad + \eta^d||\widetilde{\mathcal{X}}^{2k}||_{L^2(\eta^{-1}P\backslash T)}||\widetilde{\mathcal{X}}^{2k+2}||_{L^2(\eta^{-1}P\backslash T)} \\ &\le (C'\eta^{d-2}+C'')||\nabla\widetilde{\mathcal{X}}^{2k+2}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}, \end{aligned} \tag{7.3.81}$$

$$\begin{aligned} ||\nabla\widetilde{\mathcal{X}}^{2k+3}||^2_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} &\le 2\eta||\nabla\widetilde{\mathcal{X}}^{2k+2}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}||\widetilde{\mathcal{X}}^{2k+3} - <\widetilde{\mathcal{X}}^{2k+3}>||_{L^2(\eta^{-1}P\backslash T)} \\ &\quad + \eta^d||\widetilde{\mathcal{X}}^{2k+1}||_{L^2(\eta^{-1}P\backslash T)}||\widetilde{\mathcal{X}}^{2k+3}||_{L^2(\eta^{-1}P\backslash T)} \\ &\le C||\nabla\widetilde{\mathcal{X}}^{2k+3}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}. \end{aligned} \tag{7.3.82}$$

This implies (7.3.70) at rank $k+1$. Still up to extracting a subsequence, there exists tensor functions $\widehat{\Psi}^{2k+2}$, $\widehat{\Psi}^{2k+3}$ and constant tensors c^{2k+2} and c^{2k+3} such that

$$\begin{aligned} &<\widetilde{\mathcal{X}}^{2k+2}> \to c^{2k+2} \text{ and } <\widetilde{\mathcal{X}}^{2k+3}> \to c^{2k+3} \text{ when } \eta\to 0 \\ &\widetilde{\mathcal{X}}^{2k+2} \rightharpoonup \widehat{\Psi}^{2k+2} \text{ and } \widetilde{\mathcal{X}}^{2k+3} \rightharpoonup \widehat{\Psi}^{2k+3} \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T) \text{ when } \eta\to 0. \end{aligned}$$

The very same previous arguments yield eventually $\widehat{\Psi}^{2k+2} = c^{2k+2}\Psi$ and $\widehat{\Psi}^{2k+3} = c^{2k+3}\Psi$. Finally, setting $\Phi=1$ in (7.3.79) and (7.3.80) entails

$$-\int_{\partial T}\nabla\widehat{\Psi}^{2k+2}\cdot\boldsymbol{n}\mathrm{d}s = \lim_{\eta\to 0} -\int_{\partial T}\nabla\widetilde{\mathcal{X}}^{2k+2}\cdot\boldsymbol{n}\mathrm{d}s = \lim_{\eta\to 0} <\widetilde{\mathcal{X}}^{2k}>\otimes I = \frac{1}{|\Psi^*|^{k+1}}I^{2k+2}$$

$$-\int_{\partial T}\nabla\widehat{\Psi}^{2k+3}\cdot\boldsymbol{n}\mathrm{d}s = \lim_{\eta\to 0} -\int_{\partial T}\nabla\widetilde{\mathcal{X}}^{2k+3}\cdot\boldsymbol{n}\mathrm{d}s = \lim_{\eta\to 0} <\widetilde{\mathcal{X}}^{2k+1}>\otimes I = 0.$$

This implies $c^{2k+2} = I^{2k+2}/|\Psi^*|^{k+2}$ and $c^{2k+3}=0$, which concludes the proof.

□

Remark 7.15. This result seems to indicate that we may have not found the best scaling for the odd order tensors $\mathcal{X}^{2k+1}$ since we were unable to identify a non zero weak limit. However the established bounds are sufficient for our purpose and extend to the vectorial case.

We are now able to identify the asymptotic behavior of the constant tensors M^k. Recall we already know that $M^{2k+1} = 0$ in this scalar context (corollary 7.1).

Corollary 7.2. *Assume $d \geq 3$. The following convergences hold for the tensors M^k as $\eta \to 0$:*

$$M^0 \sim \eta^{d-2}|\Psi^*|, \tag{7.3.83}$$

$$M^2 \to -I, \tag{7.3.84}$$

$$\forall k > 1,\ M^{2k} = o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right). \tag{7.3.85}$$

Proof. We replace the asymptotics of proposition 7.14 in the explicit formula (7.3.35) for the tensor M^k. (7.3.83) is a consequence of the definition $M^0 = (\mathcal{X}^{0*})^{-1}$. (7.3.84) is obtained by writing

$$M^2 = -((\mathcal{X}^{0*})^{-1})^2\mathcal{X}^{2*} \sim -\frac{\eta^{2(d-2)}|\Psi^*|^2}{\eta^{2(d-2)}|\Psi^*|^2}I = -I.$$

Finally, by eliminating terms of odd orders in (7.3.35), we may write for any $k \geq 1$,

$$\begin{aligned}
M^{2k} &= \sum_{p=1}^{2k} \frac{(-1)^p}{(\mathcal{X}^{0*})^{p+1}} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1\ldots i_p\leq k}} \mathcal{X}^{2i_1*}\otimes\cdots\otimes\mathcal{X}^{2i_p*} \\
&= \sum_{p=1}^{2k}(-1)^p\eta^{(p+1)(d-2)}|\Psi^*|^{p+1} \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1\ldots i_p\leq k}} \frac{I^{2i_1}}{\eta^{(d-2)(i_1+1)}|\Psi^*|^{i_1+1}}\otimes\cdots\otimes\frac{I^{2i_p}}{\eta^{(d-2)(i_p+1)}|\Psi^*|^{i_p+1}} \\
&\quad + o\left(\frac{1}{\eta^{(k-1)(d-2)}}\right) \\
&= \frac{I^{2k}}{\eta^{(k-1)(d-2)}|\Psi^*|^{k-1}}\left(\sum_{p=1}^{2k}(-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1\ldots i_p\leq k}} 1\right) + o\left(\frac{1}{\eta^{(k-1)(d-2)}}\right).
\end{aligned}$$

Then (7.3.85) results from the last summation being zero:

$$\forall k > 1, \sum_{p=1}^{k}(-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1\ldots i_p\leq k}} 1 = 0. \tag{7.3.86}$$

There are several ways to obtain the latter formula. A rather direct argument in the spirit of the proof of proposition 7.7 is to apply the identity (7.3.36) to the power series $1/(1-z) = \sum_{k\in\mathbb{N}} z^k$ which yields

$$1 - z = 1 + \sum_{k=1}^{+\infty}\left(\sum_{p=1}^{k}(-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1\ldots i_p\leq k}} 1\right) z^k,$$

from where (7.3.86) follows by identifying the powers in z^k. □

Remark 7.16. We can retrieve formally in corollary 7.2 the different classical asymptotic regimes for the perforated problem (7.3.1) depending on the size of the hole $a_\varepsilon = \eta\varepsilon$ [13, 11, 103, 273]:

1. If $1 \geq a_\varepsilon >> \varepsilon^{d/(d-2)}$, that is η goes to zero at a smaller rate than $\varepsilon^{2/(d-2)}$, then the zeroth order term $\varepsilon^{-2}M^0 \sim \varepsilon^{-2}\eta^{d-2}|\Psi^*|$ is dominant in the infinite order homogenized equation (7.3.34), which is analogous to the "Darcy" regime
$$\varepsilon^{-2}M^0 u_\varepsilon^* \simeq f.$$

2. If $a_\varepsilon = \varepsilon^{d/(d-2)}$, the coefficients of the infinite order homogenized equation (7.3.34) converge to those of
$$-\Delta u^* + \Psi^* u^* = f.$$
This corresponds to the "Brinkman" regime with the well-known "strange" reaction term [103].

3. If $a_\varepsilon = o(\varepsilon^{d/(d-2)})$, that is the size of the hole η goes to zero at a rate faster than $\varepsilon^{2/(d-2)}$, then there is no strange term and the coefficients of (7.3.34) converge to those of the initial Poisson problem (7.3.1):
$$-\Delta u^* = f.$$

7.3.6 Simplifications for the tensors $\mathcal{X}^{k*}$ and M^k in case of symmetries

In this last subsection, we analyze how the homogenized tensors $\mathcal{X}^{k*}$ and M^k simplify when the obstacle ηT is symmetric with respect to the cell axes. This is a classical step in the theory of homogenization; our methodology follows e.g. section 6 in [35].

In all what follows, we denote by $S := (S_{ij})_{1\le i,j\le d}$ an arbitrary orthogonal symmetry (satisfying $S = S^T$ and $SS = I$). We shall in corollary 7.4 below specialize S to either of the two kinds of following cell symmetries:

- for $1 \le l \le d$, we consider the symmetry S^l with respect to the hyperplane orthogonal to $\boldsymbol{e}_l$:
$$S^l := I - 2\boldsymbol{e}_l \boldsymbol{e}_l^T; \tag{7.3.87}$$
- for $1 \le m \ne l \le d$, we consider the symmetry S^{lm} with respect to the diagonal hyperplane orthogonal to $\boldsymbol{e}_l - \boldsymbol{e}_m$:
$$S^{lm} := I - \boldsymbol{e}_l \boldsymbol{e}_l^T - \boldsymbol{e}_m \boldsymbol{e}_m^T + \boldsymbol{e}_l \boldsymbol{e}_m^T + \boldsymbol{e}_m \boldsymbol{e}_l^T. \tag{7.3.88}$$

Recall the Laplace operator is invariant under orthogonal symmetries S (satisfying $SS = I$ and $S^T = S$): for any smooth scalar field $\mathcal{X}$,
$$-\Delta(\mathcal{X} \circ S) = -(\Delta \mathcal{X}) \circ S. \tag{7.3.89}$$

Proposition 7.15. *If the cell $Y = P\backslash \eta T$ is invariant with respect to a symmetry S,* i.e. *$S(Y) = Y$, then the following identity holds for the cell solutions $\mathcal{X}^k$ of eqn. (7.3.15):*
$$\mathcal{X}^k_{i_1\dots i_k} \circ S = S_{i_1 j_1} \dots S_{i_k j_k} \mathcal{X}^k_{j_1\dots j_k}. \tag{7.3.90}$$

Proof. The result is proved by induction on k. For $k = 0$, it holds
$$-\Delta_{yy}(\mathcal{X}^0 \circ S) = 1 \circ S = 1$$
and the symmetry of Y implies that $\mathcal{X}^0 \circ S$ also satisfies the boundary conditions of (7.4.7). This implies $\mathcal{X}^0 \circ S = \mathcal{X}^0$. For $k = 1$, we write
$$-\Delta_{yy}(\mathcal{X}^1_{i_1} \circ S) = 2(\partial_{i_1} \mathcal{X}^0) \circ S = 2\partial_{j_1}(\mathcal{X}^0 \circ S) S_{i_1 j_1} = 2\partial_{j_1} \mathcal{X}^0 S_{i_1 j_1}.$$
This implies similarly $\mathcal{X}^1_{i_1} \circ S = S_{i_1 j_1} \mathcal{X}^1_{j_1}$. Finally, if the result holds till rank $k+1$ with $k \ge 0$, then
$$\begin{aligned}
-\Delta_{yy}(\mathcal{X}^{k+2}_{i_1\dots i_{k+2}} \circ S) &= 2(\partial_{i_{k+2}} \mathcal{X}^{k+1}_{i_1\dots i_{k+1}}) \circ S + \delta_{i_{k+1} i_{k+2}} \mathcal{X}^k_{i_1\dots i_k} \circ S \\
&= 2S_{i_{k+2} j_{k+2}} \partial_{j_{k+2}}(\mathcal{X}^{k+1}_{i_1\dots i_{k+1}} \circ S) + S_{i_{k+1} j_{k+1}} S_{i_{k+2} j_{k+2}} \delta_{j_{k+1} j_{k+2}} \mathcal{X}^k_{i_1\dots i_k} \circ S \\
&= -S_{i_1 j_1} \dots S_{i_k j_{k+2}} \Delta_{yy} \mathcal{X}^{k+2}_{j_1\dots j_{k+2}},
\end{aligned}$$
whence the result at rank $k+2$. □

Corollary 7.3. *If the cell $Y = P\backslash \eta T$ is invariant with respect to a symmetry S, then the constant tensors $\mathcal{X}^{k*}$ and M^k satisfy, for any set of indices $1 \le i_1, \dots, i_k \le d$:*
$$\mathcal{X}^{k*}_{i_1\dots i_k} = S_{i_1 j_1} \dots S_{i_k j_k} \mathcal{X}^{k*}_{j_1\dots j_k} \tag{7.3.91}$$
$$M^k_{i_1\dots i_k} = S_{i_1 j_1} \dots S_{i_k j_k} M^k_{j_1\dots j_k} \tag{7.3.92}$$
with implicit summation over the repeated indices $j_1 \dots j_k$.

Proof. Equality (7.3.91) results from the previous proposition and the following change of variables:

$$\mathcal{X}^{k*}_{i_1\dots i_k} = \int_Y \mathcal{X}^k_{i_1\dots i_k}\,\mathrm{d}y = \int_Y \mathcal{X}^k_{i_1\dots i_k} \circ S\mathrm{d}y.$$

Equality (7.3.92) can be obtained by applying (7.3.91) in the formula (7.3.35). □

Corollary 7.4. *1. If the cell Y is symmetric with respect to all cell axes $\boldsymbol{e}_l$, i.e. $S^l(Y) = Y$ for any $1 \le l \le d$, then*

$$\mathcal{X}^{k*}_{i_1\dots i_k} = 0 \text{ and } M^k_{i_1\dots i_k} = 0$$

whenever there exists a number r occurring with an odd multiplicity in the indices $i_1 \dots i_k$, i.e. whenever

$$\exists r \in \{1,\dots,d\},\ \mathrm{Card}\{j \in \{1,\dots,k\} \,|\, i_j = r\} \text{ is odd}\,.$$

2. If the cell Y is symmetric with respect to all diagonal axes orthogonal to $(\boldsymbol{e}_l - \boldsymbol{e}_m)$, i.e. $S^{l,m}(Y) = Y$ for any $1 \le l < m \le d$, then for any permutation $\sigma \in \mathfrak{S}_d$,

$$\mathcal{X}^{k*}_{\sigma(i_1)\dots\sigma(i_k)} = \mathcal{X}^{k*}_{i_1\dots i_k}.$$

$$M^k_{\sigma(i_1)\dots\sigma(i_k)} = M^k_{i_1\dots i_k}.$$

Proof. 1. The symmetry S^l is a diagonal matrix satisfying $S^l\boldsymbol{e}_l = -\boldsymbol{e}_l$ and $S^l\boldsymbol{e}_q = \boldsymbol{e}_q$ for $q \ne l$. Hence, replacing S by S^l in (7.3.91), it holds

$$\mathcal{X}^{k*}_{i_1\dots i_k} = (-1)^{\delta_{i_1 l}+\dots+\delta_{i_k l}}\mathcal{X}^{k*}_{i_1\dots i_k},$$

which implies the result.

2. Applying (7.3.91) to the symmetry $S^{l,m}$ yields the result for $\sigma = \boldsymbol{\tau}$ where $\boldsymbol{\tau}$ is the transposition exchanging l and m. Since this holds for any transposition, this implies the statement for any permutation σ.

□

Let us illustrate how the previous corollary translates for the tensors M^2 and M^4:

- if Y is symmetric with respect to the cell axes $(\boldsymbol{e}_l)_{1\le l\le d}$, then only the coefficients of the form M^2_{ii}, M^4_{iijj}, M^4_{iiii} with $1 \le i,j \le d$ and $i \ne j$ are non zeros (in particular M^2 is diagonal).
- if in addition Y is symmetric with respect to the hyperplane orthogonal to $\boldsymbol{e}_l - \boldsymbol{e}_m$, then these coefficients do not depend on the chosen distinct indices i and j: M^2 is a multiple of the identity and M^4 reduces to two effective coefficients.

7.4 High order homogenization for the perforated elasticity system

This section extends the previous analysis to the elasticity system. For a periodic, smooth, *vectorial* right-hand side $\boldsymbol{f} \in \mathcal{C}^\infty(D, \mathbb{R}^d)$, we consider now $\boldsymbol{u}_\varepsilon$ to be the solution of the elasticity system

$$\begin{cases} -\mathrm{div}(A\nabla\boldsymbol{u}_\varepsilon) = \boldsymbol{f} \text{ in } D_\varepsilon \\ \qquad\qquad \boldsymbol{u}_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon \\ \qquad\qquad \boldsymbol{u}_\varepsilon \text{ is } D \text{ periodic.} \end{cases} \tag{7.4.1}$$

The displacement $\boldsymbol{u}_\varepsilon$ is assumed to be zero on $\partial\omega_\varepsilon$: physically, (7.4.1) models a mechanical structure clamped with many screws.

In (7.4.1), $A := (A_{ijkl})_{1\le i,j,k,l\le d}$ is a 4th order tensor written in standard notation, which means that for any $d \times d$ matrix $P := (P_{kl})_{1\le k,l\le d}$, AP is the $d \times d$ matrix defined by $(AP)_{ij} := A_{ijkl}P_{kl}$. Hence, it holds

$$A\nabla\boldsymbol{u} = A_{ijkl}\partial_l u_k$$

for a vector field $\boldsymbol{u} := (u_k)_{1\le k\le d} \in \mathbb{R}^d$. Note that this notation differs slightly from our notational conventions of section 7.2.4, whereby $A \equiv A^2$ would rather be considered a second order matrix valued

tensor, see (7.4.40) below. The fourth order tensor A is assumed symmetric, positive definite in the following sense:

$$\forall P, Q \in \mathbb{R}^{d\times d},\ P : AQ = AP : Q$$

$$\exists \nu > 0,\ \forall P \in \mathbb{R}^{d\times d},\ P : AP \geq \nu P : P,$$

where ":" denotes the Frobenius scalar product on $d \times d$ matrices: $P : Q = P_{ij}Q_{ij}$.
In a more classical setting, A can be considered to be the Hooke's tensor for isotropic homogeneous materials, defined for two given Lamé constants $\lambda, \mu > 0$ as follows:

$$A\nabla \boldsymbol{v} = \mu(\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T) + \lambda \mathrm{Tr}(\nabla \boldsymbol{v}), \tag{7.4.2}$$

although we do not rely on this assumption in what follows.

Hereafter, we reproduce and extend the analysis of section 7.3 to the elasticity system (7.4.1). The main difference with the analysis of section 7.3 lies in the vectorial nature of the problem. This reflects in the following substantial changes with respect to the scalar context:

- tensors $\mathcal{X}^k$, M^k, N^k become matrix valued (e.g. $\mathcal{X}^k_{i_1\dots i_k}$ is a $d \times d$ matrix for any set of indices $i_1 \dots i_k$);
- odd order tensors $\mathcal{X}^{2k+1*}$ and M^{2k+1} do not, in general, vanish anymore (we provide numerical evidence for this fact in section 7.4.5 below) but are antisymmetric matrix valued. However, these odd order tensors vanish in the case where the obstacle ηT is symmetric with respect to the cell axes;
- even order tensors $\mathcal{X}^{2k*}$ and M^{2k} are symmetric matrix valued;
- low volume fraction asymptotics remain valid, and it turns out that *strange* odd order operators $\varepsilon^{2k+1} M^{2k+1} \cdot \nabla^{2k+1}$ disappear as $\eta \to 0$.

7.4.1 Formal infinite order two-scale asymptotics and matrix valued tensors $\mathcal{X}^k$

Again, we consider the traditional two-scale ansatz

$$\boldsymbol{u}_\varepsilon = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \boldsymbol{u}_i(x, x/\varepsilon) \tag{7.4.3}$$

where the unknowns $\boldsymbol{u}_i(x, y)$ are now vector functions $\boldsymbol{u}_i : D \times Y \to \mathbb{R}^d$ D and P–periodic. Similarly as before and this is not compulsory, we assume that $\boldsymbol{f}$ itself can depend on ε in the form of a formal power series in ε featuring only non-oscillating terms:

$$\forall x \in D,\ \boldsymbol{f}(x) = \sum_{i=0}^{+\infty} \varepsilon^i \boldsymbol{f}^i(x). \tag{7.4.4}$$

Inserting formally the ansatz (7.4.3) into the elasticity system (7.4.1) yields the cascade of equations

$$\begin{cases} \mathcal{A}_{yy}\boldsymbol{u}_{i+2} = -\mathcal{A}_{xy}\boldsymbol{u}_{i+1} - \mathcal{A}_{xx}\boldsymbol{u}_i + \boldsymbol{f}_{i+2}, \quad \forall i \geq -2 \\ \boldsymbol{u}_{-2}(x, y) = \boldsymbol{u}_{-1}(x, y) = 0, \\ \quad \boldsymbol{u}_i(x, \cdot) = 0 \text{ on } \partial(\eta T)\, \forall i \geq 0, \\ \quad \boldsymbol{u}_i(x, \cdot) \text{ is } P\text{–periodic } \forall i \geq 0, \end{cases} \tag{7.4.5}$$

where the operators $\mathcal{A}_{xx}, \mathcal{A}_{xy}$ and $\mathcal{A}_{yy}$ are defined by

$$\mathcal{A}_{yy} := -\mathrm{div}_y(A\nabla_y \cdot), \quad \mathcal{A}_{xy} := -\mathrm{div}_x(A\nabla_y \cdot) - \mathrm{div}_y(A\nabla_x \cdot), \quad \mathcal{A}_{xx} := -\mathrm{div}_x(A\nabla_x \cdot).$$

We introduce an additional differential operator ∂_l^A acting on differentiable vector fields $\boldsymbol{v}$ as follows:

$$\partial_l^A \boldsymbol{v} := \frac{1}{2}(A\nabla \boldsymbol{v} \cdot \boldsymbol{e}_l + \mathrm{div}(A[\boldsymbol{v}\boldsymbol{e}_l^T])). \tag{7.4.6}$$

The operator ∂_l^A behaves like a partial derivative in the direction $\boldsymbol{e}_l$. It can be verified that ∂_l^A is an antisymmetric operator: if $\boldsymbol{v}$ and $\boldsymbol{w}$ are two smooth P–periodic functions vanishing on the hole $\partial(\eta T)$, then it holds

$$\begin{aligned}\int_Y \boldsymbol{v}\cdot\partial_l^A\boldsymbol{w}\mathrm{d}y &= \frac{1}{2}\int_Y (\boldsymbol{v}\cdot A\nabla\boldsymbol{w} - \boldsymbol{w}\cdot A\nabla\boldsymbol{v})\cdot\boldsymbol{e}_l\mathrm{d}y\\ &= -\int_Y \boldsymbol{w}\cdot\partial_l^A\boldsymbol{v}\mathrm{d}y.\end{aligned}$$

The system (7.4.5) can be solved by introducing matrix valued tensors $\mathcal{X}^k : Y\to\mathbb{R}^{d\times d}$: they are defined from their column vectors $(\boldsymbol{\mathcal{X}}_j^k)_{1\le j\le d}$ by a recurrence analogous to (7.3.15):

$$\left\{\begin{array}{rl}\mathcal{A}_{yy}\boldsymbol{\mathcal{X}}_j^0 = \boldsymbol{e}_j \text{ in } Y,\\ \mathcal{A}_{yy}\boldsymbol{\mathcal{X}}_j^1 = 2\partial_l^A\boldsymbol{\mathcal{X}}_j^0\otimes e_l \text{ in } Y,\\ \mathcal{A}_{yy}\boldsymbol{\mathcal{X}}_j^{k+2} = 2\partial_l^A\boldsymbol{\mathcal{X}}_j^{k+1}\otimes e_l + (A[\boldsymbol{\mathcal{X}}_j^k\boldsymbol{e}_l^T]\cdot\boldsymbol{e}_m)e_l\otimes e_m \text{ in } Y, & \forall k\ge 0,\ \forall 1\le j\le d.\\ \boldsymbol{\mathcal{X}}_j^k = 0 \text{ on } \partial(\eta T),\\ \boldsymbol{\mathcal{X}}_j^k \text{ is } P\text{–periodic}\end{array}\right. \tag{7.4.7}$$

The coefficients of the matrix valued tensor $\mathcal{X}^k := [\boldsymbol{\mathcal{X}}_1^k\ \dots\ \boldsymbol{\mathcal{X}}_d^k]$ are then given by

$$\mathcal{X}_{ij}^k = \boldsymbol{\mathcal{X}}_j^k\cdot\boldsymbol{e}_i,\ 1\le i,j\le d.$$

Finally, we associate to these tensors a family of differential operators $\mathcal{X}^k\cdot\nabla^k$ defined for any smooth vector field $\boldsymbol{v} := (v_q)_{1\le q\le d}\in\mathcal{C}^\infty(D,\mathbb{R}^d)$ by $(\mathcal{X}^k\cdot\nabla^k\boldsymbol{v})_p := \mathcal{X}^k_{i_1\dots i_k,pq}\partial^k_{i_1\dots i_k}v_q$, following the conventions of section 7.2.4.

Remark 7.17. With our notational conventions, the non bold symbols $\otimes e_l$ and $\otimes e_m$ refer to the tensor part of $\boldsymbol{\mathcal{X}}^k$ and indicates the occurrence of additional indices l and m, which is consistent with the definition (7.2.22). For instance the last line of (7.4.7) must be read as

$$\mathcal{A}_{yy}\boldsymbol{\mathcal{X}}^{k+2}_{i_1\dots i_{k+2},j} = 2\partial^A_{i_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},j} + A[\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}\boldsymbol{e}^T_{i_{k+1}}]\cdot\boldsymbol{e}_{i_{k+2}}.$$

Remark 7.18. There is no coupling between $\boldsymbol{\mathcal{X}}_j^k$ and $\boldsymbol{\mathcal{X}}_l^k$ for $j\neq l$: the columns of $\mathcal{X}^k$ are defined from uncorrelated recurrences.

Using a methodology similar to section 7.3, we may prove

Proposition 7.16. *The solutions $\boldsymbol{u}_k$ of the cascade of equations (7.4.5) are given by*

$$\boldsymbol{u}_i(x,y) = \sum_{k=0}^{i}\mathcal{X}^k(y)\cdot\nabla^k\boldsymbol{f}^{i-k}(x), \tag{7.4.8}$$

which can be formally written as the following double series product

$$\boldsymbol{u}_\varepsilon(x) = \sum_{i=0}^{+\infty}\varepsilon^{i+2}\mathcal{X}^i(x/\varepsilon)\cdot\nabla^i\boldsymbol{f}(x) = \left(\sum_{i=0}^{+\infty}\varepsilon^{i+2}\mathcal{X}^i(x/\varepsilon)\cdot\nabla^i\right)\left(\sum_{i=0}^{+\infty}\varepsilon^i\boldsymbol{f}^i(x)\right),\qquad x\in D_\varepsilon.$$

Proof. The result is obtained as in proposition 7.1, by using the following two identities which hold for a scalar function $f\in\mathcal{C}^\infty(D)$ and a vector field $\boldsymbol{a}\in\mathcal{C}^\infty(Y,\mathbb{R}^d)$:

$$\mathcal{A}_{xy}(f(x)\boldsymbol{a}(y)) = -2\partial_{x_j}f(x)\partial^A_{y_j}\boldsymbol{a}(y),$$

$$\mathcal{A}_{xx}(f(x)\boldsymbol{a}(y)) = -\partial^2_{x_lx_m}f(x)A[\boldsymbol{a}(y)\boldsymbol{e}_l^T]\cdot\boldsymbol{e}_m.$$

□

Following section 7.3, we denote by $\mathcal{X}^{i*}$ and $\boldsymbol{u}_i^*$ the averages of $\mathcal{X}^i(y)$ and $\boldsymbol{u}_i(x,y)$ with respect to the y variable:

$$\mathcal{X}^{i*} := \int_Y\mathcal{X}^i(y)\mathrm{d}y,\qquad \boldsymbol{u}_i^*(x) := \int_Y\boldsymbol{u}_i(x,y)\mathrm{d}y,\quad\forall x\in D,\qquad\forall i\in\mathbb{N}. \tag{7.4.9}$$

Our main objective is the derivation of high order homogenized equations yielding an approximation of the "infinite order" homogenized solution formally defined by

$$\boldsymbol{u}_\varepsilon^*(x) := \sum_{i=0}^{+\infty} \varepsilon^{i+2} \boldsymbol{u}_i^*(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i f(x), \quad x \in D. \tag{7.4.10}$$

Our first step is to extend the identities of proposition 7.2 to the present context: we show that $\mathcal{X}^{2k*}$ and $\mathcal{X}^{2k+1*}$ are respectively symmetric and antisymmetric matrix valued tensors.

Proposition 7.17. *For any $0 \leq p \leq k$, the following identity holds for the matrix valued tensor $\mathcal{X}^{k*}$:*

$$\forall 1 \leq i,j \leq d,\ \mathcal{X}_{ij}^{k*} = (-1)^p \int_Y (\mathcal{A}_{yy} \boldsymbol{\mathcal{X}}_i^p \cdot \boldsymbol{\mathcal{X}}_j^{k-p} - A[\boldsymbol{\mathcal{X}}_i^{p-1} \boldsymbol{e}_l^T] : [\boldsymbol{\mathcal{X}}_j^{k-p-1} \boldsymbol{e}_m^T] \otimes e_l \otimes e_m) \mathrm{d}y, \tag{7.4.11}$$

where the convention $\boldsymbol{\mathcal{X}}_i^{-1} = 0$ is assumed. In particular, for any $k \geq 0$:

- *$\mathcal{X}^{2k*}$ takes values in the set of $d \times d$ symmetric matrices:*

$$\forall 1 \leq i,j \leq d,\ \mathcal{X}_{ij}^{2k*} = \mathcal{X}_{ji}^{2k*}.$$

- *$\mathcal{X}^{2k+1*}$ takes values in the set of $d \times d$ antisymmetric matrices:*

$$\forall 1 \leq i,j \leq d,\ \mathcal{X}_{ij}^{2k+1*} = -\mathcal{X}_{ji}^{2k+1*}.$$

Furthermore, $\mathcal{X}^{2k*}$ and $\mathcal{X}^{2k+1*}$ depend only on $(\boldsymbol{\mathcal{X}}_j^k)_{1\leq j\leq d}$ and $(\boldsymbol{\mathcal{X}}_j^{k-1})_{1\leq j\leq d}$ as implied by the following identities:

$$\mathcal{X}_{ij}^{2k*} = (-1)^k \int_Y (A \nabla \boldsymbol{\mathcal{X}}_i^k : \nabla \boldsymbol{\mathcal{X}}_j^k - A[\boldsymbol{\mathcal{X}}_i^{k-1} \boldsymbol{e}_l^T] : [\boldsymbol{\mathcal{X}}_j^{k-1} \boldsymbol{e}_m^T] \otimes e_l \otimes e_m) \mathrm{d}y. \tag{7.4.12}$$

$$\begin{aligned} \mathcal{X}_{ij}^{2k+1*} = (-1)^k &\int_Y (\boldsymbol{\mathcal{X}}_i^k \cdot A \nabla \boldsymbol{\mathcal{X}}_j^k - \boldsymbol{\mathcal{X}}_j^k \cdot A \nabla \boldsymbol{\mathcal{X}}_i^k) \cdot \boldsymbol{e}_l \otimes e_l \mathrm{d}y \\ &+ (-1)^k \int_Y (A[\boldsymbol{\mathcal{X}}_j^{k-1} \boldsymbol{e}_l^T] : [\boldsymbol{\mathcal{X}}_i^k \boldsymbol{e}_m^T] - A[\boldsymbol{\mathcal{X}}_i^{k-1} \boldsymbol{e}_l^T] : [\boldsymbol{\mathcal{X}}_j^k \boldsymbol{e}_m^T]) \otimes e_l \otimes e_m \mathrm{d}y. \end{aligned} \tag{7.4.13}$$

Remark 7.19.

1. There is no reason for the odd order tensors $\mathcal{X}^{2k+1*}$ being zero for all $k \geq 0$. Although it does not seem straightforward to exhibit a particular shape of hole for which it could be proved that $\mathcal{X}^{2k+1*} \neq 0$, numerical experiments tend to confirm this conjecture (see section 7.4.5). This fact is an important difference with the scalar case.

2. We shall see in section 7.4.4 below that odd order tensors $\mathcal{X}^{2k+1*}$ and M^{2k+1} vanish under symmetry assumptions of the periodic pattern Y.

3. Identity (7.4.12) is similar to (7.3.21): $\mathcal{X}^{2k*}$ is a difference betweent two positive tensors.

It can also be shown that the tensors $\boldsymbol{\mathcal{X}}_j^k$ take linearly independent vector values in $\mathbb{R}^d$ (in particular, there must exist non zero components of $\boldsymbol{\mathcal{X}}_j^k$ for any $1 \leq j \leq d$):

Proposition 7.18. *For any $1 \leq j \leq d$, the following identity holds:*

$$\forall k \geq 0,\ \mathcal{A}_{yy}(\partial_{i_1 \ldots i_k}^k \boldsymbol{\mathcal{X}}_{i_1 \ldots i_k, j}^k) = (-1)^k (k+1) \boldsymbol{e}_j.$$

Proof. The result is obtained by adapting the proof of proposition 7.3 and by using the following formulas which hold for any smooth vector field $\boldsymbol{\mathcal{X}} \in \mathcal{C}^\infty(Y)$:

$$-\partial_j \partial_j^A \boldsymbol{\mathcal{X}} = \mathcal{A}_{yy} \boldsymbol{\mathcal{X}}, \tag{7.4.14}$$

$$-\partial_{lm}^2 (A[\boldsymbol{\mathcal{X}} \boldsymbol{e}_l^T] \cdot \boldsymbol{e}_m) = \mathcal{A}_{yy} \boldsymbol{\mathcal{X}}.$$

□

7.4.2 High order homogenized equations: tensors M^k, N^k, $\mathbb{B}_K$ and $\mathbb{D}_K$

We now derive high order homogenized equations for the perforated elasticity problem (7.4.1) following the methodology described in the beginning of section 7.3.2.

Let us start by recalling the matrix $\mathcal{X}^{0*}$ is symmetric positive definite as a consequence of the identity

$$\mathcal{X}^0_{ij} = \int_Y A\nabla\boldsymbol{\mathcal{X}}^0_i : \nabla\boldsymbol{\mathcal{X}}^0_j \mathrm{d}y, \quad 1 \leq i,j \leq d,$$

and of the linear independence of the vector fields $(\boldsymbol{\mathcal{X}}^0_j)_{1\leq j\leq d}$ (proposition 7.18).

Proposition 7.19. *We define a family (M^k) of tensors of order k by induction as follows:*

$$\begin{cases} M^0 = (\mathcal{X}^{0*})^{-1} \\ M^k = -(\mathcal{X}^{0*})^{-1} \displaystyle\sum_{p=0}^{k-1} \mathcal{X}^{k-p*} \otimes M^p \quad \forall k \geq 1. \end{cases} \tag{7.4.15}$$

Then the source terms $\boldsymbol{f}_i$ (eqn. (7.4.4)) are given in terms of the averaged ansatz terms $\boldsymbol{u}^*_i$ through the following identity:

$$\forall i \geq 0,\ \boldsymbol{f}_i(x) = \sum_{k=0}^{i} M^k \cdot \nabla^k \boldsymbol{u}^*_{i-k}(x). \tag{7.4.16}$$

Recognizing a Cauchy product, this formula can be rewritten formally as the following "infinite order homogenized equation" for the averaged ansatz $\boldsymbol{u}^*_\varepsilon$ (eqn. (7.4.10)):

$$\sum_{i=0}^{+\infty} \varepsilon^{i-2} M^k \cdot \nabla^k \boldsymbol{u}^*_\varepsilon = \boldsymbol{f}. \tag{7.4.17}$$

The explicit formula for the tensors M^k of proposition 7.7 extends to the vectorial setting:

Proposition 7.20. *For any $k \geq 1$, the tensor M^k of (7.4.15) reads explicitly*

$$M^k = \sum_{p=1}^{k} (-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1,\ldots,i_p\leq k}} (\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_1*} \otimes (\mathcal{X}^{0*})^{-1} \otimes \cdots \otimes (\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_p*} \otimes (\mathcal{X}^{0*})^{-1}. \tag{7.4.18}$$

We now introduce the tensors N^k which allow to write the criminal ansatz expressing $\boldsymbol{u}_\varepsilon$ in terms of $\boldsymbol{u}^*_\varepsilon$:

Proposition 7.21. *Let N^k be the k-th order matrix valued tensor defined by*

$$N^k(y) := \sum_{p=0}^{k} \mathcal{X}^{k-p}(y) \otimes M^p, \quad y \in Y.$$

Then the terms $\boldsymbol{u}_i(x,y)$ of the oscillating ansatz (7.4.3) can be written in terms of their averages $\boldsymbol{u}^*_i(x)$ (eqn. (7.4.9)) as follows:

$$\forall i \geq 0,\ \boldsymbol{u}_i(x,y) = \sum_{k=0}^{i} N^k(y) \cdot \nabla^k \boldsymbol{u}^*_{i-k}(x). \tag{7.4.19}$$

Recognizing a Cauchy product, this formally rewrites in terms of the following "criminal" ansatz expressing the oscillating solution $\boldsymbol{u}_\varepsilon$ in terms of its formal average $\boldsymbol{u}^*_\varepsilon$:

$$\boldsymbol{u}_\varepsilon(x) = \sum_{k=0}^{+\infty} \varepsilon^k N^k(x/\varepsilon) \cdot \nabla \boldsymbol{u}^*_\varepsilon(x), \qquad x \in D_\varepsilon. \tag{7.4.20}$$

The next proposition summarizes the properties of the tensors N^k and of their columns vectors $(\boldsymbol{N}^k_j)_{1\leq j\leq d}$

Proposition 7.22. *Let $(\boldsymbol{N}_j^k)_{1\leq j\leq d}$ be the column vectors of the matrix tensors N^k:*

$$\forall 1\leq i,j\leq d,\ \boldsymbol{N}_j^k := N^k\boldsymbol{e}_j.$$

The tensors N^k and vector fields $\boldsymbol{N}_j^k$ satisfy:

1. *$\int_Y N^0(y)\mathrm{d}y = I$ and $\int_Y N^k(y)\mathrm{d}y = 0$ for any $k\geq 1$;*
2. *for any $1\leq j\leq d$, it holds*

$$\begin{cases} \mathcal{A}_{yy}\boldsymbol{N}_j^0 = M^0\boldsymbol{e}_j \\ \mathcal{A}_{yy}\boldsymbol{N}_j^1 = 2\partial_l^A\boldsymbol{N}_j^0\otimes e_l + M^1\boldsymbol{e}_j \\ \mathcal{A}_{yy}\boldsymbol{N}_j^{k+2} = 2\partial_l^A\boldsymbol{N}^{k+1}\otimes e_l + (A[\boldsymbol{N}_j^k\boldsymbol{e}_l^T]\cdot\boldsymbol{e}_m)\otimes e_l\otimes e_m + M^{k+2}\boldsymbol{e}_j; \end{cases}$$

3. *$\mathcal{A}_{yy}(\partial^k_{i_1\dots i_k}\boldsymbol{N}^k_{i_1\dots i_k,j}) = (-1)^k(k+1)M^0\boldsymbol{e}_j$, for any $k\geq 0$;*
4. *For any $1\leq p\leq k-1$,*

$$M_{ij}^k = (-1)^{p+1}\int_Y(\mathcal{A}_{yy}\boldsymbol{N}_i^p\cdot\boldsymbol{N}_j^{k-p} - A[\boldsymbol{N}_i^{p-1}\boldsymbol{e}_l^T]:[\boldsymbol{N}_j^{k-p-1}\boldsymbol{e}_m^T]\otimes e_l\otimes e_m)\mathrm{d}y.$$

In particular, M^{2k} and M^{2k+1} depend only on the tensors N^k and N^{k-1}, which depend themselves only on the first $k+1$ tensors $\mathcal{X}^0\dots\mathcal{X}^k$.

Importantly, the next proposition states that the tensors M^{2k} and M^{2k+1} are also respectively symmetric and antisymmetric matrix valued (which could also be seen from (7.4.18)):

Corollary 7.5. *For any $k\geq 0$,*

- *M^{2k} is a symmetric matrix valued tensor, and the following identities hold:*

$$M_{ij}^0 = \int_Y A\nabla\boldsymbol{N}_i^0:\nabla\boldsymbol{N}_j^0\mathrm{d}y, \tag{7.4.21}$$

$$\forall k\geq 1,\ M_{ij}^{2k} = (-1)^{k+1}\int_Y(A\nabla\boldsymbol{N}_i^k:\nabla\boldsymbol{N}_j^k - A[\boldsymbol{N}_i^{k-1}\boldsymbol{e}_l^T]:[\boldsymbol{N}_j^{k-1}\boldsymbol{e}_m^T]\otimes e_l\otimes e_m)\mathrm{d}y. \tag{7.4.22}$$

- *M^{2k+1} is an antisymmetric matrix valued tensor, and the following identities hold:*

$$\begin{aligned}\forall k\geq 0,\ M_{ij}^{2k+1} &= (-1)^{k+1}\int_Y(\boldsymbol{N}_i^k\cdot A\nabla\boldsymbol{N}_j^k - \boldsymbol{N}_j^k\cdot A\nabla\boldsymbol{N}_i^k)\cdot\boldsymbol{e}_l\otimes e_l\mathrm{d}y \\ &\quad + (-1)^{k+1}\int_Y(A[\boldsymbol{N}_j^{k-1}\boldsymbol{e}_l^T]:[\boldsymbol{N}_i^k\boldsymbol{e}_m^T] - A[\boldsymbol{N}_i^{k-1}\boldsymbol{e}_l^T]:[\boldsymbol{N}_j^k\boldsymbol{e}_m^T])\otimes e_l\otimes e_m\mathrm{d}y,\end{aligned}$$

  where the convention $N^{-1}=0$ is assumed.

Remark 7.20. It may appear quite surprising that odd order differential operators $\varepsilon^{2k-1}M^{2k+1}\cdot\nabla^{2k+1}$ arise in the "infinite order" homogenized equation (7.4.17), while the original operator $-\mathrm{div}(A\nabla\cdot)$ is symmetric. In fact, there is no contradiction because the antisymmetry of M^{2k+1} compensates the one induced by odd order derivatives which makes $M^{2k+1}\cdot\nabla^{2k+1}$ be a symmetric operator. Indeed, for two vector fields $\boldsymbol{u} := (u_i)_{1\leq i\leq d}, \boldsymbol{v} = (v_i)_{1\leq i\leq d}$, it holds

$$\begin{aligned}\int_Y\boldsymbol{v}\cdot M^{2k+1}\cdot\nabla^{2k+1}\boldsymbol{u}\mathrm{d}y &= \int_Y(M_{ij}^{2k+1}\cdot\nabla^{2k+1}u_j)v_i\mathrm{d}y = \int_Y(-1)^{2k+1}(M_{ij}^{2k+1}\cdot\nabla^{2k+1}v_i)u_j\mathrm{d}y \\ &= \int_Y(-1)^{2k+1}(-1)(M_{ji}^{2k+1}\cdot\nabla^{2k+1}v_i)u_j\mathrm{d}y = \int_Y\boldsymbol{u}\cdot M^{2k+1}\cdot\nabla^{2k+1}\boldsymbol{v}\mathrm{d}y.\end{aligned}$$

Remark 7.21. If the original tensor A acts on symmetric matrices (e.g. if A is given by the Hooke's Law (7.4.2)), which means $A_{ijkl} = A_{ijlk}$ for $1\leq i,j,k,l\leq d$, the same property is *a priori* not necessarily verified for the homogenized tensor M^2, which would mean $M^2_{i_1i_2,pq} = M^2_{i_1q,pi_2}$ (see the formulas (7.4.18) and (7.4.13)). We interpret this physically by the fact that inhomogeneities in the obstacle may translate in M^2 being sensitive to the antisymmetric part of the gradient. The same remark applies as well for the other tensors M^k: partial derivative indices cannot *a priori* be permuted with spatial indices. Such occurs however in case of symmetries of the obstacle, see the dedicated section 7.4.4.

We now derive well-posed homogenized equations of higher but finite order, which are in principle amenable to numerical computations. Following section 7.3.3, recall $\boldsymbol{u}_\varepsilon$ is the solution to the energy minimization problem

$$\begin{aligned} \min_{\boldsymbol{w}\in H^1(D_\varepsilon,\mathbb{R}^d)} \quad & J(\boldsymbol{u},\boldsymbol{f}) := \int_D \left(\frac{1}{2}A\nabla\boldsymbol{u}:\nabla\boldsymbol{u} - \boldsymbol{f}\cdot\boldsymbol{u}\right)\mathrm{d}x \\ s.t. \quad & \begin{cases} \boldsymbol{w} = 0 \text{ on } \partial\omega_\varepsilon \\ \boldsymbol{w} \text{ is } D\text{–periodic.} \end{cases} \end{aligned}$$

We consider truncated criminal ansatz $\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v})$ depending on a non oscillating vector function $\boldsymbol{v} \in H^{K+1}(D)$ obtained by truncation of (7.4.20):

$$\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v})(x) := \sum_{k=0}^{K} \varepsilon^k N^k(x/\varepsilon)\cdot\nabla^k\boldsymbol{v}(x),\ x\in D_\varepsilon. \tag{7.4.23}$$

This allows us to consider a minimization problem for the function $\boldsymbol{v}$ which is sought to approximate the formal infinite average $\boldsymbol{u}_\varepsilon^*$ of (7.4.10):

$$\begin{aligned} \min_{\boldsymbol{v}\in H^{K+1}(D)} \quad & J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f}) \\ s.t. \quad & \boldsymbol{v} \text{ is } D\text{–periodic.} \end{aligned} \tag{7.4.24}$$

Still following (7.3.3), we eliminate the fast variable x/ε in (7.4.24) by applying lemma 7.3 which yields an approximate energy $J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon)$ which we now define.

For a given $K\in\mathbb{N}$, $1\le j\le d$ and $0\le k\le K+1$, we denote by $(\widetilde{\boldsymbol{N}}^l_{p,j})_{1\le j\le d}$ the d vector tensors defined for $1\le j\le d$ by

$$\widetilde{\boldsymbol{N}}^l_{p,j}(y) := \begin{cases} \partial_p \boldsymbol{N}^0_j(y) \text{ if } l=0 \\ \partial_p \boldsymbol{N}^l_j(y) + \boldsymbol{N}^{l-1}_j(y)\otimes e_p \text{ if } 1\le l\le l, \\ \boldsymbol{N}^K_j(y)\otimes e_p \text{ if } l = K+1. \end{cases} \tag{7.4.25}$$

Then we define a constant, bilinear matrix valued tensor $\mathbb{B}_K^{k,l}$ of order $k+l$:

$$\mathbb{B}^{l,m}_{K,ij} := \int_Y A[\widetilde{\boldsymbol{N}}^l_{p,i}\boldsymbol{e}_p^T] : [\widetilde{\boldsymbol{N}}^m_{q,j}\boldsymbol{e}_q^T]\mathrm{d}y, \qquad \forall 1\le i,j\le d, \tag{7.4.26}$$

where the summation over the repeated indices p and q and a tensor product in the double contraction is implicitly assumed, following the convention of section 7.2.4. Recall that for a vector field $\boldsymbol{v} := (v_p)_{1\le p\le d}$ we denote by $\mathbb{B}^{l,m}_K\nabla^l\boldsymbol{v}\nabla^m\boldsymbol{v}$ the contraction

$$\mathbb{B}^{l,m}_K\nabla^l\boldsymbol{v}\nabla^m\boldsymbol{v} := (\mathbb{B}^{l,m}_{K,pq})_{i_1\dots i_l j_1\dots j_m}\partial^l_{i_1\dots i_l}v_p\partial^m_{j_1\dots j_m}v_q.$$

The tensors $\mathbb{B}^{l,m}$ allow us to formulate a new energy $J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon)$:

$$J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon) := \int_D \left(\frac{1}{2}\sum_{k,l=0}^{K+1}\varepsilon^{k+l-2}\mathbb{B}^{k,l}_K\nabla^k\boldsymbol{v}\nabla^l\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v}\right)\mathrm{d}x. \tag{7.4.27}$$

The definition of the energy J_K^* is motivated by the following result, which can be obtained by a straightforward adaptation of the proof of proposition 7.10:

Proposition 7.23. *Assume $\boldsymbol{f}\in\mathcal{C}^\infty(D,\mathbb{R}^d)$ and D–periodic. For any smooth D–periodic vector field $\boldsymbol{v}\in\mathcal{C}^\infty(D,\mathbb{R}^d)$ and truncated ansatz $\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v})\in\mathcal{C}^\infty(D_\varepsilon)$ of the form of (7.4.23), the following energy asymptotic holds true with $m\ge 0$ arbitrarily high:*

$$|J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f}) - J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon)| \le C_{K,m}(||\nabla^{m+2}\boldsymbol{v}||^2_{L^\infty(D,\mathbb{R}^{d\times d})} + ||\boldsymbol{f}||^2_{H^m(D,\mathbb{R}^d)}).$$

Homogenized equations are obtained by considering the new minimization problem where $J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f})$ is replaced with $J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon)$ in (7.4.24).

Definition 7.7. For any $K \in \mathbb{N}$, we call homogenized equation of order $2K+2$ the Euler-Lagrange equation associated with the minimization problem

$$\min_{\substack{\boldsymbol{v}\in H^{K+1}(D,\mathbb{R}^d),\\ D \text{ periodic}}} J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon). \tag{7.4.28}$$

This equation reads explicitly in terms of a higher order homogenized solution $\boldsymbol{v}_K^* \in H^{K+1}(D,\mathbb{R}^d)$ as

$$\begin{cases} \displaystyle\sum_{k=0}^{2K+2} \varepsilon^{k-2}\mathbb{D}_K^k \cdot \nabla^k \boldsymbol{v}_K^* = \boldsymbol{f} \\ \boldsymbol{v}_K^* \text{ is } D\text{–periodic,} \end{cases} \tag{7.4.29}$$

where the constant (matrix valued) tensors $\mathbb{D}_K^k$ are defined by the following formula for any $0 \le k \le 2K+2$:

$$\forall 1 \le i,j \le d,\ \mathbb{D}_{K,ij}^k := \begin{cases} \displaystyle\sum_{K=0}^{k} (-1)^l \frac{1}{2}(\mathbb{B}_{K,ij}^{l,k-l} + \mathbb{B}_{K,ji}^{l,k-l}), \text{ if } k \text{ is even} \\ \displaystyle\sum_{K=0}^{k} (-1)^l \frac{1}{2}(\mathbb{B}_{K,ij}^{l,k-l} - \mathbb{B}_{K,ji}^{l,k-l}), \text{ if } k \text{ is odd,} \end{cases} \tag{7.4.30}$$

where the convention $\mathbb{B}_K^{l,m} = 0$ whenever $l > K+1$ or $m > K+1$ is assumed.

Again, an important difference with the scalar case (eqn. (7.3.50)) is the occurrence of differential operators $\varepsilon^{2k-1}\mathbb{D}_K^{2k+1} \cdot \nabla^{2k+1}$ of odd orders in the homogenized equation (7.4.29). The k-th order tensor $\mathbb{D}_K^k$ is also symmetric matrix valued for even values of k, and antisymmetric otherwise, which was to be expected from remark 7.20.

Remark 7.22. Let us examine (7.4.29) at order $K = 0$. It reads

$$\mathbb{D}_0^2 \cdot \nabla^2 \boldsymbol{v}_0^* + \varepsilon^{-1}\mathbb{D}_0^1 \cdot \nabla \boldsymbol{v}_0^* + \varepsilon^{-2}\mathbb{D}_0^0 \boldsymbol{v}_0^* = \boldsymbol{f}$$

where one can verify that the tensors $\mathbb{D}_0^0$, $\mathbb{D}_0^1$ and $\mathbb{D}_0^2$ are respectively given by

$$\mathbb{D}_0^0 = M^0, \tag{7.4.31}$$

$$\mathbb{D}_0^1 = M^1, \tag{7.4.32}$$

$$\mathbb{D}_{0,ij}^2 = (\mathcal{X}^{0*})_{ip}^{-1}(\mathcal{X}^{0*})_{jq}^{-1} \int_Y A[\boldsymbol{\mathcal{X}}_p^0 \boldsymbol{e}_l^T] : [\boldsymbol{\mathcal{X}}_q^0 \boldsymbol{e}_m^T] \otimes e_l \otimes e_m \mathrm{d}y, \qquad 1 \le i,j \le d. \tag{7.4.33}$$

The physical interpretation of the rather *very strange* term $\varepsilon^{-1}\mathbb{D}_0^1 \cdot \nabla \boldsymbol{v}_0^*$ does not seem obvious to us.

Observing that the dominant tensor $\mathbb{B}_K^{K+1,K+1}$ of (7.4.26) is symmetric and nonnegative, the well-posedness result of proposition 7.11 extends without difficulty:

Proposition 7.24. *Assume further the dominant tensor $\mathbb{D}_K^{2K+2} = (-1)^{K+1}\mathbb{B}_K^{K+1,K+1}$ is non-degenerate, that is there exists a constant ν such that*

$$\forall \boldsymbol{\xi} = \boldsymbol{\xi}_{i_1 \dots i_{K+1},j} \in \mathbb{R}^{d^{K+1}} \times \mathbb{R}^d,\ \mathbb{B}_K^{K+1,K+1}\boldsymbol{\xi}\boldsymbol{\xi} \ge \nu|\boldsymbol{\xi}|^2. \tag{7.4.34}$$

Then there exists a unique solution $\boldsymbol{v}_K^* \in H^{K+1}(D,\mathbb{R}^d)$ to the homogenized equation (7.4.29).

Finally, it remains to verify that (7.4.29) is a truncation in some sense of the infinite order homogenized equation (7.4.17):

Proposition 7.25. *The first $K+1$ homogenized coefficients of the homogenized equation (7.4.29) coincide with those of the formal infinite order homogenized equation (7.4.17):*

$$\forall 0 \le k \le K,\ \mathbb{D}_K^k = M^k.$$

Proof. We follow the proof of proposition 7.12. For $0 \leq k,l \leq K$ and $1 \leq i,j \leq d$, the coefficient $\mathbb{B}^{l,k-l}_{K,ij}$ is given by

$$\mathbb{B}^{l,k-l}_{K,ij} = \int_Y A[(\partial_p \boldsymbol{N}^l_i + \boldsymbol{N}^{l-1}_i \otimes e_p)\boldsymbol{e}^T_p] : [(\partial_q \boldsymbol{N}^{k-l}_j + \boldsymbol{N}^{k-l-1}_j \otimes e_q)\boldsymbol{e}^T_q]\mathrm{d}y.$$

After integration by parts, the following identities hold:

$$\int_Y A[\partial_p \boldsymbol{N}^l_i \boldsymbol{e}^T_p] : [\partial_q \boldsymbol{N}^{k-l}_j \boldsymbol{e}^T_q]\mathrm{d}y = \int_Y A\nabla \boldsymbol{N}^l_i : \nabla \boldsymbol{N}^{k-l}_j \mathrm{d}y = \int_Y \mathcal{A}_{yy}\boldsymbol{N}^l_i \cdot \boldsymbol{N}^{k-l}_j \mathrm{d}y,$$

$$\int_Y A[\boldsymbol{N}^{l-1}_i \otimes e_p \boldsymbol{e}^T_p] : [\partial_q \boldsymbol{N}^{k-l}_j \boldsymbol{e}^T_q]\mathrm{d}y = -\int_Y \mathrm{div}(A[\boldsymbol{N}^{l-1}_i \boldsymbol{e}^T_p]) \cdot \boldsymbol{N}^{k-l}_j \otimes e_p \mathrm{d}y.$$

This allows to rewrite $\mathbb{B}^{l,k-l}_{K,ij}$ as follows:

$$\mathbb{B}^{l,k-l}_{K,ij} = \int_Y (\mathcal{A}_{yy}\boldsymbol{N}^l_i - 2\partial^A_p \boldsymbol{N}^{l-1}_i \otimes e_p - A[\boldsymbol{N}^{l-2}_i \boldsymbol{e}^T_p] \cdot \boldsymbol{e}_q \otimes e_p \otimes e_q) \cdot \boldsymbol{N}^{k-l}_j \mathrm{d}y + B^{k,l}_{ij} + B^{k,l+1}_{ij}$$

where $B^{k,l}_{ij}$ is the k-th order tensor defined by

$$B^{k,l}_{ij} := \int_Y (A\nabla \boldsymbol{N}^{l-1}_i : [\boldsymbol{N}^{k-l}_j] + A[\boldsymbol{N}^{l-2}_i \boldsymbol{e}^T_p] : [\boldsymbol{N}^{k-l}_j \boldsymbol{e}^T_q] \otimes e_p \otimes e_q)\mathrm{d}y.$$

Hence we obtain, recognizing a telescopic series:

$$\begin{aligned}\sum_{l=0}^{k}(-1)^l \mathbb{B}^{l,k-l}_{K,ij} &= \sum_{l=0}^{k-l}(-1)^l \int_Y (M^l \boldsymbol{e}_i) \cdot \boldsymbol{N}^{k-l}_j \mathrm{d}y + \sum_{l=0}^{k}((-1)^l B^{k,l}_{ij} - (-1)^{l+1} B^{k,l+1}_{ij}) \\ &= (-1)^k (M^k \boldsymbol{e}_i) \cdot \boldsymbol{e}_j + B^{k,0}_{ij} - (-1)^{k+1} B^{k,k+1}_{ij} \\ &= (-1)^k M^k_{ji}.\end{aligned}$$

From there, we obtain

$$\mathbb{D}^k_{ij} = \frac{1}{2}((-1)^k M^k_{ji} + M^k_{ij})$$

which concludes the proof since we know that both M^k and $\mathbb{D}^k$ are symmetric for even k and antisymmetric otherwise. □

Remark 7.23. This result may not be optimal, i.e. $\mathbb{D}^k_K = M^k$ may hold up to $k = K+1$ under some circumstances, because equality (7.4.32) is not explained by the previous proposition.

The proof of the error estimates in proposition 7.13 for the scalar case relied essentially on the coercivity of the operator $-\Delta$ on the space $H^1(D)$. The results extends therefore without difficulty to the vectorial context due to the coercivity of the operator $-\mathrm{div}(A\nabla\cdot)$ on $H^1(D,\mathbb{R}^d)$:

Proposition 7.26. *Let $\boldsymbol{v}^*_K$ the solution to the homogenized equation (7.4.29) of order $2K+2$. There exists a constant $C_K(\boldsymbol{f})$ independent of ε such that the following error estimates hold:*

$$\left|\left|\boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}^*_K\right|\right|_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f})\varepsilon^{K+2}. \tag{7.4.35}$$

$$\left|\left|\boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}^*_K\right|\right|_{L^2(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f})\varepsilon^{K+3}. \tag{7.4.36}$$

7.4.3 Low volume fraction limits when the size of the obstacle tends to 0

We now extend the results of section 7.3.5 to the vectorial context: we establish low volume fraction asymptotics for the tensors $\mathcal{X}^{k*}$ and M^k when the size η of the obstacle (ηT) converges to zero. Again, we assume

$$d \geq 3, \tag{7.4.37}$$

although analogous results could be stated for $d = 2$. For $1 \leq j \leq d$, we introduce the solution $\boldsymbol{\Psi}_j$ to the exterior problem

$$\begin{cases} -\mathrm{div}(A\nabla\boldsymbol{\Psi}_j) = 0 \text{ in } \mathbb{R}^d\backslash T \\ \boldsymbol{\Psi}_j = 0 \text{ on } \partial T \\ \boldsymbol{\Psi}_j \to \boldsymbol{e}_j \text{ at } \infty, \end{cases} \tag{7.4.38}$$

where the boundary condition $\boldsymbol{\Psi}_j \to \boldsymbol{e}_j$ at infinity means $\boldsymbol{\Psi}_j - \boldsymbol{e}_j \in \mathcal{D}^{1,2}(\mathbb{R}^d\backslash T)$. We denote by $\Psi^* := (\Psi^*_{ij})_{1\leq i,j\leq d}$ the matrix collecting the normal stress components:

$$\Psi^*_{ij} := \int_{\mathbb{R}^d\backslash T} A\nabla\boldsymbol{\Psi}_i : \nabla\boldsymbol{\Psi}_j \mathrm{d}x = -\int_{\partial T} \boldsymbol{e}_j \cdot A\nabla\boldsymbol{\Psi}_i \cdot \boldsymbol{n}\mathrm{d}s, \tag{7.4.39}$$

where we recall that the normal $\boldsymbol{n}$ is assumed to point *inward* T. In this section, we write A^2 in order to refer to the 4-th order elasticity tensor A seen as a second order matrix valued tensor: for given derivative indices $1 \leq i_1, i_2 \leq d$ and spatial indices $1 \leq l, m \leq d$,

$$A^2_{i_1 i_2, lm} := A_{l i_1 m i_2} \tag{7.4.40}$$

which also means

$$A^2_{ij} = A_{iljm} e_l \otimes e_m = A[\boldsymbol{e}_i \boldsymbol{e}_l^T] : [\boldsymbol{e}_j \boldsymbol{e}_m^T] e_l \otimes e_m, \qquad 1 \leq i, j \leq d. \tag{7.4.41}$$

Proposition 7.14 extends as follows:

Proposition 7.27. *Assume $d \geq 3$. For any $k \geq 0$, and $1 \leq j \leq d$, let $\widetilde{\boldsymbol{\mathcal{X}}}^{2k}_j$ and $\widetilde{\boldsymbol{\mathcal{X}}}^{2k+1}_j$ be the rescaled vector tensors in $\eta^{-1}P\backslash T$ defined by*

$$\forall x \in \eta^{-1}P\backslash T,\ \widetilde{\boldsymbol{\mathcal{X}}}^{2k}_j(x) := \eta^{(d-2)(k+1)} \boldsymbol{\mathcal{X}}^{2k}_j(\eta x) \text{ and } \widetilde{\boldsymbol{\mathcal{X}}}^{2k+1}_j(x) := \eta^{(d-2)(k+1)} \boldsymbol{\mathcal{X}}^{2k+1}_j(\eta x).$$

Then:

1. *there exists a constant $C > 0$ independent of $\eta > 0$ such that:*

$$\forall \eta > 0,\ ||\nabla\widetilde{\boldsymbol{\mathcal{X}}}^{2k}_j||_{L^2(\eta^{-1}P\backslash T, \mathbb{R}^{d\times d})} \leq C \text{ and } ||\nabla\widetilde{\boldsymbol{\mathcal{X}}}^{2k+1}_j||_{L^2(\eta^{-1}P\backslash T, \mathbb{R}^{d\times d})} \leq C, \tag{7.4.42}$$

2. *the following convergences hold as $\eta \to 0$:*

$$\widetilde{\boldsymbol{\mathcal{X}}}^{2k}_i \rightharpoonup c^{2k}_{ij} \boldsymbol{\Psi}_j \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T, \mathbb{R}^d), \tag{7.4.43}$$

$$\widetilde{\boldsymbol{\mathcal{X}}}^{2k+1}_i \rightharpoonup 0 \text{ weakly in } H^1_{loc}(\mathbb{R}^d\backslash T, \mathbb{R}^d), \tag{7.4.44}$$

$$\mathcal{X}^{2k*} \sim \frac{1}{\eta^{(d-2)(k+1)}} c^{2k}, \tag{7.4.45}$$

$$\mathcal{X}^{2k+1*} = o\left(\frac{1}{\eta^{(d-2)(k+1)}}\right), \tag{7.4.46}$$

where c^{2k}_{ij} denote the coefficients of the $2k$-th order matrix valued tensor $c^{2k} := (c^{2k}_{ij})_{1\leq i,j\leq d}$ given by:

$$c^{2k} := (\Psi^*)^{-1} \otimes \overbrace{A^2 \otimes (\Psi^*)^{-1} \cdots \otimes A^2 \otimes (\Psi^*)^{-1}}^{k \text{ times}}. \tag{7.4.47}$$

Proof. The proof follows the lines of that of proposition 7.10, so we content ourselves with highlighting only the main differences that are due to the vectorial setting. The results is proved by induction on k.

1. *Case $2k$ for $k = 0$.* The vector valued tensor $\widetilde{\boldsymbol{\mathcal{X}}}^0_i$ satisfies $\mathcal{A}_{yy}\widetilde{\boldsymbol{\mathcal{X}}}^0_i = \eta^d \boldsymbol{e}_i$ in $\eta^{-1}P\backslash T$, hence for any test function $\boldsymbol{\Phi} \in H^1(\eta^{-1}P\backslash T, \mathbb{R}^d)$ which is $\eta^{-1}P$–periodic, it holds

$$\int_{\eta^{-1}P\backslash T} A\nabla\widetilde{\boldsymbol{\mathcal{X}}}^0_i : \nabla\boldsymbol{\Phi}\mathrm{d}x = \eta^d \int_{\eta^{-1}P\backslash T} \boldsymbol{\Phi} \cdot \boldsymbol{e}_i \mathrm{d}x + \int_{\partial T} \boldsymbol{\Phi} \cdot A\nabla\widetilde{\boldsymbol{\mathcal{X}}}^0_i \cdot \boldsymbol{n}\mathrm{d}s. \tag{7.4.48}$$

Substituting $\boldsymbol{\Phi} = \widetilde{\boldsymbol{\mathcal{X}}}_i^0$ in (7.4.48) and using lemma 7.7 yields $||\nabla \widetilde{\boldsymbol{\mathcal{X}}}_i^0||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \leq C$. Then (7.3.66) implies that $< \widetilde{\boldsymbol{\mathcal{X}}}_i^0 >$ is bounded in $\mathbb{R}^d$ as $\eta \to 0$. Up to the extraction of a subsequence, there exists a matrix $c^0 = (c^0_{ij})_{1\leq i,j\leq d}$ and vector fields $(\widehat{\boldsymbol{\Psi}}^0_i)_{1\leq i\leq d}$ such that

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^0 > \cdot \boldsymbol{e}_j \to c^0_{ij} \text{ and } \widetilde{\boldsymbol{\mathcal{X}}}_i^0 \rightharpoonup \widehat{\boldsymbol{\Psi}}^0_i \text{ weakly in } H^1_{loc}(\eta^{-1}P\backslash T,\mathbb{R}^d) \text{ as } \eta \to 0.$$

Passing (7.4.48) to the limit when $\eta \to 0$ with a compactly supported $\boldsymbol{\Phi} \in \mathcal{C}^\infty(\eta^{-1}P\backslash T,\mathbb{R}^d)$ yields then $-\mathrm{div}(A\nabla\widehat{\Psi}^0_i) = 0$ in $\mathbb{R}^d\backslash T$. From (7.3.68) and the lower semi-continuity of the $H^1_{loc}(\eta^{-1}P\backslash T,\mathbb{R}^d)$ norm, we know that $\widehat{\boldsymbol{\Phi}}^0_i - c^0_{ij}\boldsymbol{e}_j$ belongs to $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T,\mathbb{R}^d)$, which readily implies $\widehat{\boldsymbol{\Psi}}^0_i = c^0_{ij}\boldsymbol{\Psi}_j$. Setting now $\boldsymbol{\Phi} = \boldsymbol{e}_j$ in (7.4.48), using the continuity of the normal force (7.4.39) with respect to the weak convergence in $H^1_{loc}(\mathbb{R}^d\backslash T,\mathbb{R}^d)$, we obtain

$$-\int_{\partial T} \boldsymbol{e}_j \cdot A\nabla\widehat{\boldsymbol{\Psi}}^0_i \cdot \boldsymbol{n}\mathrm{d}s = c^0_{ip}\Psi^*_{pj} = \delta_{ij}.$$

This means exactly $c^0 = (\Psi^*)^{-1}$, from where the results follows for $\widetilde{\boldsymbol{\mathcal{X}}}_i^0$.

2. *Case* $2k+1$ *for* $k = 0$. The tensor $\widetilde{\boldsymbol{\mathcal{X}}}_i^1$ satisfies $\mathcal{A}_{yy}\widetilde{\boldsymbol{\mathcal{X}}}_i^1 = 2\eta\partial^A_l\widetilde{\boldsymbol{\mathcal{X}}}_i^1 \otimes e_l$ which can be rewritten

$$\int_{\eta^{-1}P\backslash T} A\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^1 : \nabla\boldsymbol{\Phi}\mathrm{d}x = \int_{\eta^{-1}P\backslash T} 2\eta\partial^A_l\widetilde{\boldsymbol{\mathcal{X}}}_i^0 \cdot \boldsymbol{\Phi}\otimes e_l\mathrm{d}x + \int_{\partial T} \boldsymbol{\Phi}\cdot A\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^1 \cdot \boldsymbol{n}\mathrm{d}s \tag{7.4.49}$$

for any periodic $\boldsymbol{\Phi}$ in $H^1(\eta^{-1}P\backslash T,\mathbb{R}^d)$. Again, arguments analogous to those of the proof of proposition 7.14 imply $||\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^1||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \leq C$ and $|< \widetilde{\boldsymbol{\mathcal{X}}}_i^1 >| \leq C$, from where, up to the extraction of a subsequence, we obtain as above the existence of a matrix valued tensor c^1 such that

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^1 > \cdot \boldsymbol{e}_j \to c^1_{ij} \text{ and } \widetilde{\boldsymbol{\mathcal{X}}}_i^1 \rightharpoonup c^1_{ij}\boldsymbol{\Psi}_j \text{ weakly in } H^1_{loc}(\eta^{-1}P\backslash T,\mathbb{R}^d) \text{ as } \eta \to 0.$$

Passing (7.4.49) to the limit with $\boldsymbol{\Phi} = \boldsymbol{e}_i$ implies in this case

$$c^1_{ip}\Psi^*_{pj} = 0,$$

from where we infer $c^1 = 0$ and the convergence results for $\widetilde{\boldsymbol{\mathcal{X}}}_i^1$.

3. *General case.* Assuming the results hold till rank k, it holds

$$\mathcal{A}_{yy}\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} = 2\eta^{d-1}\partial_l\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1}\otimes e_l + \eta^d A[\widetilde{\boldsymbol{\mathcal{X}}}^{2k}\boldsymbol{e}_l^T]\cdot\boldsymbol{e}_m\otimes e_l\otimes e_m, \tag{7.4.50}$$

$$\mathcal{A}_{yy}\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3} = 2\eta\partial_l\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}\otimes e_l + \eta^d A[\widetilde{\boldsymbol{\mathcal{X}}}^{2k+1}\boldsymbol{e}_l^T]\cdot\boldsymbol{e}_m\otimes e_l\otimes e_m. \tag{7.4.51}$$

Writing their associated variational formulations as above and adapting the arguments of the proof of proposition 7.14, we obtain $||\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \leq C$ and $||\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \leq C$. Repeating the above arguments, we obtain, up to the extraction of a subsequence, the existence of matrix valued tensors c^{2k+1} and c^{2k+3} such that

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} > \cdot \boldsymbol{e}_j \to c^{2k+2}_{ij} \text{ and } \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} \rightharpoonup c^{2k+2}_{ij}\boldsymbol{\Psi}_j \text{ weakly in } H^1_{loc}(\eta^{-1}P\backslash T,\mathbb{R}^d) \text{ as } \eta \to 0.$$

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3} > \cdot \boldsymbol{e}_j \to c^{2k+3}_{ij} \text{ and } \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3} \rightharpoonup c^{2k+3}_{ij}\boldsymbol{\Psi}_j \text{ weakly in } H^1_{loc}(\eta^{-1}P\backslash T,\mathbb{R}^d) \text{ as } \eta \to 0.$$

The last step consists in passing the variational formulations for $\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}$ and $\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3}$ to the limit with the test function $\boldsymbol{\Phi} = \boldsymbol{e}_j$ in order to identify c^{2k+2}_{ij} and c^{2k+3}_{ij}. Performing this computation yields

$$c^{2k+2}_{ip}\Psi^*_{pj} = A[c^{2k}_{ip}\boldsymbol{e}_p\boldsymbol{e}_l^T] : [\boldsymbol{e}_j\boldsymbol{e}_m^T]\otimes e_l\otimes e_m,$$

$$c^{2k+3}_{ip}\Psi^*_{pj} = A[c^{2k+1}_{ip}\boldsymbol{e}_p\boldsymbol{e}_l^T] : [\boldsymbol{e}_j\boldsymbol{e}_m^T]\otimes e_l\otimes e_m.$$

Rewriting A into the matrix valued tensor A^2 following the notation (7.4.40), the above two equations read $c^{2k+2}\Psi^* = c^{2k}\otimes A^2$ and $c^{2k+3}\Psi^* = c^{2k+1}\otimes A^2$. The result follows by using (7.4.46) and (7.4.47) at rank k.

□

We are finally able to obtain the full asymptotics for the coefficients of the homogenized equation (7.4.29) when the size η of the hole converges to zero:

Corollary 7.6. *Assume $d \geq 3$. The following convergence hold for the matrix valued tensors M^k as $\eta \to 0$:*

$$M^0 \sim \eta^{d-2}\Psi^*, \tag{7.4.52}$$

$$M^1 = o\left(\eta^{d-2}\right), \tag{7.4.53}$$

$$M^2 \to -A^2, \tag{7.4.54}$$

$$\forall k > 1,\ M^{2k} = o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right), \tag{7.4.55}$$

$$\forall k > 1,\ M^{2k+1} = o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right). \tag{7.4.56}$$

Proof. We reproduce the proof of corollary 7.2 with small adaptations due to the vectorial context. (7.4.52) is immediate with $M^0 = (\mathcal{X}^{0*})^{-1}$. As for M^2, we write according to (7.4.18):

$$\begin{aligned} M^2 &= -(\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{2*} \otimes (\mathcal{X}^{0*})^{-1} + (\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{1*} \otimes (\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{1*} \otimes (\mathcal{X}^{0*})^{-1} \\ &= -\frac{\eta^{2(d-2)}}{\eta^{2(d-2)}}\Psi^* \otimes c^2 \otimes \Psi^* + o\left(\frac{\eta^{3(d-2)}}{\eta^{2(d-2)}}\right) \\ &= -\Psi^* \otimes (\Psi^*)^{-1} \otimes A^2 \otimes (\Psi^*)^{-1} \otimes \Psi^* + o(\eta^{d-2}) \\ &= -A^2 + o(\eta^{d-2}). \end{aligned}$$

The identity (7.4.56) for M^{2k+1} is obtained by using (7.4.18), where we observe that, for any $0 \leq p \leq 2k+1$ and indices $1 \leq i_1 \dots i_p \leq 2k+1$ such that $i_1 + \dots + i_p = 2k+1$, there exists at least one odd index i_q with $1 \leq q \leq p$. Using (7.4.46), we obtain therefore

$$(\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_1*} \otimes \dots \otimes \mathcal{X}^{i_p*} \otimes (\mathcal{X}^{0*})^{-1} = o\left(\frac{\eta^{(p+1)(d-2)}}{\eta^{(p+\lfloor i_1/2\rfloor+\dots+\lfloor i_p/2\rfloor)(d-2)}}\right) = o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right),$$

which eventually implies (7.4.56). Finally, in order to obtain the result for M^{2k} with $k > 1$, we separate the summands of (7.4.18) into two categories. For a given p such that $1 \leq p \leq 2k$ and indices $1 \leq i_1, \dots, i_p \leq d$ such that $i_1 + \dots + i_p = 2k$, there are only two possibilities:

1. there exists at least one odd index i_q, in that case the above reasoning implies as well
$$(\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_1*} \otimes \dots \otimes \mathcal{X}^{i_p*} \otimes (\mathcal{X}^{0*})^{-1} = o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right);$$
2. all indices $i_1 + \dots + i_p$ are even, in that case we may write, as $\eta \to 0$:
$$\begin{aligned} (\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_1*} \otimes \dots \otimes \mathcal{X}^{i_p*} \otimes (\mathcal{X}^{0*})^{-1} &\sim \frac{1}{\eta^{(d-2)(k-1)}}\Psi^* \otimes c^{i_1} \otimes \dots \Psi^* \otimes c^{i_p} \otimes \Psi^* \\ &\sim \frac{1}{\eta^{(d-2)(k-1)}} \overbrace{A^2 \otimes (\Psi^*)^{-1} \otimes \dots \otimes A^2 \otimes (\Psi^*)^{-1}}^{k-1 \text{ times}} \otimes A^2. \end{aligned}$$

In both case, the asymptotics obtained for $(\mathcal{X}^{0*})^{-1} \otimes \mathcal{X}^{i_1*} \otimes \dots \otimes \mathcal{X}^{i_p*} \otimes (\mathcal{X}^{0*})^{-1}$ do not depend on the choice of indices $1 \leq i_1, \dots, i_p \leq d$. This allows to eventually obtain, by isolating even indices $i_1 := 2j_1, \dots, i_p := 2j_p$ in the formula (7.4.18):

$$\begin{aligned} M^{2k} =& \frac{1}{\eta^{(d-2)(k-1)}} \overbrace{A^2 \otimes (\Psi^*)^{-1} \otimes \dots \otimes A^2 \otimes (\Psi^*)^{-1}}^{k-1 \text{ times}} \otimes A^2 \left(\sum_{p=1}^{2k} (-1)^p \sum_{\substack{2j_1+\dots+2j_p=2k \\ 1 \leq j_1,\dots,j_p \leq k}} 1\right) \\ &+ o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right). \end{aligned}$$

The result follows by using the identity (7.3.86). □

Remark 7.24. We have retrieved formally in the previous corollary the classical three regimes of homogenized equations [11, 105] depending on the hole size $a_\varepsilon = \eta\varepsilon$:

1. if $1 \geq a_\varepsilon >> \varepsilon^{d/(d-2)}$, that is η goes to zero at a smaller rate than $\varepsilon^{2/(d-2)}$, then the zeroth order term $\varepsilon^{-2}M^0 \sim \varepsilon^{-2}\eta^{d-2}\Psi^*$ is dominant in the infinite order homogenized equation (7.4.29);

2. if $a_\varepsilon = \varepsilon^{d/(d-2)}$, the coefficients of the infinite order homogenized equation (7.4.29) converge to those of
$$-\mathrm{div}(A\nabla \boldsymbol{u}^*) + \Psi^*\boldsymbol{u}^* = \boldsymbol{f},$$
which corresponds to a "Brinkman" regime with *strange* reaction term $\Psi^*\boldsymbol{u}^*$. Interestingly, the *very strange* first order term $\varepsilon^{-1}M^1 \cdot \nabla$ disappears because $M^1\varepsilon = o(\varepsilon^1)$, although this term dominates the second order term $M^2 \cdot \nabla$ for a fixed size $\eta > 0$. All other contributions $\varepsilon^k M^k \cdot \nabla^k$ for $k > 2$ vanish equally, at rates $o(1)$ and $o(\varepsilon)$ for respectively even and odd values of k;

3. if $a_\varepsilon = o(\varepsilon^{d/(d-2)})$, then the hole are "too small" to be seen by the homogenized model and the coefficients of (7.4.29) converge to those of the elasticity problem (7.4.1) in the homogeneous domain D (without holes):
$$-\mathrm{div}(A\nabla \boldsymbol{u}^*) = \boldsymbol{f}.$$

7.4.4 Simplifications for the tensors $\mathcal{X}^{k*}$ and M^k in case of symmetries

This last section generalizes the results of section 7.3.6 to the elasticity context: our main result is corollary 7.7 where we obtain symmetry properties for the tensors $\mathcal{X}^{k*}$ and M^k in case where the obstacle (ηT) is symmetric with respect to the cell axes.

Throughout this part, it is assumed that the elasticity tensor A commute with orthogonal symmetries $S := (S_{ij})_{1\leq i,j\leq d}$ in the following sense:

$$A(SMS) = S(AM)S. \tag{7.4.57}$$

This property is notably satisfied by the Hooke's law (7.4.2). We use below the following technical lemma:

Lemma 7.8. *Under the assumption (7.4.57), the following identities hold for any smooth vector field $\boldsymbol{\mathcal{X}}$:*

$$\mathcal{A}_{yy}(S\boldsymbol{\mathcal{X}} \circ S) = S(\mathcal{A}_{yy}\boldsymbol{\mathcal{X}}) \circ S, \tag{7.4.58}$$
$$\partial_i^A(S\boldsymbol{\mathcal{X}} \circ S) = S_{ij}S(\partial_j^A\boldsymbol{\mathcal{X}}) \circ S. \tag{7.4.59}$$

Proof. We write

$$\begin{aligned}\mathcal{A}_{yy}(S\boldsymbol{\mathcal{X}} \circ S) &= -\mathrm{div}(A\nabla(S\boldsymbol{\mathcal{X}} \circ S)) = -\mathrm{div}(A[S(\nabla\boldsymbol{\mathcal{X}}) \circ SS])\\ &= -S\mathrm{div}(A(\nabla\boldsymbol{\mathcal{X}}) \circ S)S = -SA[(\nabla\partial_i\boldsymbol{\mathcal{X}}) \circ S] \cdot S\boldsymbol{e}_j S_{ij}\\ &= -SA[(\nabla\partial_i\boldsymbol{\mathcal{X}}) \circ S]\boldsymbol{e}_i = -S\mathrm{div}(A\nabla\boldsymbol{\mathcal{X}}) \circ S,\end{aligned}$$

which proves (7.4.58). Equation (7.4.59) follows from

$$\begin{aligned}\partial_i^A(S\boldsymbol{\mathcal{X}} \circ S) &= \frac{1}{2}(A\nabla(S\boldsymbol{\mathcal{X}} \circ S) \cdot \boldsymbol{e}_i + \mathrm{div}(A[S\boldsymbol{\mathcal{X}} \circ S\boldsymbol{e}_i^T]))\\ &= \frac{1}{2}(A[S(\nabla\boldsymbol{\mathcal{X}}) \circ SS] \cdot \boldsymbol{e}_i + \mathrm{div}(SA[\boldsymbol{\mathcal{X}} \circ S(S\boldsymbol{e}_i)^T]S))\\ &= \frac{1}{2}(SA(\nabla\boldsymbol{\mathcal{X}}) \circ S \cdot S\boldsymbol{e}_i + SA[S_{lj}(\partial_l\boldsymbol{\mathcal{X}}) \circ S(S\boldsymbol{e}_i)^T]S \cdot \boldsymbol{e}_j)\\ &= \frac{1}{2}(SA(\nabla\boldsymbol{\mathcal{X}}) \circ S \cdot S\boldsymbol{e}_i + SA[(\partial_l\boldsymbol{\mathcal{X}}) \circ S(S\boldsymbol{e}_i)^T]SS\boldsymbol{e}_l)\\ &= \frac{S}{2}(A(\nabla\boldsymbol{\mathcal{X}}) \circ S \cdot (S_{ij}\boldsymbol{e}_j) + \mathrm{div}(A[\boldsymbol{\mathcal{X}}(S_{ij}\boldsymbol{e}_j)^T]) \circ S).\end{aligned}$$

□

Proposition 7.28. *If the cell $Y = P\backslash\eta T$ is invariant with respect to a symmetry S,* i.e. *$S(Y) = Y$, then the following identity holds for the vector valued tensors $(\boldsymbol{\mathcal{X}}^k_l)_{1\le l\le d}$ of eqn. (7.4.7):*

$$S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l}\circ S = S_{i_1j_1}\dots S_{i_kj_k}S_{lm}\boldsymbol{\mathcal{X}}^k_{j_1\dots j_k,m}. \tag{7.4.60}$$

As a consequence, the following identities hold for the constant matrix valued tensors $\mathcal{X}^{k*}$ and M^k:

$$\mathcal{X}^{k*}_{i_1\dots i_k,lm} = S_{i_1j_1}\dots S_{i_kj_k}S_{lp}S_{mq}\mathcal{X}^{k*}_{j_1\dots j_k,pq} \tag{7.4.61}$$

$$M^k_{i_1\dots i_k,pq} = S_{i_1j_1}\dots S_{i_kj_k}S_{lp}S_{mq}M^k_{j_1\dots j_k,pq}. \tag{7.4.62}$$

Proof. We start by proving (7.4.60) by induction. It holds $\mathcal{A}_{yy}(S\boldsymbol{\mathcal{X}}^0_l\circ S) = S\boldsymbol{e}_l\circ S = S_{ml}\boldsymbol{e}_m$. If the cell is S symmetric, then $S\boldsymbol{\mathcal{X}}^0_l\circ S$ also satisfies the boundary conditions of (7.4.7), we obtain therefore

$$S\boldsymbol{\mathcal{X}}^0_l\circ S = S_{lm}\boldsymbol{\mathcal{X}}^0_m,$$

which is the result at rank $k=0$. For $k=1$, we write

$$\mathcal{A}_{yy}(S\boldsymbol{\mathcal{X}}^1_{i_1,l}\circ S) = 2S(\partial^A_{i_1}\boldsymbol{\mathcal{X}}^0_l)\circ S = 2S_{i_1j_1}\partial^A_{j_1}(S\boldsymbol{\mathcal{X}}^0_l\circ S) = 2S_{i_1j_1}\partial^A_{j_1}(S_{lm}\boldsymbol{\mathcal{X}}^0_m),$$

which implies $S\boldsymbol{\mathcal{X}}^1_{i_1,l}\circ S = S_{i_1j_1}S_{lm}\boldsymbol{\mathcal{X}}^1_{j_1,m}$ and the result at rank $k=1$. Assuming now that the result holds till rank $k+1$ with $k\ge 0$, we write

$$\mathcal{A}_{yy}(S\boldsymbol{\mathcal{X}}^{k+2}_{i_1\dots i_{k+2},l}\circ S) = 2S(\partial^A_{i_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l})\circ S + SA[\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l}\circ S\boldsymbol{e}^T_{k+1}]\cdot\boldsymbol{e}_{k+2}.$$

The previous arguments and the result at rank $k+1$ imply the following equality for the first term:

$$2S(\partial^A_{i_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l})\circ S = S_{i_1j_1}\dots S_{i_{k+2}j_{k+2}}S_{lm}\partial^A_{j_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m}.$$

Furthermore, using the result at rank k, we can rewrite the second term as follows:

$$\begin{aligned}SA[\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l}\circ S\boldsymbol{e}^T_{i_{k+1}}]\cdot\boldsymbol{e}_{i_{k+2}} &= A[S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l}\circ S\boldsymbol{e}^T_{i_{k+1}}S]\cdot S\boldsymbol{e}_{i_{k+2}}\\ &= A[S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l}\circ S(S\boldsymbol{e}_{i_{k+1}})^T]\cdot S\boldsymbol{e}_{i_{k+2}}\\ &= S_{i_1j_1}\dots S_{i_{k+2}j_{k+2}}S_{lm}A[\boldsymbol{\mathcal{X}}^k_{j_1\dots j_k,m}\boldsymbol{e}_{j_{k+1}}]\cdot\boldsymbol{e}_{j_{k+2}}.\end{aligned}$$

The two above equalities imply the result at rank $k+2$.

Eqn. (7.4.61) then follows by the following change of variables:

$$\mathcal{X}^{k*}_{i_1\dots i_k,lm} = \int_Y \boldsymbol{e}_l\cdot\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,m}\mathrm{d}y = \int_Y (S\boldsymbol{e}_l)\cdot(S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,m}\circ S)\mathrm{d}y.$$

Remarking that (7.4.61) rewrites also as

$$\mathcal{X}^{k*}_{i_1\dots i_k} = S_{i_1j_1}\dots S_{i_kj_k}(S\mathcal{X}^{k*}_{j_1\dots j_k}S),$$

eqn. (7.4.62) follows by rewriting the summand of (7.4.18) as

$$\begin{aligned}&(\mathcal{X}^{0*})^{-1}\otimes\mathcal{X}^{i_1*}\otimes\dots\otimes(\mathcal{X}^{0*})^{-1}\otimes\mathcal{X}^{i_p*}\otimes(\mathcal{X}^{0*})^{-1}\\ &\qquad = (\mathcal{X}^{0*})^{-1}S\otimes S\mathcal{X}^{i_1*}S\otimes\dots\otimes S(\mathcal{X}^{0*})^{-1}S\otimes S\mathcal{X}^{i_p*}S\otimes S(\mathcal{X}^{0*})^{-1}\end{aligned} \tag{7.4.63}$$

before using identity (7.4.61). □

Following the proof of corollary 7.4, we obtain simplifications for the tensors $\mathcal{X}^{2k*}$ and M^{2k} as well as the vanishing of the tensors $\mathcal{X}^{2k+1*}$ and M^{2k+1} in case the obstacle ηT is symmetric with respect to the cell axes. Recall (7.3.87) and (7.3.88) for the definition of S^l and $S^{l,m}$.

Corollary 7.7. *1. If the cell Y is symmetric with respect to all cell axes $\boldsymbol{e}_l$, i.e. $S^l(Y) = Y$ for any $1\le l\le d$, then*

$$\mathcal{X}^{k*}_{i_1\dots i_k,pq} = 0 \text{ and } M^k_{i_1\dots i_k,pq} = 0$$

whenever a given integer $1\le r\le d$ occurs with an odd multiplicity in the indices $i_1\dots i_k,pq$.

In particular, this implies $\mathcal{X}^{2k+1*} = 0$ and $M^{2k+1} = 0$.

2. *If the cell Y is symmetric with respect to all diagonal axes orthogonal to $(\boldsymbol{e}_l - \boldsymbol{e}_m)$, i.e. $S^{l,m}(Y) = Y$ for any $1 \le l < m \le d$, then for any permutation $\sigma \in \mathfrak{S}_d$,*

$$\mathcal{X}^{k*}_{\sigma(i_1)\dots\sigma(i_k),\sigma(p)\sigma(q)} = \mathcal{X}^{k*}_{i_1\dots i_k,pq}.$$

$$M^{k}_{\sigma(i_1)\dots\sigma(i_k),\sigma(p)\sigma(q)} = M^{k}_{i_1\dots i_k,pq}.$$

Let us illustrate how the previous properties translate for the tensors M^2 and M^4:

- if the cell Y is symmetric with respect to all cell axes $(\boldsymbol{e}_l)_{1\le l\le d}$, only the coefficients of the form

$$M^2_{ii,jj}, M^2_{ij,ij}, M^2_{ii,ii}$$

with $i \neq j$ are non zero. For M^4, only the coefficients of the form

$$M^4_{iijj,kk}, M^4_{iijk,jk}, M^4_{iiii,jj}, M^4_{iijj,ii}, M^4_{iiij,ij}, M^4_{iiii,ii}$$

are non zero with distinct integers i, j, k.

- If in addition the obstacle is symmetric with respect to all diagonal axes, then the values of the above coefficients do not depend on the choice of the distinct integers i, j, k. As a result, M^2 reduces to at most three coefficients (the material is said to be orthotropic), and M^4 reduces to at most 6 coefficients for $d \ge 3$, and to 4 coefficients for $d = 2$.

7.4.5 Appendix: numerical evidences for the "very strange" tensors $\mathcal{X}^{1*}$ and M^1 being nonzero

In this appendix, we report on some numerical computations performed in 2-d in order to assess the magnitude of the tensors $\mathcal{X}^1$ and M^1 for a particular (non symmetric) shape of hole ηT. We consider a cell $P = [0,1] \times [0,1]$ perforated with a "boomerang" shaped hole (see Figure 7.4), parameterized by the following system of equations

$$\begin{cases} x(t) = 0.5 + r(\cos(t) + 2\cos(2t)) \\ y(t) = 0.5 + 2r\sin(t), \end{cases} \tag{7.4.64}$$

where $t \in [0, 2\pi]$ and $r = 0.15$. The cell $Y = P \backslash (\eta T)$ is discretized into a triangular mesh (represented on Figure 7.4a) whose maximum edge size was `hmax`=0.007. The elasticity tensor is given by the Hooke's law

$$A\nabla \boldsymbol{u} = \mu(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T) + \lambda \mathrm{div}(\boldsymbol{u}),$$

where the Lamé parameters are $\mu = 0.1$ and $\lambda = 1$.
By using the finite element method with $\mathbb{P}1$ finite elements, we solve the first cell problem of (7.4.7) determining $\boldsymbol{\mathcal{X}}^0_j$ for $j = 1, 2$:

$$\begin{cases} \mathcal{A}_{yy}\boldsymbol{\mathcal{X}}^0_j = \boldsymbol{e}_j, \\ \quad \boldsymbol{\mathcal{X}}^0_j \text{ is } P\text{–periodic}, \end{cases} j = 1, 2.$$

The numerical solutions obtained for $\boldsymbol{\mathcal{X}}^0_j$ with $j = 1, 2$ are represented on Figs. 7.4 and 7.5 below. They are used to compute the constant tensors $\mathcal{X}^{0*}$, $\mathcal{X}^{1*}$, M^0 and M^1 from the formulas (7.4.13) and (7.4.18):

$$\mathcal{X}^{0*}_{ij} = \int_Y \boldsymbol{\mathcal{X}}^0_j \cdot \boldsymbol{e}_i \mathrm{d}y, \tag{7.4.65}$$

$$\mathcal{X}^{1*}_{l,ij} = \int_Y (\boldsymbol{\mathcal{X}}^0_i \cdot A\nabla \boldsymbol{\mathcal{X}}^0_j - \boldsymbol{\mathcal{X}}^0_j \cdot A\nabla \boldsymbol{\mathcal{X}}^0_i) \cdot \boldsymbol{e}_l \mathrm{d}y, \tag{7.4.66}$$

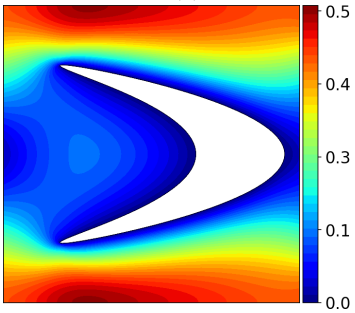
$$M^0 = (\mathcal{X}^{0*})^{-1}, \tag{7.4.67}$$

$$M^1 = -(\mathcal{X}^{0*})^{-1}\mathcal{X}^{1*}(\mathcal{X}^{0*})^{-1}. \tag{7.4.68}$$

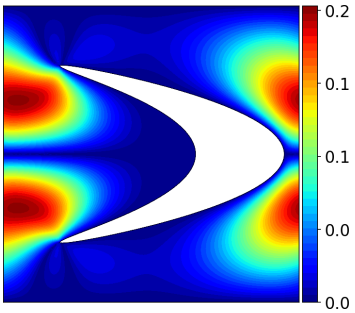
These computations are performed in `FreeFEM` [183]; we obtain the following numerical values:

$$\mathcal{X}^{0*} = \begin{pmatrix} \texttt{0.145} & \texttt{-4.30e-7} \\ \texttt{-4.30e-7} & \texttt{0.0765} \end{pmatrix} \tag{7.4.69}$$

(a) Mesh considered for the unit cell Y.



(b) Norm field $||\boldsymbol{\mathcal{X}}_1^0||$.



(c) Norm field $||\boldsymbol{\mathcal{X}}_2^0||$.

Figure 7.4: The unit cell considered for the numerical example of section 7.4.5 and the first two vector valued tensors $\boldsymbol{\mathcal{X}}_i^0$ for $i = 1, 2$.

(a) Deformation of Y by $\boldsymbol{\mathcal{X}}_1^0$.

(b) Deformation of Y by $\boldsymbol{\mathcal{X}}_2^0$.

Figure 7.5: Deformation of the unit cell Y according to the first two vector valued tensors $\boldsymbol{\mathcal{X}}_i^0$ for $i = 1, 2$.

$$\mathcal{X}_1^{1*} = \begin{pmatrix} 1.05e\text{-}21 & 1.52e\text{-}6 \\ -1.52e\text{-}6 & -2.98e\text{-}22 \end{pmatrix}, \ \mathcal{X}_2^{1*} = \begin{pmatrix} -7.16e\text{-}22 & 0.00237 \\ -0.00237 & -1.670e\text{-}21 \end{pmatrix}, \tag{7.4.70}$$

$$M^0 = \begin{pmatrix} 6.90 & 3.88e\text{-}05 \\ 3.88e\text{-}05 & 13.1 \end{pmatrix}, \tag{7.4.71}$$

$$M_1^1 = \begin{pmatrix} 1.074e\text{-}18 & 1.37e\text{-}4 \\ -1.37e\text{-}4 & 1.93e\text{-}18 \end{pmatrix}, \ M_2^1 = \begin{pmatrix} 1.67e\text{-}15 & 0.214 \\ -0.214 & 3.15e\text{-}15 \end{pmatrix}, \tag{7.4.72}$$

where we have denoted by $\mathcal{X}_i^{1*}$ and M_i^1 the i-th (matrix valued) component of these tensors of order 1. These results strongly suggest that the tensor M^1 is not zero (with e.g. the value $M^1_{1,12} = 0.214$). As a result, we expect the odd order differential operator $\varepsilon^{-1} M^1 \cdot \nabla$ to have an impact when solving the high order homogenized equation (7.4.17).

7.5 High order homogenization for the Stokes system in a porous medium

This last section is devoted to the high order homogenization of the Stokes system, which was our initial motivation:

$$\begin{cases} -\Delta \boldsymbol{u}_\varepsilon + \nabla p_\varepsilon = \boldsymbol{f} \text{ in } D_\varepsilon \\ \quad \operatorname{div}(\boldsymbol{u}_\varepsilon) = 0 \text{ in } D_\varepsilon \\ \quad \displaystyle\int_{D_\varepsilon} p_\varepsilon \mathrm{d}x = 0, \\ \quad \boldsymbol{u}_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon \\ \quad \boldsymbol{u}_\varepsilon = \text{ is } D\text{–periodic.} \end{cases} \tag{7.5.1}$$

In the sequel, we consider the following two classical assumptions for the distributions of the holes ω_ε, following [12]:

(H4) $Y \subset P$, as a subset of the unit torus (opposite matching faces are identified) is a smooth connected set.

(H5) The fluid component D_ε is a smooth connected set.

Remark 7.25. Assumption (H5) does not necessarily imply (H4), see [7] for a counterexample. Assumption (H4) is not very restrictive and can easily be generalized to the case where the subset Y has m connected components with $m \in \mathbb{N}$ (see section 7.5.6 for a more precise discussion). Assumption (H5) is stronger, but is also more physical. It forbids the existence of isolated fluid inclusions. Most of our derivations of homogenized tensors and homogenized equations only assume (H4). However, we rely on both assumption (H4) and (H5) when stating error bounds in section 7.5.3, because we use some technical results of [12] (namely Theorem 2.3 of this reference).

We start as previously by considering formal two-scale ansatz $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ of the form

$$\boldsymbol{u}_\varepsilon = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \boldsymbol{u}_i(x, x/\varepsilon), \quad p_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i (p_i^*(x) + \varepsilon p_i(x, x/\varepsilon)), \quad x \in D_\varepsilon. \tag{7.5.2}$$

where the functions $\boldsymbol{u}_i(x,y)$ and $p_i(x,y)$ are periodic in the variable $y \in P$. The decomposition (7.5.2) for the pressure p_ε is not the one commonly assumed in the literature, where it is more usually written only in terms of $p_i(x, x/\varepsilon)$. Of course, there is no loss of generality in adding the non oscillating function p_i^*, which turns to be very convenient in the sequel. In order to fix the value of p_i^* and $p_i(x,y)$, p_i^* is required to have zero average with respect to the variable x, and the oscillating functions p_i are required to have zero average with respect to the variable y:

$$\int_D p_i^*(x) \mathrm{d}x = 0, \qquad \int_Y p_i(x,y) \mathrm{d}y = 0, \quad \forall i \geq 0.$$

Assuming these conventions, our goal is to derive and analyze high order equations for the formal homogenized averages $\boldsymbol{u}_\varepsilon^*$ and p_ε^* defined by

$$\boldsymbol{u}_\varepsilon^*(x) := \sum_{i=0}^{+\infty} \varepsilon^{i+2} \int_Y \boldsymbol{u}_i(x,y)\mathrm{d}y, \quad p_\varepsilon^*(x) := \sum_{i=0}^{+\infty} \varepsilon^i p_i^*(x). \tag{7.5.3}$$

In the following, the methodologies of the previous sections 7.3 and 7.4 are applied to derive higher order homogenized equations for the perforated Stokes system (7.5.1). Most of the ingredients remain analogous to the case of the elasticity system, however additional technical difficulties occur in the proof of error estimates due to the incompressibility constraint $\mathrm{div}(\boldsymbol{u}_\varepsilon) = 0$.

This part is organized as follows. In section 7.5.1, we introduce cell problems and their solution tensors $(\mathcal{X}^k, \boldsymbol{\alpha}^k)$ which allow to identify the functions $\boldsymbol{u}_i$, p_i^* and p_i of the ansatz (7.5.2). Similarly to what occurs in the context of the elasticity system investigated in section 7.4, we prove that the averaged tensors $\mathcal{X}^{k*}$ are symmetric matrix valued matrices for even values of k, and antisymmetric valued matrices otherwise. In section 7.5.2, we show that the formal averages $\boldsymbol{u}_\varepsilon^*$ and p_ε^* are the solution of a formal, "infinite order" homogenized equation involving tensors M^k. Introducing new tensors N^k and β^k, we obtain "criminal" ansatz expressing $\boldsymbol{u}_\varepsilon$ and p_ε in terms of the derivatives of $\boldsymbol{u}_\varepsilon^*$ and p_ε^*. These tensors are used to build well-posed homogenized equations (eqn. (7.5.44)) of finite order, by using a minimization principle.

Section 7.5.3 is concerned with the error analysis of the homogenized approximations of $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ generated by our procedure: our main result is stated in proposition 7.39 where we show that the solution $(\boldsymbol{v}_K^*, q_K^*)$ of the homogenized equation of order $2K+2$ yield approximations of $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ in the $L^2(D_\varepsilon)$ norm of orders $K+3$ and $K+1$ for the velocity and the pressure respectively. In section 7.5.4, we investigate low volume fraction limits where the obstacle size η converges to zero: we prove the "coefficient-wise" convergence of the infinite order homogenized equation towards either of the three classical homogenized regimes depending on how η scales with respect to the critical size $\varepsilon^{d/(d-2)}$. We then establish in section 7.5.5 simplification properties for the tensors $\mathcal{X}^{k*}$ and M^k in case the obstacle ηT is symmetric with respect to the axes of the unit cell Y, which includes the vanishing of odd order tensors $\mathcal{X}^{2k+1*}$ and M^{2k+1} in this case. Finally, section 7.5.6 briefly discusses the extension of our results to domains featuring multiple connected components in view of future applications to heat exchangers.

Remark 7.26. Following section 7.2.4, remember that the integers $p, q \in \mathbb{N}$ may be used as a subscript or superscript indices when writing tensors in a limited number of places. Naturally, they should not to be confused with the pressure variables denoted by p_ε, p_i or q_K^*.

7.5.1 Formal infinite order two scale expansions: tensors $(\mathcal{X}^k, \boldsymbol{\alpha}^k)$

As in the previous sections, we still assume (for the derivation of two-scale asymptotic expansions only) that the right-hand side $\boldsymbol{f}$ can be formally decomposed into a power series in ε:

$$\forall x \in D,\ \boldsymbol{f}(x) = \sum_{i=0}^{+\infty} \varepsilon^i \boldsymbol{f}^i(x). \tag{7.5.4}$$

Inserting (7.5.2) into the Stokes system (7.5.1) yields the following cascade of equations

$$\begin{cases} -\Delta_{yy}\boldsymbol{u}_{i+2} + \nabla_y p_{i+2} = \boldsymbol{f}_{i+2} - \nabla_x p_{i+2}^* - \nabla_x p_{i+1} + \Delta_{xy}\boldsymbol{u}_{i+1} + \Delta_{xx}\boldsymbol{u}_i, & \forall i \geq -2, \\ \mathrm{div}_y(\boldsymbol{u}_{i+2}) = -\mathrm{div}_x(\boldsymbol{u}_{i+1}), & \forall i \geq -2, \\ \boldsymbol{u}_{-2} = \boldsymbol{u}_{-1} = 0, \\ \displaystyle\int_D p_i^* \mathrm{d}x = 0, & \forall i \geq 0, \\ p_{-1} = 0, \\ \boldsymbol{u}_i(x,\cdot) = 0 \text{ on } \partial(\eta T) \\ \boldsymbol{u}_i(x,\cdot) \text{ is } P\text{–periodic for any } x \in D,\ \forall i \geq 0, \\ \boldsymbol{u}_i(\cdot,y) \text{ is } D\text{–periodic for any } y \in P,\ \forall i \geq 0, \end{cases} \tag{7.5.5}$$

where the operators $-\Delta_{yy}$, $-\Delta_{xy}$, $-\Delta_{yy}$ are defined by

$$-\Delta_{xx} = -\mathrm{div}_x(\nabla_x \cdot), \quad -\Delta_{xy} = -\mathrm{div}_x(\nabla_y \cdot) - \mathrm{div}_y(\nabla_x \cdot), \quad -\Delta_{yy} := -\mathrm{div}_y(\nabla_y \cdot).$$

We introduce a family of respectively vector valued tensors $(\boldsymbol{\mathcal{X}}_j^k(y))_{1\leq j\leq d}$ and scalar valued tensors $(\alpha_j^k(y))_{1\leq j\leq d}$ defined by induction as the unique solutions in $H^1(Y)\times L^2(Y)/\mathbb{R}$ to the following cell problems:

$$\begin{cases} -\Delta_{yy}\boldsymbol{\mathcal{X}}_j^0 + \nabla_y \alpha_j^0 = \boldsymbol{e}_j \text{ in } Y, \\ \qquad \operatorname{div}_y(\boldsymbol{\mathcal{X}}_j^0) = 0 \text{ in } Y \end{cases} \tag{7.5.6}$$

$$\begin{cases} -\Delta_{yy}\boldsymbol{\mathcal{X}}_j^1 + \nabla_y \alpha_j^1 = (2\partial_l \boldsymbol{\mathcal{X}}_j^0 - \alpha_j^0 \boldsymbol{e}_l)\otimes e_l \text{ in } Y \\ \qquad \operatorname{div}_y(\boldsymbol{\mathcal{X}}_j^1) = -(\boldsymbol{\mathcal{X}}_j^0 - < \boldsymbol{\mathcal{X}}_j^0 >)\cdot \boldsymbol{e}_l \otimes e_l \text{ in } Y, \end{cases} \tag{7.5.7}$$

$$\begin{cases} -\Delta_{yy}\boldsymbol{\mathcal{X}}_j^{k+2} + \nabla_y \alpha_j^{k+2} = (2\partial_l \boldsymbol{\mathcal{X}}_j^{k+1} - \alpha_j^{k+1} \boldsymbol{e}_l)\otimes e_l + \boldsymbol{\mathcal{X}}_j^k \otimes I \text{ in } Y \\ \qquad \operatorname{div}_y(\boldsymbol{\mathcal{X}}_j^{k+2}) = -(\boldsymbol{\mathcal{X}}_j^{k+1} - < \boldsymbol{\mathcal{X}}_j^{k+1} >)\cdot \boldsymbol{e}_l \otimes e_l \text{ in } Y, \end{cases} \qquad \forall k \geq 0 \tag{7.5.8}$$

where $< \boldsymbol{\mathcal{X}}_j^k >$ denotes the cell average of the vector field $\boldsymbol{\mathcal{X}}_j^k$:

$$< \boldsymbol{\mathcal{X}}_j^k > := \int_Y \boldsymbol{\mathcal{X}}_j^k(y)\mathrm{d}y.$$

Equations (7.5.6) to (7.5.8) are supplemented with the following boundary conditions:

$$\begin{cases} \displaystyle\int_Y \alpha_j^k \mathrm{d}y = 0 \\ \qquad \boldsymbol{\mathcal{X}}_j^k = 0 \text{ on } \partial(\eta T) \\ \qquad \boldsymbol{\mathcal{X}}_j^k \text{ is } P\text{–periodic}, \end{cases} \qquad \forall k \geq 0. \tag{7.5.9}$$

Similarly as before, we introduce the k-th order matrix valued tensors $\mathcal{X}^k$ whose columns are the vector valued tensors $(\boldsymbol{\mathcal{X}}_j^k)$:

$$(\mathcal{X}_{ij}^k(y))_{1\leq i,j\leq d} := \left[\boldsymbol{\mathcal{X}}_1^k(y) \quad \dots \quad \boldsymbol{\mathcal{X}}_d^k(y)\right], \forall y \in Y, \quad \forall k \geq 0.$$

We also denote by $\boldsymbol{\alpha}^k$ the k-th order vector valued tensor whose coordinates are the scalar tensors α_j^k:

$$\boldsymbol{\alpha}^k(y) := (\alpha_j^k(y))_{1\leq j\leq d}, \forall y \in Y, \quad \forall k \geq 0.$$

From the definition (7.5.9), the tensors α_j^k are of zero average. Following the previous sections, we use a star notation to denote the averages of the tensor $\mathcal{X}^k$ and of the functions $\boldsymbol{u}_i$:

$$\mathcal{X}^{k*} := \int_Y \mathcal{X}^k(y)\mathrm{d}y, \qquad \forall k \geq 0, \tag{7.5.10}$$

$$\boldsymbol{u}_i^*(x) := \int_Y \boldsymbol{u}_i(x,y)\mathrm{d}y, \qquad \forall x \in D, \forall i \geq 0. \tag{7.5.11}$$

The tensors $\mathcal{X}^k$ and $\boldsymbol{\alpha}^k$ allow to solve the cascade of equations (7.5.5):

Proposition 7.29. *Assume (H4). The solutions $\boldsymbol{u}_i(x,y)$, $p_i(x,y)$ of the cascade of equations (7.5.5) are given by*

$$\boldsymbol{u}_i(x,y) = \sum_{k=0}^{i} \mathcal{X}^k(y)\cdot\nabla^k(\boldsymbol{f}_{i-k}(x) - \nabla p_{i-k}^*(x)), \quad p_i(x,y) = \sum_{k=0}^{i} \boldsymbol{\alpha}^k(y)\cdot\nabla^k(\boldsymbol{f}_{i-k}(x) - \nabla p_{i-k}^*(x)), \tag{7.5.12}$$

where the functions p_i^* are uniquely determined recursively as the solutions to the following elliptic system:

$$\begin{cases} -\operatorname{div}_x(\mathcal{X}^{0*}\nabla_x p_i^*) = \operatorname{div}_x(\mathcal{X}^{0*}\boldsymbol{f}_i) - \displaystyle\sum_{k=1}^{i} \operatorname{div}(\mathcal{X}^{k*}\cdot\nabla^k(\boldsymbol{f}_{i-k} - \nabla_x p_{i-k}^*)) \text{ in } D_\varepsilon, \\ \qquad \displaystyle\int_D p_i^* \mathrm{d}x = 0 \end{cases} \qquad \forall i \geq 0. \tag{7.5.13}$$

Since the average functions $\boldsymbol{u}_i^*$ (eqn. (7.5.11)) are given by

$$\boldsymbol{u}_i^*(x) = \sum_{k=0}^{i} \mathcal{X}^{k*} \cdot \nabla^k (\boldsymbol{f}_{i-k}(x) - \nabla p_{i-k}^*(x)),$$

recognizing Cauchy products allows to rewrite identities (7.5.12) and (7.5.13) in terms of formal equality of power series:

$$\boldsymbol{u}_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^i(x/\varepsilon) \cdot \nabla^i (\boldsymbol{f}(x) - \nabla p_\varepsilon^*(x)), \tag{7.5.14}$$

$$\boldsymbol{u}_\varepsilon^*(x) = \sum_{i=0}^{+\infty} \varepsilon^{i+2} \mathcal{X}^{i*} \cdot \nabla^i (\boldsymbol{f}(x) - \nabla p_\varepsilon^*(x)), \tag{7.5.15}$$

$$p_\varepsilon(x) = p_\varepsilon^*(x) + \sum_{i=0}^{+\infty} \varepsilon^{i+1} \boldsymbol{\alpha}^i(x/\varepsilon) \cdot \nabla^i (\boldsymbol{f}(x) - \nabla p_\varepsilon^*(x)), \tag{7.5.16}$$

$$\mathrm{div}(\boldsymbol{u}_\varepsilon^*(x)) = 0. \tag{7.5.17}$$

Proof. The result is proved by induction. The case $i = -1$ is straightforward thanks to the convention $\boldsymbol{u}_{-1} = p_{-1} = 0$. Assuming these results hold till rank $i+1$ with $i \geq -2$, we compute, replacing (7.5.12) into (7.5.5):

$$\begin{cases} (-\Delta_{yy} \boldsymbol{u}_{i+2} + \nabla_y p_{i+2})(x,y) = (f_{i+2,j}(x) - \partial_j p_{i+2}^*(x)) \boldsymbol{e}_j \\ \qquad + \sum_{k=0}^{i+1} (-\alpha_j^k(y) \boldsymbol{e}_l \otimes e_l + 2\partial_l \boldsymbol{\mathcal{X}}_j^k(y) \otimes e_l) \cdot \nabla^{k+1} (f_{i+1-k,j}(x) - \partial_j p_{i+1-k}^*(x)) \\ \qquad + \sum_{k=0}^{i} (\boldsymbol{\mathcal{X}}_j^k(y) \otimes I) \cdot \nabla^{k+2} (f_{i-k,j}(x) - \partial_j p_{i-k}(x)) \\ \qquad = (f_{i+2,j}(x) - \partial_j p_{i+2}^*(x)) \boldsymbol{e}_j + (2\partial_l \boldsymbol{\mathcal{X}}_j^0(y) \otimes e_l - \alpha_j^0(y) \boldsymbol{e}_l \otimes e_l) \cdot \nabla (f_{i+1,j}(x) - \partial_j p_{i+1}^*(x)) \\ \qquad + \sum_{k=0}^{i} (2\partial_l \boldsymbol{\mathcal{X}}_j^{k+1}(y) \otimes e_l - \alpha_j^{k+1}(y) \boldsymbol{e}_l \otimes e_l + \boldsymbol{\mathcal{X}}_j^k(y) \otimes I) \cdot \nabla^{k+2} (f_{i-k,j}(x) - \partial_j p_{i-k}^*(x)) \\ \mathrm{div}_y(\boldsymbol{u}_{i+2})(x,y) = -\sum_{k=0}^{i+1} (\boldsymbol{\mathcal{X}}_j^k(y) \cdot \boldsymbol{e}_l \otimes e_l) \cdot \nabla^{k+1} (f_{i+1-k,j}(x) - \partial_j p_{i+1-k}^*(x)). \end{cases} \tag{7.5.18}$$

The system (7.5.18) admits a unique solution $(\boldsymbol{u}_{i+2}, p_{i+2})$ with $\int_Y p_{i+2}(x,y)\mathrm{d}y = 0$ if and only if the following compatibility condition (the so-called "Fredholm alternative") holds:

$$\forall i \geq -1, \int_Y \mathrm{div}_y(\boldsymbol{u}_{i+2})(x,y)\mathrm{d}y = -\sum_{k=0}^{i+1} [< \boldsymbol{\mathcal{X}}_j^k > \cdot \boldsymbol{e}_l \otimes e_l] \cdot \nabla^{k+1} (f_{i+1-k,j}(x) - \partial_j p_{i+1-k}^*(x)) = 0.$$

The above equation can be rewritten in terms of an equation determining p_{i+1}^* given the values of p_k^* for $0 \leq k \leq i$:

$$< \boldsymbol{\mathcal{X}}_j^0 > \cdot \boldsymbol{e}_l \partial_l (f_{i+1,j} - \partial_j p_{i+1}^*) = -\sum_{k=1}^{i+1} [< \boldsymbol{\mathcal{X}}_j^k > \cdot \boldsymbol{e}_l \otimes e_l] \cdot \nabla^{k+1} (f_{i+1-k,j} - \partial_j p_{i+1-k}^*),$$

which is a rewriting of (7.5.13) (at order $i+1$ with any $i \geq -1$). In particular, this identity allows to rewrite the second equation of (7.5.18) as

$$\mathrm{div}_y(\boldsymbol{u}_{i+2})(x,y) = -\sum_{k=0}^{i+1} [(\boldsymbol{\mathcal{X}}_j^k(y) - < \boldsymbol{\mathcal{X}}_j^k >) \cdot \boldsymbol{e}_l \otimes e_l] \cdot \nabla^{k+1} (f_{i+1-k,j}(x) - \partial_j p_{i+1-k}^*(x)). \tag{7.5.19}$$

By linearity, (7.5.18) and (7.5.19) and the cell problems (7.5.6) to (7.5.8) imply

$$\begin{aligned}\boldsymbol{u}_{i+2}(x,y) &= (f_{i+2,j}(x) - \partial_j p^*_{i+2}(x))\boldsymbol{\mathcal{X}}^0_j(y) + \boldsymbol{\mathcal{X}}^1_j(y)\cdot\nabla(f_{i+1,j}(x) - \partial_j p^*_{i+1}(x)) \\ &\quad + \sum_{k=0}^{i} \boldsymbol{\mathcal{X}}^{k+2}_j(y)\cdot\nabla^{k+2}(f_{i-k,j}(x) - \partial_j p^*_{i-k}(x)) \\ p_{i+2}(x,y) &= (f_{i+2,j}(x) - \partial_j p^*_{i+2}(x))\alpha^0_j + \alpha^1_j(y)\cdot\nabla(f_{i+1,j}(x) - \partial_j p^*_{i+1}(x)) \\ &\quad + \sum_{k=0}^{i} \alpha^{k+2}_j(y)\cdot\nabla^{k+2}(f_{i-k,j}(x) - \partial_j p^*_{i-k}(x)),\end{aligned}$$

which is exactly the result (7.5.12) at rank $i+2$. □

Remark 7.27. The truncation of the series (7.5.15) at first order yields the well-known Darcy's law The next terms of the series have been obtained in [228, 74], at least up to the order $i = 1$.

Remark 7.28. The ansatz (7.5.14) involves a bit of asymptotic *crime* because it features p^*_ε which is itself a formal power series in ε (recall its definition (7.5.3)). Although this ansatz is criminal, it remains close to the classical ansatz (7.5.2). Therefore, in the subsequent sections, the denomination "criminal ansatz for the Stokes system (7.5.1)" is used exclusively for referring to the more *severe* asymptotic crime (7.5.32) introduced hereafter.

The identities of propositions 7.2 and 7.17 for the average of the tensors $\mathcal{X}^{k*}$ extend as follows in the context of the fluid system (7.5.1):

Proposition 7.30. *For any $k \geq 0$ and $0 \leq p \leq k$, the following identity holds for the matrix valued tensor $\mathcal{X}^{k*}$:*

$$\forall 1 \leq i,j \leq d,\ \mathcal{X}^{k*}_{ij} = (-1)^p \int_Y ((-\Delta_{yy}\boldsymbol{\mathcal{X}}^p_i + \nabla\alpha^p_i)\cdot\boldsymbol{\mathcal{X}}^{k-p}_j + \nabla\alpha^{k-p}_j\cdot\boldsymbol{\mathcal{X}}^p_i - \boldsymbol{\mathcal{X}}^{k-p-1}_j\cdot\boldsymbol{\mathcal{X}}^{p-1}_i \otimes I)\mathrm{d}y \quad (7.5.20)$$

with $\boldsymbol{\mathcal{X}}^{-1}_i = 0$ by convention. In particular, for any $k \geq 0$:

- *$\mathcal{X}^{2k*}$ takes values in the set of $d\times d$ symmetric matrices:*

$$\mathcal{X}^{2k*}_{ij} = (-1)^k \int_Y (\nabla\boldsymbol{\mathcal{X}}^k_i : \nabla\boldsymbol{\mathcal{X}}^k_j + \nabla\alpha^k_i\cdot\boldsymbol{\mathcal{X}}^k_j + \nabla\alpha^k_j\cdot\boldsymbol{\mathcal{X}}^k_i - \boldsymbol{\mathcal{X}}^{k-1}_i\cdot\boldsymbol{\mathcal{X}}^{k-1}_j \otimes I)\mathrm{d}y \quad (7.5.21)$$

- *$\mathcal{X}^{2k+1*}$ takes values in the set of $d\times d$ antisymmetric matrices:*

$$\begin{aligned}\mathcal{X}^{2k+1*}_{ij} &= (-1)^k \int_Y (\boldsymbol{\mathcal{X}}^k_i\cdot\nabla\boldsymbol{\mathcal{X}}^k_j - \boldsymbol{\mathcal{X}}^k_j\cdot\nabla\boldsymbol{\mathcal{X}}^k_i + \alpha^k_i\boldsymbol{\mathcal{X}}^k_j - \alpha^k_j\boldsymbol{\mathcal{X}}^k_i)\cdot\boldsymbol{e}_l\otimes e_l\mathrm{d}y \\ &\quad + (-1)^k \int_Y (\boldsymbol{\mathcal{X}}^{k-1}_j\cdot\boldsymbol{\mathcal{X}}^k_j - \boldsymbol{\mathcal{X}}^{k-1}_i\cdot\boldsymbol{\mathcal{X}}^k_j)\mathrm{d}y.\end{aligned} \quad (7.5.22)$$

Proof. The result holds for $p = 0$ because

$$\mathcal{X}^{k*}_{ij} = \int_Y \boldsymbol{\mathcal{X}}^k_j\cdot\boldsymbol{e}_i\mathrm{d}y = \int_Y \boldsymbol{\mathcal{X}}^k_j\cdot(-\Delta_{yy}\boldsymbol{\mathcal{X}}^0_i + \nabla\alpha^0_i)\mathrm{d}y.$$

Assume now that (7.5.20) holds till rank p with $k > p \geq 0$, and let us show that it implies the result at rank $p+1$. We write, after an integration by parts and by using (7.5.6) to (7.5.8):

$$\begin{aligned}\mathcal{X}^{k*}_{ij} &= (-1)^p \int_Y (-\boldsymbol{\mathcal{X}}^p_i\cdot\Delta\boldsymbol{\mathcal{X}}^{k-p}_j - \alpha^p_i\mathrm{div}(\boldsymbol{\mathcal{X}}^{k-p}_j) - \alpha^{k-p}_j\mathrm{div}(\boldsymbol{\mathcal{X}}^p_i) - \boldsymbol{\mathcal{X}}^{k-p-1}\cdot\boldsymbol{\mathcal{X}}^{p-1}_i\otimes I)\mathrm{d}y \\ &= (-1)^p \int_Y \big[(\boldsymbol{\mathcal{X}}^p_i\cdot(2\partial_l\boldsymbol{\mathcal{X}}^{k-p-1}_j - \alpha^{k-p-1}_j\boldsymbol{e}_l)\otimes e_l + \boldsymbol{\mathcal{X}}^{k-p-2}_j\otimes I - \nabla\alpha^{k-p}_j)\cdot\boldsymbol{\mathcal{X}}^p_i \\ &\quad + \alpha^p_i\boldsymbol{\mathcal{X}}^{k-p-1}_j\cdot\boldsymbol{e}_l\otimes e_l + \alpha^{k-p}_j\boldsymbol{\mathcal{X}}^{p-1}_i\cdot\boldsymbol{e}_l\otimes e_l - \boldsymbol{\mathcal{X}}^{k-p-1}_j\cdot\boldsymbol{\mathcal{X}}^{p-1}_i\otimes I\big]\mathrm{d}y \\ &= (-1)^p \int_Y \big[-\boldsymbol{\mathcal{X}}^{k-p-1}_j\cdot((2\partial_l\boldsymbol{\mathcal{X}}^p_i - \alpha^p_i\boldsymbol{e}_l)\otimes e_l + \boldsymbol{\mathcal{X}}^{p-1}_i\otimes I) + \alpha^{k-p-1}_j\mathrm{div}(\boldsymbol{\mathcal{X}}^{p+1}_i) \\ &\quad - \nabla\alpha^{k-p}_j\cdot\boldsymbol{\mathcal{X}}^p_i - \alpha^{k-p}_j\mathrm{div}(\boldsymbol{\mathcal{X}}^p_i) + \boldsymbol{\mathcal{X}}^{k-p-2}_j\cdot\boldsymbol{\mathcal{X}}^p_i\otimes I\big]\mathrm{d}y \\ &= (-1)^p \int_Y [-\boldsymbol{\mathcal{X}}^{k-p-1}_j\cdot(-\Delta_{yy}\boldsymbol{\mathcal{X}}^{p+1}_i + \nabla\alpha^{p+1}_i) - \nabla\alpha^{k-p-1}_j\cdot\boldsymbol{\mathcal{X}}^{p+1}_i + \boldsymbol{\mathcal{X}}^{k-p-2}_j\cdot\boldsymbol{\mathcal{X}}^p_i\otimes I]\mathrm{d}y,\end{aligned}$$

whence (7.5.20) at rank $p+1$.

The expression (7.5.21) for $\mathcal{X}_{ij}^{2k*}$ is obtained by setting $k \leftarrow 2k$ and $p \leftarrow k$ in (7.5.20). The expression for $\mathcal{X}_{ij}^{2k+1*}$ is obtained by setting $k \leftarrow 2k+1$ and $p \leftarrow k$ and performing the following integration by part:

$$\begin{aligned}\mathcal{X}_{ij}^{2k+1*} &= (-1)^k \int_Y ((-\Delta_{yy}\boldsymbol{\mathcal{X}}_j^{k+1} + \nabla\alpha_j^{k+1}) \cdot \boldsymbol{\mathcal{X}}_i^k + \nabla\alpha_i^k \cdot \boldsymbol{\mathcal{X}}_j^{k+1} - \boldsymbol{\mathcal{X}}_j^k \cdot \boldsymbol{\mathcal{X}}_i^{k-1} \otimes I)\mathrm{d}y \\ &= (-1)^k \int_Y [((2\partial_l \boldsymbol{\mathcal{X}}_j^k - \alpha_j^k \boldsymbol{e}_l) \otimes e_l + \boldsymbol{\mathcal{X}}_j^{k-1} \otimes I) \cdot \boldsymbol{\mathcal{X}}_i^k + \alpha_i^k \boldsymbol{\mathcal{X}}_i^k \cdot \boldsymbol{e}_l \otimes e_l - \boldsymbol{\mathcal{X}}_j^k \cdot \boldsymbol{\mathcal{X}}_i^{k-1} \otimes I]\mathrm{d}y,\end{aligned}$$

from where one recognizes (7.5.22). □

We end this part with a result analogous to that of proposition 7.18 which yields as a corollary the linear independence of the tensors $(\mathcal{X}_j^k)_{1\le j\le d}$ and the positive definiteness of the permeability tensor $\mathcal{X}^{0*}$.

Proposition 7.31. *The following identities hold for any $1 \le j \le d$:*

$$\forall k \ge 0,\ -\Delta_{yy}(\partial^k_{i_1\dots i_k} \boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}) + \nabla\left(\sum_{p=0}^{k} (-1)^p \partial^{k-p}_{i_1\dots i_{k-p}} \alpha^{k-p}_{i_1\dots i_{k-p},j}\right) = (-1)^k (k+1)\boldsymbol{e}_j, \tag{7.5.23}$$

$$\mathrm{div}(\partial^k_{i_1\dots i_k} \boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}) = 0. \tag{7.5.24}$$

Proof. We prove the following identity by induction on k,

$$\begin{aligned}\forall k \ge 0,\ -\Delta_{yy}(\partial^k_{i_1\dots i_k} \boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}) + \nabla(\partial^k_{i_1\dots i_k} \alpha^k_{i_1\dots i_k,j}) \\ = (-1)^k (k+1)\boldsymbol{e}_j - (-1)^k \sum_{p=0}^{k-1} (-1)^p \nabla(\partial^p_{i_1\dots i_p} \alpha^p_{i_1\dots i_p,j}),\end{aligned} \tag{7.5.25}$$

which is equivalent to (7.5.23). The identity clearly holds at rank $k=0$ and $k=-1$ by assuming the convention $\boldsymbol{\mathcal{X}}_j^{-1} = 0$. Assume it to be true up to rank $k-1$, we write

$$\begin{aligned}-\Delta_{yy}(\partial^k_{i_1\dots i_k} \boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}) + \nabla(\partial^k_{i_1\dots i_k} \alpha^k_j) &= \partial^k_{i_1\dots i_k}\big((2\partial_{i_k} \boldsymbol{\mathcal{X}}^{k-1}_{i_1\dots i_{k-1},j} - \alpha^{k-1}_{i_1\dots i_{k-1},j}\boldsymbol{e}_{i_k}) + \delta_{i_k i_{k-1}} \boldsymbol{\mathcal{X}}^{k-2}_{i_1\dots,i_{k-2},j}\big) \\ &= 2\Delta_{yy}(\partial^{k-1}_{i_1\dots i_{k-1}} \boldsymbol{\mathcal{X}}^{k-1}_{i_1\dots i_{k-1},j}) + \Delta_{yy}(\partial^{k-2}_{i_1\dots i_{k-2}} \boldsymbol{\mathcal{X}}^{k-2}_{i_1\dots i_{k-2},j}) - \nabla(\partial^{k-1}_{i_1\dots i_{k-1}} \alpha^{k-1}_{i_1\dots i_{k-1},j}) \\ &= 2\left(-(-1)^{k-1} k\boldsymbol{e}_j + (-1)^{k-1} \sum_{p=0}^{k-2} (-1)^p \nabla(\partial^p_{i_1\dots i_p} \alpha^p_{i_1\dots i_p,j}) + \nabla(\partial^{k-1}_{i_1\dots i_{k-1}} \alpha^{k-1}_{i_1\dots i_{k-1},j})\right) \\ &\quad + \left(-(-1)^{k-2}(k-1)\boldsymbol{e}_j + (-1)^{k-2} \sum_{p=0}^{k-3} (-1)^p \nabla(\partial^p_{i_1\dots i_p} \alpha^p_{i_1\dots i_p,j}) + \nabla(\partial^{k-2}_{i_1\dots i_{k-2}} \alpha^{k-2}_{i_1\dots i_{k-2},j})\right) \\ &\quad - \nabla(\partial^{k-1}_{i_1\dots i_{k-1}} \alpha^{k-1}_{i_1\dots i_{k-1},j}) \\ &= (-1)^k (k+1)\boldsymbol{e}_j - (-1)^k \sum_{p=0}^{k-3} (-1)^p \nabla(\partial^p_{i_1\dots i_p} \alpha^p_{i_1\dots i_p,j}) + \underbrace{(2(-1)^{2k-3}+1)}_{-1=-(-1)^k(-1)^{k-2}} \nabla(\partial^{k-2}_{i_1\dots i_{k-2}} \alpha^{k-2}_{i_1\dots i_{k-2},j}) \\ &\quad \underbrace{-(-1)}_{-(-1)=-(-1)^k(-1)^{k-1}} \nabla(\partial^{k-1}_{i_1\dots i_{k-1}} \alpha^{k-1}_{i_1\dots i_{k-1},j}),\end{aligned}$$

from where (7.5.25) follows at rank k.
The second equality is obtained similarly by writing

$$\begin{aligned}\mathrm{div}(\partial^k_{i_1\dots i_k} \boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,j}) &= -\partial^k_{i_1\dots i_k}(\boldsymbol{\mathcal{X}}^{k-1}_{i_1\dots i_{k-1},j} - < \boldsymbol{\mathcal{X}}^{k-1}_{i_1\dots i_{k-1},j} >) \cdot \boldsymbol{e}_{i_k} \\ &= -\mathrm{div}(\partial^{k-1}_{i_1\dots i_{k-1}} \boldsymbol{\mathcal{X}}^{k-1}_{i_1\dots i_{k-1},j}).\end{aligned}$$

□

Corollary 7.8. *Assume (H4). The family of k-th order vector valued tensors $(\boldsymbol{\mathcal{X}}_j^k)_{1\le j\le d}$ is linearly independent. In particular, the matrix*

$$\mathcal{X}^{0*} = (\mathcal{X}_{ij}^{0*})_{1\le i,j\le d} \text{ with } \mathcal{X}_{ij}^{0*} = \int_Y \nabla\boldsymbol{\mathcal{X}}_i^0 : \nabla\boldsymbol{\mathcal{X}}_j^0 \mathrm{d}y, \quad \forall 1 \le i,j \le d,$$

is symmetric positive definite.

Proof. (see also [272]) Let $(\lambda_j)_{1\leq j\leq d}$ be some coefficients such that $\sum_{i=1}^d \lambda_j \boldsymbol{\mathcal{X}}_j^k = 0$. Then proposition 7.31 implies the existence of a periodic function ϕ satisfying

$$\nabla\phi = (-1)^k(k+1)\sum_{j=1}^d \lambda_j \boldsymbol{e}_j. \tag{7.5.26}$$

From the fact that Y is connected, there exist constant coefficients $(C_i)_{1\leq i\leq d}$ such that

$$\forall x \in Y,\ \phi(x) = (-1)^k(k+1)\sum_{j=1}^d (\lambda_j x_j + C_j).$$

Remembering that ϕ must be a P–periodic function, this is possible only if $\lambda_j = 0$ for all $1 \leq j \leq d$. □

7.5.2 Higher order homogenized equations: tensors M^k, N^k, β^k and $\mathbb{D}_K$

In this section, we derive an "infinite order" homogenized equation (eqn. (7.5.29) below) for the formal average $\boldsymbol{u}_\varepsilon^*$, before outlining how to truncate so as to obtain well-posed homogenized equations of finite order (eqn. (7.5.39) below). Since the ansatz (7.5.14) has the same structure than (7.4.8) in the elasticity case considered in section 7.4, the construction of the tensors M^k and $\boldsymbol{N}_j^k$ follows similarly. The following definition makes sense because we recalled in corollary 7.8 that the Darcy permeability tensor $\mathcal{X}^{0*}$ is invertible.

Proposition 7.32. *Let M^k be the tensor of order k defined by induction as follows:*

$$\begin{cases} M^0 = (\mathcal{X}^{0*})^{-1} \\ M^k = -(\mathcal{X}^{0*})^{-1}\displaystyle\sum_{p=0}^{k-1} \mathcal{X}^{k-p*} \otimes M^p, \quad \forall k \geq 1. \end{cases} \tag{7.5.27}$$

The source terms $\boldsymbol{f}_i$ (eqn. (7.5.4)) are given in terms of the averaged ansatz terms $\boldsymbol{u}_i^*(x)$ and $p_i^*(x)$ through the following identity:

$$\forall i \geq 0,\ \boldsymbol{f}_i(x) - \nabla p_i^*(x) = \sum_{k=0}^{i} M^k \cdot \nabla^k \boldsymbol{u}_{i-k}^*(x). \tag{7.5.28}$$

The above identity together with (7.5.13) can be rewritten formally as the following "infinite order" homogenized equation for the homogenized averages $\boldsymbol{u}_\varepsilon^*$ and $p_\varepsilon^*(x)$ of (7.5.3):

$$\begin{cases} \displaystyle\sum_{i=0}^{+\infty} \varepsilon^{i-2} M^i \cdot \nabla^i \boldsymbol{u}_\varepsilon^* + \nabla p_\varepsilon^* = \boldsymbol{f} \quad \text{in } D \\ \qquad\qquad \operatorname{div}(\boldsymbol{u}_\varepsilon^*) = 0 \quad \text{in } D \\ \qquad\qquad \displaystyle\int_D p_\varepsilon^* \mathrm{d}x = 0 \\ \qquad\qquad \boldsymbol{u}_\varepsilon^* \text{ is } D\text{–periodic.} \end{cases} \tag{7.5.29}$$

Since the family of tensors $(M^k)_{k\in\mathbb{N}}$ is defined by the same recurrence than (7.4.15) in the elasticity case, the explicit formula for M^k given in proposition 7.20 holds without modification:

$$M^k = \sum_{p=1}^{k}(-1)^p \sum_{\substack{i_1+\cdots+i_p=k \\ 1\leq i_1,\ldots,i_p\leq k}} (\mathcal{X}^{0*})^{-1}\otimes\mathcal{X}^{i_1*}\otimes(\mathcal{X}^{0*})^{-1}\otimes\cdots\otimes(\mathcal{X}^{0*})^{-1}\otimes\mathcal{X}^{i_p*}\otimes(\mathcal{X}^{0*})^{-1}, \quad \forall k \geq 1. \tag{7.5.30}$$

We now introduce matrix valued tensors N^k and vector valued tensors $\boldsymbol{\beta}^k$ which allow to obtain "criminal ansatz" expressing the velocity and pressure $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ in terms of their formal average $(\boldsymbol{u}_\varepsilon^*, p_\varepsilon^*)$.

Proposition 7.33. *Let N^k and $\boldsymbol{\beta}^k$ be respectively the $k-$th order matrix valued and vector valued tensors defined by*

$$N^k(y) := \sum_{p=0}^{k} \mathcal{X}^{k-p}(y) \otimes M^p, \qquad \boldsymbol{\beta}^k(y) := \sum_{p=0}^{k}(-1)^p M^p \cdot \boldsymbol{\alpha}^{k-p}(y), \qquad \forall y \in Y.$$

Then the terms $\boldsymbol{u}_i(x,y)$ and $p_i(x,y)$ of the oscillating ansatz (7.5.14) and (7.5.16) can be rewritten in terms of the averages $\boldsymbol{u}_i^*$ (eqn. (7.5.11)) and p_i^* as follows:

$$\forall i \geq 0,\ \boldsymbol{u}_i(x,y) = \sum_{k=0}^{i} N^k(y) \cdot \nabla^k \boldsymbol{u}_{i-k}^*(x), \qquad p_i(x,y) = \sum_{k=0}^{i} \boldsymbol{\beta}^k(y) \cdot \nabla^k \boldsymbol{u}_{i-k}^*(x). \tag{7.5.31}$$

These equations can be rewritten formally in terms of the following "criminal ansatz" for $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$:

$$\forall x \in D_\varepsilon,\ \boldsymbol{u}_\varepsilon(x) = \sum_{i=0}^{+\infty} \varepsilon^i N^i(x/\varepsilon) \cdot \nabla^i \boldsymbol{u}_\varepsilon^*(x), \qquad p_\varepsilon(x) = p_\varepsilon^*(x) + \sum_{i=0}^{+\infty} \varepsilon^{i-1} \boldsymbol{\beta}^i(x/\varepsilon) \cdot \nabla^i \boldsymbol{u}_\varepsilon^*(x). \tag{7.5.32}$$

Proof. The proof is identical to that of proposition 7.8: the key ingredient is to substitute (7.5.28) into (7.5.12). Performing the same change of indices as in (7.3.38) yields

$$\boldsymbol{u}_i(x,y) = \sum_{k=0}^{i} \sum_{p=0}^{k} (\boldsymbol{\mathcal{X}}^p(y) \otimes M^{p-k}) \cdot \nabla^k \boldsymbol{u}_{i-k}^*(x),$$

$$p_i(x,y) = \sum_{k=0}^{i} \sum_{p=0}^{k} ((M^{p-k})^T \cdot \boldsymbol{\alpha}^p(y)) \cdot \nabla^k \boldsymbol{u}_{i-k}^*(x).$$

The result follows, where we use $(M^{p-k})^T = (-1)^{p-k} M^{p-k}$ as stated in corollary 7.9 below. □

In what follows, we denote by $(\boldsymbol{N}_j^k)_{1\leq j\leq d}$ and by $(\beta_j^k)_{1\leq j\leq d}$ respectively the column vectors and the coefficients of $N^k(y)$ and $\boldsymbol{\beta}^k(y)$:

$$\forall 1 \leq i,j \leq d,\ \boldsymbol{N}_j^k := N^k \boldsymbol{e}_j \ \text{ and } \beta_j^k := \boldsymbol{\beta}^k \cdot \boldsymbol{e}_j.$$

In addition, the convention $\boldsymbol{N}_j^{-1} = 0$ is assumed. The proposition 7.22 and its corollary 7.5 extend as follows in the context of the Stokes system:

Proposition 7.34. *The k-th order tensors N^k, $(\boldsymbol{N}_j^k)_{1\leq j\leq d}$, $\boldsymbol{\beta}^k$ and $(\beta_j^k)_{1\leq j\leq d}$ satisfy:*

(i) $\int_Y N^0(y)\mathrm{d}y = I$ *and* $\int_Y N^k(y)\mathrm{d}y = 0$ *for any* $k \geq 1$;

(ii) $\int_Y \boldsymbol{\beta}^k(y)\mathrm{d}y = 0$ *for any* $k \geq 0$;

(iii) $\forall k \geq -2$, $\forall 1 \leq j \leq d$,

$$\begin{cases} -\Delta_{yy} \boldsymbol{N}_j^{k+2} + \nabla \beta_j^{k+2} = (2\partial_l \boldsymbol{N}_j^{k+1} - \beta_j^{k+1} \boldsymbol{e}_l) \otimes e_l + \boldsymbol{N}_j^k \otimes I + M^{k+2} \boldsymbol{e}_j, \\ \mathrm{div}(\boldsymbol{N}_j^{k+2}) = -(\boldsymbol{N}_j^{k+1} - < \boldsymbol{N}_j^{k+1} >) \cdot \boldsymbol{e}_l \otimes e_l; \end{cases} \tag{7.5.33}$$

(iv) for any $k \geq 0$,

$$-\Delta_{yy}(\partial^k_{i_1\ldots i_k} \boldsymbol{N}^k_{i_1\ldots i_k,j}) + \nabla \left(\sum_{p=0}^{k} (-1)^p \partial^{k-p}_{i_1\ldots i_{k-p}} \beta^{k-p}_{i_1\ldots i_k,j} \right) = (-1)^{k+1}(k+1) M^0 \boldsymbol{e}_j; \tag{7.5.34}$$

(v) for any $1 \leq p \leq k-1$,

$$M_{ij}^k = (-1)^{p+1} \int_Y ((-\Delta_{yy} \boldsymbol{N}_i^p + \nabla \beta_i^p) \cdot \boldsymbol{N}_j^{k-p} + \nabla \beta_j^{k-p} \cdot \boldsymbol{N}_i^p - \boldsymbol{N}_i^{p-1} \cdot \boldsymbol{N}_j^{k-p-1} \otimes I)\mathrm{d}y. \tag{7.5.35}$$

Proof. (i) and (ii) are straightforward consequences of (7.5.27).

(iii) is obtained by writing, for $k \geq 0$ (implicit summation on the repeated index j assumed):

$$\begin{aligned} -\Delta_{yy} \boldsymbol{N}_j^{k+2} + \nabla \beta_j^{k+2} &= -\Delta_{yy} \left(\sum_{p=0}^{k+2} \boldsymbol{\mathcal{X}}_i^{k+2-p}(y) \otimes M_{ij}^p \right) + \nabla \left(\sum_{p=0}^{k+2} \alpha_i^{k+2-p}(y) \otimes M_{ij}^p \right) \\ &= \sum_{p=0}^{k} \left[(2\partial_l \boldsymbol{\mathcal{X}}_i^{k+1-p} - \alpha_i^{k+1-p} \boldsymbol{e}_l) \otimes e_l + \boldsymbol{\mathcal{X}}_i^{k-p} \otimes I \right] M_{ij}^p + (2\partial_l \boldsymbol{\mathcal{X}}_i^0 - \alpha_i^0 \boldsymbol{e}_l) M_{ij}^{k+1} + M_{ij}^{k+2} \boldsymbol{e}_i \\ &= (2\partial_l \boldsymbol{N}_j^{k+1} - \beta_i^{k+1} \boldsymbol{e}_l) \otimes e_l + \boldsymbol{N}_j^k \otimes I + M^{k+2} \boldsymbol{e}_j. \end{aligned}$$

$$\operatorname{div}(\boldsymbol{N}_j^{k+2}) = \sum_{p=0}^{k+2} \operatorname{div}(\boldsymbol{\mathcal{X}}_i^{k+2-p}) M_{ij}^p = -\sum_{p=0}^{k+1} M_{ij}^p (\boldsymbol{\mathcal{X}}_i^{k+1-p} - < \boldsymbol{\mathcal{X}}_i^{k+1-p} >) \cdot \boldsymbol{e}_l \otimes e_l.$$

The proof is identical for $k = -1$ and $k = -2$.

(iv) is obtained by following the proof of proposition 7.31 and using (7.5.33).

(v) is obtained by induction following the proof of proposition 7.9 and by using (7.5.33). The case $p = 1$ with $k \geq 2$ is treated by writing

$$\begin{aligned}
M_{ij}^k &= \int_Y \boldsymbol{N}_i^0 \cdot M^k \boldsymbol{e}_j \mathrm{d}y \\
&= \int_Y \boldsymbol{N}_i^0 \cdot (-\Delta \boldsymbol{N}_j^k + \nabla \beta_j^k - (2\partial_l \boldsymbol{N}_j^{k-1} - \beta_j^{k-1} \boldsymbol{e}_j) \otimes e_l - \boldsymbol{N}_j^{k-2} \otimes I) \mathrm{d}y \\
&= \int_Y (-\Delta \boldsymbol{N}_i^0 \cdot \boldsymbol{N}_j^k + 2\partial_l \boldsymbol{N}_i^0 \cdot \boldsymbol{N}_j^{k-1} \otimes e_l + \nabla \beta_j^{k-1} \cdot \boldsymbol{N}_i^1 - \boldsymbol{N}_i^0 \cdot \boldsymbol{N}_j^{k-2} \otimes I) \mathrm{d}y \\
&= \int_Y (\boldsymbol{N}_j^k \cdot M^0 \boldsymbol{e}_i - \nabla \beta_i^0 \cdot \boldsymbol{N}_j^k + (-\Delta \boldsymbol{N}_i^1 + \nabla \beta_i^1 + \beta_i^0 \boldsymbol{e}_l \otimes e_l - M^1 \boldsymbol{e}_i) \cdot \boldsymbol{N}_j^{k-1}) \mathrm{d}y \\
&\quad + \int_Y (\nabla \beta_j^{k-1} \cdot \boldsymbol{N}_i^1 - \boldsymbol{N}_i^0 \cdot \boldsymbol{N}_j^{k-2} \otimes I) \mathrm{d}y \\
&= \int_Y ((-\Delta \boldsymbol{N}_i^1 + \nabla \beta_i^1) \cdot \boldsymbol{N}_j^{k-1} + \nabla \beta_j^{k-1} \cdot \boldsymbol{N}_i^1 - \boldsymbol{N}_i^0 \cdot \boldsymbol{N}_j^{k-2} \otimes I) \mathrm{d}y.
\end{aligned}$$

We now assume the result holds till rank $1 \leq p < k-1$ and we prove it at rank $p+1$:

$$\begin{aligned}
M_{ij}^k &= (-1)^{p+1} \int_Y (-\Delta \boldsymbol{N}_j^{k-p} \cdot \boldsymbol{N}_i^p + \beta_i^p \boldsymbol{N}_j^{k-p-1} \cdot \boldsymbol{e}_l \otimes e_l + \nabla \beta_j^{k-p} \cdot \boldsymbol{N}_i^p - \boldsymbol{N}_i^{p-1} \cdot \boldsymbol{N}_j^{k-p-1} \otimes I) \mathrm{d}y \\
&= (-1)^{p+1} \int_Y (-\nabla \beta_j^{k-p} + (2\partial_l \boldsymbol{N}_j^{k-p-1} - \beta_j^{k-p-1} \boldsymbol{e}_l) \otimes e_l + \boldsymbol{N}_j^{k-p-2} \otimes I + M^{k-p} \boldsymbol{e}_j) \cdot \boldsymbol{N}_i^p \mathrm{d}y \\
&\quad + (-1)^{p+1} \int_Y (\beta_i^p \boldsymbol{N}_j^{k-p-1} \cdot \boldsymbol{e}_l \otimes e_l + \nabla \beta_j^{k-p} \cdot \boldsymbol{N}_i^p - \boldsymbol{N}_i^{p-1} \cdot \boldsymbol{N}_j^{k-p-1} \otimes I) \mathrm{d}y \\
&= (-1)^{p+1} \int_Y (-\boldsymbol{N}_j^{k-p-1} \cdot ((2\partial_l \boldsymbol{N}_i^p - \beta_i^p \boldsymbol{e}_l) \otimes e_l + \boldsymbol{N}_i^{p-1} \otimes I + M^{p+1} \boldsymbol{e}_i)) \mathrm{d}y \\
&\quad + (-1)^{p+1} \int_Y (-\nabla \beta_j^{k-p-1} \cdot \boldsymbol{N}_i^{p+1} + \boldsymbol{N}_j^{k-p-2} \cdot \boldsymbol{N}_i^p \otimes I) \mathrm{d}y \\
&= (-1)^{p+1} \int_Y (-\boldsymbol{N}^{k-p-1} \cdot (-\Delta \boldsymbol{N}_i^{p+1} + \nabla \beta_i^{p+1}) - \nabla \beta_j^{k-p-1} \cdot \boldsymbol{N}_i^{p+1} + \boldsymbol{N}_j^{k-p-2} \cdot \boldsymbol{N}_i^p \otimes I) \mathrm{d}y.
\end{aligned}$$

□

Corollary 7.9. *For any $k \geq 0$,*

- *M^{2k} is a symmetric matrix valued tensor, and the following identities hold:*

$$M_{ij}^0 = \int_Y \nabla \boldsymbol{N}_i^0 : \nabla \boldsymbol{N}_j^0 \mathrm{d}y,$$

$$\forall k \geq 1,\ M_{ij}^{2k} = (-1)^{k+1} \int_Y (\nabla \boldsymbol{N}_i^k : \nabla \boldsymbol{N}_j^k + \nabla \beta_i^k \cdot \boldsymbol{N}_j^k + \nabla \beta_j^k \cdot \boldsymbol{N}_i^k - \boldsymbol{N}_i^{k-1} \cdot \boldsymbol{N}_j^{k-1} \otimes I) \mathrm{d}y.$$

- *M^{2k+1} is an antisymmetric matrix valued tensor, and the following identities hold:*

$$\begin{aligned}
\forall k \geq 0,\ M_{ij}^{2k+1} = (-1)^{k+1} \int_Y (\boldsymbol{N}_i^k \cdot \nabla \boldsymbol{N}_j^k - \boldsymbol{N}_j^k \cdot \nabla \boldsymbol{N}_i^k + \beta_i^k \boldsymbol{N}_j^k - \beta_j^k \boldsymbol{N}_i^k) \cdot \boldsymbol{e}_l \otimes e_l \mathrm{d}y \\
+ (-1)^{k+1} \int_Y (\boldsymbol{N}_j^{k-1} \cdot \boldsymbol{N}_i^k - \boldsymbol{N}_i^{k-1} \cdot \boldsymbol{N}_j^k) \otimes I \mathrm{d}y.
\end{aligned}$$

We now use the criminal ansatz to derive well-posed homogenized equations of (finite) order $2K+2$. The formal identities (7.5.32) lead us to introduce the following truncated ansatz $\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v})$ and $q_{\varepsilon,K}(\boldsymbol{v},\phi)$ for the velocity and pressure:

$$\forall \boldsymbol{v} \in H^{K+1}(D,\mathbb{R}^d),\ \boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v})(x) := \sum_{k=0}^{K} \varepsilon^k N^k(x/\varepsilon)\cdot\nabla^k \boldsymbol{v}(x),\ x\in D_\varepsilon, \tag{7.5.36}$$

$$\forall q \in L^2(D),\ \forall \boldsymbol{v} \in H^{K+1}(D),\ q_{\varepsilon,K}(\boldsymbol{v},\phi)(x) := \phi + \sum_{i=0}^{K} \varepsilon^{i-1}\boldsymbol{\beta}^i(x/\varepsilon)\cdot\nabla^i \boldsymbol{v}(x), \tag{7.5.37}$$

where $\boldsymbol{v}$ and ϕ are functions that are sought to approximate the homogenized averages $\boldsymbol{u}_\varepsilon^*$ and p_ε^* respectively. Following the methodology of sections 7.3.3 and 7.4.2, we build well-posed homogenized equations of higher–but finite–order from a minimization principle.
Recall that the solution $\boldsymbol{u}_\varepsilon$ to the Stokes problem (7.5.1) is the unique minimizer of the constrained problem

$$\begin{aligned} \boldsymbol{u}_\varepsilon = \arg \min_{\boldsymbol{w}\in H^1(D_\varepsilon,\mathbb{R}^d)} \quad & J(\boldsymbol{w},\boldsymbol{f}) := \int_D \left(\frac{1}{2}\nabla\boldsymbol{w} : \nabla\boldsymbol{w} - \boldsymbol{f}\cdot\boldsymbol{w}\right) \mathrm{d}y \\ \text{s.t.} \quad & \begin{cases} \operatorname{div}(\boldsymbol{w}) = 0 \text{ in } D_\varepsilon, \\ \boldsymbol{w} = 0, \text{ on } \partial\omega_\varepsilon, \\ \boldsymbol{w} \text{ is } D\text{–periodic.} \end{cases} \end{aligned} \tag{7.5.38}$$

We consider the following minimization problem for the function $\boldsymbol{v}\in H^{K+1}(D,\mathbb{R}^d)$ sought to approximate $\boldsymbol{u}_\varepsilon^*$:

$$\begin{aligned} \min_{\boldsymbol{v}\in H^{K+1}(D,\mathbb{R}^d)} \quad & J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f}) \\ s.t. \quad & \begin{cases} \operatorname{div}(\boldsymbol{v}) = 0 \text{ in } D, \\ \boldsymbol{v} \text{ is } D\text{–periodic.} \end{cases} \end{aligned} \tag{7.5.39}$$

Applying lemma 7.3 to (7.5.39) in order to pass to the limit in the terms of $J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f})$ which depend on the oscillating variable x/ε , we obtain as in proposition 7.23, the existence of a functional J_K^* such that

$$\forall \boldsymbol{v}\in\mathcal{C}^\infty(D,\mathbb{R}^d),\ J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f}) = J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon) + o(\varepsilon^m)$$

holds with $m\in\mathbb{N}$ arbitrarily large. Following the derivations of section 7.4.2, the approximate energy $J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon)$ reads

$$J_K^*(\boldsymbol{v},\boldsymbol{f},\varepsilon) := \int_D \left(\frac{1}{2}\sum_{l,m=0}^{l+1} \varepsilon^{l+m-2}\mathbb{B}_K^{l,m}\nabla^l\boldsymbol{v}\nabla^m\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v}\right)\mathrm{d}x, \tag{7.5.40}$$

and the constant, matrix valued tensors $\mathbb{B}_K^{l,m}$ of order $l+m$ are defined by the formula

$$\mathbb{B}_{K,ij}^{l,m} := \int_Y \widetilde{\boldsymbol{N}}_{p,i}^l \cdot \widetilde{\boldsymbol{N}}_{p,j}^m \mathrm{d}y, \quad 1\le i,j\le d, \quad 0\le l,m\le l+1. \tag{7.5.41}$$

The vector valued tensors $\widetilde{\boldsymbol{N}}_{p,j}^l$ are still given as in (7.4.25) by

$$\widetilde{\boldsymbol{N}}_{p,j}^l(y) := \begin{cases} \partial_p \boldsymbol{N}_j^0(y) \text{ if } l=0 \\ \partial_p \boldsymbol{N}_j^l(y) + \boldsymbol{N}_j^{l-1}(y)\otimes e_p \text{ if } 1\le l\le K, \quad 1\le j,p\le d. \\ \boldsymbol{N}_j^K(y)\otimes e_p \text{ if } l=K+1. \end{cases} \tag{7.5.42}$$

Replacing $J(\boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}),\boldsymbol{f})$ by J_K^* in (7.5.39) allows us to derive a well-posed homogenized equation of order $2K+2$ for the Stokes problem:

Definition 7.8. For any $K \in \mathbb{N}$, we call homogenized equation of order $2K+2$ the Euler-Lagrange equation associated with the minimization problem

$$\begin{array}{ll} \min & J_K^*(\boldsymbol{v}, \boldsymbol{f}, \varepsilon) \\ \text{s.t.} & \begin{cases} \boldsymbol{v} \in H^{K+1}(D, \mathbb{R}^d), \\ \operatorname{div}(\boldsymbol{v}) = 0 \text{ in } D, \\ \boldsymbol{v} \text{ is } D\text{–periodic.} \end{cases} \end{array} \tag{7.5.43}$$

This equation reads explicitly in terms of a higher order homogenized solution $\boldsymbol{v}_K^*$ and a higher order homogenized pressure $q_K^* \in L^2(D)$ as:

$$\begin{cases} \displaystyle\sum_{k=0}^{2K+2} \varepsilon^{k-2} \mathbb{D}_K^k \cdot \nabla^k \boldsymbol{v}_K^* + \nabla q_K^* = \boldsymbol{f}, \\ \operatorname{div}(\boldsymbol{v}_K^*) = 0, \\ \displaystyle\int_D q_K^* \mathrm{d}x = 0, \\ \boldsymbol{v}_K^* \text{ is } D\text{–periodic,} \end{cases} \tag{7.5.44}$$

where the constant (matrix valued) tensors $\mathbb{D}_K^k$ are defined by the following formula for any $k \geq 0$:

$$\forall 1 \leq i,j \leq d,\ \mathbb{D}_{K,ij}^k := \begin{cases} \displaystyle\sum_{l=0}^{k} (-1)^l \frac{1}{2} (\mathbb{B}_{K,ij}^{l,k-l} + \mathbb{B}_{K,ji}^{l,k-l}), \text{ if } k \text{ is even} \\ \displaystyle\sum_{l=0}^{k} (-1)^l \frac{1}{2} (\mathbb{B}_{K,ij}^{l,k-l} - \mathbb{B}_{K,ji}^{l,k-l}), \text{ if } k \text{ is odd,} \end{cases} \tag{7.5.45}$$

where the convention $\mathbb{B}_K^{l,m} = 0$ for $l > K+1$ or $m > K+1$ is assumed.

Remembering $\mathbb{B}_K^{K+1,K+1}$ is nonnegative according to (7.5.41), adapting the proof of proposition 7.11 yields that the bilinear form associated with the energy (7.5.40) is symmetric and positive definite, under the additional non-degeneracy assumption

$$\exists \nu > 0,\ \forall \boldsymbol{\xi} = \boldsymbol{\xi}_{i_1 \dots i_{K+1}, j} \in \mathbb{R}^{d^{K+1}} \times \mathbb{R}^d,\ \mathbb{B}_K^{K+1,K+1} \boldsymbol{\xi}\boldsymbol{\xi} \geq \nu |\boldsymbol{\xi}|^2. \tag{7.5.46}$$

From there, standard theory for saddle point problems based on the zero divergence constraint (see e.g. the textbooks [304, 301, 171, 144]) yields existence and uniqueness of a solution for (7.5.44):

Proposition 7.35. *Assume the dominant tensor $\mathbb{D}_K^{2K+2} = (-1)^{K+1} \mathbb{B}_K^{K+1,K+1}$ is non-degenerate, that is there exists a constant $\nu > 0$ such that (7.5.46) holds. Then there exists a unique velocity and pressure couple $(\boldsymbol{v}_K^*, q_K^*) \in H^{K+1}(D, \mathbb{R}^d) \times L^2(D)/\mathbb{R}$ solving the higher order homogenized equation (7.5.44).*

Finally, our next proposition (which we shall use in the proof of our error estimates in the next section) states that (7.5.44) is indeed a "truncation" of the infinite order homogenized equation (7.5.29).

Proposition 7.36. *The first $K+1$ homogenized coefficients of the homogenized equation (7.4.29) coincide with those of the formal infinite order homogenized equation (7.4.17):*

$$\forall 0 \leq k \leq K,\ \mathbb{D}_K^k = M^k.$$

Proof. We follow the proof of proposition 7.12. For $0 \leq k,l \leq K$ and $1 \leq i,j \leq d$, the coefficient $\mathbb{B}_{K,ij}^{l,k-l}$ is given by

$$\begin{aligned} \mathbb{B}_{K,ij}^{l,k-l} &= \int_Y (\partial_p \boldsymbol{N}_i^l + \boldsymbol{N}_i^{l-1} \otimes e_p) \cdot (\partial_p \boldsymbol{N}_j^{k-l} + \boldsymbol{N}_j^{k-l-1} \otimes e_p) \mathrm{d}y \\ &= \int_Y (-\Delta_{yy} \boldsymbol{N}_i^l - 2\partial_p \boldsymbol{N}_i^{l-1} \otimes e_p - \boldsymbol{N}_i^{l-2} \otimes I) \cdot \boldsymbol{N}_j^{k-l} \mathrm{d}y \\ &\quad + \int_Y (\partial_p \boldsymbol{N}_i^{l-1} \cdot \boldsymbol{N}_j^{k-l} \otimes e_p + \partial_p \boldsymbol{N}_i^l \otimes \boldsymbol{N}_j^{k-l-1} \otimes e_p) \mathrm{d}y \\ &\quad + \int_Y (\boldsymbol{N}_i^{l-2} \cdot \boldsymbol{N}_j^{k-l} \otimes I + \boldsymbol{N}_i^{l-1} \cdot \boldsymbol{N}_j^{k-l-1} \otimes I) \mathrm{d}y. \end{aligned}$$

Using the identities (7.5.33) and performing an integration by part, we may write

$$\int_Y (-\Delta_{yy}\boldsymbol{N}_i^l - 2\partial_p \boldsymbol{N}_i^{l-1}\otimes e_p - \boldsymbol{N}_i^{l-2}\otimes I)\cdot \boldsymbol{N}_j^{k-l}\mathrm{d}y = \int_Y (-\nabla\beta_i^l - \beta_i^{l-1}\boldsymbol{e}_p\otimes e_p + M^l\boldsymbol{e}_i)\cdot \boldsymbol{N}_j^{k-l}\mathrm{d}y$$
$$= \int_Y [M^l\boldsymbol{e}_i\cdot \boldsymbol{N}_j^{k-l} - (\beta_i^l \boldsymbol{N}_j^{k-l-1} + \beta_i^{l-1}\boldsymbol{N}_j^{k-l})\cdot \boldsymbol{e}_p\otimes e_p]\mathrm{d}y.$$

This allows to rewrite $\mathbb{B}_{K,ij}^{l,k-l}$ as follows:

$$\mathbb{B}_{K,ij}^{l,k-l} = \int_Y M^l\boldsymbol{e}_i\cdot \boldsymbol{N}_j^{k-l}\mathrm{d}y + B_{ij}^{k,l} + B_{ij}^{k,l+1}$$

where $B^{k,l}$ is the k-th order tensor defined by

$$B^{k,l} := \int_Y \left[(\partial_p \boldsymbol{N}_i^{l-1}\cdot \boldsymbol{N}_j^{k-l} - \beta_i^{l-1}\boldsymbol{N}_j^{k-l}\cdot \boldsymbol{e}_p)\otimes e_p + \boldsymbol{N}_i^{l-2}\cdot \boldsymbol{N}_j^{k-l}\otimes I\right]\mathrm{d}y.$$

Recognizing a telescopic series, this allows to obtain

$$\sum_{l=0}^{k}(-1)^l \mathbb{B}_{K,ij}^{l,k-l} = (-1)^k M_{ji}^k$$

from where the result follows as in the proof of proposition 7.25. □

7.5.3 Error estimates and justifications of the higher order homogenization process

This section is devoted to establishing error estimates for the approximation of the oscillating solution $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ by means of the truncated two-scale ansatz (7.5.36) and (7.5.37) and of the higher order homogenized equation (7.5.44). We obtain error bounds for the classical and criminal truncated ansatz (eqn. (7.5.14), (7.5.16) and (7.5.32) respectively) in propositions 7.37 and 7.38. We then prove approximation results for the solution of the homogenized equation of order $2K+2$ (eqn. (7.5.39)) in proposition 7.39, which is our main result.

The mathematical analysis of the present situation is more difficult than the one of the perforated scalar and elasticity systems of the previous sections 7.3 and 7.4 because of the zero-divergence constraint (see e.g. [301, 304, 171, 12] for an overview of topics related to this matter); as a consequence, several of the previous arguments need to be adapted. Therefore, we start by stating a few technical results (more particularly, corollary 7.11) that we need in our proofs.

In this section, we use the following notation when referring to Sobolev spaces of D–periodic functions:

$$H^1_{per}(D_\varepsilon,\mathbb{R}^d) := \{\boldsymbol{v}\in H^1(D_\varepsilon,\mathbb{R}^d)\,|\,\boldsymbol{v} \text{ is } D\text{–periodic}\},$$

$$H^1_{per}(D,\mathbb{R}^d) := \{\boldsymbol{v}\in H^1(D,\mathbb{R}^d)\,|\,\boldsymbol{v} \text{ is } D\text{–periodic}\}.$$

The following lemma is due to [152], it establishes the existence of a continuous right inverse for the divergence B_ε—so-called a Bogovskii's operator—with a bound explicit in ε on the uniform continuity constant:

Lemma 7.9. *Assume (H4) and (H5). There exists a linear operator $B_\varepsilon : L^2(D_\varepsilon)\to H^1_{per}(D_\varepsilon,\mathbb{R}^d)$ and a constant C independent of ε satisfying, for any $\phi\in L^2(D_\varepsilon)$ such that $\int_{D_\varepsilon}\phi\mathrm{d}x = 0$:*

(i) $\mathrm{div}(B_\varepsilon\phi) = \phi$ *in* D_ε,

(ii) $B_\varepsilon\phi = 0$ *on* $\partial\omega_\varepsilon$,

(iii) $||\nabla(B_\varepsilon\phi)||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})}\leq C\varepsilon^{-1}||\phi||_{L^2(D_\varepsilon)}$, *for a constant $C>0$ independent of ϕ and ε.*

Proof. We follow the proof of [152], Lemma 2.1., which is also inspired by [14], Lemma 2.2.4. Let $\phi\in L^2(D_\varepsilon)$ and consider the extension $\widetilde{\phi}\in L^2(D_\varepsilon)$ defined by

$$\widetilde{\phi} = \begin{cases}\phi \text{ in } D_\varepsilon\\ 0 \text{ in } \omega_\varepsilon.\end{cases}$$

Let $B : L^2(D) \to H^1_{per}(D, \mathbb{R}^d)$ be the "classical" Bogovskii's operator in the non perforated domain D: there exists a constant $C > 0$ independent of ε and ϕ such that

$$\mathrm{div}(B\widetilde{\phi}) = \widetilde{\phi} \text{ in } D \tag{7.5.47}$$

$$||\nabla(B\widetilde{\phi})||_{L^2(D,\mathbb{R}^{d\times d})} \leq C||\widetilde{\phi}||_{L^2(D)} = C||\phi||_{L^2(D_\varepsilon)}. \tag{7.5.48}$$

From [12] (Theorem 2.3) (by adapting the proof to the periodicity with respect to D), there exists a linear *restriction* operator $q_\varepsilon : H^1_{per}(D, \mathbb{R}^d) \to H^1_{per}(D_\varepsilon, \mathbb{R}^d)$ satisfying the following properties:

1. $\forall \boldsymbol{v} \in H^1_{per}(D, \mathbb{R}^d)$, $q_\varepsilon \boldsymbol{v} = 0$ on $\partial\omega_\varepsilon$;
2. $\forall \boldsymbol{v} \in H^1_{per}(D, \mathbb{R}^d)$, $\boldsymbol{v} = 0$ in $\omega_\varepsilon \Rightarrow q_\varepsilon \boldsymbol{v} = \boldsymbol{v}$ in D_ε;
3. $\forall \boldsymbol{v} \in H^1_{per}(D, \mathbb{R}^d)$, $||q_\varepsilon \boldsymbol{v}||_{L^2(D_\varepsilon,\mathbb{R}^d)} + \varepsilon||\nabla(R_\varepsilon \boldsymbol{v})||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})} \leq C(||\boldsymbol{v}||_{L^2(D,\mathbb{R}^d)} + \varepsilon||\nabla \boldsymbol{v}||_{L^2(D,\mathbb{R}^{d\times d})})$ for a constant $C > 0$ independent of ε;
4. $\forall \boldsymbol{v} \in H^1_{per}(D, \mathbb{R}^d)$, $\mathrm{div}(\boldsymbol{v}) = 0$ in $D \Rightarrow \mathrm{div}(q_\varepsilon \boldsymbol{v}) = 0$ in D_ε.

Following an observation of [152] (see the proof of Lemma 2.1) and looking closely to the construction of the operator R_ε in [12], it can be shown that R_ε satisfies in fact the following property, which is stronger than (4):

(4*) $\forall \boldsymbol{v} \in H^1_{per}(D, \mathbb{R}^d)$, $\mathrm{div}(\boldsymbol{v}) = 0$ in $\omega_\varepsilon \Rightarrow \mathrm{div}(q_\varepsilon \boldsymbol{v}) = \mathrm{div}(\boldsymbol{v})$ in D_ε.

Following [152], we set

$$B_\varepsilon \phi := q_\varepsilon(B\widetilde{\phi}).$$

The resulting operator B_ε satisfies (i) to (iii). Indeed:

(i) $\mathrm{div}(B\widetilde{\phi}) = 0$ in ω_ε yields, according to (4*):

$$\mathrm{div}(q_\varepsilon(B\widetilde{\phi})) = \mathrm{div}(B\widetilde{\phi}) = \widetilde{\phi} = \phi \text{ in } D^\varepsilon;$$

(ii) is a straightforward consequence of the property (2);

(iii) is obtained by writing

$$\begin{aligned}
\varepsilon||\nabla(B_\varepsilon\phi)||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})} &= \varepsilon||\nabla(q_\varepsilon(B\widetilde{\phi}))||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})} \\
&\leq C(||B\widetilde{\phi}||_{L^2(D,\mathbb{R}^d)} + \varepsilon||\nabla(B\widetilde{\phi})||_{L^2(D,\mathbb{R}^{d\times d})}) \\
&\leq C||\nabla(B\widetilde{\phi})||_{L^2(D,\mathbb{R}^{d\times d})} \quad \text{(from the Poincaré inequality in } D) \\
&\leq C||\phi||_{L^2(D)} \text{ (from (7.5.48))}.
\end{aligned}$$

□

Let us remark that the previous lemma implies an estimation of the inf-sup constant associated with the perforated Stokes problem in terms of ε:

Corollary 7.10 (inf-sup constant for the perforated Stokes problem)**.** *There exists a constant $C > 0$ independent of ε such that*

$$\inf_{\substack{\phi\in L^2(D_\varepsilon) \\ \int_{D_\varepsilon}\phi \mathrm{d}x=0}} \sup_{\substack{\boldsymbol{v}\in H^1_{per}(D_\varepsilon,\mathbb{R}^d) \\ \boldsymbol{v}=0 \text{ on } \partial\omega_\varepsilon}} \int_{D_\varepsilon} \frac{\phi\,\mathrm{div}(\boldsymbol{v})}{||\phi||_{L^2(D_\varepsilon)}||\nabla\boldsymbol{v}||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})}}\mathrm{d}x \geq C\varepsilon.$$

Proof. The result is obtained by writing, for any $\phi \in L^2(D_\varepsilon)$ satisfying $\int_{D_\varepsilon}\phi \mathrm{d}x = 0$:

$$\begin{aligned}
\sup_{\substack{\boldsymbol{v}\in H^1_{per}(D_\varepsilon,\mathbb{R}^d) \\ \boldsymbol{v}=0 \text{ on } \partial\omega_\varepsilon}} \int_{D_\varepsilon} \frac{\phi\,\mathrm{div}(\boldsymbol{v})}{||\phi||_{L^2(D_\varepsilon)}||\nabla\boldsymbol{v}||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})}}\mathrm{d}x &\geq \int_{D_\varepsilon} \frac{\phi\,\mathrm{div}(B_\varepsilon\phi)}{||\phi||_{L^2(D)}||\nabla(B_\varepsilon\phi)||_{L^2(D_\varepsilon,\mathbb{R}^{d\times d})}}\mathrm{d}x \\
&\geq \int_{D_\varepsilon} \frac{\phi^2}{||\phi||_{L^2(D_\varepsilon)}C\varepsilon^{-1}||\phi||_{L^2(D_\varepsilon)}}\mathrm{d}x = \frac{1}{C}\varepsilon,
\end{aligned}$$

where the constant C above is that of the point (iii) of lemma 7.9. □

Corollary 7.11. *Assume (H4) and (H5). For any $\boldsymbol{h} \in L^2(D_\varepsilon, \mathbb{R}^d)$ and $g \in L^2(D_\varepsilon)$ satisfying $\int_{D_\varepsilon} g\mathrm{d}x = 0$, let $(\boldsymbol{v}, \phi) \in H^1(D_\varepsilon, \mathbb{R}^d) \times L^2(D_\varepsilon)$ be the unique solution to the Stokes problem*

$$\begin{cases} -\Delta \boldsymbol{v} + \nabla \phi = \boldsymbol{h} \text{ in } D_\varepsilon \\ \quad \mathrm{div}(\boldsymbol{v}) = g \text{ in } D_\varepsilon \\ \quad \displaystyle\int_{D_\varepsilon} \phi \mathrm{d}x = 0 \\ \quad\quad \boldsymbol{v} = 0 \text{ on } \partial\omega_\varepsilon \\ \quad\quad \boldsymbol{v} \text{ is } D\text{–periodic.} \end{cases} \tag{7.5.49}$$

There exists a constant C independent of ε, $\boldsymbol{h}$ and g such that

$$||\nabla \boldsymbol{v}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} + \varepsilon ||\phi||_{L^2(D_\varepsilon)} \leq C(\varepsilon ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + \varepsilon^{-1} ||g||_{L^2(D_\varepsilon)}), \tag{7.5.50}$$

Proof. We use the operator B_ε of lemma 7.9 to lift the divergence of $\boldsymbol{v}$. Let $\boldsymbol{w} := \boldsymbol{v} - B_\varepsilon g \in H^1_{per}(D_\varepsilon, \mathbb{R}^d)$, then it holds:

$$\begin{cases} \mathrm{div}(\boldsymbol{w}) = 0 \text{ in } D_\varepsilon, \\ \quad \boldsymbol{w} = 0 \text{ on } \partial\omega_\varepsilon. \end{cases}$$

After an integration by part, we obtain:

$$\begin{aligned} ||\nabla \boldsymbol{w}||^2_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} &= \int_{D_\varepsilon} \boldsymbol{h} \cdot \boldsymbol{w} \mathrm{d}x - \int_{D_\varepsilon} \nabla(B_\varepsilon g) : \nabla \boldsymbol{w} \mathrm{d}x \\ &\leq ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} ||\boldsymbol{w}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + ||\nabla(B_\varepsilon g)||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} ||\nabla \boldsymbol{w}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} \\ &\leq C(\varepsilon ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + ||\nabla(B_\varepsilon g)||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})}) ||\nabla \boldsymbol{w}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})}, \end{aligned}$$

where the last inequality is a consequence of lemma 7.2. Therefore, simplifying by $||\nabla \boldsymbol{w}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})}$ and using the point (iii) of lemma 7.9, we obtain

$$\begin{aligned} ||\nabla \boldsymbol{v}||_{L^2(D, \mathbb{R}^{d\times d})} &\leq ||\nabla \boldsymbol{w}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} + ||\nabla(B_\varepsilon g)||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} \\ &\leq C(\varepsilon ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + \varepsilon^{-1} ||g||_{L^2(D_\varepsilon)}), \end{aligned}$$

which proves the first part of the bound (7.5.50) on $\nabla \boldsymbol{v}$. The bound on the pressure is then classically obtained thanks to a reasoning analogous to that of the one of corollary 7.10. Using $B_\varepsilon \phi$ a test function, we obtain:

$$\begin{aligned} ||\phi||^2_{L^2(D_\varepsilon)} &= \int_{D_\varepsilon} \phi \mathrm{div}(B_\varepsilon \phi) \mathrm{d}x = -\int_{D_\varepsilon} \nabla \phi \cdot B_\varepsilon \phi \mathrm{d}x \\ &= \int_{D_\varepsilon} (-\Delta \boldsymbol{v} - \boldsymbol{h}) \cdot B_\varepsilon \phi \mathrm{d}x = \int_{D_\varepsilon} (\nabla \boldsymbol{v} \cdot \nabla(B_\varepsilon \phi) - \boldsymbol{h} \cdot B_\varepsilon \phi) \mathrm{d}x \\ &\leq ||\nabla \boldsymbol{v}||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} ||\nabla(B_\varepsilon \phi)||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} + ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} ||B_\varepsilon \phi||_{L^2(D_\varepsilon, \mathbb{R}^d)} \\ &\leq C(\varepsilon ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + \varepsilon^{-1} ||g||_{L^2(D_\varepsilon)}) ||\nabla(B_\varepsilon \phi)||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} \\ &\leq C\varepsilon^{-1}(\varepsilon ||\boldsymbol{h}||_{L^2(D_\varepsilon, \mathbb{R}^d)} + \varepsilon^{-1} ||g||_{L^2(D_\varepsilon)}) ||\phi||_{L^2(D_\varepsilon)}, \end{aligned}$$

whence the result. □

The previous lemma allows to establish a convergence result analogous to proposition 7.4 for the truncation of the "classical" ansatz (7.5.2), which are defined for any $K \in \mathbb{N}$ by

$$\boldsymbol{u}_{\varepsilon,K}(x) := \sum_{i=0}^{K} \varepsilon^{i+2} \boldsymbol{u}_i(x, x/\varepsilon), \quad p_{\varepsilon,K}(x) := \sum_{i=0}^{K} \varepsilon^i (p_i^*(x) + \varepsilon p_i(x, x/\varepsilon)). \tag{7.5.51}$$

We recall equations (7.5.5) and (7.5.12) for the definition of $\boldsymbol{u}_i$ and p_i. The next result is the extension of proposition 7.4 to the context of the Stokes system:

Proposition 7.37. *Assume (H4) and (H5). Let $(\boldsymbol{u}_\varepsilon, p_\varepsilon) \in H^1_{per}(D_\varepsilon, \mathbb{R}^d) \times L^2(D_\varepsilon)$ the solution to (7.5.1). Let $(\boldsymbol{u}_{\varepsilon,K}, p_{\varepsilon,K})$ the truncated ansatz at rank K defined by (7.5.51). There exists a constant $C_K > 0$ independent of ε and $\boldsymbol{f}$ such that*

$$||\boldsymbol{u}_\varepsilon - \boldsymbol{u}_{\varepsilon,K}||_{L^2(D, \mathbb{R}^d)} + \varepsilon ||\nabla(\boldsymbol{u}_\varepsilon - \boldsymbol{u}_{\varepsilon,K})||_{L^2(D_\varepsilon, \mathbb{R}^{d\times d})} \leq C_K \varepsilon^{K+3} ||\boldsymbol{f}||_{H^{K+3}(D, \mathbb{R}^d)}, \tag{7.5.52}$$

$$||p_\varepsilon - p_{\varepsilon,K}||_{L^2(D_\varepsilon)} \leq C_K \varepsilon^{K+1} ||\boldsymbol{f}||_{H^{K+3}(D, \mathbb{R}^d)}. \tag{7.5.53}$$

Proof. By using that $\boldsymbol{u}_i(x,y)$ and $p_i(x,y)$ are the solutions to the cascade of equations (7.5.5), we obtain, for any $K' \in \mathbb{N}$,

$$\begin{aligned}
&\forall x \in D_\varepsilon,\ (-\Delta \boldsymbol{u}_{\varepsilon,K'} + \nabla p_{\varepsilon,K'})(x) \\
&= \sum_{i=-2}^{K'-2} \varepsilon^{i+2}(-\Delta_{xx}\boldsymbol{u}_i - \Delta_{xy}\boldsymbol{u}_{i+1} - \Delta_{yy}\boldsymbol{u}_{i+2})(x,x/\varepsilon) + \sum_{i=-2}^{K'-2} \varepsilon^{i+2}(\nabla_x p^*_{i+2} + \nabla_x p_{i+1} + \nabla_y p_{i+2})(x,x/\varepsilon) \\
&\quad - (\varepsilon^{K'+1}\Delta_{xx}\boldsymbol{u}_{K'-1} + \varepsilon^{K'+2}\Delta_{xx}\boldsymbol{u}_{K'} + \varepsilon^{K'+1}\Delta_{xy}\boldsymbol{u}_{K'})(x,x/\varepsilon) + \varepsilon^{K'+1}(\nabla_x p_K)(x,x/\varepsilon) \\
&= \boldsymbol{f} - (\varepsilon^{K'+1}\Delta_{xx}\boldsymbol{u}_{K'-1} + \varepsilon^{K'+2}\Delta_{xx}\boldsymbol{u}_{K'} + \varepsilon^{K'+1}\Delta_{xy}\boldsymbol{u}_{K'} + \varepsilon^{K'+1}\nabla_x p_K)(x,x/\varepsilon).
\end{aligned}$$

$$\begin{aligned}
\operatorname{div}(\boldsymbol{u}_{\varepsilon,K'})(x) &= \sum_{i=-2}^{K'-2} \varepsilon^{i+3}(\operatorname{div}_x \boldsymbol{u}_{i+1} + \operatorname{div}_y \boldsymbol{u}_{i+2})(x,x/\varepsilon) + \varepsilon^{K'+2}(\operatorname{div}_x \boldsymbol{u}_{K'})(x,x/\varepsilon) \\
&= \varepsilon^{K'+2}(\operatorname{div}_x \boldsymbol{u}_{K'})(x,x/\varepsilon).
\end{aligned}$$

From standard regularity theory, it can be easily shown that

$$\begin{aligned}
\left|\left|(\varepsilon^{K'+1}\Delta_{xx}\boldsymbol{u}_{K'-1} + \varepsilon^{K'+2}\Delta_{xx}\boldsymbol{u}_{K'} + \varepsilon^{K'+1}\Delta_{xy}\boldsymbol{u}_{K'} + \varepsilon^{K'+1}\nabla_x p_K)(\cdot,\cdot/\varepsilon)\right|\right|_{L^2(D_\varepsilon,\mathbb{R}^d)} \\
\leq C_K \varepsilon^{K'+1} ||\boldsymbol{f}||_{H^{K'+2}(D,\mathbb{R}^d)},
\end{aligned}$$

$$||\varepsilon^{K'+2}(\operatorname{div}_x \boldsymbol{u}_{K'})(\cdot,\cdot/\varepsilon)||_{L^2(D_\varepsilon)} \leq C_K \varepsilon^{K'+2} ||\boldsymbol{f}||_{H^{K'+1}(D,\mathbb{R}^d)}.$$

Therefore, applying the result of corollary 7.11 to $(\boldsymbol{v}_\varepsilon, q_\varepsilon) := (\boldsymbol{u}_\varepsilon - \boldsymbol{u}_{\varepsilon,K'}, p_\varepsilon - p_{\varepsilon,K'})$ with $K' := K+1$, we infer the existence of a constant $C_K > 0$ such that

$$||\boldsymbol{u}_\varepsilon - \boldsymbol{u}_{\varepsilon,K+1}||_{H^1(D_\varepsilon,\mathbb{R}^d)} \leq C_K(\varepsilon^{K+3}||\boldsymbol{f}||_{H^{K+3}(D,\mathbb{R}^d)} + \varepsilon^{K+2}||\boldsymbol{f}||_{H^{K+2}(D,\mathbb{R}^d)})$$

$$||p_\varepsilon - p_{\varepsilon,K+1}||_{L^2(D_\varepsilon)} \leq C_K(\varepsilon^{K+2}||\boldsymbol{f}||_{H^{K+3}(D,\mathbb{R}^d)} + \varepsilon^{K+1}||\boldsymbol{f}||_{H^{K+2}(D,\mathbb{R}^d)}).$$

Finally, let us remark that $\boldsymbol{u}_{\varepsilon,K+1}$ and $p_{\varepsilon,K+1}$ are high order corrections of $\boldsymbol{u}_{\varepsilon,K}$ and $p_{\varepsilon,K}$:

$$||\boldsymbol{u}_{\varepsilon,K+1} - \boldsymbol{u}_{\varepsilon,K}||_{H^1(D_\varepsilon,\mathbb{R}^d)} = ||\varepsilon^{K+3}\boldsymbol{u}_{K+1}(\cdot,\cdot/\varepsilon)||_{H^1(D_\varepsilon,\mathbb{R}^d)} \leq C_K \varepsilon^{K+2}||\boldsymbol{f}||_{H^{K+1}(D,\mathbb{R}^d)}$$

$$||p_{\varepsilon,K+1} - p_{\varepsilon,K}||_{L^2(D)} = \left|\left|\varepsilon^{K+1}\left(p^*_{K+1} + \varepsilon p_{K+1}(\cdot,\cdot/\varepsilon)\right)\right|\right|_{L^2(D_\varepsilon)} \leq C_K \varepsilon^{K+1}||\boldsymbol{f}||_{H^{K+1}(D,\mathbb{R}^d)}.$$

The result follows by using the triangle inequality. □

Remark 7.29. The term $\varepsilon^{K+1}p_K(x,x/\varepsilon)$ can be removed in the truncated ansatz $p_{\varepsilon,K}$ of (7.5.51) because it is of order ε^{K+1} in the L^2 norm.

Remark 7.30. As a result of the scaling ε^{-1} in corollary 7.11, we pay a factor ε^{-1} in the error induced by the non zero divergence constraint. However we are able to obtain the right order of ε in the error estimates (7.5.52) and (7.5.53) thanks to the use of higher order terms. This phenomenon is quite classical in the truncation analysis of two-scale expansions where the higher order terms of the ansatz are used to establish the estimate and removed at the end (see e.g. [74, 19]).

Remark 7.31. In [12], the following convergence is obtained (in a setting involving much lower regularity for $\boldsymbol{f}$ and D):

$$P_\varepsilon \to p^*_0 \text{ in } L^2(D),$$

where P_ε is the extension of the pressure p_ε to the whole domain D defined as follows. Denote by $(\mathcal{P}^\varepsilon_\rangle)_\rangle$ and the collection of cells of size ε and by (Y^ε_i) their corresponding fluid part. The extension of the pressure P_ε is defined by:

$$P_\varepsilon = \begin{cases} p_\varepsilon \text{ in } D_\varepsilon, \\ \dfrac{1}{|Y^\varepsilon_i|}\displaystyle\int_{Y^i_\varepsilon} p_\varepsilon \mathrm{d}x \text{ in } \mathcal{P}^\varepsilon_\rangle \backslash \mathcal{Y}^\varepsilon_\rangle \text{ for each } \rangle. \end{cases} \tag{7.5.54}$$

This convergence results can be retrieved formally very naturally in view of the ansatz (7.5.2). Indeed, our estimates of proposition 7.37 yield

$$||p_\varepsilon - p^*_0||_{L^2(D_\varepsilon)} \leq C\varepsilon$$

for a constant independent of C. Furthermore, if $x \in Y_i^\varepsilon$, it can also be expected, with $q \geq 0$ as large as one likes:

$$\frac{1}{|Y_i^\varepsilon|}\int_{Y_\varepsilon^i} p_\varepsilon \mathrm{d}x \simeq \sum_{i=0}^{+\infty} \varepsilon^i \left(\frac{1}{|Y|}\int_Y p_i^*(x)\mathrm{d}y + \varepsilon \frac{1}{|Y|}\int_Y p_i(x,y)\mathrm{d}y \right) + o(\varepsilon^q) = p_0^*(x) + o(\varepsilon),$$

which suggests indeed that $P_\varepsilon \to p_0^*$ in the whole domain $D = D_\varepsilon \cup \omega_\varepsilon$.

We now show that $\boldsymbol{u}_\varepsilon$ and p_ε can be indeed approximated by criminal expansions of the form of (7.5.14) and (7.5.16); the following result is the analogous of lemma 7.5 for the present Stokes context.

Proposition 7.38. *Let $\boldsymbol{u}_{\varepsilon,K}^*, p_{\varepsilon,K}^*$ the averages of the truncated expansions $\boldsymbol{u}_{\varepsilon,K}$ and $p_{\varepsilon,K}$ of (7.5.51):*

$$\forall x \in D,\ \boldsymbol{u}_{\varepsilon,K}^*(x) := \sum_{k=0}^{K} \varepsilon^{k+2} \boldsymbol{u}_i^*(x), \quad p_{\varepsilon,K}^*(x) := \sum_{k=0}^{K} \varepsilon^k p_k^*(x).$$

There exists a constant $C_K(\boldsymbol{f})$ independent of ε such that

$$\left|\left| \boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{u}_{\varepsilon,K}^* \right|\right|_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+2} \tag{7.5.55}$$

$$\left|\left| p_\varepsilon - \left(p_{\varepsilon,K}^* + \sum_{k=0}^{K-1} \varepsilon^{k-1} \boldsymbol{\beta}^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{u}_{\varepsilon,K}^* \right) \right|\right|_{L^2(D)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+1}. \tag{7.5.56}$$

Proof. The proof of the error bound for (7.5.55) follows exactly the same lines of lemma 7.5.The result is obtained by using the estimates (7.5.52) and (7.5.53) together with the identities (7.5.31). □

Remark 7.32. We need only $K-1$ derivatives in the truncated criminal ansatz (7.5.56) for the pressure, because the term of highest order has a norm of order ε^K (recall $\boldsymbol{u}_{\varepsilon,K}^*$ has a norm of order ε^2).

We now turn to our main result which states that the solution $(\boldsymbol{v}_K^*, q_K^*)$ of the homogenized equation (7.5.39) of order $2K+2$ yields higher order approximations of $(\boldsymbol{u}_\varepsilon, p_\varepsilon)$ with error bounds similar to those of (7.5.55) and (7.5.56). Our proof is slightly different from that of proposition 7.13 because the variational framework is different and the treatment of the divergence constraint is more delicate. In order to prove our result, we need beforehand the following regularity estimate for the solution $(\boldsymbol{v}_K^*, q_K^*)$:

Lemma 7.10. *The solution $(\boldsymbol{v}_K^*, q_K^*)$ of (7.5.44) is smooth and for any $m \in \mathbb{N}$, there exists a constant $C_m(\boldsymbol{f})$ depending only on m and $\boldsymbol{f}$ such that*

$$||\boldsymbol{v}_{K^*}||_{H^m(D,\mathbb{R}^d)} \leq C_m(\boldsymbol{f}) \varepsilon^2.$$

Proof. This can be obtained by solving (7.5.44) explicitly with Fourier series. □

Proposition 7.39. *Let $(\boldsymbol{v}_K^*, q_K^*)$ be the unique solution to the homogenized equation (7.5.44) of order $2K+2$. There exists a constant $C_K(\boldsymbol{f})$ depending only on K and $\boldsymbol{f}$ (and a priori on the shape of the hole (ηT)) such that the following error estimates hold:*

$$\left|\left| \boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}_K^* \right|\right|_{L^2(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+3},$$

$$\left|\left| \boldsymbol{u}_\varepsilon - \sum_{k=0}^{K} \varepsilon^k N^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}_K^* \right|\right|_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+2},$$

$$\left|\left| p_\varepsilon - \left(q_K^* + \sum_{k=0}^{K-1} \varepsilon^{k-1} \boldsymbol{\beta}^k(\cdot/\varepsilon) \cdot \nabla^k \boldsymbol{v}_K^* \right) \right|\right|_{L^2(D)} \leq C_K(\boldsymbol{f}) \varepsilon^{K+1}.$$

Proof. Let us compute

$$\begin{aligned}-\Delta \boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*) = \sum_{k=0}^{K} \varepsilon^{k-2}(-\Delta \boldsymbol{N}_j^k - 2\partial_l \boldsymbol{N}_j^{k-1}\otimes e_l - \boldsymbol{N}_j^{k-2}\otimes I)(\cdot/\varepsilon)\cdot\nabla^k v_{K,j}^* \\ - \varepsilon^{K-1}(2\partial_l \boldsymbol{N}_j^K \otimes e_l + \boldsymbol{N}_j^{K-1}\otimes I)(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^* - \varepsilon^K \boldsymbol{N}_j^K(\cdot/\varepsilon)\otimes I\cdot\nabla^{K+2} v_{K,j}^* \\ - \varepsilon^{K+1}\Delta(\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^*),\end{aligned}$$

$$\nabla q_{\varepsilon,K}(q_K^*, \boldsymbol{v}_K^*) = \nabla q_K^* + \sum_{k=0}^{K} \varepsilon^{k-2}\boldsymbol{e}_l(\partial_l \beta_j^k + \beta_j^{k-1}\otimes e_l)(\cdot/\varepsilon)\cdot\nabla^k v_{K,j}^* + \varepsilon^{K-1}\boldsymbol{e}_l(\beta_j^K\otimes e_l)(\cdot/\varepsilon)\cdot\nabla^K v_{K,j}^*.$$

Summing the two equations and using (7.5.33), we obtain

$$\begin{aligned}-\Delta \boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}_K^*) + \nabla q_{\varepsilon,K}(q_K^*, \boldsymbol{v}_K^*) = \sum_{k=0}^{K}\varepsilon^{k-2} M^k\cdot\nabla^k \boldsymbol{v}_K^* + \nabla q_K^* \\ - \varepsilon^{K-1}(2\partial_l \boldsymbol{N}_j^K\otimes e_l + \boldsymbol{N}_j^{K-1}\otimes I)(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^* - \varepsilon^K \boldsymbol{N}_j^K(\cdot/\varepsilon)\otimes I\cdot\nabla^{K+2} v_{K,j}^* \\ - \varepsilon^{K+1}\Delta(\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^*).\end{aligned} \tag{7.5.57}$$

We now use proposition 7.36 and (7.5.44) to rewrite

$$\sum_{k=0}^{K}\varepsilon^{k-2} M^k\cdot\nabla^k \boldsymbol{v}_K^* + \nabla q_K^* = \boldsymbol{f} - \sum_{k=K+1}^{2K+2}\varepsilon^{k-2}\mathbb{D}_K^k\cdot\nabla^k \boldsymbol{v}_K^*. \tag{7.5.58}$$

Let us also compute the divergence of $\boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*)$:

$$\begin{aligned}\operatorname{div}(\boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*)) &= \sum_{k=0}^{K+1}\varepsilon^{k-1}(\operatorname{div}(\boldsymbol{N}_j^k)(\cdot/\varepsilon)\cdot\nabla^k v_{K,j}^* + (\boldsymbol{N}_j^{k-1}(\cdot/\varepsilon)\cdot\boldsymbol{e}_l\otimes e_l)\cdot\nabla^k v_{K,j}^*) \\ &\quad + \varepsilon^{K+1}\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\boldsymbol{e}_l\otimes\cdot\nabla^{K+2} v_{K,j}^* \\ &= \sum_{k=0}^{K+1}\varepsilon^{k-1}(-(\boldsymbol{N}_j^{k-1}(\cdot/\varepsilon) - <\boldsymbol{N}_j^{k-1}>)\cdot\boldsymbol{e}_l\otimes e_l\cdot\nabla^k v_{K,j}^* + (\boldsymbol{N}_j^{k-1}(\cdot/\varepsilon)\cdot\boldsymbol{e}_l\otimes e_l)\cdot\nabla^k v_{K,j}^*) \\ &\quad + \varepsilon^{K+1}\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\boldsymbol{e}_l\otimes e_l\cdot\nabla^{K+2} v_{K,j}^* \\ &= \operatorname{div}(\boldsymbol{v}_K^*) + \varepsilon^{K+1}\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\boldsymbol{e}_l\otimes e_l\cdot\nabla^{K+2} v_{K,j}^* = \varepsilon^{K+1}\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\otimes e_l\cdot\nabla^{K+2} v_{K,j}^*.\end{aligned} \tag{7.5.59}$$

We now apply corollary 7.11 to estimate $(\boldsymbol{v}_\varepsilon, q_\varepsilon) := (\boldsymbol{u}_\varepsilon - \boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*), p_\varepsilon - q_{\varepsilon,K}(\boldsymbol{v}_K^*, q_K^*))$. Equations (7.5.1) and (7.5.57) to (7.5.59) imply that $(\boldsymbol{v}_\varepsilon, q_\varepsilon)$ solves the following Stokes system:

$$\begin{cases} -\Delta \boldsymbol{v}_\varepsilon + \nabla q_\varepsilon = \boldsymbol{h}_\varepsilon \text{ in } D_\varepsilon \\ \operatorname{div}(\boldsymbol{v}_\varepsilon) = g_\varepsilon \in D_\varepsilon \\ \boldsymbol{v}_\varepsilon = 0 \text{ on } \partial\omega_\varepsilon, \\ \displaystyle\int_D q_\varepsilon = 0, \\ \boldsymbol{v}_\varepsilon \text{ is } D\text{–periodic}, \end{cases}$$

where the source functions $\boldsymbol{h}_\varepsilon$ and g_ε are given by

$$\begin{aligned}\boldsymbol{h}_\varepsilon := -\sum_{k=K+1}^{2K+2}\varepsilon^{k-2}\mathbb{D}_K^k\cdot\nabla^k \boldsymbol{v}_k^* - \varepsilon^{K-1}(2\partial_l \boldsymbol{N}_j^K\otimes e_l + \boldsymbol{N}_j^{K-1}\otimes I)(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^* \\ - \varepsilon^K \boldsymbol{N}_j^K(\cdot/\varepsilon)\otimes I\cdot\nabla^{K+2} v_{K,j}^* - \varepsilon^{K+1}\Delta(\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\cdot\nabla^{K+1} v_{K,j}^*),\end{aligned} \tag{7.5.60}$$

$$g_\varepsilon := \varepsilon^{K+1}\boldsymbol{N}_j^{K+1}(\cdot/\varepsilon)\otimes e_l\cdot\nabla^{K+2} v_{K,j}^*. \tag{7.5.61}$$

Using the result of lemma 7.10, we infer the existence of a constant $C_K(\boldsymbol{f})$ independent of ε such that

$$||\boldsymbol{h}_\varepsilon||_{L^2(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f})\varepsilon^{K+1},$$

$$||g_\varepsilon||_{L^2(D)} \leq C_K(\boldsymbol{f})\varepsilon^{K+3}.$$

Therefore, the estimates of corollary 7.11 yield

$$||\boldsymbol{u}_\varepsilon - \boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*)||_{H^1(D,\mathbb{R}^d)} = ||\boldsymbol{v}_\varepsilon||_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f})\varepsilon^{K+2},$$

$$||p_\varepsilon - q_{\varepsilon,K}(\boldsymbol{v}_K^*, q_K^*)||_{L^2(D)} = ||q_\varepsilon||_{L^2}(D) \leq C_K(\boldsymbol{f})\varepsilon^{K+1}.$$

The final result follows from the triangle's and the following estimates:

$$||\boldsymbol{w}_{\varepsilon,K+1}(\boldsymbol{v}_K^*) - \boldsymbol{w}_{\varepsilon,K}(\boldsymbol{v}_K^*)||_{H^1(D,\mathbb{R}^d)} = ||\varepsilon^{K+1}N^{K+1}\cdot\nabla^{K+1}\boldsymbol{v}_K^*||_{H^1(D,\mathbb{R}^d)} \leq C_K(\boldsymbol{f})\varepsilon^{K+2}.$$

$$||q_{\varepsilon,K}(\boldsymbol{v}_K^*, p_K^*) - q_{\varepsilon,K-1}(\boldsymbol{v}_K^*, q_K^*)||_{L^2(D)} = ||\varepsilon^{K-1}\boldsymbol{\beta}^K\cdot\nabla^K\boldsymbol{v}_K^*||_{L^2(D)} \leq C_K(\boldsymbol{f})\varepsilon^{K+1}.$$

□

Remark 7.33. The above proof further highlights that there is not a unique way to derive well-posed higher order homogenized equations such as (7.5.44): what is important is the fact that the first $K+1$ coefficients of the equation satisfied by $(\boldsymbol{v}_K^*, q_K^*)$ are those of the "infinite order" homogenized equation (7.5.29): $\mathbb{D}_K^k = M^k$ for $0 \leq k \leq K+1$. More precisely, the approximation result of proposition 7.39 holds provided lemma 7.10 holds and

$$\left|\left|\sum_{k=0}^{K}\varepsilon^{k-2}M^k\cdot\boldsymbol{v}_K^* + \nabla q_K^* - \boldsymbol{f}\right|\right| \leq C_K(\boldsymbol{f})\varepsilon^{K+1}.$$

The coefficients $\mathbb{D}_K^k$ for $K+1 \leq k \leq 2K+2$ are selected in such a way that (7.5.39) is well posed, which is the case when defining $\mathbb{D}_K^k$ from the energy minimization principle (7.5.43). In the next section 7.5.4, we provide evidences that (7.5.44) is a "well-behaved" equation when it is obtained from the minimization principle, because it contains all three classical homogenized regimes in the low-volume fraction limit where the obstacle's size $\eta \to 0$.

7.5.4 Low volume fraction limits when the size of the obstacle tends to 0

This section is devoted to the study of the asymptotics of the tensors $\mathcal{X}^{k*}$, M^k and $\mathbb{D}_K^k$ in the low volume fraction limit where the obstacle's size vanishes, i.e. $\eta \to 0$. In this whole subsection, we assume again, for simplicity, that the space dimension is greater than 3:

$$d \geq 3.$$

The notation convention is that of sections 7.3.5 and 7.4.3. We recall the inequalities in lemma 7.7 that we are going to use extensively. We also need the following technical result which yields estimates of Stokes solutions in the domain $\eta^{-1}P\backslash T$ uniform in $\eta \to 0$.

Lemma 7.11. *Consider $\boldsymbol{h} \in L^2(\eta^{-1}P\backslash T, \mathbb{R}^d)$ and $g \in L^2(\eta^{-1}P\backslash T)$ satisfying $\int_{\eta^{-1}P\backslash T} g\mathrm{d}x = 0$. Let $(\boldsymbol{v}, \phi) \in H^1(\eta^{-1}P\backslash T, \mathbb{R}^d) \times L^2(\eta^{-1}P\backslash T)$ be the unique solution to the following Stokes system:*

$$\begin{cases} -\Delta\boldsymbol{v} + \nabla\phi = \boldsymbol{h} \text{ in } \eta^{-1}P\backslash T \\ \mathrm{div}(\boldsymbol{v}) = g \text{ in } \eta^{-1}P\backslash T \\ \displaystyle\int_{\eta^{-1}P\backslash T}\phi\mathrm{d}x = 0 \\ \boldsymbol{v} = 0 \text{ on } \partial T \\ \boldsymbol{v} \text{ is } \eta^{-1}P\text{–periodic.} \end{cases} \tag{7.5.62}$$

There exists a constant $C > 0$ independent of η,$\boldsymbol{h}$ and g such that

$$||\nabla\boldsymbol{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} + ||\phi||_{L^2(\eta^{-1}P\backslash T)} \leq C(\eta^{-1}||\boldsymbol{h} - <\boldsymbol{h}>||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} + \eta^{-d}|<\boldsymbol{h}>| + ||g||_{L^2(\eta^{-1}P\backslash T)}). \tag{7.5.63}$$

Proof. From Lemma 2.2.4 in [14], for any $\eta > 0$, there exists a linear "Bogovskii's" operator $B_\eta : L^2(P\backslash(\eta T)) \to H^1(P\backslash(\eta T), \mathbb{R}^d)$ satisfying for any $\phi \in L^2(P\backslash(\eta Y))$ such that $\int_{P\backslash(\eta T)}\phi\mathrm{d}y = 0$:

(i) $\mathrm{div}(B_\eta\phi) = \phi$,

(ii) $B_\eta\phi = 0$ on $\partial(\eta T)$,

(iii) $B_\eta\phi$ is P–periodic,

(iv) $||\nabla(B_\eta\phi)||_{L^2(P\backslash(\eta T),\mathbb{R}^{d\times d})} \leq C||\phi||_{L^2(P\backslash(\eta T))}$ for a constant C independent of η and ϕ.

For any $\widetilde{\phi} \in L^2(\eta^{-1}P\backslash T)$ such that $\int_{\eta^{-1}P\backslash T}\widetilde{\phi}\mathrm{d}y = 0$, we define

$$\widetilde{B}_\eta(\widetilde{\phi}) := \eta^{-1}\big[B_\eta(\widetilde{\phi}(\eta^{-1}\,\cdot))(\eta\,\cdot)\big].$$

The operator $\widetilde{B}_\eta : L^2(\eta^{-1}P\backslash T) \to H^1(\eta^{-1}P\backslash T,\mathbb{R}^d)$ satisfies the following properties: for any $\widetilde{\phi} \in L^2(\eta^{-1}P\backslash T)$ such that $\int_{\eta^{-1}P\backslash T}\widetilde{\phi}\mathrm{d}x = 0$,

(i) $\mathrm{div}(\widetilde{B}_\eta\widetilde{\phi}) = \widetilde{\phi}$ in $\eta^{-1}P\backslash T$,

(ii) $\widetilde{B}_\eta\widetilde{\phi} = 0$ on ∂T,

(iii) $\widetilde{B}_\eta\widetilde{\phi}$ is $\eta^{-1}P$–periodic,

(iv) $||\nabla(\widetilde{B}_\eta\widetilde{\phi})||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \leq C||\widetilde{\phi}||_{L^2(\eta^{-1}P\backslash T)}$ for a constant C independent of η and ϕ.

The proof follows then classically along the lines of corollary 7.11. Upon an integration by parts and by using lemma 7.7, it is readily obtained with $\boldsymbol{w} := \boldsymbol{v} - \widetilde{B}_\eta g$:

$$\begin{aligned}
||\nabla\boldsymbol{w}||^2_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} &= \int_{\eta^{-1}P\backslash T}\boldsymbol{h}\cdot\boldsymbol{w}\mathrm{d}y \\
&= \int_{\eta^{-1}P\backslash T}(\boldsymbol{h}- <\boldsymbol{h}>)\cdot(\boldsymbol{w}- <\boldsymbol{w}>)\mathrm{d}y + \int_{\eta^{-1}P\backslash T} <\boldsymbol{h}> \cdot <\boldsymbol{w}> \mathrm{d}y \\
&\leq C\big(||\boldsymbol{h}- <\boldsymbol{h}> ||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)}||\boldsymbol{w}- <\boldsymbol{w}> ||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} + \eta^{-d}|<\boldsymbol{h}>||<\boldsymbol{w}>|\big) \\
&\leq C\big(\eta^{-1}||\boldsymbol{h}- <\boldsymbol{h}> ||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^d)} + \eta^{-d}|<\boldsymbol{h}>|\big)||\nabla\boldsymbol{w}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})}.
\end{aligned} \tag{7.5.64}$$

for a constant $C > 0$ independent of η and $\boldsymbol{h}$. This implies

$$\begin{aligned}
||\nabla\boldsymbol{v}||_{L^2(\eta^{-1}P\backslash T)} &\leq ||\nabla\boldsymbol{w}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} + ||\nabla(\widetilde{b}_\eta g||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} \\
&\leq C\big(||\nabla\boldsymbol{w}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})} + ||g||_{L^2(\eta^{-1}P\backslash T)}\big),
\end{aligned}$$

whence the bound on $||\nabla\boldsymbol{v}||_{L^2(\eta^{-1}P\backslash T,\mathbb{R}^{d\times d})}$ by using (7.5.64). The bound for the pressure is obtained by writing

$$\int_{\eta^{-1}P\backslash T}\phi^2\mathrm{d}x = \int_{\eta^{-1}P\backslash T}\phi\mathrm{div}(B_\eta\phi)\mathrm{d}x = -\int_{\eta^{-1}P\backslash T}\nabla\phi\cdot B_\eta\phi\mathrm{d}x = \int_{\eta^{-1}P\backslash T}(\nabla\boldsymbol{v} : \nabla(B_\eta\phi) - \boldsymbol{h}\cdot(B_\eta\phi))\mathrm{d}x,$$

from where (7.5.63) follows analogously. □

For any $1 \leq j \leq d$, let us introduce the unique solution $(\boldsymbol{\Psi}_j, \sigma_j)$ to the exterior Stokes problem

$$\begin{cases}
-\Delta\boldsymbol{\Psi}_j + \nabla\sigma_j = 0 \text{ in } \mathbb{R}^d\backslash T \\
\mathrm{div}(\boldsymbol{\Psi}_j) = 0 \text{ in } \mathbb{R}^d\backslash T \\
\boldsymbol{\Psi}_j = 0 \text{ on } \partial T \\
\boldsymbol{\Psi}_j \to \boldsymbol{e}_j \text{ at } \infty \\
\sigma_j \in L^2(\mathbb{R}^d\backslash T).
\end{cases} \tag{7.5.65}$$

The boundary condition $\boldsymbol{\Psi}_j \to \boldsymbol{e}_j$ at infinity must be understood in the sense that $\boldsymbol{\Psi}_j - \boldsymbol{e}_j$ belongs to the Deny-Lions space $\mathcal{D}^{1,2}(\mathbb{R}^d\backslash T,\mathbb{R}^d)$ (see definition 7.6). Similarly, the pressures σ_j are uniquely determined by the condition $\sigma_j \in L^2(\mathbb{R}^d\backslash T)$ (see e.g. Lemma 1.1, chapter V. of [163]). We denote by $\Psi^* := (\Psi^*_{ij})_{1\leq i,j\leq d}$ the matrix collecting the drag force components:

$$\Psi^*_{ij} := \int_{\mathbb{R}^d\backslash T}\nabla\boldsymbol{\Psi}_i : \nabla\boldsymbol{\Psi}_j\mathrm{d}x = -\int_{\partial T}\boldsymbol{e}_j\cdot(\nabla\boldsymbol{\Psi}_i - \sigma_i I)\cdot\boldsymbol{n}\mathrm{d}s, \tag{7.5.66}$$

where the normal $\boldsymbol{n}$ is pointing *inward* T. The asymptotics of $\boldsymbol{\mathcal{X}}^0$ and $\boldsymbol{\mathcal{X}}^{0*}$ have been obtained in of Theorem 3.1 in [13]. The following proposition extends propositions 7.14 and 7.27 to the Stokes system: it improves the result of [13] by providing asymptotics for the whole family of tensors $(\boldsymbol{\mathcal{X}}^k)_{k\in\mathbb{N}}$ and $(\mathcal{X}^{k*})_{k\in\mathbb{N}}$.

Proposition 7.40. *Assume $d \geq 3$. For any $k \geq 0$ and $1 \leq j \leq d$, denote by $(\widetilde{\boldsymbol{\mathcal{X}}}_j^{2k}, \widetilde{\alpha}_j^{2k})$ and $(\widetilde{\boldsymbol{\mathcal{X}}}_j^{2k+1}, \widetilde{\alpha}_j^{2k+1})$ the rescaled tensors in $\eta^{-1}P\backslash T$ defined as follows:*

$$\forall x \in \eta^{-1}P\backslash T, \begin{cases} \widetilde{\boldsymbol{\mathcal{X}}}_j^{2k}(x) := \eta^{(d-2)(k+1)}\boldsymbol{\mathcal{X}}_j^{2k}(\eta x) \\ \widetilde{\alpha}_j^{2k}(x) := \eta^{(d-2)(k+1)-1}\alpha_j^{2k}(\eta x) \end{cases}, \qquad \begin{cases} \widetilde{\boldsymbol{\mathcal{X}}}_j^{2k+1}(x) := \eta^{(d-2)(k+1)}\boldsymbol{\mathcal{X}}_j^{2k+1}(\eta x) \\ \widetilde{\alpha}_j^{2k+1}(x) := \eta^{(d-2)(k+1)-1}\alpha_j^{2k+1}(\eta x). \end{cases}$$

Then:

1. *there exists a constant C independent of $\eta > 0$ such that*

$$\forall \eta > 0,\ ||\nabla \widetilde{\boldsymbol{\mathcal{X}}}_j^{2k}||_{L^2(\eta^{-1}P\backslash T, \mathbb{R}^{d\times d})} + ||\widetilde{\alpha}_j^{2k}||_{L^2(\eta^{-1}P\backslash T)} \leq C,$$

$$\forall \eta > 0,\ ||\nabla \widetilde{\boldsymbol{\mathcal{X}}}_j^{2k+1}||_{L^2(\eta^{-1}P\backslash T, \mathbb{R}^{d\times d})} + ||\widetilde{\alpha}_j^{2k+1}||_{L^2(\eta^{-1}P\backslash T)} \leq C;$$

2. *the following convergences hold as $\eta \to 0$:*

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k}, \widetilde{\alpha}_i^{2k}) \rightharpoonup (c_{ij}^{2k}\boldsymbol{\Psi}_j, c_{ij}^{2k}\sigma_j) \qquad \text{weakly in } H^1_{loc}(\mathbb{R}^d\backslash T, \mathbb{R}^d) \times L^2_{loc}(\mathbb{R}^d\backslash T), \tag{7.5.67}$$

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1}, \widetilde{\alpha}_i^{2k+1}) \rightharpoonup (0,0) \qquad \text{weakly in } H^1_{loc}(\mathbb{R}^d\backslash T, \mathbb{R}^d) \times L^2_{loc}(\mathbb{R}^d\backslash T), \tag{7.5.68}$$

$$\mathcal{X}^{2k*} \sim \frac{1}{\eta^{(d-2)(k+1)}} c^{2k}, \tag{7.5.69}$$

$$\mathcal{X}^{2k+1*} = o\left(\frac{1}{\eta^{(d-2)(k+1)}}\right), \tag{7.5.70}$$

where c_{ij}^{2k} denote the coefficients of the $2k$-th order matrix valued tensor $c^{2k} := (c_{ij}^{2k})_{1\leq i,j\leq d}$ given by

$$c^{2k} := (\Psi^*)^{-(k+1)} I^{2k} \text{ with } I^{2k} = \overbrace{I \otimes I \otimes \cdots \otimes I}^{k \text{ times}}.$$

Proof. Following proposition 7.27, the result is proved by induction on k.

1. *Case $2k$ with $k = 0$.* The tensor $(\widetilde{\boldsymbol{\mathcal{X}}}_i^0, \widetilde{\alpha}_i^0)$ satisfies

$$\begin{cases} -\Delta \widetilde{\boldsymbol{\mathcal{X}}}_i^0 + \nabla \widetilde{\alpha}_i^0 = \eta^d \boldsymbol{e}_i \text{ in } \mathbb{R}^d\backslash T \\ \operatorname{div}(\widetilde{\boldsymbol{\mathcal{X}}}_i^0) = 0 \text{ in } \mathbb{R}^d\backslash T, \end{cases} \tag{7.5.71}$$

as well as the other boundary conditions of (7.5.62). Therefore lemma 7.11 yields

$$||\nabla \widetilde{\boldsymbol{\mathcal{X}}}_i^0||_{L^2(\eta^{-1}P\backslash T, \mathbb{R}^{d\times d})} + ||\widetilde{\alpha}_i^0||_{L^2(\eta^{-1}P\backslash T)} \leq C\eta^{-d}\eta^d |< \boldsymbol{e}_i >| \leq C.$$

From (7.3.66), we also obtain that $< \widetilde{\boldsymbol{\mathcal{X}}}_i^0 >$ is bounded. Hence, up to extracting a subsequence, there exists a constant matrix $c^0 := (c_{ij}^0)_{1\leq i,j\leq d}$, and fields $(\widehat{\boldsymbol{\Psi}}_i^0, \widehat{\sigma}_i^0)_{1\leq i\leq d}$ such that

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^0 > \cdot \boldsymbol{e}_j \to c_{ij}^0,$$

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^0, \widetilde{\alpha}_i^0) \rightharpoonup (\widehat{\boldsymbol{\Psi}}_i^0, \widehat{\sigma}_i^0) \text{ weakly in } H^1_{loc}(\eta^{-1}P\backslash T, \mathbb{R}^d) \times L^2_{loc}(\eta^{-1}P\backslash T).$$

Multiplying (7.5.71) by a compactly supported test function $\boldsymbol{\Phi} \in \mathcal{C}_c^\infty(\mathbb{R}^d\backslash T)$ and integrating by parts yields

$$\int_{\eta^{-1}P\backslash T} (\nabla \widetilde{\boldsymbol{\mathcal{X}}}_i^0 : \nabla \boldsymbol{\Phi} - \widetilde{\alpha}_i^0 \operatorname{div}(\boldsymbol{\Phi}))\mathrm{d}x = \int_{\eta^{-1}P\backslash T} \eta^d \boldsymbol{\Phi} \cdot \boldsymbol{e}_i \mathrm{d}x.$$

Passing to the limit as $\eta \to 0$ then implies

$$\begin{cases} -\Delta \widehat{\boldsymbol{\Psi}}_i^0 + \nabla \widehat{\sigma}_i^0 = 0 \text{ in } \mathbb{R}^d \backslash T \\ \operatorname{div}(\boldsymbol{\Psi}_i^0) = 0 \text{ in } \mathbb{R}^d \backslash T \\ \boldsymbol{\Psi}_i^0 = 0 \text{ on } \partial T. \end{cases}$$

Furthermore, by applying the point (7.3.68) of lemma 7.7 and by using the lower semi-continuity of the Lebesgue space norms, we infer $(\widehat{\boldsymbol{\Phi}}_i^0 - c_{ij}^0 \boldsymbol{e}_j, \widehat{\sigma}_i^0) \in \mathcal{D}^{1,2}(\mathbb{R}^d \backslash T) \times L^2(\mathbb{R}^d \backslash T)$. Hence, by linearity, it is necessary that $(\widehat{\Phi}_i^0, \widehat{\sigma}_i^0) = (c_{ij}^0 \boldsymbol{\Psi}_j, c_{ij}^0 \sigma_j)$ where $(\boldsymbol{\Psi}_j, \sigma_j)$ are the solution to the exterior problem (7.5.65). Integrating (7.5.65) by parts against the test function $\boldsymbol{\Phi} = \boldsymbol{e}_j$ then yields

$$0 = \eta^d \int_{\eta^{-1} P \backslash T} \delta_{ij} \mathrm{d}x + \int_{\partial T} \boldsymbol{e}_j \cdot (\nabla \widetilde{\boldsymbol{\mathcal{X}}}_i^0 - \widetilde{\alpha}_i^0 I) \cdot \boldsymbol{n} \mathrm{d}x.$$

Passing to the limit as $\eta \to 0$ by using the continuity of the drag force with respect to the weak convergence and (7.5.66) yields then

$$0 = \delta_{ij} + \int_{\partial T} \boldsymbol{e}_j \cdot (\nabla \widehat{\boldsymbol{\Phi}}_i^0 - \widehat{\sigma}_i^0) \cdot \boldsymbol{n} \mathrm{d}x = \delta_{ij} - c_{ip}^0 \Psi_{pj}^*.$$

This implies $c^0 = (\Psi^*)^{-1}$ as claimed, and the convergence of the whole sequence by uniqueness of the limit. The asymptotic for $\mathcal{X}^{0*}$ as $\eta \to 0$ is obtained by a simple change of variable $y = \eta x$:

$$\mathcal{X}_{ij}^{0*} = \boldsymbol{e}_i \cdot \int_{P \backslash (\eta T)} \boldsymbol{\mathcal{X}}_j^0 \mathrm{d}y = \eta^{2-d} \eta^d \boldsymbol{e}_i \cdot \int_{\eta^{-1} P \backslash T} \widetilde{\boldsymbol{\mathcal{X}}}_j^0 \mathrm{d}y \sim \eta^{2-d} < \widetilde{\boldsymbol{\mathcal{X}}}_j^0 > \cdot \boldsymbol{e}_i \sim \eta^{2-d} c_{ji}^0.$$

2. *Case* $2k+1$ *with* $k = 0$. The tensor $(\widetilde{\boldsymbol{\mathcal{X}}}_i^1, \widetilde{\alpha}_i^1)$ satisfies

$$\begin{cases} -\Delta \widetilde{\boldsymbol{\mathcal{X}}}_i^1 + \nabla \widetilde{\alpha}_i^1 = \eta (2 \partial_l \widetilde{\boldsymbol{\mathcal{X}}}_i^0 - \widetilde{\alpha}_j^0 \boldsymbol{e}_l) \otimes e_l \text{ in } \eta^{-1} P \backslash T \\ \operatorname{div}(\widetilde{\boldsymbol{\mathcal{X}}}_i^1) = -\eta (\widetilde{\boldsymbol{\mathcal{X}}}_j^0 - < \widetilde{\boldsymbol{\mathcal{X}}}_j^0 >) \cdot \boldsymbol{e}_l \otimes e_l \text{ in } \eta^{-1} P \backslash T. \end{cases} \tag{7.5.72}$$

Applying lemma 7.11 and remarking that $< 2 \partial_l \widetilde{\boldsymbol{\mathcal{X}}}_i^0 - \widetilde{\alpha}_i^0 \boldsymbol{e}_l >= 0$, we obtain

$$||\nabla \widetilde{\boldsymbol{\mathcal{X}}}_i^1||_{L^2(\eta^{-1} P \backslash T, \mathbb{R}^{d \times d})} + ||\widetilde{\alpha}_i^1||_{L^2(\eta^{-1} P \backslash T)} \leq C.$$

Integrating (7.5.72) by parts against a compactly supported test function $\boldsymbol{\Phi} \in \mathcal{C}^c(\mathbb{R}^d \backslash T)$ and passing to the limit as $\eta \to 0$, we obtain with similar arguments, the existence of a matrix valued tensor $c^1 := (c_{ij}^1))_{1 \leq i,j \leq d}$ (of order 1) such that, up to the extraction of a subsequence:

$$< \widetilde{\boldsymbol{\mathcal{X}}}_i^1 > \cdot \boldsymbol{e}_j \to c_{ij}^1,$$

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^1, \widetilde{\alpha}_i^1) \rightharpoonup (c_{ij}^1 \Psi_j, c_{ij}^1 \sigma_j) \text{ weakly in } H^1_{loc}(\eta^{-1} P \backslash T, \mathbb{R}^d) \times L^2_{loc}(\eta^{-1} P \backslash T).$$

Integrating (7.5.72) by parts against the test function $\boldsymbol{e}_j$ and passing to the limit as $\eta \to 0$ yields in this context

$$0 = c_{ij}^1 \Psi_{pj}^*$$

whence $c^1 = 0$.

3. *General case.* Assuming the result holds till rank k, we write the differential equations satisfied by the rescaled tensors:

$$\begin{cases} -\Delta \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} + \nabla \widetilde{\alpha}_i^{2k+2} = \eta^{d-1} (\partial_l \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1} - \widetilde{\alpha}_i^{2k+1} \boldsymbol{e}_l) \otimes e_l + \eta^d \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k} \otimes I & \text{in } \eta^{-1} P \backslash T \\ \operatorname{div}(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}) = -\eta^{d-1} (\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1} - < \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1} >) \cdot \boldsymbol{e}_l \otimes e_l & \text{in } \eta^{-1} P \backslash T. \end{cases} \tag{7.5.73}$$

$$\begin{cases} -\Delta \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3} + \nabla \widetilde{\alpha}_i^{2k+3} = \eta (\partial_l \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} - \widetilde{\alpha}_i^{2k+2} \boldsymbol{e}_l) \otimes e_l + \eta^d \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+1} \otimes I & \text{in } \eta^{-1} P \backslash T \\ \operatorname{div}(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3}) = -\eta (\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} - < \widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2} >) \cdot \boldsymbol{e}_l \otimes e_l & \text{in } \eta^{-1} P \backslash T. \end{cases} \tag{7.5.74}$$

Using lemma 7.7, lemma 7.11 and the point (1) of the proposition at rank k, we readily obtain

$$||\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}||_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^{d\times d})}+||\widetilde{\alpha}_i^{2k+2}||_{L^2(\eta^{-1}P\setminus T)}\leq C,$$

$$||\nabla\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3}||_{L^2(\eta^{-1}P\setminus T,\mathbb{R}^{d\times d})}+||\widetilde{\alpha}_i^{2k+3}||_{L^2(\eta^{-1}P\setminus T)}\leq C.$$

Repeating the above arguments, we obtain, up to the extraction of a subsequence, the existence of matrix valued tensors c^{2k+2} and c^{2k+3} such that

$$<\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2}>\cdot\boldsymbol{e}_j\rightharpoonup c_{ij}^{2k+2},\text{ and }<\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3}>\cdot\boldsymbol{e}_j\rightharpoonup c_{ij}^{2k+3},$$

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+2},\widetilde{\alpha}_i^{2k+2})\rightharpoonup(c_{ij}^{2k+2}\boldsymbol{\Psi}_j,c_{ij}^{2k+2}\sigma_j)\text{ weakly in }H^1_{loc}(\mathbb{R}^d\setminus T,\mathbb{R}^d)\times L^2_{loc}(\mathbb{R}^d\setminus T),$$

$$(\widetilde{\boldsymbol{\mathcal{X}}}_i^{2k+3},\widetilde{\alpha}_i^{2k+3})\rightharpoonup(c_{ij}^{2k+3}\boldsymbol{\Psi}_j,c_{ij}^{2k+3}\sigma_j)\text{ weakly in }H^1_{loc}(\mathbb{R}^d\setminus T,\mathbb{R}^d)\times L^2_{loc}(\mathbb{R}^d\setminus T).$$

The last step consists in integrating (7.5.73) and (7.5.74) by part against the test function $\boldsymbol{e}_j$ and to pass to the limit as $\eta\to 0$ to identify c_{ij}^{2k+2} and c_{ij}^{2k+3}. Performing this computation as above, we obtain

$$0=c_{ij}^{2k}\otimes I-c_{ip}^{2k+2}\Psi^*_{pj},$$

$$0=c_{ij}^{2k+1}\otimes I-c_{ip}^{2k+3}\Psi^*_{pj}$$

from where we infer $c^{2k+2}=c^{2k}(\Psi^*)^{-1}\otimes I$, $c^{2k+3}=c^{2k+1}(\Psi^*)^{-1}\otimes I$, hence the result (recall $c^1=0$ from the point (2) of the proof).

□

Using the identity (7.5.30), we obtain the asymptotics for the coefficients M^k of the infinite order homogenized equation (7.5.29).

Corollary 7.12. *Assume $d\geq 3$. The following convergences hold for the matrix valued tensors M^k as $\eta\to 0$:*

$$M^0\sim\eta^{d-2}\Psi^*, \tag{7.5.75}$$

$$M^1=o\left(\eta^{d-2}\right), \tag{7.5.76}$$

$$M^2\to -I, \tag{7.5.77}$$

$$\forall k\geq 1,\ M^{2k}=o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right), \tag{7.5.78}$$

$$\forall k\geq 1,\ M^{2k+1}=o\left(\frac{1}{\eta^{(d-2)(k-1)}}\right). \tag{7.5.79}$$

Proof. The proof is identical to that of corollary 7.6 by substituting A^2 with I. □

7.5.5 Simplifications for the tensors $\mathcal{X}^{k*}$ and M^k under symmetries

This final subsection generalizes the results of sections 7.3.6 and 7.4.4 to the Stokes system: we examine how the symmetries of the obstacle ηT with respect to the cell axes reflect into the coefficients of the matrix valued tensors $\mathcal{X}^{k*}$ and M^k. Our final result is stated in corollary 7.13, it is based on the following elementary lemma:

Lemma 7.12. *Let $S\in\mathbb{R}^{d\times d}$ an orthogonal symmetry, i.e. $S=S^T$ and $SS=I$. The following identities hold for any smooth vector field $\boldsymbol{\mathcal{X}}$ and scalar field α:*

$$-\Delta(S\boldsymbol{\mathcal{X}}\circ S)+\nabla(\alpha\circ S)=S(-\Delta\boldsymbol{\mathcal{X}}+\nabla\alpha)\circ S, \tag{7.5.80}$$

$$\operatorname{div}(S\boldsymbol{\mathcal{X}}\circ S)=\operatorname{div}(\boldsymbol{\mathcal{X}})\circ S, \tag{7.5.81}$$

$$\partial_i(S\boldsymbol{\mathcal{X}}\circ S)=S_{ij}S(\partial_j\boldsymbol{\mathcal{X}})\circ S. \tag{7.5.82}$$

Proof. The first two identities are obtained by writing

$$\begin{aligned}-\Delta(S\boldsymbol{\mathcal{X}}\circ S)+\nabla(\alpha\circ S)&=-S\partial_{ij}\boldsymbol{\mathcal{X}}\circ SS_{il}S_{jl}+S(\nabla\alpha)\circ S\\&=-S(\Delta\boldsymbol{\mathcal{X}}+\nabla\alpha)\circ S,\end{aligned}$$

$$\operatorname{div}(S\boldsymbol{\mathcal{X}}\circ S)=\operatorname{Tr}(\nabla(S\boldsymbol{\mathcal{X}}\circ S))=\operatorname{Tr}(S(\nabla\boldsymbol{\mathcal{X}})\circ SS)=\operatorname{Tr}((\nabla\boldsymbol{\mathcal{X}})\circ S)=\operatorname{div}(\boldsymbol{\mathcal{X}})\circ S.$$

Identity (7.5.82) is an elementary consequence of the chain rule. □

Proposition 7.41. *If the cell $Y = P\backslash(\eta T)$ is invariant with respect to a symmetry S, i.e. $S(Y) = Y$, then the following identity holds for the tensors $(\boldsymbol{\mathcal{X}}^k_l, \alpha^k_l)$ (defined in (7.5.6) to (7.5.8)):*

$$S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l} \circ S = S_{i_1j_1}\dots S_{i_kj_k}S_{lm}\boldsymbol{\mathcal{X}}^k_{j_1\dots j_k,m}, \tag{7.5.83}$$

$$\alpha^k_{i_1\dots i_k,l} \circ S = S_{i_1j_1}\dots S_{i_kj_k}S_{lm}\alpha^k_{j_1\dots j_k,m}. \tag{7.5.84}$$

As a consequence, the following identities hold for the constant matrix valued tensors $\mathcal{X}^{k*}$ and M^k:

$$\mathcal{X}^{k*}_{i_1\dots i_k,lm} = S_{i_1j_1}\dots S_{i_kj_k}S_{lp}S_{mq}\mathcal{X}^{k*}_{j_1\dots j_k,pq} \tag{7.5.85}$$

$$M^k_{i_1\dots i_k,lm} = S_{i_1j_1}\dots S_{i_kj_k}S_{lp}S_{mq}M^k_{j_1\dots j_k,pq}. \tag{7.5.86}$$

Proof. We follow the proof of proposition 7.28, that is we prove (7.5.83) and (7.5.84) by induction. Using section 7.5.5, we easily obtain

$$\begin{cases} -\Delta_{yy}(S\boldsymbol{\mathcal{X}}^0_l \circ S) + \nabla_y(\alpha^0_l \circ S) = S\boldsymbol{e}_l \circ S = S\boldsymbol{e}_l = S_{mj}\boldsymbol{e}_m, \\ \operatorname{div}(S\boldsymbol{\mathcal{X}}^0_l \circ S) = 0. \end{cases}$$

Since the cell is symmetric with respect to S, $(S\boldsymbol{\mathcal{X}}^0_l \circ S, \alpha^0_l \circ S)$ satisfies the same boundary conditions (7.5.9) than $S_{mj}(\boldsymbol{\mathcal{X}}^0_m, \alpha^0_m)$, therefore these are equal and we infer (7.5.83) and (7.5.84) at rank $k = 0$. We then write, for a given $1 \le i_1 \le d$:

$$\begin{cases} -\Delta_{yy}(S\boldsymbol{\mathcal{X}}^1_{i_1,l} \circ S) + \nabla_y(\alpha^1_{i_1,l} \circ S) = S(2\partial_{i_1}\boldsymbol{\mathcal{X}}^0_l - \alpha^0_l\boldsymbol{e}_{i_1}) \circ S \\ \qquad\qquad = S_{i_1j_1}(2\partial_{j_1}(S\boldsymbol{\mathcal{X}}^0_l \circ S) - \alpha^0_l \circ S\boldsymbol{e}_{j_1}) \\ \qquad\qquad = S_{i_1j_1}S_{lm}(2\partial_{j_1}\boldsymbol{\mathcal{X}}^0_m - \alpha^0_m\boldsymbol{e}_{j_1}), \\ \operatorname{div}_y(S\boldsymbol{\mathcal{X}}^1_{i_1,l} \circ S) = -(\boldsymbol{\mathcal{X}}^0_l \circ S - <\boldsymbol{\mathcal{X}}^0_l>) \cdot \boldsymbol{e}_{i_1} \\ \qquad\qquad = -S_{lm}S(\boldsymbol{\mathcal{X}}^0_m - <\boldsymbol{\mathcal{X}}^0_m>) \cdot \boldsymbol{e}_{i_1} \\ \qquad\qquad = -S_{i_1j_1}S_{lm}(\boldsymbol{\mathcal{X}}^0_m - <\boldsymbol{\mathcal{X}}^0_m>) \cdot \boldsymbol{e}_{j_1}, \end{cases}$$

where we used $<\boldsymbol{\mathcal{X}}^0_l> = <\boldsymbol{\mathcal{X}}^0_l \circ S>$. This implies similarly (7.5.83) and (7.5.84) at rank $k = 1$. Assuming now the result holds till rank $k+1$ with $k \ge 0$, we prove with the same arguments that it holds at rank $k+2$:

$$\begin{cases} -\Delta_{yy}(S\boldsymbol{\mathcal{X}}^{k+2}_{i_1\dots i_{k+2},l} \circ S) + \nabla_y(\alpha^{k+2}_{i_1\dots i_{k+2},l} \circ S) \\ \quad = S(2\partial_{i_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l} - \alpha^{k+1}_{i_1\dots i_{k+1},l}\boldsymbol{e}_{i_{k+2}}) \circ S + S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l} \circ S\delta_{i_{k+1}i_{k+2}} \\ \quad = S_{i_{k+2}j_{k+2}}(2\partial_{j_{k+2}}(S\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l} \circ S) - \alpha^{k+1}_{i_1\dots i_{k+1},l} \circ S\boldsymbol{e}_{j_{k+2}}) \\ \qquad + S_{i_{k+1}j_{k+1}}S_{i_{k+2}j_{k+2}}\delta_{j_{k+1}j_{k+2}}S\boldsymbol{\mathcal{X}}^k_{i_1\dots i_k,l} \circ S \\ \quad = S_{i_1j_1}\dots S_{i_{k+2}i_{k+2}}S_{lm}[(2\partial_{j_{k+2}}\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m} - \alpha^{k+1}_{j_1\dots j_{k+1},m}\boldsymbol{e}_{j_{k+2}}) + \delta_{j_{k+1}j_{k+2}}\boldsymbol{\mathcal{X}}^k_{j_1\dots j_k,m}] \\ \operatorname{div}_y(S\boldsymbol{\mathcal{X}}^{k+2} \circ S) = -(\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l} \circ S - <\boldsymbol{\mathcal{X}}^{k+1}_{i_1\dots i_{k+1},l}>)\dots\boldsymbol{e}_{i_{k+2}} \\ \quad = -S_{i_1j_1}\dots S_{i_{k+1}j_{k+1}}S_{lm}S(\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m} - <\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m}>) \cdot \boldsymbol{e}_{i_{k+2}} \\ \quad = -S_{i_1j_1}\dots S_{i_{k+2}j_{k+2}}S_{lm}(\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m} - <\boldsymbol{\mathcal{X}}^{k+1}_{j_1\dots j_{k+1},m}>) \cdot \boldsymbol{e}_{j_{k+2}}. \end{cases}$$

The identities (7.5.85) and (7.5.86) follow as in the proof of proposition 7.28. □

Corollary 7.13. *1. If the cell Y is symmetric with respect to all cell axes $(\boldsymbol{e}_l)_{1\le l\le d}$, then*

$$\mathcal{X}^{k*}_{i_1\dots i_k,pq} = 0 \text{ and } M^k_{i_1\dots i_k,pq} = 0$$

whenever any given integer $1 \le l \le d$ occurs an odd number of times in the indices $i_1\dots i_k, p, q$. In particular, this implies $\mathcal{X}^{2k+1*} = 0$ and $M^{2k+1} = 0$.

2. If the cell Y is symmetric with respect to all diagonal axes orthogonal to $(\boldsymbol{e}_l - \boldsymbol{e}_m)$, i.e. $S^{l,m}(Y) = Y$ for any $1 \le l < m \le d$, then for any permutation $\sigma \in \mathfrak{S}_d$,

$$\mathcal{X}^{k*}_{\sigma(i_1)\dots\sigma(i_k),\sigma(p)\sigma(q)} = \mathcal{X}^{k*}_{i_1\dots i_k,pq}.$$

$$M^k_{\sigma(i_1)\dots\sigma(i_k),\sigma(p)\sigma(q)} = M^k_{i_1\dots i_k,pq}.$$

Proof. The proof is identical to that of corollary 7.7. □

7.5.6 Appendix: extension to multicomponent fluid domains

Let us outline in this appendix how the previous methodologies could be extended to the case where the unit cell Y (a subset of the torus, i.e. opposite matching boundaries are identified) has m connected components $(Y_l)_{1\leq q\leq m}$ instead of one as assumed in (H4). Related models could be of interest for topology and shape optimization of multicomponent fluid systems, such as heat exchangers [255, 303].

In such a case the ansatz for the pressure $p_\varepsilon(x)$ of (7.5.2) must be modified as follows:

$$\boldsymbol{u}_\varepsilon(x)=\sum_{i=0}^{+\infty}\varepsilon^{i+2}\boldsymbol{u}_i(x,x/\varepsilon),\quad p_\varepsilon(x)=\sum_{i=0}^{+\infty}\varepsilon^i\left(\sum_{l=1}^{m}p_i^{*,l}(x)\mathbf{1}_{Y_l}(x/\varepsilon)+\varepsilon p_i(x,x/\varepsilon)\right) \tag{7.5.87}$$

where m homogenized pressures $p_i^{*,l}(x)$ are now involved, with $1\leq l\leq m$ and $i\in\mathbb{N}$. The uniqueness of the decomposition is obtained by imposing the average of p_i to be zero in every connected component:

$$\forall 1\leq l\leq m,\ \forall x\in D,\ \int_{Y_l}p_i(x,y)\mathrm{d}y=0,\qquad \int_D p_i^{*,l}(x)\mathrm{d}x=0.$$

This ansatz yields m homogenized pressures $(p_\varepsilon^{*,l})_{1\leq l\leq m}$ and velocity fields $(\boldsymbol{u}_\varepsilon^{*,l})_{1\leq l\leq m}$ for each of the corresponding connected fluid components:

$$\forall 1\leq l\leq m,\ \boldsymbol{u}_\varepsilon^{*,l}(x):=\sum_{i=0}^{+\infty}\varepsilon^{i+2}\int_{Y_l}\boldsymbol{u}_i(x,y)\mathrm{d}y,\quad p_\varepsilon^{*,l}(x):=\sum_{i=0}^{+\infty}\varepsilon^i p_i^{*,l}(x).$$

The analysis of section 7.5.1 can be easily adapted to this new context. A Fredholm alternative, analogous to (7.5.13), yields an elliptic second order equation determining the value of the m pressures $p_i^{*,l}$:

$$\mathrm{div}_x(\mathcal{X}^{0*,l}\nabla_x(\boldsymbol{f}_i-\nabla_x p_i^{*,l}))=-\sum_{k=1}^{i}\mathrm{div}(\mathcal{X}^{k*,l}\cdot\nabla^k(\boldsymbol{f}_{i-k}-\nabla_x p_{i-k}^{*,l})),\ \forall i\geq 0,\forall 1\leq l\leq m,$$

where $\mathcal{X}_{ij}^{0*,l}:=\int_{Y_l}\boldsymbol{\mathcal{X}}_j^0\cdot\boldsymbol{e}_i\mathrm{d}y$. Then the velocity and pressure $\boldsymbol{u}_i(x,y)$ and $p_i(x,y)$ of the ansatz (7.5.87) read

$$\boldsymbol{u}_i(x,y)=\sum_{k=0}^{i}\mathcal{X}^k(y)\cdot\nabla^k(\boldsymbol{f}_{i-k}(x)-\mathbf{1}_{Y_l}(y)\nabla p_{i-k}^{*,l}(x))$$

$$p_i(x,y)=\sum_{k=0}^{i}\boldsymbol{\alpha}^k(y)\cdot\nabla^k(\boldsymbol{f}_{i-k}(x)-\mathbf{1}_{Y_l}(y)\nabla p_{i-k}^{*,l}(x)),$$

where the summation over the repeated index l is assumed. These conclusions are not surprising: we have just verified that the homogenization theory for a fluid system with m components is that of m homogenized systems of single component (obtained e.g. by replacing $m-1$ fluid components with solid).

REFERENCES

[1] N. Aage, E. Andreassen, B. S. Lazarov, and O. Sigmund, *Giga-voxel computational morphogenesis for structural design*, Nature, 550 (2017), p. 84.

[2] T. Abatzoglou, *The metric projection on C2 manifolds in Banach spaces*, Journal of Approximation Theory, 26 (1979), pp. 204–211.

[3] A. Abdulle and T. Pouchon, *Effective models for the multidimensional wave equation in heterogeneous media over long time and numerical homogenization*, Mathematical Models and Methods in Applied Sciences, 26 (2016), pp. 2651–2684.

[4] A. Abdulle and T. Pouchon, *Effective models and numerical homogenization for wave propagation in heterogeneous media on arbitrary timescales*, arXiv preprint arXiv:1905.09062, (2019).

[5] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[6] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158.

[7] E. Acerbi, V. ChiadoPiat, G. Dal Maso, and D. Percivale, *An extension theorem from connected sets, and homogenization in general periodic domains*, Nonlinear Analysis: Theory, Methods & Applications, 18 (1992), pp. 481–496.

[8] M. F. Adams, H. H. Bayraktar, T. M. Keaveny, and P. Papadopoulos, *Ultrascalable implicit finite element analyses in solid mechanics with over a half a billion degrees of freedom*, in Proceedings of the 2004 ACM/IEEE conference on Supercomputing, IEEE Computer Society, 2004, p. 34.

[9] S.-H. Ahn and S. Cho, *Level set–based topological shape optimization of heat conduction problems considering design-dependent convection boundary*, Numerical Heat Transfer, Part B: Fundamentals, 58 (2010), pp. 304–322.

[10] J. Alexandersen, O. Sigmund, and N. Aage, *Large scale three-dimensional topology optimisation of heat sinks cooled by natural convection*, International Journal of Heat and Mass Transfer, 100 (2016), pp. 876–891.

[11] G. Allaire, *Homogénéisation des équations de Stokes et de Navier-Stokes*, PhD thesis, Université Paris 6, 1989.

[12] ——, *Homogenization of the Stokes flow in a connected porous medium*, Asymptotic Analysis, 2 (1989), pp. 203–222.

[13] G. Allaire, *Continuity of the Darcy's law in the low-volume fraction limit*, Annali della Scuola Normale Superiore di Pisa. Classe di Scienze. Serie IV, 18 (1991), pp. 475–499.

[14] G. Allaire, *Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes I. Abstract framework, a volume distribution of holes*, Archive for Rational Mechanics and Analysis, 113 (1991), pp. 209–259.

[15] G. Allaire, *Homogenization of the Navier-Stokes equations with a slip boundary condition*, Communications on pure and applied mathematics, 44 (1991), pp. 605–641.

[16] G. Allaire, *Homogenization and two-scale convergence*, SIAM Journal on Mathematical Analysis, 23 (1992), pp. 1482–1518.

[17] G. Allaire, *Conception optimale de structures*, vol. 58 of Mathématiques & Applications (Berlin), Springer-Verlag, Berlin, 2007.

[18] G. Allaire, *Shape optimization by the homogenization method*, vol. 146, Springer Science & Business Media, 2012.

[19] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM: Control, Optimisation and Calculus of Variations, 4 (1999), pp. 209–243.

[20] G. Allaire, M. Briane, and M. Vanninathan, *A comparison between two-scale asymptotic expansions and Bloch wave expansions for the homogenization of periodic structures*, SEMA journal, 73 (2016), pp. 237–259.

[21] G. Allaire, E. Cancès, and J.-L. Vié, *Second-order shape derivatives along normal trajectories, governed by Hamilton-Jacobi equations*, Structural and Multidisciplinary Optimization, 54 (2016), pp. 1245–1266.

[22] G. Allaire and C. Conca, *Bloch wave homogenization and spectral asymptotic analysis*, Journal de mathématiques pures et appliquées, 77 (1998), pp. 153–208.

[23] G. Allaire, C. Dapogny, G. Delgado, and G. Michailidis, *Multi-phase structural optimization via a level set method*, ESAIM: Control, Optimisation and Calculus of Variations, 20 (2014), pp. 576–611.

[24] G. Allaire, C. Dapogny, and P. Frey, *A mesh evolution algorithm based on the level set method for geometry and topology optimization*, Structural and Multidisciplinary Optimization, 48 (2013), pp. 711–715.

[25] ——, *Shape optimization with a level set based mesh evolution method*, Computer Methods in Applied Mechanics and Engineering, 282 (2014), pp. 22–53.

[26] G. Allaire, F. De Gournay, F. Jouve, and A.-M. Toader, *Structural optimization using topological and shape sensitivity via a level set method*, Control and cybernetics, 34 (2005), p. 59.

[27] G. Allaire, P. Geoffroy-Donders, and O. Pantz, *Topology optimization of modulated and oriented periodic microstructures by the homogenization method*, Computers & Mathematics with Applications, (2018).

[28] G. Allaire and L. Jakabcin, *Taking into account thermal residual stresses in topology optimization of structures built by additive manufacturing*, Mathematical Models and Methods in Applied Sciences, 28 (2018), pp. 2313–2366.

[29] G. Allaire, F. Jouve, and G. Michailidis, *Molding direction constraints in structural optimization via a level-set method*, in Variational Analysis and Aerospace Engineering, Springer, 2016, pp. 1–39.

[30] G. Allaire, F. Jouve, and G. Michailidis, *Thickness control in structural optimization via a level set method*, Structural and Multidisciplinary Optimization, 53 (2016), pp. 1349–1382.

[31] G. Allaire, F. Jouve, and A. Toader, *A level-set method for shape optimization*, Comptes Rendus Mathematique, 334 (2002), pp. 1125–1130.

[32] G. Allaire, F. Jouve, and A.-M. Toader, *Structural optimization using sensitivity analysis and a level-set method*, Journal of Computational Physics, 194 (2004), pp. 363–393.

[33] G. Allaire, A. Lamacz, and J. Rauch, *Crime Pays; Homogenized Wave Equations for Long Times*, arXiv preprint arXiv:1803.09455, (2018).

[34] G. Allaire and O. Pantz, *Structural optimization with FreeFem++*, Structural and Multidisciplinary Optimization, 32 (2006), pp. 173–181.

[35] G. Allaire and T. Yamada, *Optimization of dispersive coefficients in the homogenization of the wave equation in periodic structures*, Numerische Mathematik, 140 (2018), pp. 265–326.

[36] L. Ambrosio, *Geometric evolution problems, distance function and viscosity solutions*, in Calculus of variations and partial differential equations, Springer, 2000, pp. 5–93.

[37] N. Amenta, S. Choi, and R. K. Kolluri, *The power crust, unions of balls, and the medial axis transform*, Computational Geometry, 19 (2001), pp. 127–153.

[38] P. R. Amestoy, I. S. Duff, J. Koster, and J.-Y. L'Excellent, *A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling*, SIAM Journal on Matrix Analysis and Applications, 23 (2001), pp. 15–41.

[39] O. Amir, N. Aage, and B. S. Lazarov, *On multigrid-CG for efficient topology optimization*, Structural and Multidisciplinary Optimization, 49 (2014), pp. 815–829.

[40] S. Amstutz and H. Andrä, *A new algorithm for topology optimization using a level-set method*, Journal of Computational Physics, 216 (2006), pp. 573–588.

[41] M. Andersen, J. Dahl, and L. Vandenberghe, *CVXOPT: A Python package for convex optimization*, Available at http://cvxopt.org/, (2012).

[42] C. S. Andreasen and O. Sigmund, *Topology optimization of fluid–structure-interaction problems in poroelasticity*, Computer Methods in Applied Mechanics and Engineering, 258 (2013), pp. 55–62.

[43] E. Andreassen, A. Clausen, M. Schevenels, B. S. Lazarov, and O. Sigmund, *Efficient topology optimization in MATLAB using 88 lines of code*, Structural and Multidisciplinary Optimization, 43 (2011), pp. 1–16.

[44] S. Arguillère, E. Trélat, A. Trouvé, and L. Younes, *Shape deformation analysis from the optimal control viewpoint*, Journal de Mathématiques Pures et Appliquées. Neuvième Série, 104 (2015), pp. 139–178.

[45] D. Attali, J.-D. Boissonnat, and H. Edelsbrunner, *Stability and computation of medial axes-a state-of-the-art report*, in Mathematical foundations of scientific visualization, computer graphics, and massive data exploration, Springer, 2009, pp. 109–125.

[46] J.-L. Auriault, *On the domain of validity of brinkman?s equation*, Transport in Porous Media, 79 (2009), pp. 215–223.

[47] J.-L. Auriault, C. Geindreau, and C. Boutin, *Darcy's law, brinkman's law and poor separation of scales*, Poromechanics III: Biot Centennial (1905-2005) - Proceedings of the 3rd Biot Conference on Poromechanics, (2005), pp. 553–558.

[48] H. Azegami, S. K. M. Shimoda, and E. Katamine, *Irregularity of shape optimization problems and an improvement technique*, WIT Transactions on The Built Environment, 31 (1997).

[49] H. Azegami and Z. C. Wu, *Domain optimization analysis in linear elastic problems: approach using traction method*, JSME international journal. Ser. A, Mechanics and material engineering, 39 (1996), pp. 272–278.

[50] P. Azérad, *Equations de Navier-Stokes en bassin peu profond*, PhD thesis, Université de Neuchâtel, 1995.

[51] P. Azérad and J. Pousin, *Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport*, Comptes rendus de l'Académie des sciences. Série 1, Mathématique, 322 (1996), pp. 721–727.

[52] C. Bacuta, *A unified approach for Uzawa algorithms*, SIAM Journal on Numerical Analysis, 44 (2006), pp. 2633–2649.

[53] N. Bakhvalov and G. Panasenko, *Homogenisation: averaging processes in periodic media*, vol. 36 of Mathematics and its Applications (Soviet Series), Kluwer Academic Publishers Group, Dordrecht, 1989.

[54] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, V. Eijkhout, W. D. Gropp, D. Karpeyev, D. Kaushik, M. G. Knepley, D. A. May, L. C. McInnes, R. T. Mills, T. Munson, K. Rupp, P. Sanan, B. F. Smith, S. Zampini, H. Zhang, and H. Zhang, *PETSc Web page*. https://www.mcs.anl.gov/petsc, 2019.

[55] ——, *PETSc users manual*, Tech. Rep. ANL-95/11 - Revision 3.11, Argonne National Laboratory, 2019.

[56] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith, *Efficient management of parallelism in object oriented numerical software libraries*, in Modern Software Tools in Scientific Computing, E. Arge, A. M. Bruaset, and H. P. Langtangen, eds., Birkhäuser Press, 1997, pp. 163–202.

[57] N. V. Banichuk, V. G. Belcprime skiui, and V. V. Kobelev, *Optimization in problems of elasticity with unknown boundaries*, Izvestiya Akademii Nauk SSSR. Mekhanika Tverdogo Tela, 19 (1984), pp. 46–52.

[58] C. Barbarosie, *Shape optimization of periodic structures*, Computational Mechanics, 30 (2003), pp. 235–246.

[59] C. Barbarosie and S. Lopes, *A gradient-type algorithm for optimization with constraints*, submitted for publication, see also Pre-Print CMAF Pre-2011-001 at http://cmaf.fc.ul.pt/preprints.html, (2011).

[60] C. Barbarosie, S. Lopes, and A.-M. Toader, *A gradient-type algorithm for constrained optimization with applications to multi-objective optimization of auxetic materials*, arXiv preprint arXiv:1711.04863, (2017).

[61] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, vol. 17 of Mathématiques & Applications (Berlin), Springer-Verlag, Paris, 1994.

[62] V. A. Barry, B. A. Gregory, and N. Abuaf, *Turbine blade with enhanced cooling and profile optimization*, Nov. 9 1999. US Patent 5,980,209.

[63] R. Becker, V. Heuveline, and R. Rannacher, *An optimal control approach to adaptivity in computational fluid mechanics*, International journal for numerical methods in fluids, 40 (2002), pp. 105–120.

[64] G. Bellettini, *Lecture notes on mean curvature flow, barriers and singular perturbations*, vol. 12 of Appunti. Scuola Normale Superiore di Pisa (Nuova Serie), Edizioni della Normale, Pisa, 2013.

[65] M. P. Bendsøe and N. Kikuchi, *Generating optimal topologies in structural design using a homogenization method*, Computer methods in applied mechanics and engineering, 71 (1988), pp. 197–224.

[66] M. P. Bendsoe and O. Sigmund, *Topology optimization*, Springer-Verlag, Berlin, 2003. Theory, methods and applications.

[67] A. Bensalah, *Une approche nouvelle de la modélisation mathématique et numérique en aéroacoustique par les équations de Goldstein et applications en aéronautique*, PhD thesis, Université Paris Saclay, 2018.

[68] S. Bertoluzza, V. Chabannes, C. Prud'Homme, and M. Szopos, *Boundary conditions involving pressure for the Stokes problem and applications in computational hemodynamics*, Computer Methods in Applied Mechanics and Engineering, 322 (2017), pp. 58–80.

[69] L. T. Biegler, *Nonlinear programming*, vol. 10 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), 2010.

[70] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: theoretical and practical aspects*, Springer Science & Business Media, 2006.

[71] T. Borrvall and J. Petersson, *Large-scale topology optimization in 3D using parallel computing*, Computer methods in applied mechanics and engineering, 190 (2001), pp. 6201–6229.

[72] ——, *Topology optimization of fluids in Stokes flow*, International Journal for Numerical Methods in Fluids, 41 (2003), pp. 77–107.

[73] B. Bourdin and A. Chambolle, *Design-dependent loads in topology optimization*, ESAIM: Control, Optimisation and Calculus of Variations, 9 (2003), pp. 19–48.

[74] A. Bourgeat, E. Maruvsic-Paloka, and A. Mikelic, *Weak nonlinear corrections for Darcy's law*, Mathematical Models and Methods in Applied Sciences, 6 (1996), pp. 1143–1155.

[75] F. Boyer et al., *Trace theorems and spatial continuity properties for the solutions of the transport equation*, Differential and integral equations, 18 (2005), pp. 891–934.

[76] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*, Springer Science & Business Media, 2010.

[77] R. Bro and S. De Jong, *A fast non-negativity-constrained least squares algorithm*, Journal of Chemometrics: A Journal of the Chemometrics Society, 11 (1997), pp. 393–401.

[78] D. Bucur and J. Zolesio, *Shape analysis of eigenvalues on manifolds*, in Homogenization and applications to material sciences (Nice, 1995), vol. 9 of GAKUTO Internat. Ser. Math. Sci. Appl., Gakkotosho, Tokyo, 1995, pp. 67–79.

[79] D. Bucur and J.-P. Zolésio, *Optimisation de forme sous contrainte capacitaire*, Comptes Rendus de l'Académie des Sciences. Série I. Mathématique, 318 (1994), pp. 795–800.

[80] C. Bui, C. Dapogny, and P. Frey, *An accurate anisotropic adaptation method for solving the level set advection equation*, International Journal for Numerical Methods in Fluids, 70 (2012), pp. 899–922.

[81] M. Burger, *A framework for the construction of level set methods for shape optimization and reconstruction*, Interfaces and Free boundaries, 5 (2003), pp. 301–329.

[82] P. Cannarsa and P. Cardaliaguet, *Representation of equilibrium solutions to the table problem of growing sandpiles*, Journal of the European Mathematical Society, 6 (2004), pp. 435–464.

[83] P. Cannarsa, G. Da Prato, and J.-P. Zolésio, *Dynamical shape control of the heat equation*, Systems & Control Letters, 12 (1989), pp. 103–109.

[84] J. Céa, *Conception optimale ou identification de formes, calcul rapide de la dérivée directionnelle de la fonction coût*, ESAIM: Mathematical Modelling and Numerical Analysis, 20 (1986), pp. 371–402.

[85] J. Céa, S. Garreau, P. Guillaume, and M. Masmoudi, *The shape and topological optimizations connection*, Computer Methods in Applied Mechanics and Engineering, 188 (2000), pp. 713–726. IV WCCM, Part II (Buenos Aires, 1998).

[86] J. Céa, A. Gioan, and J. Michel, *Quelques résultats sur l'identification de domaines*, Calcolo. A Quarterly on Numerical Analysis and Theory of Computation, 10 (1973), pp. 207–232.

[87] J. Céa, A. Gioan, and J. Michel, *Adaptation de la méthode du gradient à un problème d'identification de domaine*, in Computing methods in applied sciences and engineering (Proc. Internat. Sympos., Versailles, 1973), Part 2, 1974, pp. 391–402. Lecture Notes in Comput. Sci., Vol. 11.

[88] S. Chen, M. Y. Wang, and A. Q. Liu, *Shape feature control in structural topology optimization*, Computer-Aided Design, 40 (2008), pp. 951–962.

[89] D. Chenais, *Optimal design of midsurface of shells: differentiability proof and sensitivity computation*, Applied Mathematics and Optimization, 16 (1987), pp. 93–133.

[90] D. Chenais, J. Monnier, and J.-P. Vila, *Shape optimal design for a fluid-heat coupled system*, Applied Mathematics and Computer Science, 6 (1996), pp. 245–261. Shape optimization and scientific computations (Warsaw, 1994).

[91] D. Chenais, J. Monnier, and J. P. Vila, *Shape optimal design problem with convective and radiative heat transfer: analysis and implementation*, Journal of Optimization Theory and Applications, 110 (2001), pp. 75–117.

[92] G. Cheng, Y. Mei, and X. Wang, *A feature-based structural topology optimization method*, in IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials, Springer, 2006, pp. 505–514.

[93] A. Cherkaev, *Variational methods for structural optimization*, vol. 140, Springer Science & Business Media, 2012.

[94] J. Chetboun, *Conception de formes aérodynamiques en présence d'écoulements décollés: contrôle et optimisation*, PhD thesis, École polytechnique, 2010.

[95] C. Chicone, *Ordinary Differential Equations with Applications*, Springer New York, 1999.

[96] D. L. Chopp, *Computing minimal surfaces via level set curvature flow*, Journal of Computational Physics, 106 (1993), pp. 77–91.

[97] D. L. Chopp, *Another look at velocity extensions in the level set method*, SIAM Journal on Scientific Computing, 31 (2009), pp. 3255–3273.

[98] A. N. Christiansen, J. A. Bærentzen, M. Nobel-Jørgensen, N. Aage, and O. Sigmund, *Combined shape and topology optimization of 3D structures*, Computers & Graphics, 46 (2015), pp. 25–35.

[99] A. N. Christiansen, M. Nobel-Jørgensen, N. Aage, O. Sigmund, and J. A. Bærentzen, *Topology optimization using an explicit interface representation*, Structural and Multidisciplinary Optimization, 49 (2014), pp. 387–399.

[100] C. P. Chukwudozie, *Shape Optimization for Drag Minimization Using the Navier-Stokes Equation*, Master's thesis, Louisiana State University, 2015.

[101] H. Chung, O. Amir, and H. A. Kim, *Nonlinear Thermoelastic Topology Optimization with the Level-Set Method*, in AIAA Scitech 2019 Forum, 2019, p. 1470.

[102] D. Cioranescu, A. Damlamian, and G. Griso, *The periodic unfolding method in homogenization*, SIAM Journal on Mathematical Analysis, 40 (2008), pp. 1585–1620.

[103] D. Cioranescu and F. Murat, *A strange term coming from nowhere*, in Topics in the mathematical modelling of composite materials, vol. 31 of Progr. Nonlinear Differential Equations Appl., Birkhäuser Boston, Boston, MA, 1997, pp. 45–93.

[104] P. Coffin and K. Maute, *Level set topology optimization of cooling and heating devices using a simplified convection model*, Structural and Multidisciplinary Optimization, 53 (2016), pp. 985–1003.

[105] C. Conca, F. Murat, and O. Pironneau, *The Stokes and Navier-Stokes equations with boundary conditions involving the pressure*, Japanese journal of mathematics. New series, 20 (1994), pp. 279–318.

[106] C. Conca, C. Pares, O. Pironneau, and M. Thiriet, *Navier-Stokes equations with imposed pressure and velocity fluxes*, International journal for numerical methods in fluids, 20 (1995), pp. 267–287.

[107] C. Dapogny, *Optimisation de formes, méthode des lignes de niveaux sur maillages non structurés et évolution de maillages*, PhD thesis, Université Pierre et Marie Curie-Paris VI, 2013.

[108] C. Dapogny, C. Dobrzynski, and P. Frey, *Three-dimensional adaptive domain remeshing, implicit domain meshing, and applications to free and moving boundary problems*, Journal of Computational Physics, 262 (2014), pp. 358–378.

[109] C. Dapogny, R. Estevez, A. Faure, and G. Michailidis, *Shape and topology optimization considering anisotropic features induced by additive manufacturing processes*, Hal preprint: `https://hal.archives-ouvertes.fr/hal-01660850/`, (2017).

[110] C. Dapogny, A. Faure, G. Michailidis, G. Allaire, A. Couvelas, and R. Estevez, *Geometric constraints for shape and topology optimization in architectural design*, Computational Mechanics, 59 (2017), pp. 933–965.

[111] C. Dapogny and P. Frey, *Computation of the signed distance function to a discrete contour on adapted triangulation*, Calcolo, 49 (2012), pp. 193–219.

[112] C. Dapogny, P. Frey, F. Omnès, and Y. Privat, *Geometrical shape optimization in fluid mechanics using FreeFem++*. HAL preprint hal-01481707, Mar. 2017.

[113] ——, *Geometrical shape optimization in fluid mechanics using FreeFem++*, Structural and Multidisciplinary Optimization, (2017), pp. 1–28.

[114] C. Dapogny, N. Lebbe, and E. Oudet, *Optimization of the shape of regions supporting boundary conditions*. working paper or preprint, Mar. 2019.

[115] G. David and S. Semmes, *Uniform rectifiability and singular sets*, Annales de l'Institut Henri Poincare (C) Non Linear Analysis, 13 (1996), pp. 383–443.

[116] T. Dbouk, *A review about the engineering design of optimal heat transfer systems using topology optimization*, Applied Thermal Engineering, 112 (2017), pp. 841–854.

[117] F. de Gournay, *Velocity extension for the level-set method and multiple eigenvalues in shape optimization*, SIAM journal on control and optimization, 45 (2006), pp. 343–367.

[118] E. M. Dede, *Multiphysics topology optimization of heat transfer and fluid flow systems*, in proceedings of the COMSOL Users Conference, 2009.

[119] M. Delfour and J. Zolésio, *Shape analysis via distance functions*, J. Funct. Anal, 123 (1994), pp. 129–201.

[120] M. Delfour and J.-P. Zolésio, *Shape identification via metrics constructed from the oriented distance function*, Control and Cybernetics, 34 (2005), pp. 137–164.

[121] M. C. Delfour, G. Payre, and J.-P. Zolésio, *Shape optimal design of a radiating fin*, in System modelling and optimization (Copenhagen, 1983), vol. 59 of Lect. Notes Control Inf. Sci., Springer, Berlin, 1984, pp. 810–818.

[122] M. C. Delfour and J.-P. Zolésio, *Shapes and geometries: metrics, analysis, differential calculus, and optimization*, SIAM, 2011.

[123] A. Dervieux, F. Courty, M. Vázquez, and B. Koobus, *Additive multilevel optimization and its application to sonic boom reduction*, in Numerical methods for scientific computing. Variational problems and applications, Internat. Center Numer. Methods Eng. (CIMNE), Barcelona, 2003, pp. 31–44.

[124] T. K. Dey and W. Zhao, *Approximate medial axis as a voronoi subcomplex*, in Proceedings of the seventh ACM symposium on Solid modeling and applications, ACM, 2002, pp. 356–366.

[125] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69, Springer Science & Business Media, 2011.

[126] L. Dieci and L. Lopez, *A survey of numerical methods for IVPs of ODEs with discontinuous right-hand side*, Journal of Computational and Applied Mathematics, 236 (2012), pp. 3967–3991.

[127] J. Dieudonné, *Foundations of modern analysis*, Academic press, New York and London, 1960.

[128] C. Dilgen, S. Dilgen, D. Fuhrman, O. Sigmund, and B. Lazarov, *Topology optimization of turbulent flows*, Computer Methods in Applied Mechanics and Engineering, 331 (2018), pp. 363–393.

[129] S. B. Dilgen, C. B. Dilgen, D. R. Fuhrman, O. Sigmund, and B. S. Lazarov, *Density based topology optimization of turbulent flow heat transfer systems*, Structural and Multidisciplinary Optimization, 57 (2018), pp. 1905–1918.

[130] Q. V. Dinh, G. Rogé, C. Sevin, and B. Stoufflet, *Shape optimization in computational fluid dynamics*, Revue Européenne des Éléments Finis. European Journal of Finite Elements, 5 (1996), pp. 569–594.

[131] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Inventiones mathematicae, 98 (1989), pp. 511–547.

[132] V. Dolean, P. Jolivet, and F. Nataf, *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation*, vol. 144, SIAM, 2015.

[133] X. Duan, Y. Ma, and R. Zhang, *Optimal shape control of fluid flow using variational level set method*, Physics letters A, 372 (2008), pp. 1374–1379.

[134] F. Dugast, Y. Favennec, C. Josset, Y. Fan, and L. Luo, *Topology optimization of thermal fluid flows with an adjoint Lattice Boltzmann Methodi*, Journal of Computational Physics, 365 (2018), pp. 376–404.

[135] P. D. Dunning and H. A. Kim, *Introducing the sequential linear programming level-set method for topology optimization*, Structural and Multidisciplinary Optimization, 51 (2015), pp. 631–643.

[136] J. Duoandikoetxea, *Forty years of Muckenhoupt weights*, Function Spaces and Inequalities, (2013), pp. 23–75.

[137] T. Dupont, M. F. Wheeler, et al., *A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems*, Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, 8 (1974), pp. 47–59.

[138] R. G. Durán and M. A. Muschietti, *An explicit right inverse of the divergence operator which is continuous in weighted norms*, Univ., 2000.

[139] M. C. Duta, S. Shahpar, and M. B. Giles, *Turbomachinery design optimization using automatic differentiated adjoint code*, in ASME Turbo Expo 2007: Power for Land, Sea, and Air, American Society of Mechanical Engineers, 2007, pp. 1435–1444.

[140] P. Duysinx, L. Van Miegroet, T. Jacobs, and C. Fleury, *Generalized shape optimization using X-FEM and level set methods*, in IUTAM Symposium on Topological Design Optimization of Structures, Machines and Materials, Springer, 2006, pp. 23–32.

[141] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM journal on Matrix Analysis and Applications, 20 (1998), pp. 303–353.

[142] B. Epstein, A. Jameson, S. Peigin, D. Roman, N. Harrison, and J. Vassberg, *Comparative study of three-dimensional wing drag minimization by different optimization techniques*, Journal of Aircraft, 46 (2009), pp. 526–541.

[143] B. Erem and D. H. Brooks, *Differential geometric approximation of the gradient and hessian on a triangulated manifold*, in Proceedings/IEEE International Symposium on Biomedical Imaging: from nano to macro, NIH Public Access, 2011, p. 504.

[144] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, vol. 159, Springer Science & Business Media, 2013.

[145] H. A. Eschenauer, V. V. Kobelev, and A. Schumacher, *Bubble method for topology and shape optimization of structures*, Structural optimization, 8 (1994), pp. 42–51.

[146] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[147] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, Studies in Advanced Mathematics. CRC Press, 1992.

[148] W. D. Evans, *Weighted Sobolev Spaces*, Bulletin of the London Mathematical Society, 18 (1986), pp. 220–221.

[149] R. D. Falgout and U. M. Yang, *hypre: A library of high performance preconditioners*, in International Conference on Computational Science, Springer, 2002, pp. 632–641.

[150] A. Faure, *Optimisation de forme de matériaux et structures architecturés par la méthode des lignes de niveaux avec prise en compte des interfaces graduées*, PhD thesis, Université Grenoble Alpes, 2017.

[151] A. Faure, G. Michailidis, G. Parry, N. Vermaak, and R. Estevez, *Design of thermoelastic multi-material structures with graded interfaces using topology optimization*, Structural and Multidisciplinary Optimization, 56 (2017), pp. 823–837.

[152] E. Feireisl and Y. Lu, *Homogenization of stationary Navier–Stokes equations in domains with tiny holes*, Journal of Mathematical Fluid Mechanics, 17 (2015), pp. 381–392.

[153] F. Feppon, G. Allaire, F. Bordeu, J. Cortial, and C. Dapogny, *Shape optimization of a coupled thermal fluid–structure problem in a level set mesh evolution framework*, SeMA Journal, (2019), pp. 1–46.

[154] F. Feppon, G. Allaire, and C. Dapogny, *A variational formulation for computing shape derivatives of geometric constraints along rays*, To appear in M2AN, (2019).

[155] F. Feppon, G. Allaire, and C. Dapogny, *Null space gradient flows for constrained optimization with applications to shape optimization*, Submitted, (2019).

[156] F. Feppon and P. Lermusiaux, *The Extrinsic Geometry of Dynamical Systems Tracking Nonlinear Matrix Projections*, SIAM Journal on Matrix Analysis and Applications, 40 (2019), pp. 814–844.

[157] F. Feppon and P. F. Lermusiaux, *A geometric approach to dynamical model order reduction*, SIAM Journal on Matrix Analysis and Applications, 39 (2018), pp. 510–538.

[158] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides: Control Systems*, vol. 18 of Mathematics and its applications, Springer Netherlands, 1988.

[159] M. Firdaouss, J.-L. Guermond, and P. Le Quéré, *Nonlinear corrections to darcy's law at low reynolds numbers*, Journal of Fluid Mechanics, 343 (1997), pp. 331–350.

[160] Y. Fischer, B. Marteau, and Y. Privat, *Some inverse problems around the Tokamak Tore Supra*, Communications on Pure and Applied Analysis, 11 (2012), pp. 2327–2349.

[161] R. Fletcher, *Practical methods of optimization*, John Wiley & Sons, 2013.

[162] R. Fransen, *LES based aerothermal modeling of turbine blade cooling systems*, PhD thesis, Université de Toulouse, 2013.

[163] G. P. Galdi, *Steady Stokes Flow in Exterior Domains*, Springer New York, New York, NY, 1994, pp. 244–303.

[164] H. Garcke, C. Hecht, M. Hinze, C. Kahle, and K. F. Lam, *Shape optimization for surface functionals in Navier-Stokes flow using a phase field approach*, Interfaces and Free Boundaries. Mathematical Analysis, Computation and Applications, 18 (2016), pp. 219–261.

[165] S. Garreau, P. Guillaume, and M. Masmoudi, *The topological asymptotic for PDE systems: the elasticity case*, SIAM Journal on Control and Optimization, 39 (2001), pp. 1756–1778.

[166] P. Geoffroy Donders, *Homogenization method for topology optmization of structures built with lattice materials.*, theses, Université Paris-Saclay, Dec. 2018.

[167] C. Geuzaine and J.-F. Remacle, *Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities*, International journal for numerical methods in engineering, 79 (2009), pp. 1309–1331.

[168] M. Giacomini, O. Pantz, and K. Trabelsi, *Volumetric expressions of the shape gradient of the compliance in structural shape optimization*, arXiv preprint arXiv:1701.05762, (2017).

[169] R. Giannoni, *Gas-liquid heat exchanger and method for its manufacture*, Oct. 7 1999. European Patent 1 098 156 A1.

[170] F. Gibou, R. Fedkiw, and S. Osher, *A review of level-set methods and some recent applications*, Journal of Computational Physics, 353 (2018), pp. 82 – 109.

[171] V. Girault and P.-A. Raviart, *Finite element approximation of the Navier-Stokes equations*, vol. 749, Springer Science & Business Media, 1979.

[172] R. Glowinski and O. Pironneau, *Towards the computation of minimum drag profiles in viscous laminar flow*, Applied Mathematical Modelling, 1 (1976/77), pp. 58–66.

[173] G. H. Golub and C. F. Van Loan, *Matriz computations*, The Johns Hopkins University Press, Baltimore, USA, (1989).

[174] F. A. Gomes and T. A. Senne, *An SLP algorithm and its application to topology optimization*, Computational & Applied Mathematics, 30 (2011).

[175] J. S. Gray, J. T. Hwang, J. R. R. A. Martins, K. T. Moore, and B. A. Naylor, *OpenMDAO: an open-source framework for multidisciplinary design, analysis, and optimization*, Structural and Multidisciplinary Optimization, 59 (2019), pp. 1075–1104.

[176] P. Grisvard, *Elliptic problems in nonsmooth domains*, vol. 69 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[177] J. P. Groen and O. Sigmund, *Homogenization-based topology optimization for high-resolution manufacturable microstructures*, International Journal for Numerical Methods in Engineering, 113 (2018), pp. 1148–1163.

[178] J. K. Guest, *Imposing maximum length scale in topology optimization*, Structural and Multidisciplinary Optimization, 37 (2009), pp. 463–473.

[179] P. Guillaume and M. Masmoudi, *Calcul numérique des dérivées d'ordre supérieur en conception optimale de formes*, Comptes Rendus de l'Académie des Sciences. Série I. Mathématique, 316 (1993), pp. 1091–1096.

[180] J. Hadamard, *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, vol. 33, Imprimerie nationale, 1908.

[181] S. Harries, C. Abt, and M. Brenner, *Upfront CAD—Parametric modeling techniques for shape optimization*, in Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences, Springer, 2019, pp. 191–211.

[182] J. Haslinger and P. Neittaanmäki, *Shape optimization in contact problems. Approximation and numerical realization*, RAIRO Modélisation Mathématique et Analyse Numérique, 21 (1987), pp. 269–291.

[183] F. Hecht, *New development in FreeFem++*, Journal of Numerical Mathematics, 20 (2012), pp. 251–265.

[184] A. Henrot and M. Pierre, *Variation et optimisation de formes: une analyse géométrique*, vol. 48, Springer Science & Business Media, 2006.

[185] ——, *Shape variation and optimization*, vol. 28 of EMS Tracts in Mathematics, European Mathematical Society (EMS), Zürich, 2018.

[186] A. Henrot and Y. Privat, *What is the optimal shape of a pipe?*, Archive for rational mechanics and analysis, 196 (2010), pp. 281–302.

[187] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM Journal on Optimization, 13 (2002), pp. 865–888.

[188] R. Hiptmair, A. Paganini, and S. Sargheini, *Comparison of approximate shape gradients*, BIT Numerical Mathematics, 55 (2015), pp. 459–485.

[189] G. P. Huang, D. Doman, M. Oppenheimer, A. Tipton, and D. Sigthorsson, *Topology optimization of a fuel thermal management system*, in AIAA Aviation 2019 Forum, 2019, p. 3471.

[190] K. Ito, H. T. Tran, and J. S. Scroggs, *Mathematical issues in optimal design of a vapor transport reactor*, in Flow control (Minneapolis, MN, 1992), vol. 68 of IMA Vol. Math. Appl., Springer, New York, 1995, pp. 197–218.

[191] A. Jameson, *Aerodynamic design via control theory*, in Recent advances in computational fluid dynamics (Princeton, NJ, 1988), vol. 43 of Lecture Notes in Engrg., Springer, Berlin, 1989, pp. 377–401.

[192] ——, *Computational algorithms for aerodynamic analysis and design*, Applied Numerical Mathematics. An IMACS Journal, 13 (1993), pp. 383–422.

[193] I. G. Jang and B. M. Kwak, *Evolutionary topology optimization using design space adjustment based on fixed grid*, International Journal for Numerical Methods in Engineering, 66 (2006), pp. 1817–1840.

[194] N. Jenkins and K. Maute, *Level set topology optimization of stationary fluid-structure interaction problems*, Structural and Multidisciplinary Optimization, 52 (2015), pp. 179–195.

[195] M. Jensen, *Discontinuous Galerkin methods for Friedrichs systems with irregular solutions*, PhD thesis, University of Oxford, 2005.

[196] W. Jing, *A unified homogenization approach for the Dirichlet problem in Perforated Domains*, arXiv preprint arXiv:1901.08251, (2019).

[197] C. Jog, *Distributed-parameter optimization and topology design for non-linear thermoelasticity*, Computer Methods in Applied Mechanics and Engineering, 132 (1996), pp. 117–134.

[198] P. Jolivet, *FreeFEM tutorial*. http://jolivet.perso.enseeiht.fr/FreeFem-tutorial/, Apr 2019.

[199] P. Jolivet, F. Hecht, F. Nataf, and C. Prud'Homme, *Scalable domain decomposition preconditioners for heterogeneous elliptic problems*, Scientific Programming, 22 (2014), pp. 157–171.

[200] H. T. Jongen and O. Stein, *On the complexity of equalizing inequalities*, Journal of Global Optimization, 27 (2003), pp. 367–374.

[201] H. T. Jongen and O. Stein, *Constrained global optimization: adaptive gradient flows*, in Frontiers in global optimization, vol. 74 of Nonconvex Optim. Appl., Kluwer Acad. Publ., Boston, MA, 2004, pp. 223–236.

[202] S. Kambampati, C. Jauregui, K. Museth, and H. A. Kim, *Fast level set topology optimization using a hierarchical data structure*, in 2018 Multidisciplinary Analysis and Optimization Conference, 2018, p. 3881.

[203] D.-H. Kim, S.-H. Lee, I.-H. Park, and J.-H. Lee, *Derivation of a general sensitivity formula for shape optimization of 2-D magnetostatic systems by continuum approach*, IEEE Transactions on Magnetics, 38 (2002), pp. 1125–1128.

[204] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifolds*, Proceedings of the national academy of Sciences, 95 (1998), pp. 8431–8435.

[205] T. Kondoh, T. Matsumori, and A. Kawamoto, *Drag minimization and lift maximization in laminar flows via topology optimization employing simple objective function expressions based on body force integration*, Structural and Multidisciplinary Optimization, 45 (2012), pp. 693–701.

[206] A. Koshakji, A. Quarteroni, and G. Rozza, *Free form deformation techniques applied to 3D shape optimization problems*, Communications in Applied and Industrial Mathematics, 4 (2013), pp. e452, 26.

[207] T. Krainer and B. Schulze, *Weighted Sobolev spaces*, Springer, 1985.

[208] S. Kreissl, G. Pingen, A. Evgrafov, and K. Maute, *Topology optimization of flexible microfluidic devices*, Structural and Multidisciplinary Optimization, 42 (2010), pp. 495–516.

[209] A. Kufner and B. Opic, *How to define reasonably weighted Sobolev spaces*, Commentationes Mathematicae Universitatis Carolinae, 25 (1984), pp. 537–554.

[210] O. A. Ladyzhenskaya, *The mathematical theory of viscous incompressible flow*, vol. 2, Gordon and Breach New York, 1969.

[211] A. B. Lambe and J. R. R. A. Martins, *Matrix-free aerostructural optimization of aircraft wings*, Structural and Multidisciplinary Optimization, 53 (2016), pp. 589–603.

[212] S. Lang, *Fundamentals of differential geometry*, vol. 191, Springer Science & Business Media, 2012.

[213] N. Lebbe, C. Dapogny, E. Oudet, K. Hassan, and A. Gliere, *Robust shape and topology optimization of nanophotonic devices using the level set method*, Journal of Computational Physics, (2019).

[214] L. Leifsson and S. Koziel, *Aerodynamic shape optimization by variable-fidelity computational fluid dynamics models: a review of recent progress*, Journal of Computational Science, 10 (2015), pp. 45–54.

[215] Y. Li and L. Nirenberg, *The distance function to the boundary, finsler geometry, and the singular set of viscosity solutions of some hamilton-jacobi equations*, Communications on pure and applied mathematics, 58 (2005), pp. 85–146.

[216] J.-L. Lions, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Mathematics, Vol. 323, Springer-Verlag, Berlin-New York, 1973.

[217] J.-L. Lions, *Some methods in the mathematical analysis of systems and their control*, Beijing, Science Press, (1981).

[218] J.-L. Lions and E. Magenes, *Non-Homogenous Boundary Value Problems and Applications*, vol. 1, Springer, New York, 01 1972.

[219] W. G. Litvinov, *On the optimal shape of a hydrofoil*, Journal of Optimization Theory and Applications, 85 (1995), pp. 325–345.

[220] J. Liu and Y. Ma, *A survey of manufacturing oriented topology optimization methods*, Advances in Engineering Software, 100 (2016), pp. 161–175.

[221] C. Lundgaard, J. Alexandersen, M. Zhou, C. S. Andreasen, and O. Sigmund, *Revisiting density-based topology optimization for fluid-structure-interaction problems*, Structural and Multidisciplinary Optimization, 58 (2018), pp. 969–995.

[222] J. Luo, Z. Luo, S. Chen, L. Tong, and M. Y. Wang, *A new level set method for systematic design of hinge-free compliant mechanisms*, Computer Methods in Applied Mechanics and Engineering, 198 (2008), pp. 318–331.

[223] C. Mantegazza and A. C. Mennucci, *Hamilton-Jacobi Equations and Distance Functions on Riemannian Manifolds*, Applied Mathematics & Optimization, 47 (2003).

[224] G. Marck, M. Nemer, and J.-L. Harion, *Topology optimization of heat and mass transfer problems: Laminar flow*, Numerical Heat Transfer, Part B: Fundamentals, 63 (2013), pp. 508–539.

[225] G. Marck and Y. Privat, *On some shape and topology optimization problems in conductive and convective heat transfers*, in OPTI 2014, An International Conference on Engineering and Applied Sciences Optimization, M. Papadrakakis, M. Karlaftis, and N. Lagaros, eds., June 2014, pp. 1640–1657.

[226] N. Marco and A. Dervieux, *Multilevel parametrization for aerodynamical optimization of 3D shapes*, Finite Elements in Analysis and Design, 26 (1997), pp. 259–277.

[227] J. Martínez-Frutos and D. Herrero-Pérez, *GPU acceleration for evolutionary topology optimization of continuum structures using isosurfaces*, Computers & Structures, 182 (2017), pp. 119–136.

[228] E. Maruvsic-Paloka, *Asymptotic expansion for a flow in a periodic porous medium*, Comptes Rendus de l'Academie des Sciences-Series IIB-Mechanics-Physics-Chemistry-Astronomy, 325 (1997), pp. 369–374.

[229] M. Masmoudi and P. Guillaume, *Conception optimale de formes et applications*, in Optimisation et contrôle (Sophia-Antipolis, 1992), Cépaduès, Toulouse, 1993, pp. 201–216.

[230] A. Maury, G. Allaire, and F. Jouve, *Elasto-plastic shape optimization using the level set method*, SIAM Journal on Control and Optimization, 56 (2018), pp. 556–581.

[231] W. C. H. McLean, *Strongly elliptic systems and boundary integral equations*, Cambridge university press, 2000.

[232] L. A. M. Mello, E. De Sturler, G. H. Paulino, and E. C. N. Silva, *Recycling Krylov subspaces for efficient large-scale electrical impedance tomography*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 3101–3110.

[233] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, *Discrete differential-geometry operators for triangulated 2-manifolds*, in Visualization and mathematics III, Springer, 2003, pp. 35–57.

[234] G. Michailidis, *Manufacturing Constraints and Multi-Phase Shape and Topology Optimization via a Level-Set Method*, PhD thesis, Ecole polytechnique, 2014.

[235] B. Mohammadi and B. Mohammadi, *Optimal shape design, reverse mode of automatic differentiation and turbulence*, in 35th Aerospace Sciences Meeting and Exhibit, 1997, p. 99.

[236] B. Mohammadi and O. Pironneau, *Applied shape optimization for fluids*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, second ed., 2010.

[237] P. Morin, R. Nochetto, M. Pauletti, and M. Verani, *Adaptive SQP method for shape optimization*, in Numerical Mathematics and Advanced Applications 2009, Springer, 2010, pp. 663–673.

[238] J. Moulin, P. Jolivet, and O. Marquet, *Augmented Lagrangian preconditioner for large-scale hydrodynamic stability analysis*, Computer Methods in Applied Mechanics and Engineering, 351 (2019), pp. 718–743.

[239] C. Multiphysics, *Comsol*, Inc., Burlington, MA, www. comsol. com, (1994).

[240] D. J. Munk, T. Kipouros, G. A. Vio, G. T. Parks, and G. P. Steven, *On the effect of fluid-structure interactions and choice of algorithm in multi-physics topology optimisation*, Finite Elements in Analysis and Design, 145 (2018), pp. 32–54.

[241] F. Murat and J. Simon, *Etude de problèmes d'optimal design*, in IFIP Technical Conference on Optimization Techniques, Springer, 1975, pp. 54–62.

[242] ——, *Sur le contrôle par un domaine géométrique*, Publication du Laboratoire d'Analyse Numérique de l'Université Pierre et Marie Curie, (1976).

[243] J.-C. Nédélec, *Acoustic and electromagnetic equations: integral representations for harmonic problems*, Springer Science & Business Media, 2001.

[244] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Science, 35 (1999).

[245] Noesis Solutions, NV, *Optimus theoretical background*, Leuven, Belgium, (2010).

[246] H. Okumura and M. Kawahara, *Shape optimization of body located in incompressible Navier-Stokes flow based on optimal control theory*, CMES. Computer Modeling in Engineering & Sciences, 1 (2000), pp. 71–77.

[247] S. Osher and R. Fedkiw, *Level set methods and dynamic implicit surfaces*, vol. 153 of Applied Mathematical Sciences, Springer-Verlag, New York, 2003.

[248] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations*, Journal of computational physics, 79 (1988), pp. 12–49.

[249] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM Journal on numerical analysis, 28 (1991), pp. 907–922.

[250] S. J. Osher and F. Santosa, *Level set methods for optimization problems involving geometry and constraints. I. Frequencies of a two-density inhomogeneous drum*, J. Comput. Phys., 171 (2001), pp. 272–288.

[251] B. Palmerio and A. Dervieux, *Identification de frontière dans le cas d'un problème de Dirichlet*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B, 275 (1972), pp. A1111–A1113.

[252] ——, *Une formule de Hadamard dans des problèmes d'identification de domaines*, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B, 280 (1975), pp. Aii, A1697–A1700.

[253] O. Pantz, *Sensibilité de l'équation de la chaleur aux sauts de conductivité*, Comptes Rendus Mathematique, 341 (2005), pp. 333–337.

[254] O. Pantz and K. Trabelsi, *A post-treatment of the homogenization method for shape optimization*, SIAM Journal on Control and Optimization, 47 (2008), pp. 1380–1398.

[255] P. Papazoglou, *Topology optimization of heat exchangers*, Master's thesis, TU Delft, 2015.

[256] P.-O. Persson, *Mesh generation for implicit geometries*, PhD thesis, Massachusetts Institute of Technology, 2005.

[257] E. Peynaud, *Rayonnement sonore dans un écoulement subsonique complexe en régime harmonique: analyse et simulation numérique du couplage entre les phénomènes acoustiques et hydrodynamiques*, PhD thesis, Toulouse, INSA, 2013.

[258] M. Pietropaoli, F. Montomoli, and A. Gaymann, *Three-dimensional fluid topology optimization for heat transfer*, Structural and Multidisciplinary Optimization, 59 (2019), pp. 801–812.

[259] O. Pironneau, *On optimum profiles in Stokes flow*, Journal of Fluid Mechanics, 59 (1973), pp. 117–128.

[260] ——, *Optimal shape design for elliptic systems*, in System Modeling and Optimization, Springer, 1982, pp. 42–66.

[261] P. Plotnikov and J. Sokołowski, *Compressible Navier-Stokes equations: theory and shape optimization*, vol. 73, Springer Science & Business Media, 2012.

[262] P. I. Plotnikov and J. Sokolowski, *Optimal shape control of airfoil in compressible gas flow governed by Navier-Stokes equations*, Evolution Equations and Control Theory, 2 (2013), pp. 495–516.

[263] N. Pollini, O. Sigmund, C. S. Andreasen, and J. Alexandersen, *A "poor man's" approach for high-resolution three-dimensional topology design for natural convection problems*, Advances in Engineering Software, 140 (2020), p. 102736.

[264] T. N. Pouchon, *Effective models and numerical homogenization methods for long time wave propagation in heterogeneous media*, PhD thesis, EPFL, 2017.

[265] J. Rauch and M. Taylor, *Potential and scattering theory on wildly perturbed domains*, Journal of Functional Analysis, 18 (1975), pp. 27–59.

[266] A. Rebei, *Développement de méthode d'optimisation topologique adaptée aux écoulements en régime turbulent, application au cas des échangeurs de chaleur*, PhD thesis, Thèse de l'université PSL. Préparée à Mines ParisTech., 2019.

[267] T. Richter, *Fluid-structure interactions*, vol. 118 of Lecture Notes in Computational Science and Engineering, Springer, Cham, 2017. Models, analysis and finite elements.

[268] J. C. Robinson, *Infinite-dimensional dynamical systems: an introduction to dissipative parabolic PDEs and the theory of global attractors*, vol. 28, Cambridge University Press, 2001.

[269] W. Rudin, *Real and complex analysis*, Tata McGraw-Hill Education, 2006.

[270] S. Rusinkiewicz, *Estimating curvatures and their derivatives on triangle meshes*, in 2nd International Symposium on 3D Data Processing, Visualization and Transmission, IEEE, 2004, pp. 486–493.

[271] D. Saltzman, M. Bichnevicius, S. Lynch, T. W. Simpson, E. W. Reutzel, C. Dickman, and R. Martukanitz, *Design and evaluation of an additively manufactured aircraft heat exchanger*, Applied Thermal Engineering, 138 (2018), pp. 254–263.

[272] E. Sanchez-Palencia, *Fluid flow in porous media*, Non-homogeneous media and vibration theory, (1980), pp. 129–157.

[273] E. Sánchez-Palencia, *On the asymptotics of the fluid flow past an array of fixed obstacles*, International Journal of Engineering Science, 20 (1982), pp. 1291–1301.

[274] F. Santosa and W. W. Symes, *A dispersive effective medium for wave propagation in periodic composites*, SIAM Journal on Applied Mathematics, 51 (1991), pp. 984–1005.

[275] K. R. Saviers, R. Ranjan, and R. Mahmoudi, *Design and validation of topology optimized heat exchangers*, in AIAA Scitech 2019 Forum, 2019, p. 1465.

[276] T. Schneider and R. Klein, *Overcoming mass losses in level set-based interface tracking schemes*, in Finite volumes for complex applications II, Hermes Sci. Publ., Paris, 1999, pp. 41–50.

[277] J. Schropp and I. Singer, *A dynamical systems approach to constrained minimization*, Numerical functional analysis and optimization, 21 (2000), pp. 537–551.

[278] V. H. Schulz, *A Riemannian view on shape optimization*, Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics, 14 (2014), pp. 483–501.

[279] A. Segal, M. ur Rehman, and C. Vuik, *Preconditioners for incompressible Navier-Stokes solvers*, Numerical Mathematics: Theory, Methods and Applications, 3 (2010), pp. 245–275.

[280] J. A. Sethian, *A fast marching level set method for monotonically advancing fronts*, Proceedings of the National Academy of Sciences, 93 (1996), pp. 1591–1595.

[281] J. A. Sethian and A. Wiegmann, *Structural boundary design via level set and immersed interface methods*, Journal of computational physics, 163 (2000), pp. 489–528.

[282] V. Shikhman and O. Stein, *Constrained optimization: projected gradient flows*, Journal of optimization theory and applications, 140 (2009), pp. 117–130.

[283] H. Si and A. TetGen, *A quality tetrahedral mesh generator and three-dimensional delaunay triangulator*, Weierstrass Institute for Applied Analysis and Stochastic, Berlin, Germany, 81 (2006).

[284] O. Sigmund and K. Maute, *Topology optimization approaches*, Structural and Multidisciplinary Optimization, 48 (2013), pp. 1031–1055.

[285] O. Sigmund and S. Torquato, *Design of materials with extreme thermal expansion using a three-phase topology optimization method*, Journal of the Mechanics and Physics of Solids, 45 (1997), pp. 1037–1067.

[286] S. N. Skinner and H. Zare-Behtash, *State-of-the-art in aerodynamic shape optimisation methods*, Applied Soft Computing, 62 (2018), pp. 933–962.

[287] V. P. Smyshlyaev and K. Cherednichenko, *On rigorous derivation of strain gradient effects in the overall behaviour of periodic heterogeneous media*, Journal of the Mechanics and Physics of Solids, 48 (2000), pp. 1325–1357.

[288] J. Sobieszczanski-Sobieski and R. T. Haftka, *Multidisciplinary aerospace design optimization: survey of recent developments*, Structural optimization, 14 (1997), pp. 1–23.

[289] J. Sokolowski and A. Zochowski, *Topological derivative in shape optimization*, Encyclopedia of Optimization, (2009), pp. 3908–3918.

[290] J. Sokolowski and J.-P. Zolésio, *Shape sensitivity analysis of contact problem with prescribed friction*, Nonlinear Analysis. Theory, Methods & Applications. An International Multidisciplinary Journal, 12 (1988), pp. 1399–1411.

[291] J. Sokolowski and J.-P. Zolésio, *Introduction to shape optimization*, vol. 16 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1992. Shape sensitivity analysis.

[292] M. Sonntag, S. Schmidt, and N. R. Gauger, *Shape derivatives for the compressible Navier-Stokes equations in variational form*, Journal of Computational and Applied Mathematics, 296 (2016), pp. 334–351.

[293] M. Spivak, *A comprehensive introduction to differential geometry. Vol. III*, Publish or Perish, Inc., Wilmington, Del., second ed., 1979.

[294] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Université de Grenoble. Annales de l'Institut Fourier, 15 (1965), pp. 189–258.

[295] J. Strain, *Semi-Lagrangian methods for level set equations*, Journal of Computational Physics, 151 (1999), pp. 498–533.

[296] K. Sturm, *Shape optimization with nonsmooth cost functions: from theory to numerics*, SIAM Journal on Control and Optimization, 54 (2016), pp. 3319–3346.

[297] W. Sun and Y.-X. Yuan, *Optimization theory and methods: nonlinear programming*, vol. 1, Springer Science & Business Media, 2006.

[298] K. Svanberg, *The method of moving asymptotes—a new method for structural optimization*, International Journal for Numerical Methods in Engineering, 24 (1987), pp. 359–373.

[299] A. Takezawa, S. Nishiwaki, and M. Kitamura, *Shape and topology optimization based on the phase field method and sensitivity analysis*, Journal of Computational Physics, 229 (2010), pp. 2697–2718.

[300] K. Tanabe, *A geometric method in nonlinear programming*, Journal of Optimization Theory and Applications, 30 (1980), pp. 181–210.

[301] L. Tartar, *Topics in nonlinear analysis*, Publications mathématiques d'Orsay, 78 (1978).

[302] ——, *An introduction to Sobolev spaces and interpolation spaces*, vol. 3, Springer Science & Business Media, 2007.

[303] R. Tawk, B. Ghannam, and M. Nemer, *Topology optimization of heat and mass transfer problems in two fluids-one solid domains*, Numerical Heat Transfer, Part B: Fundamentals, 76 (2019), pp. 130–151.

[304] R. Temam, *Navier stokes equations: Theory and numerical analysis*, vol. 45, North-Holland Publishing Company, 1977.

[305] L. N. Trefethen and D. Bau III, *Numerical linear algebra*, vol. 50, Siam, 1997.

[306] B. O. Turesson, *Nonlinear potential theory and weighted Sobolev spaces*, Springer, 2007.

[307] M. G. Ukken and M. Sivapragasam, *Aerodynamic shape optimization of airfoils at ultra-low Reynolds numbers*, Sādhanā, 44 (2019), p. 130.

[308] G. N. Vanderplaats and F. Moses, *Structural optimization by methods of feasible directions*, Computers & Structures, 3 (1973), pp. 739–755.

[309] C. H. Villanueva and K. Maute, *CutFEM topology optimization of 3D laminar incompressible flow problems*, Computer Methods in Applied Mechanics and Engineering, 320 (2017), pp. 444–473.

[310] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Mathematical programming, 106 (2006), pp. 25–57.

[311] M. Y. Wang, X. Wang, and D. Guo, *A level set method for structural topology optimization*, Computer methods in applied mechanics and engineering, 192 (2003), pp. 227–246.

[312] K. Welker, *Suitable Spaces for Shape Optimization*, arXiv preprint arXiv:1702.07579, (2017).

[313] O. Widlund and M. Dryja, *An additive variant of the Schwarz alternating method for the case of many subregions*, Technical Report 339, Ultracomputer Note 131, Department of Computer Science, Courant Institute, 12 1987.

[314] Q. Xia and M. Y. Wang, *Topology optimization of thermoelastic structures using level set method*, Computational Mechanics, 42 (2008), pp. 837–857.

[315] K. Yaji, T. Yamada, S. Kubo, K. Izui, and S. Nishiwaki, *A topology optimization method for a coupled thermal–fluid problem using level set boundary expressions*, International Journal of Heat and Mass Transfer, 81 (2015), pp. 878–888.

[316] T. Yamada, K. Izui, S. Nishiwaki, and A. Takezawa, *A topology optimization method based on the level set method incorporating a fictitious interface energy*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 2876–2891.

[317] H. Yamashita, *A differential equation approach to nonlinear programming*, Mathematical Programming, 18 (1980), pp. 155–168.

[318] G. H. Yoon, *Topological layout design of electro-fluid-thermal-compliant actuator*, Computer Methods in Applied Mechanics and Engineering, 209–212 (2012), pp. 28 – 44.

[319] ——, *Stress-based topology optimization method for steady-state fluid–structure interaction problems*, Computer Methods in Applied Mechanics and Engineering, 278 (2014), pp. 499 – 523.

[320] G. H. Yoon, S. Heo, and Y. Y. Kim, *Minimum thickness control at various levels for topology optimization using the wavelet method*, International journal of solids and structures, 42 (2005), pp. 5945–5970.

[321] Y.-x. Yuan, *A review of trust region algorithms for optimization*, in ICIAM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics, 09 1999.

[322] M. Yulin and W. Xiaoming, *A level set method for structural topology optimization with multi-constraints and multi-materials*, Acta Mechanica Sinica, 20 (2004), pp. 507–518.

[323] H. Zhao, *A fast sweeping method for eikonal equations*, Mathematics of computation, 74 (2005), pp. 603–627.

[324] X. Zhao, M. Zhou, O. Sigmund, and C. Andreasen, *A "poor man's approach" to topology optimization of cooling channels based on a Darcy flow model*, International Journal of Heat and Mass Transfer, 116 (2018), pp. 1108–1123.

[325] C. Zhuang, Z. Xiong, and H. Ding, *A level set method for topology optimization of heat conduction problem under multiple load cases*, Computer Methods in Applied Mechanics and Engineering, 196 (2007), pp. 1074–1084.

[326] J.-P. Zolésio, *The material derivative (or speed) method for shape optimization*, in Optimization of distributed parameter structures, Vol. II (Iowa City, Iowa, 1980), vol. 50 of NATO Adv. Study Inst. Ser. E: Appl. Sci., Nijhoff, The Hague, 1981, pp. 1089–1151.

[327] G. Zoutendijk, *Methods of feasible directions: A study in linear and non-linear programming*, Elsevier Publishing Co., Amsterdam-London-New York-Princeton, N.J., 1960.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Optimisation de formes de systèmes multiphysiques

Mots clés : Optimisation topologique, remaillage, transfert thermique convectif, interaction fluide-structure, contraintes géométriques, modèles homogénéisés d'ordres élevés.

Résumé : Cette thèse est consacrée à l'optimisation de la topologie et de la forme de systèmes multiphysiques motivés par des applications de l'industrie aéronautique. Nous calculons les dérivées de forme de fonctions de coût arbitraires pour un modèle fluide, thermique et mécanique faiblement couplé. Nous développons ensuite un algorithme de type gradient adapté à la résolution de problèmes d'optimisation de formes sous contraintes qui ne requiert par de réglage de paramètres non physiques. Nous introduisons ensuite une méthode variationnelle qui permet de calculer des intégrales le long de rayons sur un maillage par la résolution d'un problème variationnel qui ne requiert pas la détermination explicite de ces lignes sur la discrétisation spatiale. Cette technique nous a ainsi permis d'imposer une contrainte de non-mélange de phases pour une application à l'optimisation d'échangeurs de chaleur bi-tubes. Tous ces ingrédients ont été employés pour traiter une variété de cas tests d'optimisation de formes pour des systèmes multi-physiques 2-d ou 3-d. Nous avons considéré des problèmes à une seule, deux ou bien trois physiques couplées en 2-d, et des problèmes de tailles relativement élevées en 3-d pour la mécanique, la conduction thermique, l'optimisation de profils aérodynamiques, et de la forme de systèmes en interaction fluide-structure. Un dernier chapitre d'ouverture est consacré à l'étude de modèles homogénéisées d'ordres élevés pour les équations de Stokes en milieu poreux. Ces équations d'ordres élevés englobent les trois régimes homogénéisés classiques—Stokes, Brinkman et Darcy–associés à divers rapports d'échelles pour la taille des obstacles. Elles pourraient permettre, lors de futurs travaux, de développer de nouvelles méthodes d'optimisation pour la conception de systèmes fluides caractérisés par des motifs multiéchelles, tels que les échangeurs thermiques industriels.

Title : Shape and topology optimization of multiphysics systems

Keywords : Topology optimization, remeshing, convective heat transfer, fluid-structure interaction, geometric constraints, high order homogenization.

Abstract : This work is devoted to shape and topology optimization of multiphysics systems motivated by aeronautic industrial applications. Shape derivatives of arbitrary objective functionals are computed for a weakly coupled thermal fluid-structure model. A novel gradient flow type algorithm is then developed for solving generic constrained shape optimization problems without the need for tuning non-physical metaparameters. Motivated by the need for enforcing non-mixing constraints in the design of liquid-liquid heat exchangers, a variational method is developed in order to simplify the numerical evaluation of geometric constraints: it allows to compute line integrals on a mesh by solving a variational problem without requiring the explicit knowledge of these lines on the spatial discretization. All these ingredients allowed us to implement a variety of 2-d and 3-d multiphysics shape optimization test cases: from single, double or three physics problems in 2-d, to moderately large-scale 3-d test cases for structural design, thermal conduction, aerodynamic design and a fluid-structure interacting system. A final opening chapter derives high order homogenized equations for the Stokes system in a porous medium. These high order equations encompass the three classical homogenized regimes—namely Stokes, Brinkman and Darcy—associated with different obstacle's size scalings. They could allow, in future works, to develop new topology optimization methods for the design of fluid systems characterized by multi-scale patterns such as industrial heat exchangers.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France