Mathematical modelling and numerical simulation in materials science

Sébastien Boyaval

UNIVERSITÉ PARIS EST
ÉCOLE DOCTORALE ICMS

Thèse de Doctorat
Mathématiques Appliquées

Sébastien Boyaval

Modélisation mathématique et simulation numérique en science des matériaux

Thèse soutenue le 16 décembre 2009 devant un jury composé de

Anthony T. Patera	Président
Raz Kupferman	Rapporteur
Marco Picasso	Rapporteur
Olivier Lemaître	Examinateur
Jacques Sainte-Marie	Examinateur
Claude Le Bris	Directeur
Tony Lelièvre	Directeur
Michel Benoît	Invité

Titre : Modélisation mathématique et simulation numérique en science des matériaux.

Résumé : Dans une première partie, nous étudions des schémas numériques utilisant la méthode des éléments finis pour discrétiser le système d'équations Oldroyd-B modélisant un fluide viscolélastique avec conditions de collement dans un domaine borné, en dimension deux ou trois. Le but est d'obtenir des schémas stables au sens où ils dissipent une énergie libre, imitant ainsi des propriétés thermodynamiques de dissipation similaires à celles identifiées pour des solutions régulières du modèle continu. Cette étude s'ajoute à de nombreux travaux antérieurs sur les instabilités obervées dans les simulations numériques d'équations viscoélastiques (dont celles connues comme étant des *Problèmes à Grand Nombre de Weissenberg*). A notre connaissance, c'est la première étude qui considère rigoureusement la stabilité numérique au sens de la dissipation d'une énergie pour des discrétisations de type Galerkin.

Dans une seconde partie, nous adaptons et utilisons les idées d'une méthode numérique initialement développée dans des travaux de Y. Maday, A. T. Patera *et al.*, la *méthode des bases réduites*, pour simuler efficacement divers modèles multi-échelles. Le principe est d'approcher numériquement chaque élément d'une collection paramétrée d'objets complexes dans un espace de Hilbert par la plus proche combinaison linéaire dans le meilleur sous-espace vectoriel engendré par quelques éléments bien choisis au sein de la même collection paramétrée. Nous appliquons ce principe pour des problèmes numériques liés :

- à l'homogénéisation numérique d'équations elliptiques scalaires du second-ordre, avec coefficients de diffusion oscillant à deux échelles, puis
- à la propagation d'incertitudes (calculs de moyenne et de variance) dans un problème elliptique avec coefficients stochastiques (un champ aléatoire borné dans une condition de bord du troisième type), enfin
- au calcul Monte-Carlo de l'espérance de nombreuses variables aléatoires paramétrées, en particulier des fonctionnelles de processus stochastiques d'Itô paramétrés proches de ce qu'on rencontre dans les modèles micro-macro de fluides polymériques, avec une variable de contrôle pour en réduire la variance.

Dans chaque application, le but de l'approche bases-réduites est d'accélérer les calculs sans perte de précision.

Mots-clefs : Fluides viscoélastiques, Méthode des éléments finis, Méthode des bases réduites, Homogénéisation numérique, Propagation d'incertitudes, Réduction de Variance

Title : Mathematical modelling and numerical simulation in materials science.

Summary : In a first part, we study numerical schemes using the finite-element method to discretize the Oldroyd-B system of equations, modelling a viscoelastic fluid under no flow boundary condition in a $2-$ or $3-$dimensional bounded domain. The goal is to get schemes which are stable in the sense that they dissipate a free-energy, mimicking that way thermodynamical properties of dissipation similar to those actually identified for smooth solutions of the continuous model. This study adds to numerous previous ones about the instabilities observed in the numerical simulations of viscoelastic fluids (in particular those known as *High Weissenberg Number Problems*). To our knowledge, this is the first study that rigorously considers the numerical stability in the sense of an energy dissipation for Galerkin discretizations.

In a second part, we adapt and use ideas of a numerical method initially developped in the works of Y. Maday, A.T. Patera *et al.*, the *reduced-basis method*, in order to efficiently simulate some multiscale models. The principle is to numerically approximate each element of a parametrized family of complicate objects in a Hilbert space through the closest linear combination within the best linear subspace spanned by a few elements well chosen inside the same parametrized family. We apply this principle to numerical problems linked :

- to the numerical homogenization of second-order elliptic equations, with two-scale oscillating diffusion coefficients, then
- to the propagation of uncertainty (computations of the mean and the variance) in an elliptic problem with stochastic coefficients (a bounded stochastic field in a boundary condition of third type), last
- to the Monte-Carlo computation of the expectations of numerous parametrized random variables, in particular functionals of parametrized Itô stochastic processes close to what is encountered in micro-macro models of polymeric fluids, with a control variate to reduce its variance.

In each application, the goal of the reduced-basis approach is to speed up the computations without any loss of precision.

Keywords : Viscoelastic fluids, Finite-element method, Reduced-basis method, Numerical homogenization, Uncertainty propagation, Variance reduction

Mathematical Subject Classification (MSC2010) : 76A10, 65N12, 65N30, 65D15, 90C59

Table des matières

Chapitre 0

Introduction générale et principales contributions de la thèse

Cette thèse a pour objet l'étude mathématique et la simulation numérique de quelques modèles de comportement mécanique pour des matériaux complexes dotés d'une micro-structure. L'enjeu est de décrire numériquement, à l'échelle humaine (celle qui intéresse l'ingénieur), les propriétés mécaniques de matériaux complexes en tenant compte des connaissances physiques sur la matière à une échelle plus fine. Une telle description, dite *multi-échelle*, fait l'hypothèse d'une *micro-structure* sous-jacente à la description macroscopique de la matière. La simulation de modèles micro-macro intéresse en particulier les concepteurs, les producteurs et les utilisateurs de matériaux complexes tels que, dans le cas fluide :

- le béton, les matériaux granulaires du type sable ou gravier, les pâtes et autres matériaux du génie civil,
- les polymères plastiques de l'industrie automobile,
- les suspensions particulaires de l'industrie agroalimentaire...

Les résultats de la thèse sont regroupés dans deux parties et on donne, au début de chacune d'elles, une introduction, respectivement

[Chap. 1] à la modélisation mathématique de quelques matériaux complexes dotés d'une micro-structure, ceux au comportement mécanique de fluide viscoélastique, et

[Chap. 4] à une méthode de simulation numérique par réduction de bases pour des calculs *intensifs*, dite *méthode des bases réduites (certifiée)*, avec une synthèse de son application à divers modèles multi-échelles rencontrés en science des matériaux.

Dans une première partie, on étudie des schémas numériques utilisant la méthode des éléments finis pour discrétiser des modèles mécaniques de fluides complexes *viscoélastiques*. Les modèles en question ne sont pas du type micro-macro : ils s'écrivent uniquement avec des variables représentant directement les quantités intéressantes du problème à l'échelle macroscopique. Toutefois, notre étude concerne essentiellement les modèles qui peuvent être déduits d'une approche micro-macro (éventuellement après une suite d'approximations, souvent appelée *coarse-graining* en anglais), pourvu que celle-ci soit formulée mathématiquement. L'objet de cette partie est d'améliorer la simulation de phénomènes multi-échelles avec ces modèles.

De nombreuses *équations constitutives* ont en effet été proposées pour décrire le mouvement d'un fluide comme un milieu continu sous l'effet de forces visqueuses et élastiques. Seules quelques-unes résultent de l'addition d'effets microscopiques – moléculaires – bien compris. En particulier, parmi les équations constitutives qui peuvent être obtenues par intégration d'un modèle micro-macro, celles qui s'appuient sur une description statistique rigoureuse de la matière à l'échelle microscopique permettent d'identifier (dans des écoulements suffisamment simples) les bonnes quantités thermodynamiques à l'échelle macroscopique, par exemple l'énergie libre [Lel04, JLBLO06]. Partant de cette observation, nous avons proposé, pour une équation constitutive prototypique (Oldroyd-B), des schémas numériques (par la méthode des éléments finis), qui respectent, après discrétisation, une loi de dissipation de l'énergie libre similaire à celle identifiée au niveau continu (dans un écoulement simple avec conditions de bord de type Dirichlet homogène, dites aussi de "collement").

A notre connaissance, aucune étude antérieure n'avait considéré ce critère comme condition de stabilité pour un schéma numérique discrétisant une équation constitutive. (Dans [LO03, LX06], on considère une autre loi d'énergie, qui n'est pas une loi de dissipation.) L'étude de ce critère apporte donc quelques réponses nouvelles aux nombreuses questions soulevées par les instabilités observées dans la simulation numérique des équations

constitutives (voir, par exemple, la discussion sur le *High Weissenberg Number Problem* dans [OP02], et des résultats numériques récents à ce sujet dans [FK05, HFK05]). Plus précisément, pour les équations constitutives et les écoulements tels que les deux questions suivantes sont (supposées) résolues, notre étude de la stabilité des schémas numériques contribue à améliorer la compréhension des instabilités observées par les simulateurs en répondant à la troisième :

- le problème avec conditions initiales et de bord possède-t-il un état stationnaire ?
- quelle régularité possède la solution du problème avec conditions initiales et de bord ?
- comment approcher numériquement la solution du problème avec conditions initiales et de bord ?

Elle n'est toutefois encore qu'un préalable à une réponse définitive sur l'origine des instabilités numériques observées, et à une pratique totalement rigoureuse de la simulation numérique d'équations constitutives pour un fluide viscoélastique. En effet, même dans le cas particulièrement simple du système d'équations Oldroyd-B, on ne sait pas :

- s'il existe un état stationnaire pour un problème avec conditions de bord générales (en particulier, dans les cas-tests importants du flot autour d'un cylindre et de la contraction),
- ni en quel sens le problème aux limites a éventuellement une solution (à moins de supposer les conditions initiales voisines d'un état stationnaire supposé [LLZ05, LLZ08], ou le champ de vitesse corotationnel [LM00]).

Plus précisément, nous présentons les résultats suivants, qui restranscrivent respectivement [BLM09] et [BB09b].

[Chap. 2] Il est possible de construire (par la méthode des éléments finis) des schémas numériques, stables au sens du critère de *dissipation de l'énergie libre*, pour simuler le problème de Cauchy associé aux équations d'Oldroyd-B avec des conditions de type Dirichlet homogène au bord d'un domaine régulier borné. Toutefois, cela requiert :

- un traitement particulier des termes d'advection dans l'équation d'Oldroyd (utilisant par exemple la méthode des caractéristiques, ou la méthode de Galerkin discontinue avec des termes de flux),
- un schéma implicite (résultant en un système d'équations algébrique non-linéaire), et
- une discrétisation en espace d'ordre faible pour le champ de variables associé à l'équation d'Oldroyd (dans notre cas, un tenseur de conformation plutôt qu'un tenseur de contraintes), autorisant en particulier l'emploi d'approximations constantes par morceaux pour ce champ (éléments finis de type $\mathbb{P}_0$).

On montre alors que le schéma numérique possède une unique solution stable, globale en temps si le pas de temps utilisé pour la discrétisation temporelle, de type Euler rétrograde, est suffisamment petit (inférieur à une limite fonction de la donnée initiale). Les difficultés rencontrées sont essentiellement liées à la non-linéarité de la fonction test en le champ de variables inconnu, et à une contrainte de positivité sur ce même champ. Par ailleurs, on montre qu'il est facile d'adapter nos discrétisations à une reformulation de l'équation d'Olrdoyd utilisant le logarithme des variables inconnues contraintes à la positivité, voir [FK04, FK05, HFK05], et qu'alors, une solution stable existe sans condition sur le pas de temps. Pour cette dernière reformulation, toute solution est nécessairement stable au sens du critère dissipatif.

[Chap. 3] Pour certains schémas numériques discrétisant l'équation d'Oldroyd-B et proposés au Chap. 2, on peut construire des solutions stables, globales en temps, sans contrainte sur le pas de temps. Toutefois, il n'y a alors plus unicité, et d'autres solutions éventuelles pourraient ne pas être stables au sens du critère dissipatif.

Par ailleurs, on peut aussi construire des schémas numériques stables pour l'équation d'Oldroyd-B avec une technique de discrétisation différente de celles employées au Chap. 2 pour les termes d'advection. Avec cette dernière technique, qui utilise des éléments finis continus et n'a pas recourt au calcul des lignes caractéristiques, on montre qu'une sous-suite de solutions aux schémas numériques stables converge vers une solution d'un problème de Cauchy-Dirichlet pour des équations proches du système d'Oldroyd-B (avec un terme additionnel de dissipation)[1].

Le sujet de cette première partie comporte encore de nombreuses questions ouvertes. En particulier, il appelle à de nouveaux tests numériques comparant les qualités des divers schémas que nous avons proposés : stabilité numérique (également pour d'autres problèmes que ceux de type Dirichlet homogène dans une géométrie simple), simplicité de mise en œuvre (en comparaison des nombreuses discrétisations déjà employées par les praticiens de la mécanique des fluides non-newtoniens). On peut aussi se demander l'effet des méthodes de résolution algébriques pour le système non-linéaire résultant des schémas numériques stables implicites que l'on a proposés ci-dessus (comme, par exemple, dans [Pro08] pour des schémas dissipatifs stables discrétisant les équations de la magnétohydrodynamique incompressible), ou l'existence éventuelle de bifurcations en temps long pour

[1] Pour obtenir la bonne compacité, on a ajouté un terme de dissipation en espace dans l'équation sur le champ de contraintes. Ce terme régularisant existe effectivement [SB95], mais il est en général très petit, donc habituellement négligé par les physiciens dans le modèle d'Oldroyd.

les solutions des schémas numériques. Enfin, il conviendrait encore de s'interroger sur la possibilité d'étendre nos résultats à des conditions de bord plus générales, à d'autres classes d'approximations utilisées en pratique (éléments finis quadrilatères [HFK05], discrétisations temporelles d'ordre supérieur [Emm08]), et à des équations consitutives plus générales (les travaux [JLBLO06] et [HL07] sur les lois de dissipation généralisées au niveau continu pourraient servir de base à des extensions ultérieures de nos discrétisations).

Dans une seconde partie, on étudie l'application à quelques modèles multi-échelles d'une méthode numérique visant à réduire la complexité d'un grand nombre de calculs similaires entre eux. Cette méthode dite des *bases réduites* (*certifiée*) est en effet très générale dans son principe ; nous avons travaillé pour en étendre l'application à quelques modèles multi-échelles particuliers, gourmands en calculs.

Le but de la méthode des bases réduites est de résoudre rigoureusement et efficacement (c'est-à-dire précisément et rapidement) un grand nombre de problèmes paramétrés, pour de nombreuses valeurs du paramètre. La méthode des bases réduites standard s'applique essentiellement au calcul de fonctionnelles linéaires en la solution de problèmes paramétrés, qui sont typiquement des problèmes aux limites elliptiques admettant une formulation variationnelle affine en des fonctions du paramètre. La méthode est concue en deux étapes, dites *hors ligne* puis *en ligne*. D'abord, hors ligne, on construit un espace vectoriel de petite dimension N servant à approcher les solutions paramétrées des problèmes aux limites par la méthode de Galerkin. Pour un N donné, il existe un espace vectoriel engendré par N solutions en N valeurs distinctes du paramètre qui est optimal au sens de la minimisation de la pire erreur commise par la méthode de Galerkin sur l'ensemble des solutions paramétrées avec cet espace vectoriel[2]. En pratique, la méthode utilise un estimateur *a posteriori* rigoureux pour évaluer l'erreur de Galerkin, et choisit les N valeurs du paramètre itérativement, en cherchant à minimiser la pire erreur d'approximation à chaque étape. (On qualifie de glouton, *greedy* en anglais, ce type d'algorithmes qui résolvent successivement une suite de problèmes simples pour obtenir une solution (quasi-)optimale d'un problème d'optimisation compliqué.) Ensuite, en ligne, on approche systématiquement toute solution du problème paramétré (pour n'importe quelle valeur du paramètre) par une combinaison linéaire[3] bien choisie dans l'espace vectoriel de petite dimension N, à savoir celle donnée par la méthode de Galerkin. L'erreur d'approximation qui en résulte est alors rigoureusement évaluable par un estimateur *a posteriori* similaire à celui utilisé hors-ligne pour construire l'espace d'approximation : on obtient donc une approximation *certifiée*. La méthode des bases réduites économise ainsi de nombreux calculs par rapport à une approche directe (celle qui aurait résolu aveuglément le problème paramétré dans un espace de grande dimension pour chaque valeur du paramètre, typiquement avec des élements finis), dans la mesure où on a besoin de résoudre en ligne le problème en de suffisamment nombreuses valeurs du paramètre pour compenser les calculs hors ligne.

Avec quelques adaptations de la méthode standard, on a pu appliquer la méthode des bases réduites à trois problèmes numériques dépendant d'un paramètre et rencontrés dans la simulation de modèles multi-échelles en science des matériaux. On présente ces applications dans trois chapitres, correspondant respectivement aux publications [Boy08, BBM+09, BL09a].

[Chap. 5] Pour homogénéiser un problème elliptique où les coefficients de diffusion oscillent vite (mais ne sont pas exactement périodiques) dans l'espace physique, il faut calculer les solutions de nombreux problèmes intermédiaires, dits *de cellule*, en chaque point de l'espace physique (théoriquement , une infinité !). Avec une méthode de discrétisation peu coûteuse, on peut ensuite calculer une bonne approximation de la solution du probème elliptique homogénéisé, elle-même bonne approximation – éventuellement après correction – de la solution oscillante du problème elliptique initial (pour laquelle les méthodes d'approximation usuelles par discrétisation directe sont nécessairement coûteuses en degrés de liberté). En pratique, dans l'homogénéisation numérique, c'est le calcul des solutions à de nombreux problèmes de cellule qui reste coûteux (le choix des – nombreux – points de l'espace physique où calculer ces solutions dépendant de la technique de discrétisation qu'on choisira pour le problème homogénéisé final). Dans notre étude, on

[2]Cette pire erreur a une signification proche d'une épaisseur de Kolmogorov pour la sous-variété engendrée par la famille des solutions paramétrées, concept utile en théorie de l'approximation [Pin85] qui est précisé au Chap. 4. Elle correspond à un critère de type L^∞ sur le paramètre, pourvu que l'on dote l'espace des paramètres d'une topologie, et prend tout son sens dans certains contextes précis, selon la nature de la fonctionnelle de sortie et son utilisation. D'autres méthodes de réduction sont différentes par nature : par exemple, certaines cherchent à minimiser l'erreur en moyenne par rapport aux variations du paramètres, ou en moyenne quadratique. Elles correspondent alors respectivement à des critère de type L^1 et de type L^2 sur le paramètre. En particulier, le critère de type L^2 a déjà donné naissance à de nombreuses méthodes de réduction, qui se sont avérées utiles dans certaines applications précises : l'analyse en composantes principales (*principal component analysis* en anglais, également appelée *proper orthogonal decomposition* ou POD) [KV02], la quantification (*quantization* en anglais) [PP05], la décomposition de Karhunen-Loève [New96] (que nous retrouverons au Chap. 6), la décomposition en valeurs singulières [LBLM08] (SVD pour *singular value decomposition* en anglais, qui donne lieu à des méthodes dites *low-rank approximations*)...

[3]Noter qu'une version non-linéaire est en cours de développement [EPR].

cherche à réduire ce coût, dans le cadre de l'homogénéisation à deux échelles, où les problèmes de cellule s'expriment simplement comme une famille de problèmes elliptiques paramétrés, avec pour fonctionnelle de sortie le tenseur des coefficients de diffusion homogénéisés.

On utilise donc la méthode des bases réduites usuelle pour la famille (paramétrée) des solutions des problèmes de cellule, modulo le fait que la sortie n'est pas une fonction scalaire des solutions paramétrées. Les résultats numériques sont très satisfaisants pour l'homogénéisation numérique d'une équation de diffusion scalaire, avec un tenseur de diffusion oscillant, à deux échelles, où les problèmes de cellules sont des problèmes de diffusion scalaires périodiques, engendrés par la variation en taille, position et intensité d'une inclusion homogène isotrope carrée au sein d'une matrice carrée de conductivité également homogène isotrope.

Notre résolution de ce problème est un paradigme pour de nombreux modèles multi-échelles autres que l'homogénéisation (on cherche à calculer une quantité "macro", dite *homogénéisée* ici, qui nécessite de résoudre de nombreux problèmes "micro" paramétrés). Elle est aussi susceptible de nombreuses extensions en théorie de l'homogénéisation elle-même (y compris dans ses développements les plus récents [BLL06]).

[Chap. 6] Dans de nombreuses applications numériques, il est important de tenir compte de l'influence des incertitudes de modélisation sur le résultat, qui sont par exemple des fluctuations dans les coefficients d'une équation dont on cherche la solution. C'est en particulier le cas dans la simulation de nombreux modèles multi-échelles, où l'on calcule, typiquement avec des algorithmes ségrégés, les influences réciproques d'une échelle sur l'autre, tandis que la modélisation à une échelle doit tenir compte de fluctuations (incertitudes) à une autre échelle.

Quand des coefficients incertains dans une équation oscillent vite autour d'une valeur moyenne (hypothèse de séparation des échelles), on peut par exemple utiliser la théorie de l'homogénéisation (éventuellement stochastique [BP04, BLL06]), qui donne comme solution moyenne la solution homogénéisée (et parfois une variance pour sa fiabilité, voir le travail récent [GO09]). Sinon, la physique peut aussi mener à faire des hypothèses sur la forme des variations en les coefficients incertains. (Par exemple, la théorie cinétique pour les modèles micro-macro présentés au Chap. 1 suppose des variations gaussiennes.) Enfin, en l'absence de modèle précis, l'ingénieur peut aussi tenter de calibrer les incertitudes sur des mesures expérimentales, par exemple en estimant la loi des variations en les coefficients incertains (voir par exemple [JZ08]). Dans ce dernier cas, la solution moyenne choisie est typiquement la moyenne des solutions, et sa fiabilité est directement donnée par la variance des solutions (selon le théorème de la limite centrale). Les ingénieurs ont par ailleurs développé diverses techniques numériques de *propagation d'incertitudes*, en vue de calculer une solution moyenne et d'évaluer sa fiabilité.

Nous avons utilisé la méthode des bases réduites pour calculer efficacement et simplement la moyenne et la variance d'une variable aléatoire qui est une fonction régulière de la solution d'un problème elliptique scalaire où le coefficient dans une condition de bord du troisième type (dite de Robin) est un champ aléatoire borné[4]. Pour cela on utilise la méthode de Monte-Carlo, et on considère le problème paramétré par les réalisations du champ aléatoire (qui prend ici ses valeurs sur le bord du domaine physique). Le champ aléatoire paramètre appartient toutefois à un espace de dimension infinie (les fonctions bornées sur le bord du domaine physique), et une forme variationnelle usuelle pour le problème paramétré n'a pas alors la bonne structure – affine en quelques coefficients du paramètre –, nécessaire pour l'efficacité de la méthode des bases réduites. On utilise donc une étape de discrétisation supplémentaire par rapport à la méthode des bases réduites standard, qui consiste à tronquer la décomposition de Karhunen-Loève du champ aléatoire. Pour les champs aléatoires suffisamment corrélés en espace, le spectre des valeurs propres dans la décomposition de Karhunen-Loève décroît vite, et notre approche s'avère efficace et précise.

Ainsi, notre application de la méthode des bases réduites à un problème de bord paramétré par des coefficients stochastiques suggère donc une approche numérique possible pour des problèmes multi-échelles invoquant un bruit coloré comme paramètre modélisant les fluctuations d'une échelle sur l'autre. Dans des applications de type ingénieur, après estimation empirique des variations (incertitudes) dans un coefficient, notre méthode accélère vraisemblablement diverses approches numériques déjà utilisées en propagation d'incertitudes. (On peut d'ailleurs envisager diverses extensions de notre application numérique dans cette direction, pour d'autres champs aléatoires coefficients, d'autres équations et d'autres techniques de discrétisation dans l'espace des variations stochastiques.) Toutefois, pour certaines applications en science de la matière, qui supposent par exemple que des coefficients varient comme un bruit blanc gaussien (cas des polymères dilués modélisés comme des particules browniennes), la décroissance du spectre des valeurs

[4]Noter qu'on suggère aussi une méthode pour des problèmes de fiabilité qui cherchent à calculer les quantiles d'une variable aléatoire qui est une fonction régulière de la solution paramétrée, mais qu'on n'a pas encore de résultat numérique à ce sujet.

propres dans la décomposition de Karhunen-Loève est lente, et notre approche sera probablement peu efficace. C'est pourquoi on propose dans le chapitre suivant une autre approche pour des calculs intensifs avec les solutions d'équations différentielles stochastiques au sens d'Itô paramétrées.

[Chap. 7] On a déjà évoqué ci-dessus le fait qu'en science de la matière, de nombreux modèles multi-échelles utilisent des algorithmes ségrégés pour évaluer les influences réciproques d'une échelle sur l'autre. En particulier, quand des fluctuations à une échelle donnée sont modélisées avec des outils probabilistes, les algorithmes ségrégés couplant les différentes échelles requièrent typiquement l'évaluation de l'espérance de nombreuses variables aléatoires, qui sont par exemple des fonctionnelles d'un processus stochastique d'Itô (c'est le cas des modèles micro-macro pour les polmyères).

On considère ici une famille paramétrée de variables aléatoires, solutions d'une équation différentielle stochastique au sens d'Itô où les coefficients de dérive et de diffusion sont paramétrés. Nous nous sommes inspirés de la méthode des bases réduites usuelle pour calculer efficacement l'espérance d'un grand nombre de ces variables aléatoires paramétrées. Nous proposons deux approches pratiques pour calculer par bases réduites des variables de contrôle permettant de réduire la variance d'estimateurs Monte-Carlo pour ces variables aléatoires paramétrées. Pour calculer efficacement de nombreuses variables de contrôle, qui sont simplement des variables aléatoires paramétrées de carré intégrable, on a modifié la méthode des bases réduites standard. Dans l'une des deux approches, on substitue simplement un calcul par moindres carrés à la méthode de Galerkin pour calculer les coefficients de la combinaison linéaire qui est l'approximation par base réduite. Dans l'autre approche, quand la variable aléatoire est une fonctionnelle d'un processus d'Itô et qu'une variable de contrôle s'écrit à l'aide d'une formule de Feynman-Kac, on construit en fait une base réduite pour la collection des solutions de l'équation de Kolmogorov rétrograde, puis on procède de même par moindres carrés, en utilisant une combinaison linéaire de solutions bien choisies de l'équation de Kolmogorov rétrograde comme approximation par bases réduites. Dans les deux cas, on a surtout conservé l'idée d'utiliser (en ligne) un espace vectoriel d'approximation de petite dimension pour une famille paramétrée d'éléments d'un espace de Hilbert, qui est construit en sélectionnant (hors ligne) N valeurs du paramètre avec un algorithme glouton.

Les résultats numériques obtenus sont prometteurs, pour accélérer la calibration de la volatilité dans un modèle de Black-Scholes (mathématiques financières), et pour accèlérer la simulation de l'évolution en temps d'une famille d'équations de Langevin où les coefficients des équations évoluent aussi (typiquement, de manière couplée avec les calculs d'espérances, comme dans la dynamique moléculaire des modèles micro-macro en rhéologie).

Les problèmes numériques traités ci-dessus sont tous du type paramétré *intensif*, tels que la résolution d'une même équation paramétrée, coûteuse numériquement, est nécessaire en de nombreuses valeurs du paramètre. Des modèles micro-macro pour des fluides viscoélastiques (les modèles statistiques de polymères dilués) génèrent en particulier des problèmes de ce type ; ils nous ont servi de guide dans notre démarche. Les trois étapes (Chapitres 5, 6 et 7) correspondent à l'extension progressive de la méthode des bases réduites :

(i) à un modèle multi-échelle génèrant des problèmes numériques stantards pour la méthode des bases réduites usuelle, quoique avec une fonctionnelle de sortie tensorielle (non-scalaire), puis

(ii) à des modèles multi-échelles génèrant des problèmes numériques où le paramètre a une structure stochastique de dimension infinie (on a besoin d'une représentation spectrale tronquée, qui est une nouvelle technique prolongeant la méthode des bases réduites usuelle à un certain type de problèmes non-affines en le paramètre), enfin

(iii) à des modèles multi-échelles génèrant des problèmes numériques non-standards pour la méthode des bases réduites (calculs de Monte-Carlo plutôt qu'éléments finis).

La dernière étape est donc une véritable généralisation des idées de la méthode des bases réduites à des calculs paramétrés intensifs quelconques dans un espace de Hilbert, ce qui ouvre la voie à de futures investigations (comparaison, et combinaison, avec d'autres méthodes d'approximation – éventuellement réduites –). De plus, par delà les différents thèmes traités dans cette partie, le champ des applications de la méthode des bases réduites semble encore très ouvert et à explorer, en particulier pour des problèmes de type multi-échelle, avec autant d'adaptations et d'extensions possibles de la méthode pour chaque application spécifique.

Liste des articles publiés dans des revues à comité de lecture :

[Boy08] S. Boyaval. Reduced-basis approach for homogenization beyond the periodic setting. *SIAM Multiscale Modeling & Simulation (MMS)*, 7(1) :466–494, 2008.

[BLM09] S. Boyaval, T. Lelièvre, and C. Mangoubi. Free-energy-dissipative schemes for the Oldroyd-B model. *ESAIM : Mathematical Modelling and Numerical Analysis (M2AN)*, 43(3) :523–561, may 2009.

[BBM+09] S. Boyaval, C. Le Bris, Y. Maday, N.C. Nguyen, and A.T. Patera. A reduced basis approach for variational problems with stochastic parameters : Application to heat conduction with variable robin coefficient. *Computer Methods in Applied Mechanics and Engineering (CMAME)*, 198(41-44) :3187–3206, september 2009.

[BL09a] S. Boyaval and T. Lelièvre. A variance reduction method for parametrized stochastic differential equations using the reduced basis paradigm. In Pingwen Zhang, editor, *Accepted for publication in Communication in Mathematical Sciences (CMS)*, volume Special Issue "Mathematical Issues on Complex Fluids", 2009. ARXIV :0906.3600

Liste des articles soumis à des revues à comité de lecture :

[BB09b] J. W. Barrett and S. Boyaval. Existence and approximation of a (regularized) Oldroyd-B model. *(preprint submitted for publication http ://fr.arxiv.org/abs/0907.4066)*, 2009.

Première partie

Discrétisations d'une équation constitutive dans un modèle de fluide viscoélastique par la méthode des éléments finis

Chapitre 1

Introduction à la modélisation des fluides viscoélastiques micro-structurés

1-I Problèmatique : rhéologie et modèles multi-échelles

La *rhéologie* cherche à décrire le comportement mécanique des matériaux complexes, en particulier les *fluides non-newtoniens* [BHW89]. Plus précisément, elle cherche à décrire quantitativement, pendant un intervalle de temps $t \in (0,T)$, la déformation et l'écoulement, sous l'effet de forces extérieures, d'une portion de matière condensée modélisée comme un milieu continu, sans faire ni l'approximation *newtonienne* (pour les liquides) ni l'approximation *hookéenne* (pour les solides).

La distinction de nature entre solides et liquides n'est pas évidente, elle se fait usuellement *via* leur comportement rhéologique : quand ils sont soumis à un champ de forces extérieures, les solides (élastiques) se déforment instantanément, puis adoptent une forme finale où persiste une déformation (finie) après un laps de temps fini, par opposition aux liquides (visqueux) qui changent continuellement de forme dès qu'une force extérieure leur est appliquée[1]. Mais les notions de temps et déformation finis dépendent naturellement de l'observateur, c'est-à-dire d'un temps et d'une distance caractéristiques (en particulier, tout matériau peut être liquide ou solide selon ce point de vue). Aussi, la rhéologie tente-t-elle de dépasser cette limitation et de définir la nature des matériaux complexes (en fait, tous les matériaux réels) par une formule mathématique de leur comportement. Les *fluides viscoélastiques* sont ceux dont le comportement est intermédiaire entre les approximations extrêmes du solide hookéen et du liquide newtonien.

Nous nous intéressons ici au comportement non-newtonien en tant qu'écart au comportement newtonien (idéal) des liquides, c'est pourquoi nous allons maintenant utiliser une description *eulérienne* des déplacements de la matière (par opposition à la description *lagrangienne* utilisée en mécanique des solides[2]). Si on repère l'espace par la variable $\boldsymbol{x} \in \mathbb{R}^3$ dans un référentiel galiléen (attaché au laboratoire), la portion de matière considérée ici dans la suite occupera typiquement un domaine ouvert régulier $\mathcal{D} \subset \mathbb{R}^3$, borné connexe indépendant du temps et de mesure de Lebesgue non-nulle[3]. La matière sera considérée incompressible et isotherme (c'est évidemment une grossière approximation si on veut modéliser de nombreux procédés industriels, il s'agit donc d'une approche simplifiée des liquides non-Newtoniens). Le champ (scalaire) de *masse volumique*, $\rho : (t,\boldsymbol{x}) \in \mathcal{D}_T := (0,T) \times \mathcal{D} \to \rho(t,\boldsymbol{x}) \in \mathbb{R}_{\geq 0}$, est alors constant (un même volume a toujours la même masse dans un tel fluide) :

$$\rho(t,\boldsymbol{x}) \equiv \rho \in \mathbb{R}_{>0}\,, \forall (t,\boldsymbol{x}) \in \mathcal{D}_T\,.$$

Dans la description eulérienne d'un fluide, on introduit ensuite la variable *champ de vitesse*, c'est-à-dire un champ de vecteurs $\boldsymbol{u} : (t,\boldsymbol{x}) \in \mathcal{D}_T \to \boldsymbol{u}(\boldsymbol{x}) \in \mathbb{R}^3$. Le principe de *conservation de la masse* avec l'hypothèse d'incompressibilité impose au champ de vitesse d'être solénoïdal :

$$\operatorname{div} \boldsymbol{u} = 0 \text{ dans } \mathcal{D}. \tag{1-I.1}$$

[1] En particulier, sur Terre, on définit souvent les liquides par leur capacité à prendre spontanément la forme de leur récipient (domaine dans lequel ils sont inclus et ne peuvent sortir) sous l'effet des forces de pesanteur.

[2] Toutefois, certains auteurs utilisent parfois une description lagrangienne, équivalente à notre description eulérienne pourvu que les conditions d'écoulement permettent d'établir cette équivalence, en vue d'obtenir certains résultats mathématiques [LLZ05, LX06].

[3] Néanmoins, les problèmes non-newtoniens à surface libre ne sont pas inintéressants, on consultera par exemple [SR94, LM95, BPL06].

Ensuite, le principe de *conservation de la quantité de mouvement* en présence d'un champ volumique de forces extérieures par unité de masse $\boldsymbol{f}:(t,\boldsymbol{x})\in\mathcal{D}_T\to\boldsymbol{f}(t,\boldsymbol{x})\in\mathbb{R}^3$ s'écrit (sous forme locale) :

$$\rho\left(\frac{\partial\boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)=\operatorname{div}\boldsymbol{T}+\rho\boldsymbol{f}\text{ dans }\mathcal{D}_T,\tag{1-I.2}$$

où $\boldsymbol{T}:(t,\boldsymbol{x})\in\mathcal{D}_T\to\boldsymbol{T}(t,\boldsymbol{x})\in\mathbb{R}^{3\times3}$ est le tenseur *champ de contraintes* de Cauchy qui modélise les efforts intérieurs au fluide. (Les forces internes, d'origine moléculaire, sont supposées transmises uniquement par contacts surfaciques et localement dépendantes uniquement de l'orientation de la surface de contact.) Pour fermer le système d'équations, en plus de conditions au bord de $\mathcal{D}_T$, il manque une relation liant $\boldsymbol{T}$ aux autres variables du système telles $\boldsymbol{u}$.

A chaque instant $t\in(0,T)$ et en chaque point $\boldsymbol{x}\in\mathcal{D}$, la valeur de $\boldsymbol{T}(t,\boldsymbol{x})$ dépend théoriquement de la configuration exacte des molécules au voisinage du point $\boldsymbol{x}$ dans le matériau, ce qui n'est *a priori* pas donné par les valeurs instantanées des champs $\boldsymbol{u}(t,\cdot)$ et $p(t,\cdot)$. De nombreuses expériences ont néanmoins montré que, pour un grand nombre de matériaux, on pouvait exprimer $\boldsymbol{T}(t,\cdot)$ comme une fonction de ces champs instantanés. L'approximation *newtonienne* du comportement suppose ainsi que le tenseur $\boldsymbol{T}$ s'exprime comme une fonction affine du tenseur *taux de déformation* $\boldsymbol{d}=\frac{1}{2}\left(\boldsymbol{\nabla}\boldsymbol{u}+\boldsymbol{\nabla}\boldsymbol{u}^T\right)$,

$$\boldsymbol{T}=-p\boldsymbol{I}+2\eta\boldsymbol{d},\tag{1-I.3}$$

où l'on a introduit le champ (scalaire) de pression $p:(t,\boldsymbol{x})\in\mathcal{D}_T\to p(t,\boldsymbol{x})\in\mathbb{R}$ (le multiplicateur de Lagrange associé à une contrainte isotrope permettant de satisfaire la condition d'incompressibilité (1-I.1)), le tenseur identité $\boldsymbol{I}$, le champ tensoriel *gradient de vitesse* $\boldsymbol{\nabla}\boldsymbol{u}:(t,\boldsymbol{x})\in\mathcal{D}_T\to\boldsymbol{\nabla}\boldsymbol{u}(t,\boldsymbol{x})\in\mathbb{R}^{3\times3}$ à valeurs une matrice 3×3 de composantes $\left(\frac{\partial\boldsymbol{u}_i}{\partial\boldsymbol{x}_j}\right)_{1\le i,j\le3}$, et la constante de *viscosité* η (un nombre propre au fluide newtonien en question). Le système fermé d'équations (1-I.2–1-I.3) est connu sous le nom d'équations de Navier-Stokes incompressibles, il décrit bien les comportements de l'eau et de la glycine (entres autres exemples) dans de nombreux régimes d'écoulement.

Toutefois, il est nécessaire d'améliorer l'hypothèse du comportement newtonien pour décrire des phénomènes qui ne sont pas le résultat d'un équilibre entre les forces visqueuses (la résistance à un cisaillement) et les forces d'inertie (la résistance à une accélération), par exemple :

- la montée de liquide le long d'un axe tournant dans ce liquide (dit effet Weissenberg),
- le gonflement d'un jet de liquide à la sortie d'une extrusion,
- le siphonnage sans tube d'un liquide hors de son récipient,

qu'on observe en manipulant des fluides tels la peinture, la boue, le ciment, ou le yoghourt [BCAH87a, BHW89, Ren00].

Une interprétation physique des effets non-newtoniens suppose l'existence d'une *micro-structure* à l'intérieur du fluide, qui atteint un état d'équilibre thermodynamique quand le fluide est dans un état stationnaire. Dans ces liquides micro-structurés, les effets non-newtoniens se manifesteraient quand le temps de relaxation microscopique (un temps caractéristique de l'évolution spontanée des molécules vers le nouvel état d'équilibre thermodynamique après une sollicitation) ne serait pas négligeable par rapport au temps de relaxation macroscopique (un temps caractéristique de l'évolution macroscopique du fluide vers le nouvel état stationnaire à l'échelle du continu). Pour décrire cela, les modèles mathématiques introduisent alors un nouveau nombre adimensionnel : le nombre de Deborah De , (rapport entre les échelles de temps de la relaxation spontanée locale de la micro-structure et de la relaxation globale du fluide sous sollicitation) ou le nombre de Weissenberg Wi (rapport entre les échelles de temps de la relaxation spontanée locale de la micro-structure et un temps caractéristique de l'expérience à l'échelle macroscopique). Les effets non-newtoniens apparaissent dès que ces nombres (caractéristiques d'un fluide) sont de l'ordre de 1. (Noter que cette interprétation physique fait donc l'hypothèse d'une séparation des échelles micro et macro en espace et en temps : reste à décrire précisément chacune d'entre elles.) On appelle aussi *liquides élastiques* ou *fluides à mémoire* ces fluides non-newtoniens avec une micro-structure qui se déforme élastiquement, tels que $\boldsymbol{T}(t,\cdot)$ dépend de toute l'histoire des déformations du matériau.

Pour améliorer (1-I.3) en vue de caractériser quantitativement le comportement de fluides non-newtoniens micro-structurés, deux approches sont possibles :

- l'approche purement macroscopique, qui cherche à remplacer directement (1-I.3) par une *équation constitutive* entre les variables macroscopiques $\boldsymbol{u},p,\boldsymbol{T},\ldots$ à l'échelle du continu, par exemple avec une *équation aux dérivées partielles* (EDP) ou une formulation intégrale, ou

– l'approche micro-macro, qui cherche d'abord à décrire mathématiquement la micro-structure, puis somme les effets microscopiques pour les intégrer à la description macroscopique via une relation définissant le tenseur $\boldsymbol{T}$, par exemple une moyenne statistique sur un ensemble de configurations aléatoires caractérisant la micro-structure.

Dans les deux approches, des échelles de temps et d'espace différentes interviennent *via* De ou Wi : tous ces modèles sont donc multi-échelles en ce sens. Néanmoins, seuls les modèles micro-macro figurent une description mathématique précise des différentes échelles (en fait, de la micro-structure).

Il est à noter que les deux approches sont classiques en sciences des matériaux. Aussi, dans cette thèse, nous traitons de quelques questions mathématiques (essentiellement numériques) qui sont communes à de nombreux modèles multi-échelles. Les modèles de fluides non-newtoniens sont un paradigme mathématique pour une large classe de modèles multi-échelles [LB05].

En particulier, dans une approche purement macroscopique, de nombreuses approximations conduisent classiquement à une réduction de la réalité. Il faut néanmoins prendre garde à ne pas perdre d'information dans le processus d'approximation : si les principes d'invariance galiléenne sont en général respectés par les équations constitutives, il semble que les principes thermodynamiques soient parfois difficiles à concilier avec ces approximations (*c.f.* [WH98b, Grm98, WH98a], ou plus récemment [Ö05]), alors qu'ils sont fondamentaux pour comprendre les propriétés de stabilité du modèle mathématique, notamment en temps long [JLBLO06]. D'ailleurs, nous montrons dans la première partie de cette thèse que cette notion de stabilité (existence d'une loi de dissipation en vertu des principes thermodynamiques), qui n'est donc pas facilement identifiable pour une équation constitutive, n'est pas non plus facilement imitable par les discrétisations d'une équation consitutive dont on sait pourtant exprimer la loi de dissipation (voir les chapitres 2 et 3). Ceci pourrait expliquer certains phénomènes d'instabilité observés dans la simulation numérique des équations constitutives (dont certains Problèmes à Grand Nombre de Weissenberg, voir [OP02]), et montre en tout cas qu'il peut être difficile d'utiliser rigoureusement une équation constitutive pour décrire un fluide viscoélastique (en particulier pour la simulation numérique d'écoulements complexes).

Quant à la seconde approche dite micro-macro, elle recourt à des équations pour la micro-structure qui sont souvent posées sur un espace de grande dimension. Leur simulation numérique est donc généralement difficile et il faut réfléchir à de nouvelles méthodes numériques pour ces *simulations moléculaires* en sciences de la matière[4]. C'est ce que nous ferons dans la deuxième partie de cette thèse (voir le chapitre 7, auquel on est amené après les chapitres 5 et 6).

Chacune des deux approches est abordée par l'étude d'un seul modèle : un modèle d'Oldroyd pour la première approche (partie I de cette thèse) et un modèle d'haltères pour la seconde (en fin de partie II). Ces deux modèles conviennent bien pour décrire quelques fluides non-newtoniens, dont les *liquides polymériques dilués*[5]. Certes, ils ne contiennent pas toute la richesse des comportements mécaniques que l'on peut retrouver dans la zoologie des fluides non-newtoniens, même issus des liquides polymériques dilués uniquement. Toutefois, chacun de ces deux modèles possède à la fois une formulation mathématique suffisamment simple pour être étudiée précisément et une formulation commune à de nombreux autres modèles non-newtoniens. Ils sont donc de bons candidats pour être les supports à l'étude de questions relatives à la discrétisation numérique intéressant de nombreux modèles non-newtoniens. Nous introduisons ces modèles dans les deux sections suivantes.

1-II Equations constitutives : l'exemple d'Oldroyd-B

On appelle *équation constitutive* une loi de comportement reliant directement $\boldsymbol{T}$ à $\boldsymbol{u}$ (de type viscoélastique ici), qui est obtenue par une modélisation purement macroscopique d'un fluide non-newtonien.

Rappelons que la condition d'incompressibilité implique l'existence d'un paramètre de Lagrange p, égal à une composante sphérique du tenseur des contraintes $\boldsymbol{T}$ et appelé pression hydrostatique (par similitude avec les seuls efforts isotropes qui s'exercent dans un fluide au repos). On note donc en général

$$\boldsymbol{T} = -p\boldsymbol{I} + \boldsymbol{\tau}\,, \tag{1-II.1}$$

[4]Noter que la thermodynamique joue là aussi un rôle important, puisque les simulations numériques micro-macro consistent en général à échantillonner directement l'espace des phases pour la micro-structure. De ce point de vue, la principale différence entre les modèles micro-macro et les modèles macroscopiques est que les premiers se réfèrent explicitement à un cadre thermodynamique (certes postulé) pour ensuite décrire l'état d'un fluide complexe, tandis que les seconds décrivent directement l'état d'un fluide complexe par des variables seulement supposées être issues d'un cadre thermodynamique (souvent non formulé explicitement). Pour une discussion plus physique, on renvoie par exemple à [BE94, WH98b, Grm98, WH98a, Ö05].

[5]Les *suspensions* sont une autre classe importante de fluides non-newtoniens [BHW89]. Elles sont décrites par d'autres modèles, que nous n'évoquerons pas ici, quoiqu'ils suggèrent également des questions mathématiques.

où la partie déviatorique $\boldsymbol{\tau}$ du tenseur des contraintes internes est appelée tenseur des *extra-contraintes*.

Il s'agit alors de caractériser $\boldsymbol{\tau}$ à l'aide d'une équation faisant intervenir les trois inconnues $(\boldsymbol{u},p,\boldsymbol{\tau})$. Comme $\boldsymbol{\tau}$ est un tenseur (donc un champ caractérisé en chaque point de $\mathcal{D}$ par la valeur d'une matrice à 3×3 composantes), il faut théoriquement 9 équations scalaires. En fait, on fait souvent l'hypothèse que les points matériels portant la contrainte $\boldsymbol{T}$ ont un moment angulaire nul[6], ce qui implique que $\boldsymbol{T}$ (donc aussi $\boldsymbol{\tau}$) est un tenseur symétrique à 6 composantes.

De plus, le principe d'objectivité exige en mécanique (classique) des milieux continus que les lois de comportement soient invariantes aux changements de repère par mouvements de corps rigide. Ceci restreint la forme des opérateurs intégro-différentiels utilisables pour exprimer ces lois [BHW89]. Parmi les différents choix possibles, nous étudierons ici des modèles différentiels du type Oldroyd qui n'emploient qu'une dérivée matérielle dite *sur-convectée*, car c'est exactement cette dernière que nous retrouverons par une approche micro-macro simple dans la Section 1-III. On pose alors

$$\boldsymbol{\tau} = \boldsymbol{\tau}_s + \boldsymbol{\tau}_p \tag{1-II.2}$$

où $\boldsymbol{\tau}_s = 2\eta_s \boldsymbol{d}$ sont les extra-contraintes dans le solvant pur (en l'absence de polymères), supposées newtoniennes avec une viscosité η_s, et où $\boldsymbol{\tau}_p$ sont les extra-contraintes dues à la présence de polymères dans le fluide non-newtonien. L'équation constitutive d'Oldroyd-B s'écrit :

$$\lambda\Big(\partial_t \boldsymbol{\tau}_p + (\boldsymbol{u}\cdot\nabla)\boldsymbol{\tau}_p - (\boldsymbol{\nabla u})\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\boldsymbol{\nabla u}^T)\Big) + \boldsymbol{\tau}_p = 2\eta_p \boldsymbol{d}, \tag{1-II.3}$$

où $\boldsymbol{\nabla u}^T$ désigne la matrice transposée de $\boldsymbol{\nabla u}$, λ est un temps caractéristique des polymères et η_p un coefficient de viscosité propre aux polymères.

L'EDP (1-II.3) ainsi construite devrait permettre de décrire le comportement de certains fluides dans certains écoulements : ceux tels que les paramètres λ et η_p varient peu dans $\mathcal{D}_T$[7]. Pour des fluides supposés tels, on calibre alors les paramètres de l'EDP pour qu'ils coïncident avec des observations expérimentales, typiquement mesurées grâce à des *rhéomètres* dans des flots particuliers, dits *viscométriques*. Néanmoins, on n'est pas sûr que ces valeurs restent les mêmes (ou seulement constantes) dans des flots complexes, ce qui limite la prédictibilité des équations constitutives pour un fluide donné. De plus, on aimerait qu'une bonne équation constitutive reproduise, dans un flot complexe, les divers effets non-newtoniens observés selon les fluides utilisés. Or, la plupart des équations ne peuvent reproduire qu'un nombre limité d'effets non-newtoniens quelles que soient les valeurs choisies pour les paramètres. Voire, aucune équation constitutive connue ne reproduit certains effets non-newtoniens pourtant observés en pratique. Pour un flot donné (même simple, viscométrique par exemple), la prédictibilité des équations constitutives est donc aussi très limitée[8].

Remarque 1 (Adéquation aux expériences de l'équation constitutive (1-II.3))**.** *Le modèle d'Oldroyd-B* (1-II.3) *ne peut pas reproduire certains effets non-newtoniens usuels, même dans des écoulements simples tels que le flot laminaire cisaillé dit de Couette. (La viscosité ne dépend pas du taux de cisaillement et deux contraintes normales sont nécessairement égales.) De plus, dans un flot élongationnel, son comportement est qualitativement incorrect pour tous les fluides non-newtoniens. (Les viscosités élongationnelles explosent.) On peut consulter [RH88, Ren00, TS07a, BPP08] sur les divers défauts de ce modèle. Néanmoins, on utilise en pratique de nombreuses variantes à l'EDP* (1-II.3)*, qui corrigent immédiatement ces défauts en ajoutant un terme non-linéaire. C'est pourquoi nous travaillons ici avec le modèle d'Oldroyd-B* (1-II.3)*, pierre angulaire des modèles de Giesekus, de Phan-Thien Tanner (ou PTT), FENE-P,etc. [BCAH87a]. Dans le futur, nous souhaiterions prolonger nos travaux sur des méthodes de discrétisation stables à des variantes d'Oldroyd-B telles que FENE-P [BB09a] :*

$$\lambda\Big(\partial_t \boldsymbol{\tau}_p + (\boldsymbol{u}\cdot\nabla)\boldsymbol{\tau}_p - (\boldsymbol{\nabla u})\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\boldsymbol{\nabla u}^T)\Big) + \left(\frac{1}{1-\mathrm{tr}(\boldsymbol{\tau}_p)/b}\right)\boldsymbol{\tau}_p = 2\eta_p \boldsymbol{d}. \tag{1-II.4}$$

[6] Cette hypothèse est évidemment fausse si les points sont, par exemple, chargés électriquement, et qu'un couple électromagnétique s'exerce sur chacun d'eux. Le milieu est alors dit *polaire* ou *polarisé*.

[7] Noter que ces hypothèses reposent d'ailleurs sur le cadre thermodynamique sous-jacent dont on a parlé précédemment, qui est supposé satisfait par la micro-structure – non-explicite – du fluide [BE94, WH98b, WH98a, Ö05].

[8] A noter qu'une classe très simple d'équations constitutives n'a pas été évoquée directement à travers l'exemple (1-II.3), où l'on avait mentionné des contraintes sur les opérateurs intégro-différentiels liées au principe d'objectivité matérielle. C'est celle des relations algébriques entre les différentes variables $\boldsymbol{u},\boldsymbol{T}$, sans opérateur intégro-différentiel, qui semblent convenir particulièrement bien à la description des flots "lents". En particulier, dans cette catégorie, une extension simple du modèle newtonien quand les effets dynamiques sur $\boldsymbol{\tau}$ sont peu importants est le modèle newtonien généralisé [Emm08] :

$$\boldsymbol{\tau} = 2\eta(\boldsymbol{d})\boldsymbol{d}.$$

Le choix du modèle n'est donc pas totalement décorrélé du flot pour lequel on l'utilise [CDW84]. Néanmoins, pour certains flots particuliers, comme la sortie d'une extrusion, il semble qu'il n'y ait même pas encore de modèle satisfaisant, qui décrirait par exemple certaines instabilités observées expérimentalement [Ren00].

Remarque 2 (Nonlinéarité de l'équation constitutive (1-II.3))**.** *Outre (classiquement en mécanique des fluides) la dérivée matérielle* $(\boldsymbol{u}\cdot\nabla)\boldsymbol{\tau}_p$*, le terme sur-convectif* $-(\boldsymbol{\nabla u})\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\boldsymbol{\nabla u}^T)$ *dans le modèle d'Oldroyd-B* (1-II.3) *est aussi non-linéaire en le champ d'inconnues. C'est une première manifestation d'un point générique qu'il faut garder à l'esprit : les équations constitutives décrivant correctement des effets non-newtoniens sont nécessairement non-linéaires.*

Historiquement, les premiers modèles étaient pourtant linéaires. Ainsi l'ancêtre de (1-II.3) *est le modèle (linéaire) de Maxwell, où* $\boldsymbol{\tau}_s = 0$ *et*

$$\lambda\,\partial_t\boldsymbol{\tau} + \boldsymbol{\tau} = 2\eta_p\boldsymbol{d}. \tag{1-II.5}$$

(Ce modèle s'interprète comme une extension simple du cas newtonien $\boldsymbol{\tau} = 2\eta_p\boldsymbol{d}$ *par l'introduction d'un temps caractéristique* λ *dit de relaxation : une fonction de Heaviside pour le cisaillement produit en effet un Dirac pour la contrainte newtonienne correspondante, tandis que* (1-II.5) *régularise le Dirac en une fonction du type* $e^{-\frac{t}{\lambda}}$*.) Mais même dans les flots tels que celui de Couette où la dérivée sur-convective s'annule, ainsi qu'on l'a déjà dit dans la Remarque 1, il semble que des nonlinéarités soient nécessaires pour décrire les phénomènes non-newtoniens importants.*

C'est pourquoi dans la suite, nous basons notre étude sur le premier modèle différentiel non-linéaire simple, le modèle d'Oldroyd-B (1-II.3)*, plutôt que* (1-II.5) *par exemple. Certes, puisque l'EDP* (1-II.3) *est encore affine en la seule variable* $\boldsymbol{\tau}_p$*, il faudrait encore introduire d'autres termes non-linéaires pour le rendre plus proche de la réalité non-newtonienne comme la thixotropie ou la rhéopexie : doubler les conditions initiales d'un flot ne doit pas entraîner une multiplication par deux de* $\boldsymbol{T}$ *tout au long de l'écoulement. (D'ailleurs, dans un flot laminaire cisaillé, le modèle d'Oldroyd-B* (1-II.3) *redevient même linéaire, ce qui explique en partie pourquoi il n'est alors plus correct physiquement [RH88].) Nous réservons toutefois l'étude de non-linéarités supplémentaires pour de futurs travaux [BB09a] : nos premières études, qui correspondent donc au cadre non-linéaire le plus simple, montrent que déjà de nombreuses difficultés surviennent dans la simulation numérique des schémas non-linéaires correspondants (voir la Section 2-VII).*

Paradoxalement, l'intérêt des modèles utilisant des équations constitutives réside essentiellement dans leur simplicité, car ils sont relativement peu coûteux à simuler numériquement (par opposition aux modèles micro-macro de la Section 1-III). Nous revenons néanmoins sur la discrétisation des EDPs constitutives à la Section 1-IV-A-b : contrairement aux apparences, elle n'est pas si aisée que cela.

1-III Modèles micro-macro : l'exemple des haltères

De nombreux modèles micro-macro, imaginant diverses micro-structures, ont été proposés pour décrire des liquides polymériques. L'idée de départ est l'espoir d'obtenir une bonne description des effets non-newtoniens (viscoélastiques) à l'échelle du continu, en tenant compte précisément de leur origine physique (c'est-à-dire au plus proche des molécules). Dans les liquides polymériques, cette origine est principalement les forces intramoléculaires[9] agissant sur la conformation des polymères en solution.

On peut distinguer diverses classes de modèles micro-macro pour décrire des fluides polymériques [BCAH87b, DE98] : les modèles de chaînes connectant des billes entre elles, des modèles de reptation de segments prisonniers dans un tube, des modèles de réseaux..., qui conviennent plus ou moins bien selon le fluide considéré. Pour les modèles de chaînes en particulier, des théories cinétiques de la physique statistique permettent de bien décrire les forces intramoléculaires pour un ensemble statistique de polymères dilués en solution (au moins quand les polymères sont à l'équilibre thermodynamique, par exemple au sein d'un écoulement stationnaire[10]). Dans la suite, nous considérerons uniquement le plus simple des modèles de chaînes (pour des liquides polymériques dilués), celui des haltères (deux billes connectées par un ressort, *dumbbells* en anglais).

Le modèle des haltères [BCAH87b, Ö96] réduit un polymère en solution au concept d'un vecteur matériel $\boldsymbol{X}_t$ décrivant l'orientation et l'élongation de la molécule (voir Fig. 1.1). A chaque extrémité du vecteur, on imagine ainsi une bille sphérique, qui porte une certaine masse. Dans un solvant au repos, la vitesse de chacune des deux billes massiques prise isolément à une extrémité du vecteur satisfait alors une équation de Langevin, expression de l'équilibre des forces entre (i) les collisions *browniennes* dues aux mouvements aléatoires des particules du solvant et (ii) la force de friction qui résulte du mouvement de chacune des billes dans le solvant. En fait,

[9] A noter que la principale origine des effets non-newtoniens pour l'autre grande classe de fluides non-newtoniens, les suspensions, est au contraire les forces intermoléculaires [Isr92].

[10] L'approche micro-macro pour décrire un écoulement quelconque où les polymères ne sont plus à l'équilibre thermodynamique est encore un sujet de recherche, comme la plupart des phénomènes thermodynamiques hors-équilibre. De même, on rappelle que les modèles macroscopiques n'échappent pas aux difficultés thermodynamiques, *a fortiori* celles de formalisation d'une thermodynamique hors-équilibre [WH98b, Grm98, WH98a, Ö05].

chacune des deux billes subit aussi une force d'interaction avec l'autre bille, que l'on choisit spécifiquement à une molécule pour modéliser l'action des forces intramoléculaires dans le polymère. On dit souvent que les deux billes sont *connectées* par un ressort de force $\boldsymbol{F}(\boldsymbol{X}_t)$. Puis, en introduisant un espace de probabilités Ω, diverses approximations d'échelle, essentiellement liées à la petite taille des polymères en solution par rapport aux conditions macroscopiques de l'écoulement, mènent finalement à une Equation aux Dérivées Partielles Stochastique (EDPS) pour le champ aléatoire vectoriel $\boldsymbol{X}_t : (t,\boldsymbol{x},\omega) \in \mathcal{D}_T \times \Omega \to \boldsymbol{X}_t(\boldsymbol{x},\omega) \in \mathbb{R}^3$:

$$d\boldsymbol{X}_t + \boldsymbol{u}\cdot\nabla_{\boldsymbol{x}}\boldsymbol{X}_t dt = \left((\nabla_{\boldsymbol{x}}\boldsymbol{u})\,\boldsymbol{X}_t - \frac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t) \right) dt + 2\sqrt{\frac{k_B T_B}{\zeta}}\, d\boldsymbol{B}_t\,, \tag{1-III.1}$$

qui décrit donc en chaque point $(t,\boldsymbol{x}) \in \mathcal{D}_T$ du domaine macroscopique l'équilibre (statistique) entre :

- la force d'interaction intramoléculaire $\boldsymbol{F}(\boldsymbol{X}) = \nabla\Pi(|\boldsymbol{X}|)$, dite aussi entropique car, pour des calculs précis avec une molécule de polymère donnée, on la dérive typiquement d'un pré-calcul où Π est un potentiel thermodynamique égal à l'énergie libre du polymère à l'équilibre thermodynamique[11],
- la force de friction avec le milieu fluide, supposée proportionnelle à la vitesse relative des billes par rapport au fluide avec un coefficient ζ (selon la loi de Stokes pour une sphère dure dans un milieu infini), et
- la force brownienne $2\sqrt{\frac{k_B T_B}{\zeta}}\,\frac{d}{dt}\boldsymbol{B}_t$, typiquement une collection de bruits blancs en temps, par exemple décorrélés en espace[12], d'intensité donnée selon une relation de fluctuation-dissipation par la température T_B du bain que constitue le solvant (k_B étant la constante de Boltzman).

Outre la température T_B, des paramètres physiques sont à ajuster pour que l'équation (1-III.1) décrive correctement la statistique de polymères en solution : ils sont cachés dans la formule explicite de la force $\boldsymbol{F}(\boldsymbol{X})$. Le choix générique le plus simple pour $\boldsymbol{F}$ est linéaire : $\boldsymbol{F}(\boldsymbol{X}) = H\boldsymbol{X}$. C'est le modèle hookéen[13] (*Hookean dumbbells* en anglais), avec un seul paramètre $H > 0$. En pratique, on choisit plutôt pour modèle générique des forces "explosives" (plus réalistes), qui limitent l'extensibilité du vecteur $\boldsymbol{X}$: $\mathbf{F}(\boldsymbol{X}) = \frac{H\boldsymbol{X}}{1-|\boldsymbol{X}|^2/b}$ (*FENE dumbbells* en anglais, pour *Finitely-Extensible Nonlinear Elastic*), avec un paramètre additionnel $\sqrt{b} > 0$ qui correspond à la taille maximale[14] du vecteur $\boldsymbol{X}$. Noter que ces deux forces dérivent bien d'un potentiel. Un choix naturel pour la condition initiale du champ aléatoire $\boldsymbol{X}_t$, quand $\boldsymbol{\nabla}_{\boldsymbol{x}} u(t=0,\cdot) = \kappa$ est un champ symétrique uniforme sur $\mathcal{D}$, est alors un champ $\boldsymbol{X}_0$ stationnaire, de loi proportionnelle à $e^{\frac{1}{2}\kappa\boldsymbol{X}\cdot\boldsymbol{X} - \frac{2}{\zeta}\Pi(|\boldsymbol{X}|)}$, ce qui correspond à la notion d'équilibre thermodynamique invoquée pour calculer le potentiel Π (voir aussi la Remarque 5).

Enfin, on couple la dynamique de polymères (1-III.1) avec l'évolution du champ de vitesse (1-I.2) par une relation de fermeture, complétant (1-II.1–1-II.2) avec une expression pour la contribution des polymères au champ de contrainte :

$$\boldsymbol{\tau}_p(t,\cdot) = n_p \mathbb{E}\left(\boldsymbol{X}_t \otimes \mathbf{F}(\boldsymbol{X}_t) - 2k_B T_B \boldsymbol{I}\right), \tag{1-III.2}$$

où $\mathbb{E}$ désigne l'espérance pour la mesure de probabilité $\mathbb{P}$ et n_p la concentration de polymères par unité de volume. L'équation (1-III.2) s'appelle *relation de Kramers* ; elle s'obtient rigoureusement à partir de la définition du tenseur de Cauchy comme tenseur des contraintes internes au fluide. Les contraintes sont ici les forces entre deux parties du fluide en contact qui sont exercées par les polymères à travers la surface de contact [BCAH87b].

Remarque 3. *La distribution de probabilité $\psi(t,\boldsymbol{x},\boldsymbol{X})$ du champ vectoriel aléatoire $\boldsymbol{X}_t$ vérifie pour tout $t \in (0,T)$, $\boldsymbol{x} \in \mathcal{D}$ et $\boldsymbol{X} \in \mathbb{R}^3$ l'équation maîtresse (dite aussi de Fokker-Planck, Kolmogorov ou Smoluchowsky) :*

$$\partial_t\psi + (\boldsymbol{u}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})\psi = -\operatorname{div}_{\boldsymbol{X}}\left([(\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{u})\boldsymbol{X} - \frac{2}{\zeta}\mathbf{F}(\boldsymbol{X})]\psi \right) + \frac{2k_B T_B}{\zeta}\Delta_{\boldsymbol{X}}\psi\,. \tag{1-III.3}$$

[11] La détermination de la force $\boldsymbol{F}(|\boldsymbol{X}|)$ se fait par des arguments de physique statistique. Typiquement, on peut modéliser plus finement l'architecture d'un polymère que par le concept grossier d'une haltère – en fait une échelle mésoscopique –, en utilisant par exemple une suite de petits segments rigides connectés entre eux. De la connaissance de la distribution de l'orientation de chaque segment à l'équilibre thermodynamique, on en déduit alors la statistique de l'extension totale $|\boldsymbol{X}|$ du polymère (voir par exemple [DE98]), de loi $\phi(|\boldsymbol{X}|)$, puis l'énergie libre $\Pi = -k_B T_B \ln(\phi)$. Noter que c'est dans cette étape-ci, lors de la détermination de la force $\boldsymbol{F}$, qu'on postule l'équilibre thermodynamique "local" des polymères, une limitation des modèles micro-macro déjà évoquée. Par ailleurs, noter aussi que ϕ (donc Π) est ici nécessairement une fonction radiale, dépendant uniquement de l'extension $|\boldsymbol{X}|$ du polymère, puisqu'il n'y a pas de raison pour que la statistique de l'équilibre thermodynamique "local" soit liée à l'orientation du dumbbell dans le solvant.

[12] Le statut des corrélations en espace n'est pas totalement figé [JLBL04].

[13] Le modèle hookéen, où $\Pi(\boldsymbol{X}) = \frac{1}{2}H|\boldsymbol{X}|^2$, correspond à une description fine telle que les segments rigides sont de même taille, d'orientations équidistribuées uniformes [DE98].

[14] On parvient à la formule FENE par l'approximation d'une formule complexe pour $\mathcal{F}$, résultant d'un calcul fin tenant compte des interactions à longue portée au sein de la chaîne de segments rigides. Cette approximation est connue sous le nom d'approximation de Warner, voir par exemple [BCAH87b].)

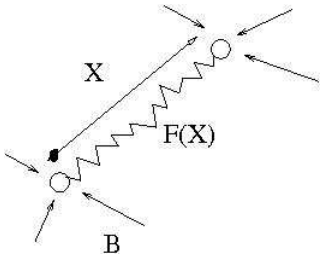


FIG. 1.1 – Concept de dumbbell.

Sa connaissance ne suffit pas à déterminer $\boldsymbol{X}_t$*, mais suffit pour calculer les moments de* $\boldsymbol{X}_t$*, par exemple :*

$$\boldsymbol{\tau}_p(t,\cdot)=n_p\int_{\mathbb{R}^3}(\boldsymbol{X}\otimes\mathbf{F}(\boldsymbol{X})-2k_BT_B\boldsymbol{I})\,\psi(t,\cdot,\boldsymbol{X})d\mathbb{P}(\boldsymbol{X}). \tag{1-III.4}$$

Par ailleurs, on peut aussi exprimer $\boldsymbol{\tau}_p$ *par une formule de Feynman-Kac. Quand* $\boldsymbol{X}_0$ *est uniforme sur* $\mathcal{D}$*, on a*

$$\boldsymbol{\tau}_p(t,\boldsymbol{x})=n_p\mathbb{E}\left(\boldsymbol{C}^{t,\boldsymbol{X}_0}(0,\boldsymbol{x},\boldsymbol{X}_0)-2k_BT_B\boldsymbol{I}\right),\ \forall\boldsymbol{x}\in\mathcal{D},$$

où le tenseur de conformation $\boldsymbol{C}^{t,\boldsymbol{X}_0}$ *satisfait pour tout* $(s,\boldsymbol{x},\boldsymbol{X})\in\mathcal{D}_T\times\mathbb{R}^3$ *l'équation de Kolmogorov rétrograde :*

$$\partial_s\boldsymbol{C}^{t,\boldsymbol{X}_0}(s,\boldsymbol{x},\boldsymbol{X})+\left[-\boldsymbol{u}(s,\boldsymbol{x})\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}+\left((\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{u})\boldsymbol{X}-\frac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X})\right)\cdot\boldsymbol{\nabla}_{\boldsymbol{X}}+\frac{2k_BT_B}{\zeta}\boldsymbol{\Delta}_{\boldsymbol{X}}\right]\boldsymbol{C}^{t,\boldsymbol{X}_0}(s,\boldsymbol{x},\boldsymbol{X})=0 \tag{1-III.5}$$

avec pour condition finale : $\boldsymbol{C}^{t,\boldsymbol{X}_0}(s=t,\boldsymbol{x},\boldsymbol{X})=\boldsymbol{X}\otimes\boldsymbol{F}(\boldsymbol{X})$.

Le calcul de (1-III.2), typiquement par des méthodes Monte-Carlo (voir la Remarque 5), nécessite de simuler un grand nombre de trajectoires pour l'EDPS (1-III.1). C'est le principal désavantage des modèles micro-macro : ils sont coûteux à simuler. Nous revenons sur ce point dans la Section 1-IV-B-b.

Les modèles micro-macro ont au contraire plusieurs avantages sur les modèles purement macroscopiques :
- ils donnent accès à plus d'informations *via* la microstructure,
- ils semblent mieux fondés physiquement (au moins par rapport aux équations constitutives qui n'ont pas de lien explicite avec un modèle micro-macro, voir la Remarque 4) et produisent effectivement, dans les simulations numériques, nombre des effets non-newtoniens observés dans la réalité (Remarque 5),
- enfin, ils semblent numériquement plus stables [MHK09].

On notera que les modèles moléculaires pour les liquides polymériques sont utilisés depuis longtemps (travaux historiques de Kirkwood, Zimm, Rouse,...) comme *concepts*, afin de dériver ensuite, grâce à une théorie cinétique et des relations de fermeture approchées, des équations constitutives (dont celles de Giesekus, PTT ou FENE-P, voir la Remarque 1). C'est le développement des ordinateurs et des méthodes numériques qui permet aujourd'hui de les utiliser en plus pour la simulation numérique d'écoulements non-newtoniens [Ö96] (au moins dans certains régimes proches de l'équilibre thermodynamique en ce qui concerne les polymères en solution).

Remarque 4. *On peut retrouver l'origine physique des équations constitutives dérivées de modèles moléculaires en retraçant mathématiquement les étapes (souvent des approximations !) dans le processus de dérivation. En fait, l'équation d'Oldroyd-B* (1-II.3) *peut-être obtenue exactement en utilisant* (1-III.1) *avec une force entropique hookéenne :* $\mathbf{F}(\boldsymbol{X})=H\boldsymbol{X}$*. Il suffit pour cela d'appliquer la formule d'Itô au tenseur* $\boldsymbol{X}_t\otimes\boldsymbol{X}_t$ *et on retrouve* (1-II.3) *si* $\mathbb{E}\left(\int_0^t\boldsymbol{X}_s\otimes d\boldsymbol{B}_s\right)=0$ *sur* $\mathcal{D}_T$ *(propriété de la martingale* $\boldsymbol{X}_t$*). C'est toutefois une exception : quand on peut retrouver l'origine micro-macro d'une équation constitutive, le processus mathématique implique presque systématiquement de faire des approximations. Ainsi, le modèle FENE-P* (1-II.4) *de la Remarque 2 est construit en utilisant* (1-III.1) *avec une approximation de la force entropique FENE dite de Peterlin (d'où "P") :*

$$\mathbf{F}(\boldsymbol{X}_t)=\frac{H\boldsymbol{X}_t}{1-\mathbb{E}(|\boldsymbol{X}_t|^2)/b},\ \textit{puis}\quad \boldsymbol{\tau}_p=\frac{n_p\mathbb{E}(H\boldsymbol{X}_t\otimes\boldsymbol{X}_t-k_BT_B\boldsymbol{I})}{1-\mathbb{E}(|\boldsymbol{X}_t|^2)/b}.$$

Remarque 5 (Extensions du modèle)**.** *Avec l'EDPS* (1-III.1) *et une force de type FENE, le modèle d'haltères reproduit bien certains effets non-newtoniens : la viscosité dépend du taux de cisaillement et des différences entre les contraintes normales apparaissent (dans le flot laminaire cisaillé de Couette en particulier, voir Remarque 1). Toutefois, il faut encore compliquer le modèle pour obtenir des valeurs quantitativement proches de la réalité. En particulier, on peut alors penser remettre en question certaines hypothèses en vue d'améliorer le modèle :*

- *Le choix d'une condition initiale pour le champ aléatoire* $\boldsymbol{X}_t$ *quand* $\boldsymbol{\nabla}_{\boldsymbol{x}} u(t=0,\cdot)$ *n'est pas un champ symétrique uniforme sur* $\mathcal{D}$ *n'est pas évident du tout [JLBLO06]. Dans ce cas, il n'est même pas sûr que l'on puisse donner un sens à la notion d'équilibre thermodynamique "local". Une piste naturelle pour que l'on retrouve cette notion quand* $\boldsymbol{\nabla}_{\boldsymbol{x}} u(t=0,\cdot)$ *n'est pas un champ symétrique uniforme sur* $\mathcal{D}$ *pourrait être de définir le potentiel* Π *tel qu'il corresponde bien à un état d'équilibre pour un champ* $\boldsymbol{\nabla}_{\boldsymbol{x}} u(t=0,\cdot)$ *stationnaire donné. De manière générale, on pourrait améliorer la formule de la force* $\boldsymbol{F}$ *pour qu'elle corresponde mieux aux forces intramoléculaires dans un écoulement donné.*
- *En établissant* (1-III.1)*, nous avons choisi pour la force de friction une formule très simple,* $-\zeta(d\boldsymbol{X}_t + \boldsymbol{u}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{X}_t dt)$*, qui correspond à la loi de Stokes avec un coefficient de friction isotrope* ζ*. Nous n'avons donc pas tenu compte des interactions hydrodynamiques : par son mouvement propre et la réaction induite dans le solvant ambiant, chaque bille a un effet dans la force de friction vue par l'ensemble des autres billes. Or, même si on se limite à l'effet d'une bille sur les autres billes de la même chaîne – connectées entre elles par des forces intramoléculaires – (en fait, l'unique autre bille du même dumbbell ici), remplacer cette formule par une meilleure approximation telle que celle d'Oseen-Burgers ou celle de Rotne-Prager-Yamakawa conduit tout de suite à une formule (non-linéaire en* $\boldsymbol{X}_t$*) très compliquée (voir (15.4-17) dans [BCAH87b]).*
- *Les polymères sont en général de longues molécules à l'architecture complexe. Même à l'échelle mésoscopique, on pourrait donc les modéliser comme une chaîne de nombreuses billes connectées entre elles par des forces plutôt que par une seule haltère. La nature des forces intramoléculaires peut alors s'enrichir et devenir très non-linéaire (dans ce cas, le choix des coordonnées* $\boldsymbol{X}$ *décrivant la molécule est très important pour décrire le plus simplement des forces-ressorts entre plus proches voisins, mais aussi tenir compte d'effets angulaires, d'exclusions de volume au sein de la chaîne...). Au final, on travaille souvent dans un espace de très grande dimension, avec des opérateurs très non-linéaires. Cela explique pourquoi, en pratique, on calcule* (1-III.2) *avec des méthodes de Monte-Carlo plutôt que* (1-III.3) *avec des éléments finis ou spectraux (et éventuellement* (1-III.4) *avec une formule de quadrature).*
- *Le bain thermique dû aux collisions avec les molécules du solvant a été modélisé comme un mouvement brownien dans* $\mathbb{R}^3$*. Or, la molécule de polymère peut être prisonnière de certaines contraintes géométriques (régime semi-dilué où la présence des molécules voisines n'est plus négligeable), ce qu'on peut modéliser en prenant un autre type de bruit. Ce type de modifications permet d'ailleurs de retrouver une autre classe de modèles, qui est plutôt adaptée aux liquides polymériques concentrés (ou fondus)*[15]*, les modèles de reptation [BCAH87b].*

1-IV Choix de modélisation et questions mathématiques

Bien que l'on se soit limité à quelques exemples simples de modélisation pour des fluides polymériques dilués, de nombreuses extensions, macroscopiques ou micro-macro, sont apparues chaque fois. Se pose donc naturellement la question : quel modèle choisir pour simuler numériquement tel fluide dans tel écoulement ?

Evidemment, le choix du modèle doit d'abord se faire en fonction de critères d'ordre physique : Le modèle peut-il décrire ce qu'on cherche à simuler (quelles sont les bonnes variables, les bons principes physiques à respecter) ? Le modèle peut-il être calibré facilement sur des données expérimentales du fluide (l'expérimentateur ne dispose souvent que de quelques expériences tests pour cela) ?

Les mathématiques peuvent aussi apporter des indications d'un autre ordre sur ce choix : Y a-t-il des solutions aux équations du modèle (de quelle nature sont-elles, quelles hypothèses requièrent-elles) ? Le modèle est-il facile d'utilisation (peut-on calculer des solutions, éventuellement approchées) ?

De manière générale, les modèles macroscopiques sont aujourd'hui ceux utilisés pour simuler des écoulements complexes (dans l'industrie par exemple) vu leur faible coût d'utilisation (complexité du calcul numérique). On utilisera les modèles micro-macro plutôt "en laboratoire" pour simuler finement la physique de petits volumes de fluides dans des géométries simples. Chaque emploi est confronté a des limitations propres : des instabilités numériques subsistent dans les diverses discrétisations proposées pour les modèles macroscopiques (en particulier dans des écoulements complexes pour un nombre de Weissenberg Wi "grand") et les simulations de modèles micro-macro sont fortement limitées en taille par leur complexité numérique. Nous essaierons d'aborder ces deux problèmes successivement, en parties I et II.

Mais auparavant, nous rappelons quelques résultats mathématiques connus pour les deux modèles que nous allons étudier. A cette fin, nous les réécrirons sous forme adimensionnée sans changer le nom des variables.

[15] Noter que pour les polymères concentrés ou fondus, une troisième classe de modèles, dits *de réseau*, propose une autre approche plus simple mathématiquement, en ne modélisant que le squelette des polymères en contact les uns avec les autres.

1-IV-A Le modèle d'Oldroyd-B

1-IV-A-a Théorie mathématique du modèle

Soit $\mathcal{D} \subset \mathbb{R}^d$ un domaine borné connexe de frontière régulière (Lipschitz). Nous aurons $d=3$ dans le cas le plus général, et $d=2$ si le problème physique possède une symétrie de translation. Considérons le problème de Cauchy-Dirichlet suivant, dit à trois champs, d'inconnue (avec $\mathbb{R}^{d\times d}_S$ les matrices $d\times d$ symétriques) :

$$(\boldsymbol{u},p,\boldsymbol{\tau}) : (t,\boldsymbol{x}) \in \mathcal{D}_T \to (\boldsymbol{u}(t,\boldsymbol{x}),p(t,\boldsymbol{x}),\boldsymbol{\tau}(t,\boldsymbol{x})) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^{d\times d}_S$$

satisfaisant sur l'ouvert $\mathcal{D}_T$ le système d'EDP

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right) = -\boldsymbol{\nabla} p + (1-\varepsilon)\Delta\boldsymbol{u} + \mathrm{div}\,\boldsymbol{\tau} + \boldsymbol{f}, \\ \mathrm{div}\,\boldsymbol{u} = 0, \\ \mathrm{Wi}\left(\dfrac{\partial \boldsymbol{\tau}}{\partial t} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\tau} - (\boldsymbol{\nabla}\boldsymbol{u})\boldsymbol{\tau} - \boldsymbol{\tau}(\boldsymbol{\nabla}\boldsymbol{u})^T\right) = \varepsilon\left[\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^T\right] - \boldsymbol{\tau}, \end{cases} \tag{1-IV.1}$$

version adimensionnelle de (1-I.2–1-I.1–1-II.1–1-II.2–1-II.3) où l'on a utilisé trois nombres : $\mathrm{Re}\geq 0$, $\mathrm{Wi}>0$, $\varepsilon\in(0,1)$, les mêmes notations pour les variables adimensionalisées (excepté $\boldsymbol{\tau}_p$, devenu $\boldsymbol{\tau}$), avec un champ de forces extérieures $\boldsymbol{f}$.

La nature des données de bord à adjoindre au système (1-IV.1) pour que le problème de Cauchy associé soit bien posé n'est pas évidente (voir par exemple [Jos90]). On adjoint ici à (1-IV.1) des conditions homogènes de type Dirichlet au bord de $\partial\mathcal{D}$ pour $\boldsymbol{u}$ (également dites "de collement") et des conditions initiales régulières pour $\boldsymbol{u}$ et $\boldsymbol{\tau}$. Le problème possède alors une solution (forte) :

- unique et locale en temps, de plus globale en temps si le terme source $\boldsymbol{f}$, les conditions initiales et ε sont petits, et enfin asymptotiquement stable L^2 si $\boldsymbol{f}$ est périodique [GS90, FCGO02] ;
- unique et locale en temps (avec $\boldsymbol{u}$ satisfaisant un critère de type Beale-Kato-Majda dans L^2), de plus globale en temps si les conditions initiales sont proches de l'état d'équilibre $\boldsymbol{u}=0$, $\boldsymbol{\tau}=\boldsymbol{I}$ [LLZ05, LLZ08],
- globale en temps quand le champ de vitesse est co-rotationnel ($\boldsymbol{\nabla}\boldsymbol{u}=-\boldsymbol{\nabla}\boldsymbol{u}^T$) [LM00],
- unique et locale en temps quand $d=3$, avec $\boldsymbol{\tau}$ satisfaisant un critère de type Beale-Kato-Majda dans L^∞ quand $\mathrm{Re}\to 0$ [KMT08].

Les résultats [GS90, FCGO02] s'étendent à de nombreuses extensions d'Oldroyd-B, dont Giesekus et PTT, et donnent aussi des résultats supplémentaires pour des problèmes stationnaires et des flots *rampants* ($\mathrm{Re}=0$). Des modèles proches d'Oldroyd-B ont aussi été étudiés par ailleurs, voir par exemple [Ren00]. En particulier, nous avons exclu le cas $\varepsilon=1$ où Oldroyd-B dégénère en le modèle de Maxwell (1-II.5) sur-convecté (UCM pour *Upper-Convected Maxwell* en anglais). Les résultats [GS90, FCGO02, BSS05, KMT08] que nous mentionnons s'appuient en effet sur la théorie des équations de Navier-Stokes (voir [Tem84] par exemple). Mais on trouvera dans [LLZ05] le cas où $\varepsilon\to 1$ pour $d=2$.

Même pour des solutions faibles, il n'y a pas (encore) de résultat d'existence globale en temps pour le problème en général. Néanmoins, des problèmes régularisés proches possèdent (au moins) une solution (faible) globale en temps, voir par exemple [BSS05] où la formulation micro-macro avec des haltères hookéennes a été régularisée (on peut normalement retrouver (1-IV.1), au moins dans un sens faible, à partir de ce système), et le chapitre 3 de cette thèse, où l'on a ajouté un terme dissipatif additionnel $\Delta\boldsymbol{\tau}$ dans l'équation d'Oldroyd-B (avec une troncature quand $d=3,2$, ou une condition sur la donnée initiale quand $d=2$).

Enfin, si l'on fait l'hypothèse qu'une solution globale en temps existe et qu'elle est suffisamment régulière, alors on montre dans [HL07] (avec des arguments de [JLBLO06], voir aussi le chapitre 2) qu'elle converge exponentiellement vite vers la solution stationnaire $\boldsymbol{u}=0$, $\boldsymbol{\tau}=\boldsymbol{I}$. (En fait, cela est vrai en général pour des conditions de bord telles qu'un état stationnaire existe, avec des conditions initiales telles qu'une solution globale suffisamment régulière existe, voir aussi la Remarque 5.)

Il y a peu de résultats pour d'autres conditions de bord. Dans [FCGO02], on trouve néanmoins quelques résultats pour l'équation d'Oldroyd-B (1-II.3) seule (découplée du système (1-IV.1)) dans des flots intéressants physiquement tels celui de Poiseuille. Voir aussi [LM95] par exemple, pour des conditions de surface libre.

1-IV-A-b Discrétisations du modèle

De nombreuses discrétisations ont été proposées pour le système (1-IV.1), avec des techniques classiques pour des équations de la mécanique des fluides, voir par exemple l'ouvrage de synthèse [OP02]. Elles utilisent

par exemple des méthodes de différences finies [FK05] dans des géométries simples (telles la cavité entraînée, *lid-driven cavity* en anglais), des méthodes de Galerkin (parfois spectrales [CO01], ou plus souvent par éléments finis [CDW84, HFK05]), ou éventuellement des méthodes de volumes finis. Néanmoins, les techniques de discrétisation qui marchent bien pour des équations de la mécanique des fluides newtoniens se sont heurtées à des difficultés en mécanique des fluides non-newtoniens, dont une large part subsiste encore.

En fait, on observe en pratique que, pour la plupart des équations constitutives utilisées et pour de nombreux écoulements physiquement intéressants (en particulier le flot autour d'un cylindre et la contraction 4 :1), les tentatives de simulation numérique *naïves* basées sur des discrétisations intuitives, sont incapables de converger quand on raffine les pas de discrétisation, voire explosives, à partir d'une valeur critique suffisamment grande du paramètre Wi (valeur néanmoins physiquement réaliste pour des comparaisons avec l'expérience). Ces instabilités numériques sont souvent désignées sous le vocable générique de *problèmes aux grands nombres de Weissenberg* (HWNP pour *High-Weissenberg Number Problem* en anglais) [Keu90, Keu00, OP02].

Les instabilités de type HWNP ont suscité de nombreuses études, théoriques et numériques. Clairement, l'origine des instabilités est parfois liée au manque de régularité des solutions du problème continu que l'on cherche à approcher numériquement [RH88, San99, TS07a, BPP08]. Mais cela n'explique vraisemblablement pas toutes les observations (même si on a encore peu de résultats mathématiques sur la régularité des solutions à approcher pour des conditions de bord générales), ainsi que le suggèrent les améliorations apportées par diverses techniques de discrétisation développées *ad hoc* pour les fluides viscoélastiques et au moins observées dans certains cas tests de simulation numérique.

D'une part, il semble clair que des difficultés numériques peuvent survenir quand le terme de convection $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\tau}$ dans l'équation d'Oldroyd-B est grand. Aussi, de nombreuses discrétisations ont été proposées, qui tentent de stabiliser ce terme par des méthodes proches de stabilisations classiques, comme par exemple la méthode Streamline Upwind / Petrov–Galerkin (SUPG) [BH82] et la méthode Galerkin Least-Square (GaLS) [FS91], ou simplement en introduisant un terme additionnel de diffusion isotrope [SB95]. D'autre part, le travail initial [MC87] a mis en évidence que la nature mixte du problème (à la fois elliptique et hyperbolique) nécessite aussi un traitement particulier[16]. Des discrétisations utilisant des méthodes de stabilisation en partie classiques, et en partie développées spécifiquement pour le couplage de Navier-Stokes avec l'équation d'Oldroyd-B, par exemple obtenues en intervenant :

- dans le choix de la formulation des équations à discrétiser (voir par exemple [CDW84] pour les formulations EEME, EVSS, ou [WKL00] pour Lagrangian tracking, parmi d'autres formulations équivalentes des équations à discrétiser),
- dans le choix des espaces de discrétisation (le choix est vaste en s'autorisant des approximations discontinues par la méthode DG inspirée de [LR74], comme il a été observé initialement dans [FF89]),
- dans le choix de la formulation discrète des équations (dans les discrétisations EVSS-G ou D-EVSS [FGP97] on remplace certaines variables inconnues par leur projection),
- voire éventuellement dans le choix de la méthode de résolution pour le système algébrique d'équations obtenu après discrétisation (utilisation de différents algorithmes de point fixe, par exemple combinés avec la méthode GMRES et divers préconditionneurs pour les systèmes linéarisés creux de grand dimension, et une méthode de continuation en le paramètre Wi [How09]),

ont ainsi abouti à l'élévation du nombre de Weissenberg critique dans de nombreux cas tests. Mais pas à la disparition du HWNP [OP02, FK05] !

De nouveaux efforts se sont concentrés sur des techniques de discrétisation qui maintiennent positives certaines quantités [LO03, LX06], en vertu d'une interprétation physique naturelle de ces quantités. D'ailleurs, une discrétisation récente [FK04], initialement proposée comme paliatif à des effets possibles de couche limite et basée sur une reformulation logarithmique des équations (1-IV.1) (voir le chapitre 2), préserve non-seulement les propriétés (physiques) de positivité mentionnées ci-dessus, mais semble aussi limiter le développement d'instabilités et réduire le nombre de modes instables. Elle a donné des résultats numériques satisfaisant – convergeant pour de grandes valeurs de Wi – dans certains cas tests [Kwo04, HFK05, HP07a]. Ce succès récent n'est toutefois pas encore généralisé à de nombreux cas tests, et il n'est pas encore bien compris non plus.

Enfin, quelques résultats mathématiques sur des discrétisations proposées ci-dessus pour (1-IV.1) existent, en général sous réserve d'hypothèses de régularité pour la solution du problème continu à approcher numériquement [BS92a, BM97]. En particulier, quand la solution du problème continu à approcher numériquement est régulière, certains problèmes de type HWNP semblent aujourd'hui bien élucidés. Par exemple, celui lié aux instabilités dans la limite $\varepsilon \to 1$. En effet, le problème elliptique-hyperbolique (1-IV.1) change alors de nature

[16] C'est ainsi qu'est interprété dans [MC87] le résultat numérique décevant produit alors par la première utilisation de SUPG pour la simulation d'équations constitutives viscoélastiques, en l'occurences le modèle UCM.

(de nouvelles caractéristiques sont introduites dans le problème), et il est alors nécessaire de respecter une nouvelle condition de compatibilité de type inf−sup entre $\boldsymbol{u}$ et $\boldsymbol{\tau}$, similaire à la célèbre condition inf−sup de Ladyshenskaya-Babuška-Brezzi (LBB) pour le couple $(\boldsymbol{u},p)$. On comprend alors assez bien l'effet des méthodes de stabilisation [BFT93, BPS01]. Néanmoins, des problèmes subsistent aux grandes valeurs du nombre de Weissenberg, qui probablement rassemblent encore des instabilités d'origines diverses [Ren00]. Ces problèmes pourraient être liés à la régularité des solutions (apparition de couches limites), comme à des problèmes de discrétisation non encore considérés, même quand la solution est régulière. Il n'y a pas encore de réponse rigoureuse et définitive au HWNP !

Dans la suite de cette thèse (en fait, la première partie), on s'intéresse spécifiquement à la discrétisation du problème de Cauchy-Dirichlet simple pour (1-IV.1). On sait alors qu'une solution régulière pour le problème continu existe, et on cherche à bien l'approcher numériquement par des techniques de discrétisation stables. Nous nous intéressons en particulier à un critère de stabilité numérique encore jamais étudié, pour diverses discrétisations des équations constitutives : peut-on reproduire discrètement les lois de dissipation d'énergie vérifiées par le modèle continu ? (Dans les travaux antérieurs [LO03, LX06], la loi d'énergie considérée n'était pas une loi de dissipation.) Cette question a guidé nos recherches dans les chapitres 2 et 3, avec pour unique support le modèle d'Oldroyd-B (1-IV.1) (ainsi que sa reformulation logarithmique dans le chapitre 2) ; notre approche n'est cependant pas limitée à ce seul modèle et nous travaillons actuellement à son extension au modèle FENE-P (qui contient plus de termes non-linéaires et produit des comportements plus physiques, voir les Remarques 1 et 2).

Les conclusions de notre étude montrent que le critère de stabilité considéré n'est pas aisément satisfait par une discrétisation des équations constitutives d'Oldroyd-B. (Des discrétisations utilisées en pratique semblent *proches* de celles pour lesquelles on montre que notre critère est satisfait, mais on n'a pas pu montrer que ce critère est bien satisfait pour la plupart des discrétisations utilisées en pratique.) Elles donnent aussi des pistes pour comprendre le succès récent de la formulation logarithmique [Kwo04, HFK05, HP07a], et suggèrent de nouvelles études numériques (approfondissant les quelques résultats de la Section 2-VII).

1-IV-B Les modèles d'haltères

1-IV-B-a Théorie mathématique du modèle

La même adimensionalisation que dans (1-IV.1) conduit à :

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)=-\boldsymbol{\nabla}p+(1-\varepsilon)\Delta\boldsymbol{u}+\operatorname{div}\boldsymbol{\tau}\,, \\ \operatorname{div}\boldsymbol{u}=0\,, \\ \boldsymbol{\tau}=\dfrac{\varepsilon}{\mathrm{Wi}}\mathbb{E}(\boldsymbol{X}_t\otimes\mathbf{F}(\boldsymbol{X}_t)-2\boldsymbol{I})\,, \\ d\boldsymbol{X}_t+\boldsymbol{u}\cdot\nabla_{\boldsymbol{x}}\boldsymbol{X}_t dt=\left((\nabla_{\boldsymbol{x}}\boldsymbol{u})\boldsymbol{X}_t-\dfrac{1}{2\mathrm{Wi}}\boldsymbol{F}(\boldsymbol{X}_t)\right)dt+\dfrac{1}{\sqrt{\mathrm{Wi}}}d\boldsymbol{B}_t\,, \end{cases} \tag{1-IV.2}$$

avec pour équation de Fokker-Planck correspondante :

$$\partial_t\psi+(\boldsymbol{u}\cdot\boldsymbol{\nabla}_{\boldsymbol{x}})\psi=-\operatorname{div}_{\boldsymbol{X}}\left([(\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{u})\boldsymbol{X}-\frac{1}{2\mathrm{Wi}}\mathbf{F}(\boldsymbol{X})]\psi\right)+\frac{1}{2\mathrm{Wi}}\Delta_{\boldsymbol{X}}\psi\,. \tag{1-IV.3}$$

La théorie mathématique sur les notions d'existence et de régularité des solutions au problème de Cauchy-Dirichlet pour (1-IV.2) avec conditions de bord homogènes sur $\boldsymbol{u}$ est essentiellement similaire à celle pour (1-IV.1), limitée par les mêmes difficultés [JLL04, ELZ04, LZZ04, LZ07, LM07, Mas08, BSS05, BS07, BS08, BS09, BCP06]. On notera toutefois les approches [JLL04, ELZ04, BCP06] qui se distinguent des autres en étudiant (1-IV.2) par des méthodes probabilistes. (La plupart des preuves sont en effet pour la loi de $(\boldsymbol{X}_t)$, solution de (1-IV.3), et non pour le processus $(\boldsymbol{X}_t)$ – après suppression du terme d'advection [BCP06], ou annulation dans le cas de flots de Couette [JLL04], le champ aléatoire se réduit à un processus en temps –.) Noter aussi [BL04], qui est une piste pour l'étude du champ aléatoire $(\boldsymbol{X}_t)$ (prise en compte du terme d'advection).

1-IV-B-b Discrétisations du modèle

Les modèles micro-macro sont généralement simulés avec des méthodes probabilistes de type Monte-Carlo [Ö96]. Ces méthodes sont en général robustes et faciles à programmer, mais gourmandes en ressources

informatiques (la convergence de la loi des grands nombres peut être très lente). Un sujet de développement important pour ces méthodes (qui ne sont pas encore opérationnelles pour des simulations complexes à l'échelle industrielle) est donc la réduction des ressources informatiques nécessaires (on notera qu'une stratégie évidente si on a suffisamment de moyens matériels est la parallélisation des calculs, qui est très simple à mettre en œuvre si on distribue les réalisations sur différentes machines). En particulier, pour approcher des estimations Monte-Carlo qui satisfont un théorème centrale limite, les méthodes de réduction de variance [HDH64, New94, MT06] suscitent beaucoup d'intérêt chez les utilisateurs de méthodes Monte-Carlo (les rhéologues en particulier [MO95, OvdBH97, BP99], mais aussi les financiers [Aro04, Jou09]...).

Nous avons proposé deux méthodes de réduction de variance *accélérées* dans le chapitre 7, basées sur une utilisation intensive optimisée de variables de contrôle. Ces méthodes s'appliquent au cas du calcul de (1-III.2). Nous présentons des résultats numériques préliminaires, pour une EDS du type (1-III.1) sans le terme d'advection (plus de dérivée partielle) et pour des variations du champ de vitesse $\boldsymbol{u}$ données *a priori*.

La résolution numérique de l'équation de Fokker-Planck (1-IV.3) a aussi été tentée par plusieurs auteurs [Loz03, CL04, LC03, KS09b, KS09a]. Elle semble toutefois limitée à des cas où l'espace des configurations polymériques est décrit par un vecteur $\boldsymbol{X}$ de faible dimension, tels les dumbbells (d coordonnées suffisent). Dans le cas des dumbbells, des techniques de bases réduites, similaires à celles que nous employons pour accélérer le calcul des variables de contrôles en Monte-Carlo, peuvent aussi accélérer la résolution numérique (voir [KP09] et la section 7-VII du chapitre 7).

Noter aussi qu'en pratique, il semble que les modèles micro-macro soient moins sujets aux instabilités du type HWNP que les modèles macroscopiques. Une analyse est proposée dans [MHK09]. Ce point reste toutefois à approfondir.

Chapitre 2

Schémas dissipant l'énergie libre pour le modèle d'Oldroyd-B

Ce chapitre est la reproduction exacte des résultats publiés dans [BLM09], obtenus en collaboration avec T. Lelièvre et C. Mangoubi.

On y analyse la stabilité de divers schémas numériques pour les équations d'Oldroyd-B, prototype des modèles différentiels de fluides viscoélastiques, pour lequel on sait montrer qu'une solution régulière du problème de Cauchy-Dirichlet avec conditions de collement au bord d'un domaine borné $\mathcal{D} \subset \mathbb{R}^d$ $(d=2,3)$ dissipe une *énergie libre*. Plus précisément, on vérifie sous quelles hypothèses les schémas numériques satisfont une propriété de dissipation d'énergie libre semblable au cas continu. Parmi les schémas numériques analysés figurent des discrétisations fondées sur la reformulation logarithmique du système Oldroyd-B telle que proposée par Fattal et Kupferman dans [FK04], reformulation qui a permis d'obtenir des résultats numériques apparemment plus stables que les formulations usuelles dans certains cas tests [FK05, HFK05]. Notre analyse tente ainsi de donner des pistes pour la compréhension de ces observations numériques.

Plusieurs discrétisations usuelles de l'équation de Navier-Stokes couplée à celle d'Oldroyd-B sont passées en revue. On choisit ainsi pour le couple $(\boldsymbol{u}_h, p_h)$ divers espaces d'éléments finis mixtes :

- l'élément de Scott-Vogelius $(\mathbb{P}_d)^d \times \mathbb{P}_{d-1,\mathrm{disc}}$, que nous étudions quand $d=2$ (soit $(\mathbb{P}_2)^d \times \mathbb{P}_{1,disc}$),
- l'élément conforme de Taylor-Hood $(\mathbb{P}_2)^d \times \mathbb{P}_1$,
- l'élément non-conforme de Crouzeix Raviart au plus bas ordre $(\mathbb{P}_1^{CR})^d \times \mathbb{P}_0$,

qui satisfont tous une condition de compatibilité inf$-$sup discrète sur des partitions de $\mathcal{D}$ en simplexes (quoique avec des restrictions sur le maillage de $\mathcal{D}$ pour Scott-Vogelius). On étudie aussi le cas d'éléments finis non compatibles, stabilisés. Ces différents choix ne sont pas exhaustifs, mais ils montrent comment adapter l'essentiel de nos idées à de nombreuses discrétisations utilisées en pratique. (Toutefois, le cas d'éléments quadrilatères n'est peut-être pas une extension directe.)

Pour obtenir une dissipation d'énergie libre discrète, nous employons des éléments discontinus pour approcher le tenseur de conformation $\boldsymbol{\sigma}$, inconnue principale de notre formulation de l'équation d'Oldroyd-B, et imposons une restriction sur le pas de la discrétisation en temps (de type Euler rétrograde), en fonction de la condition initiale. Ces limitations seront relaxées dans le chapitre 3, mais on n'obtiendra plus alors l'unicité de la solution discrète. Dans le chapitre présent, nous montrons au contraire (avec des restrictions) l'existence et *l'unicité* d'une solution au système des équations discrétisées, telle qu'une dissipation d'énergie libre discrète est satisfaite : on obtient ainsi un résultat de convergence en temps long de la discrétisation numérique vers l'(unique) état stationnaire quelles que soient les conditions initiales (stabilité asymptotique). On observe que ces restrictions sont moins fortes pour les discrétisations de la formulation logarithmique de l'équation d'Oldroyd-B proposée en [FK04] (pas de restriction sur le pas de temps en fonction de la condition initiale). Cela pourrait être une explication des bons résultats numériques observés dans [FK05, HFK05] pour cette formulation (quoique avec des discrétisations quelque peu différentes des nôtres).

Des difficultés subsistent néanmoins avec nos discrétisations. Elles sont non-linéaires, donc la solution (implicite) peut être difficile à calculer (même en complétant des méthodes de point fixe classiques, telles que celles de Picard ou Newton, avec des techniques de continuation comme dans [How09]). Elles sont clairement de bas ordre de précision en le tenseur de conformation $\boldsymbol{\sigma}$ (*idem* avec les éléments continus du chapitre 3), même si on n'a pas étudié explicitement la consistance des schémas. Et la discrétisation du terme d'advection $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}$ requiert, soit (i) une méthode de caractéristiques qui est potentiellement destructrice de la stabilité si son inté-

gration n'est pas exacte, soit (ii) une méthode de type Galerkin discontinue, qui est gourmande en ressources informatiques (plus de degrés de liberté) et parfois plus difficile à résoudre algébriquement (convergence des méthodes itératives). Les difficultés quant à la discrétisation du terme d'advection $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}$ seront en partie levées dans le chapitre 3 grâce à l'emploi d'éléments continus avec une approche *ad hoc*, c'est-à-dire adaptée à la procédure de test permettant d'obtenir une inégalité d'énergie libre.

On notera une difficulté permanente dans l'obtention de nos discrétisations, qui est le souci de maintenir la positivité du tenseur symétrique $\boldsymbol{\sigma}$ discrétisé (on montre en effet facilement que si la solution continue au problème de Cauchy est régulière, – le champ de vitesse et ses trajectoires caractéristiques en particulier –, alors le modèle d'Oldroyd-B préserve au cours du temps la positivité – physique ! – du tenseur $\boldsymbol{\sigma}$, quand la condition initiale est elle-même positive). Cette positivité est nécessaire pour établir une inégalité d'énergie libre. Et elle est "gratuite" avec la reformulation logarithmique, d'où notre résultat (théorique) meilleur[1]. D'autres reformulations, plus favorables au maintien de la positivité, pourraient donc aussi être étudiées, qui bénéficieraient vraisemblablement de qualités de stabilité importantes (voir par exemple [LX06]). Néanmoins, de telles reformulations emploieraient vraisemblablement des transformations non-linéaires (comme [LX06]), et leurs discrétisations pourraient être tout aussi difficiles à résoudre numériquement si l'on veut conserver leurs propriétés de stabilité éventuelles (voir notre remarque en Section 2-VI-E).

Dans le chapitre 3, on s'efforce de relaxer la contrainte de positivité par une méthode qui mériterait des simulations numériques. Des tests sont en cours. On présente aussi en fin de ce chapitre quelques résultats numériques fragmentaires qui n'ont pas encore été publiés et qui ont trait à l'utilisation de la reformulation logarithmique.

[1] Ce n'est toutefois pas l'explication avancée dans [FK04, FK05, HFK05] pour l'emploi avantageux d'une reformulation logarithmique ; on y fait plutôt référence à des qualités de résolution pour des couches limites en $\boldsymbol{\sigma}$.

Free-energy-dissipative schemes for the Oldroyd-B model

Sébastien Boyaval[a,b], Tony Lelièvre[a,b], Claude Mangoubi[a,b,c]

[a]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).

[b]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

[c]The Hebrew University of Jerusalem, Institute of Mathematics (Jerusalem 91904, Israel).

In this chapter, we analyze the stability of various numerical schemes for differential models of viscoelastic fluids. More precisely, we consider the prototypical Oldroyd-B model, for which a *free energy* dissipation holds, and we show under which assumptions such a dissipation is also satisfied for the numerical scheme. Among the numerical schemes we analyze, we consider some discretizations based on the *log-formulation* of the Oldroyd-B system proposed by Fattal and Kupferman in [FK04], for which solutions in some benchmark problems have been obtained beyond the limiting Weissenberg numbers for the standard scheme (see [FK05, HFK05]). Our analysis gives some tracks to understand these numerical observations.

Keywords : Viscoelastic fluids, Weissenberg number, stability, entropy, finite elements methods, discontinuous Galerkin method, characteristic method.

2-I Introduction

2-I-A The stability issue in numerical schemes for viscoelastic fluids

An abundant literature has been discussing for over twenty years the stability of numerical schemes for discretizing equations modelling *viscoelastic fluids* (see [Keu90, Keu00, FHKK07, LX06] for a small sample). Indeed, most numerical schemes for macroscopic constitutive equations are known to suffer from instabilities in some benchmark problems, especially when a parameter, the *Weissenberg number*, increases.

Many possible reasons of that so-called *high Weissenberg number problem* (HWNP) have been identified [Kei92, KL95, San99, FK05]. However, these results have not led yet to a complete understanding of the numerical instabilities [Keu00], despite some progress [FK05, HFK05]. Roughly speaking, we can distinguish between three possible causes of the HWNP :

1. *Absence of stationary state* : In many situations (flow past a cylinder, 4 :1 contraction), the existence of a stationary state for viscoelastic models is still under investigation. It may happen that the non-convergence of the numerical scheme is simply due to the fact that, for the model under consideration, there exists no stationary state while the numerical scheme implicitly assumes such a stationary state.
2. *Instabilities for the exact solution* : More generally, the instabilities observed for the numerical scheme may originate at the continuous level, for the model under consideration, if the solution to the problem indeed blows up in finite time, or if it is not sufficiently regular to be well approximated in the discretization spaces. Such situations are known to occur for the Oldroyd-B model in extensional flows, for example (see [RH88, TS07a, BPP08]).
3. *Bad numerical scheme* : It may also happen that the problem at the continuous level indeed admits a regular solution, and the instabilities are only due to the discretization method.

In this paper, we focus on the third origin of instabilities, and we propose a criterion to test the stability of numerical schemes. More precisely, we look *under which conditions a numerical scheme does not bring spurious free energy* in the system. We concentrate on the Oldroyd-B model, for which a *free energy dissipation* is known to hold at the continuous level (see Thm. 1 below and [HL07]) and we try to obtain a similar dissipation at the discrete level. It is indeed particularly important that no spurious free energy is brought to the system in long-time computations, since they are often used as a way to obtain the stationary state.

The Oldroyd-B system of equations is definitely not a good physical model for dilute polymer fluids. In particular, it can be derived from a kinetic theory, with dumbbells modeling polymer molecules that are unphysically assumed to be infinitely extensible (and this indeed seems to be the cause of some instabilities for the flow of an Oldroyd-B fluid past a cylinder, see [RH88, TS07a, BPP08]). But from the mathematical viewpoint, it is nevertheless a good first step into the study of *macroscopic constitutive equations* for viscoelastic fluids. Indeed,

it already contains mathematical difficulties common to most of the viscoelastic models, while its strict equivalence with a kinetic model allows for a deep understanding of this set of equations. Let us also emphasize that the free energy dissipation we use and the numerical schemes we consider are not restricted to the Oldroyd-B model : they can be generalized to many other models (like FENE-P for instance, see [HL07]), so that we believe that our analysis can be used as a guideline to derive "good" numerical schemes for many macroscopic models for viscoelastic fluids. In summary, our aim here is *not* to discuss the HWNP but to propose a new criterion to assess the stability of numerical schemes for viscoelastic flows.

2-I-B Mathematical setting of the problem

We consider the Oldroyd-B model for dilute polymeric fluids in d-dimensional flows ($d=2,3$). Confined to an open bounded domain $\mathcal{D}\subset\mathbb{R}^d$, the fluid is governed by the following nondimensionalized system of equations :

$$\left\{\begin{array}{c} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)=-\boldsymbol{\nabla}p+(1-\varepsilon)\Delta\boldsymbol{u}+\mathrm{div}\,\boldsymbol{\tau}, \\ \mathrm{div}\,\boldsymbol{u}=0, \\ \dfrac{\partial \boldsymbol{\tau}}{\partial t}+(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\tau}=(\boldsymbol{\nabla}\boldsymbol{u})\boldsymbol{\tau}+\boldsymbol{\tau}(\boldsymbol{\nabla}\boldsymbol{u})^T-\dfrac{1}{\mathrm{Wi}}\boldsymbol{\tau}+\dfrac{\varepsilon}{\mathrm{Wi}}\left[\boldsymbol{\nabla}\boldsymbol{u}+\boldsymbol{\nabla}\boldsymbol{u}^T\right], \end{array}\right. \tag{2-I.1}$$

where $\boldsymbol{u}:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\to\boldsymbol{u}(t,\boldsymbol{x})\in\mathbb{R}^d$ is the velocity of the fluid, $p:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\to p(t,\boldsymbol{x})\in\mathbb{R}$ is the pressure and $\boldsymbol{\tau}:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\to\boldsymbol{\tau}(t,\boldsymbol{x})\in\mathbb{R}^{d\times d}$ is the extra-stress tensor. The matrix $\boldsymbol{\nabla}\boldsymbol{u}$ is the $d\times d$ matrix with components $\left(\frac{\partial \boldsymbol{u}_i}{\partial \boldsymbol{x}_j}\right)_{i,j}$. The following parameters are dimensionless : the Reynolds number $\mathrm{Re}\in\mathbb{R}_+$ (where $\mathbb{R}_+=[0,+\infty)$), the Weissenberg number $\mathrm{Wi}\in\mathbb{R}_+^*$ (where $\mathbb{R}_+^*=(0,+\infty)$) and the elastic viscosity to total viscosity fraction $\varepsilon\in(0,1)$.

In what follows, we assume for the sake of simplicity that the system (2-I.1) is supplied with homogeneous Dirichlet boundary conditions for the velocity $\boldsymbol{u}$:

$$\boldsymbol{u}=0 \text{ on } \partial\mathcal{D}. \tag{2-I.2}$$

Therefore, we study the energy dissipation of the equations (2-I.1) as time goes, that is, the way $(\boldsymbol{u},\boldsymbol{\tau})$ converges to the stationary state $(0,0)$ (equilibrium) in the long-time limit $t\to\infty$. Let us mention that it is possible to extend the analysis to non-zero boundary conditions (or more generally non-zero forcing) in the following way : it can be shown (see [JLBLO06]) that if the stationary velocity is not too large, then exponential convergence to the stationary state is achieved, at the continuous level. The schemes we propose are likely to exhibit similar behaviour, but we have not checked all the details for such a situation.

Local-in-time existence results for the above problem have been proved in the bounded domain $[0,T)\times\mathcal{D}$ when the system is supplied with sufficiently smooth initial conditions $\boldsymbol{u}(t=0)$ and $\boldsymbol{\tau}(t=0)$ (see [GS90, FCGO02] for instance). Moreover, global-in-time smooth solutions of the system (2-I.1) are known to converge exponentially fast to equilibrium in the sense defined in [JLBLO06]. Let us also mention the work of Lin *et al.* [LLZ05] where, for Oldroyd-like models, local-in-time existence and uniqueness results are proven, but also global-in-time existence and uniqueness results for small data. Notice that more general global-in-time results have been collected only for a mollified version of the Oldroyd-B system (2-I.1) (see [BSS05]), for another system close to (2-I.1) namely the co-rotational Oldroyd-B system (see [LM00]), or in the form of a Beale-Kato-Majda criterion when $\mathcal{D}=\mathbb{R}^3$ (see [KMT08]). Even though the question of the global-in-time existence for some solutions of the Oldroyd-B system (2-I.1) is still out-of-reach, it is possible to analyze global-in-time existence for solutions to *discretizations* of that system. This will be one of the output of this article.

2-I-C Outline of the paper and results

We will show that it is possible to build numerical (time and space discretizing) schemes for the Oldroyd-B system (2-I.1)–(2-I.2) such that solutions to those discretizations satisfy a free energy estimate similar to that established in [JLBLO06, HL07] for smooth solutions to the continuous equations. Our approach bears similarity with [LO03], where the authors also derive a discretization that preserves an energy estimate satisfied at the continuous level, and with [LX06], where another discretization is proposed for the same energy estimate as in [LO03]. Yet, unlike the estimates in [LO03, LX06], our estimate, the so-called *free energy* estimate derived in [JLBLO06, HL07], ensures (free) energy dissipation and exponential convergence of the solution to equilibrium. In particular, the long-time stability of solutions is ensured. As mentioned above, long-time computations are

indeed often used to obtain a stationary state, so that such a property may be seen as an interesting feature of a numerical scheme.

We also analyze discretizations of the log-formulation presented in [FK04, FK05], where the authors suggest to rewrite the set of equations (2-I.1) after mapping the (symmetric positive definite) *conformation tensor* :

$$\boldsymbol{\sigma} = \boldsymbol{I} + \frac{\mathrm{Wi}}{\varepsilon}\boldsymbol{\tau} \tag{2-I.3}$$

to its matrix logarithm :

$$\boldsymbol{\psi} = \ln \boldsymbol{\sigma}.$$

In the following, we assume that :

$$\boldsymbol{\sigma}(t=0) \text{ is } \textit{symmetric positive definite}, \tag{2-I.4}$$

and it can be shown that this property is propagated in time (see Lem. 3 below), so that $\boldsymbol{\psi}$ is indeed well defined. The log-formulation ensures, by construction, that the conformation tensor always remains symmetric positive definite, even after discretization. This is not only an important physical characteristic of the Oldroyd-B model but also an essential feature in the free energy estimates derived beneath. Besides, in some benchmark problems [Kwo04, FK05, HFK05], discretizations of the log-formulation have indeed been reported to yield solutions beyond the limiting Weissenberg number for standard discretizations of the usual formulation (for the Oldroyd-B and the Giesekus models). It is thus interesting to investigate whether the numerical success of this log-formulation may be related to a free energy dissipation property.

The main outputs of this work are :

(i) One crucial feature of the numerical scheme to obtain free energy estimates is the appropriate discretization of the advection term $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\tau}$ (or $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\psi}$ in the log-formulation) in the equation on the extra-stress tensor. We will analyze below two types of discretization : the characteristic method, and the discontinuous Galerkin method (see Sect. 2-IV, and Appendix D for higher-order schemes).

(ii) To obtain free energy estimates, we will need the extra-stress tensor to be discretized in a (elementwise) discontinuous finite element space (Sect. 2-IV and Appendix D).

(iii) The existence of a solution to the numerical schemes that satisfies a free energy estimate will be proved whatever the time step for the log-formulation in terms of $\boldsymbol{\psi}$, while it will be shown under a CFL-like condition for the usual formulation in terms of $\boldsymbol{\tau}$ (see Sect. 2-V). Moreover, any solution to the log-formulation satisfies the free energy estimate (which is not the case for the usual formulation in terms of $\boldsymbol{\tau}$. This may be related to the fact that the log-formulation has been reported to be more stable than the formulation in terms of $\boldsymbol{\tau}$ (see [HFK05]).

We would like to mention the work in preparation [BB09b] where the existence of a solution to a numerical scheme which satisfies a free energy estimate is also obtained whatever the time step for the usual formulation of the Oldroyd-B model in terms of $\boldsymbol{\sigma}$, but only as the limit of a subsequence of regularized discretizations. This means that, in the case where the CFL condition is not fulfilled hence uniqueness not ensured, there may be many solutions to our numerical schemes for the usual formulation of the Oldroyd-B model in terms of $\boldsymbol{\sigma}$, one of which is actually shown to satisfy a free energy estimate ; on the contrary, every solution to our numerical schemes for the log-formulation necessarily satisfies a free energy estimate. Moreover, it is shown in this work [BB09b] that, using a particular discretization of the advection term, it is possible to use continuous finite element spaces to obtain a discrete analogue of the free energy bound for a regularized Oldroyd-B model. In addition, subsequence convergence, as the mesh parameters tend to zero, of such a scheme is proved, which yields existence of global-in-time solutions to this modified Oldroyd-B system.

Notice that we here concentrate on stability issues. All the schemes we analyze are of course consistent, but we do not study the order of consistency of these schemes, neither the convergence.

Let us now make precise how the paper is organized. In Section 2-II, we formally derive the free energy estimates for the Oldroyd-B set of equations and for its logarithm formulation, in the spirit of [HL07]. Then, Section 2-III is devoted to the presentation of a finite element scheme (using piecewise constant approximations of the conformation tensor and its log-formulation, and Scott-Vogelius finite elements for the velocity and pressure), that is shown to satisfy a discrete free energy estimate in Section 2-IV. Some variants of this discretization are also studied, still for piecewise constant stress tensor, and a summary of the requirements on the discretizations to satisfy a free energy estimate is provided in Tables 2.1 and 2.2 (we show in Appendix D how to use an interpolation operator so as to adapt the previous results to piecewise linear approximations of the stress tensor). Finally, in Section 2-V, we show how the previous stability results can be used to prove long-time existence results for the discrete solutions. Some numerical studies are needed to illustrate this numerical analysis, and this is a work in progress.

2-I-D Notation and auxiliary results

In the following, we will make use of the usual notation : $L^2(\mathcal{D})=\{f:\mathcal{D}\to\mathbb{R},\int_{\mathcal{D}}|f|^2<\infty\}$, $H^1(\mathcal{D})=\{f:\mathcal{D}\to\mathbb{R},\int_{\mathcal{D}}|f|^2+|\nabla f|^2<\infty\}$, $H^2(\mathcal{D})=\{f:\mathcal{D}\to\mathbb{R},\int_{\mathcal{D}}|f|^2+|\nabla f|^2+|\nabla^2 f|^2<\infty\}$, $C([0,T))$ for continuous functions on $[0,T)$ and $C^1([0,T))$ for continuously differentiable functions on $[0,T)$.

We will denote by $\boldsymbol{\tau}:\boldsymbol{\sigma}$ the double contraction between rank-two tensors (matrices) $\boldsymbol{\tau}$, $\boldsymbol{\sigma}\in\mathbb{R}^{d\times d}$:

$$\boldsymbol{\tau}:\boldsymbol{\sigma}=\mathrm{tr}(\boldsymbol{\tau}\boldsymbol{\sigma}^T)=\mathrm{tr}(\boldsymbol{\tau}^T\boldsymbol{\sigma})=\sum_{1\le i,j\le d}\boldsymbol{\tau}_{ij}\boldsymbol{\sigma}_{ij}.$$

Notice that if $\boldsymbol{\tau}$ is antisymmetric and $\boldsymbol{\sigma}$ symmetric, $\boldsymbol{\tau}:\boldsymbol{\sigma}=0$.

The logarithm of a positive definite diagonal matrix is a diagonal matrix with, on its diagonal, the logarithm of each entry. We define the logarithm of any symmetric positive definite matrix $\boldsymbol{\sigma}$ using a diagonal decomposition $\boldsymbol{\sigma}=R^T\Lambda R$ of $\boldsymbol{\sigma}$ with R an orthogonal matrix and Λ a positive definite diagonal matrix :

$$\ln\boldsymbol{\sigma}=R^T(\ln\Lambda)R. \tag{2-I.5}$$

Although the diagonal decomposition of $\boldsymbol{\sigma}$ is not unique, (2-I.5) uniquely defines $\ln\boldsymbol{\sigma}$. The matrix logarithm bijectively maps the set of symmetric positive definite matrices with real entries $\mathcal{S}_+^*(\mathbb{R}^{d\times d})$ to the vector subspace $\mathcal{S}(\mathbb{R}^{d\times d})$ of symmetric real matrices, where it is exactly the inverse function of the matrix exponential.

We will make use of the following simple algebraic formulae, which are proved in Appendices 2-VI-A-a and 2-VI-A-b.

Lemma 1. *Let $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ be two symmetric positive definite matrices. We have :*

$$\mathrm{tr}\ln\boldsymbol{\sigma}=\ln\det\boldsymbol{\sigma}, \tag{2-I.6}$$

$$\boldsymbol{\sigma}-\ln\boldsymbol{\sigma}-\boldsymbol{I}\ \textit{is symmetric positive semidefinite and thus}\ \mathrm{tr}(\boldsymbol{\sigma}-\ln\boldsymbol{\sigma}-\boldsymbol{I})\ge 0, \tag{2-I.7}$$

$$\boldsymbol{\sigma}+\boldsymbol{\sigma}^{-1}-2\boldsymbol{I}\ \textit{is symmetric positive semidefinite and thus}\ \mathrm{tr}(\boldsymbol{\sigma}+\boldsymbol{\sigma}^{-1}-2\boldsymbol{I})\ge 0, \tag{2-I.8}$$

$$\mathrm{tr}(\boldsymbol{\sigma}\boldsymbol{\tau})=\mathrm{tr}(\boldsymbol{\tau}\boldsymbol{\sigma})\ge 0, \tag{2-I.9}$$

$$\mathrm{tr}\left((\boldsymbol{\sigma}-\boldsymbol{\tau})\boldsymbol{\tau}^{-1}\right)=\mathrm{tr}(\boldsymbol{\sigma}\boldsymbol{\tau}^{-1}-\boldsymbol{I})\ge\ln\det(\boldsymbol{\sigma}\boldsymbol{\tau}^{-1})=\mathrm{tr}\left(\ln\boldsymbol{\sigma}-\ln\boldsymbol{\tau}\right), \tag{2-I.10}$$

$$\mathrm{tr}\left((\ln\boldsymbol{\sigma}-\ln\boldsymbol{\tau})\boldsymbol{\sigma}\right)\ge\mathrm{tr}\left(\boldsymbol{\sigma}-\boldsymbol{\tau}\right). \tag{2-I.11}$$

We will also use the Jacobi's formulae :

Lemma 2. *For any symmetric positive definite matrix $\boldsymbol{\sigma}(t)\in\left(C^1\left([0,T)\right)\right)^{\frac{d(d+1)}{2}}$, we have $\forall t\in[0,T)$:*

$$\left(\frac{d}{dt}\boldsymbol{\sigma}\right):\boldsymbol{\sigma}^{-1}=\mathrm{tr}\left(\boldsymbol{\sigma}^{-1}\frac{d}{dt}\boldsymbol{\sigma}\right)=\frac{d}{dt}\mathrm{tr}(\ln\boldsymbol{\sigma}), \tag{2-I.12}$$

$$\left(\frac{d}{dt}\ln\boldsymbol{\sigma}\right):\boldsymbol{\sigma}=\mathrm{tr}\left(\boldsymbol{\sigma}\frac{d}{dt}\ln\boldsymbol{\sigma}\right)=\frac{d}{dt}\mathrm{tr}\boldsymbol{\sigma}. \tag{2-I.13}$$

2-II Formal free energy estimates at the continuous level

We are going to derive free energy estimates for two formulations of the Oldroyd-B system in Theorems 1 and 2. An important corollary to these theorems is the exponential convergence of the solutions to equilibrium in the long-time limit. Throughout this section, we assume that $(\boldsymbol{u},p,\boldsymbol{\tau})$ is a sufficiently *smooth* solution of problem (2-I.1) so that all the subsequent computations are valid. For example, the following regularity is sufficient :

$$(\boldsymbol{u},p,\boldsymbol{\tau})\in\left(C^1\left([0,T),H^2(\mathcal{D})\cap C^{0,1}(\mathcal{D})\right)\right)^d\times\left(C^0\left([0,T),H^1(\mathcal{D})\right)\right)\times\left(C^1\left([0,T),C^1(\mathcal{D})\right)\right)^{d\times d}, \tag{2-II.1}$$

where we denote, for instance by $\left(C^1(\mathcal{D})\right)^d$ a vector field of dimension d with $C^1(\mathcal{D})$ components.

2-II-A Free energy estimate for the Oldroyd-B system

2-II-A-a Conformation-tensor formulation of the Oldroyd-B system

Recall that the *conformation* tensor $\boldsymbol{\sigma}$ is defined from the *extra-stress* tensor $\boldsymbol{\tau}$ through the following bijective mapping :

$$\boldsymbol{\tau} = \frac{\varepsilon}{\mathrm{Wi}}(\boldsymbol{\sigma} - \boldsymbol{I}).$$

With this mapping, it is straightforward to bijectively map the solutions of system (2-I.1) with those of the following system :

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right) = -\boldsymbol{\nabla} p + (1-\varepsilon)\Delta\boldsymbol{u} + \dfrac{\varepsilon}{\mathrm{Wi}}\operatorname{div}\boldsymbol{\sigma}, \\ \operatorname{div}\boldsymbol{u} = 0, \\ \dfrac{\partial \boldsymbol{\sigma}}{\partial t} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma} = (\boldsymbol{\nabla}\boldsymbol{u})\boldsymbol{\sigma} + \boldsymbol{\sigma}(\boldsymbol{\nabla}\boldsymbol{u})^T - \dfrac{1}{\mathrm{Wi}}(\boldsymbol{\sigma} - \boldsymbol{I}). \end{cases} \tag{2-II.2}$$

Notice that with such an affine mapping, the solution $\boldsymbol{\sigma}$ to system (2-II.2) has the same regularity as $\boldsymbol{\tau}$ solution to system (2-I.1), which is that assumed in (2-II.1) for the following manipulations.

2-II-A-b A free energy estimate

Let us first recall a free energy estimate derived in [JLBLO06, HL07]. The free energy of the fluid is defined as the sum of two terms as follows :

$$F(\boldsymbol{u},\boldsymbol{\sigma}) = \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} |\boldsymbol{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \operatorname{tr}(\boldsymbol{\sigma} - \ln\boldsymbol{\sigma} - \boldsymbol{I}). \tag{2-II.3}$$

The *kinetic* term $\displaystyle\int_{\mathcal{D}} |\boldsymbol{u}|^2$ is always non negative. Besides, we have the following lemma (see Appendix B or [Hul90] for a proof) :

Lemma 3. *Let $\boldsymbol{\sigma} \in \left(C^1\left([0,T), C^1(\mathcal{D})\right)\right)^{d\times d}$ be a smooth solution to the system* (2-II.2)*. Then, if the initial condition $\boldsymbol{\sigma}(t=0)$ is symmetric positive definite (everywhere in $\mathcal{D}$), the solution $\boldsymbol{\sigma}(t)$ remains so at all times $t \in [0,T)$ and for all $\boldsymbol{x} \in \mathcal{D}$. In particular, the matrix $\boldsymbol{\sigma}(t)$ is invertible.*

From Lemma 3 and the equation (2-I.7), the *entropic* term $\displaystyle\int_{\mathcal{D}} \operatorname{tr}(\boldsymbol{\sigma} - \ln\boldsymbol{\sigma} - \boldsymbol{I})$ is thus well defined and non negative, provided $\boldsymbol{\sigma}(t=0)$ is symmetric positive definite.

The free energy is an interesting quantity to characterize the long-time asymptotics of the solutions, and thus the stability of the system (2-II.2). *A priori* estimates using the free energy are presented in [JLBLO06] for micro-macro models (such as the Hookean or the FENE dumbbell models) and in [HL07] for macroscopic models (such as the Oldroyd-B or the FENE-P models). Similar considerations can be found in the physics literature about thermodynamic theory for viscoelastic models (see [Leo92, BE94, Ö05, WH98b]).

For the sake of consistency, we recall results from [HL07] :

Theorem 1. *Let $(\boldsymbol{u}, p, \boldsymbol{\sigma})$ be a smooth solution to system* (2-II.2) *supplied with homogeneous Dirichlet boundary conditions for $\boldsymbol{u}$, and with symmetric positive definite initial condition $\boldsymbol{\sigma}(t=0)$. The free energy satisfies :*

$$\frac{d}{dt}F(\boldsymbol{u},\boldsymbol{\sigma}) + (1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla}\boldsymbol{u}|^2 + \frac{\varepsilon}{2\,Wi^2}\int_{\mathcal{D}} \operatorname{tr}\left(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{-1} - 2\boldsymbol{I}\right) = 0. \tag{2-II.4}$$

From this estimate, we get that $F(\boldsymbol{u},\boldsymbol{\sigma})$ decreases exponentially fast in time to zero.

Proof of Theorem (1). Let $(\boldsymbol{u}, p, \boldsymbol{\sigma})$ be a smooth solution to system (2-II.2), with symmetric positive definite initial condition $\boldsymbol{\sigma}(t=0)$. We first compute the inner product of the Navier-Stokes equation with the velocity :

$$\frac{\mathrm{Re}}{2}\frac{d}{dt}\int_{\mathcal{D}} |\boldsymbol{u}|^2 = -(1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla}\boldsymbol{u}|^2 - \frac{\varepsilon}{\mathrm{Wi}}\int_{\mathcal{D}} \boldsymbol{\nabla}\boldsymbol{u} : \boldsymbol{\sigma}. \tag{2-II.5}$$

Then, taking the trace of the evolution equation for the conformation tensor, we obtain :

$$\frac{d}{dt}\int_{\mathcal{D}} \operatorname{tr}\boldsymbol{\sigma} = 2\int_{\mathcal{D}} \boldsymbol{\nabla}\boldsymbol{u} : \boldsymbol{\sigma} - \frac{1}{\mathrm{Wi}}\int_{\mathcal{D}} \operatorname{tr}(\boldsymbol{\sigma} - \boldsymbol{I}). \tag{2-II.6}$$

Last, remember that smooth solutions $\boldsymbol{\sigma}$ are invertible matrices (Lem. 3). Thus, contracting the evolution equation for $\boldsymbol{\sigma}$ with $\boldsymbol{\sigma}^{-1}$, we get :

$$\int_{\mathcal{D}} \left(\frac{\partial}{\partial t}\boldsymbol{\sigma} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma} \right) : \boldsymbol{\sigma}^{-1} = 2\int_{\mathcal{D}} \mathrm{tr}(\boldsymbol{\nabla}\boldsymbol{u}) - \frac{1}{\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}\left(\boldsymbol{I} - \boldsymbol{\sigma}^{-1}\right). \tag{2-II.7}$$

Using (2-I.12) with $\boldsymbol{\sigma} \in C^1\left(\mathcal{D}\times[0,T), \mathcal{S}_+^\star(\mathbb{R}^{d\times d})\right)$, we find :

$$\int_{\mathcal{D}} \left(\frac{\partial}{\partial t}\boldsymbol{\sigma} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma} \right) : \boldsymbol{\sigma}^{-1} = \int_{\mathcal{D}} \left(\frac{\partial}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla} \right) \mathrm{tr}(\ln\boldsymbol{\sigma}),$$

which can be combined with (2-II.7) to get, using $\mathrm{tr}(\boldsymbol{\nabla}\boldsymbol{u}) = \mathrm{div}\,\boldsymbol{u} = 0$ and $\boldsymbol{u} = 0$ on $\partial\mathcal{D}$:

$$\frac{d}{dt}\int_{\mathcal{D}} \mathrm{tr}\ln\boldsymbol{\sigma} = \frac{1}{\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}\left(\boldsymbol{\sigma}^{-1} - \boldsymbol{I}\right). \tag{2-II.8}$$

We now combine (2-II.5) $+\frac{\varepsilon}{2\mathrm{Wi}}\times$ (2-II.6) $-\frac{\varepsilon}{2\mathrm{Wi}}\times$ (2-II.8) to obtain (2-II.4) :

$$\frac{d}{dt}\left[\frac{\mathrm{Re}}{2}\int_{\mathcal{D}} |\boldsymbol{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(\boldsymbol{\sigma} - \ln\boldsymbol{\sigma} - \boldsymbol{I}) \right] + (1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla}\boldsymbol{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}} \mathrm{tr}\left(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{-1} - 2\boldsymbol{I}\right) = 0.$$

Since, by (2-I.8), we have $\mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{-1} - 2\boldsymbol{I}) \geq 0$, then $F(\boldsymbol{u},\boldsymbol{\sigma})$ decreases in time. Moreover, by (2-I.7) applied to $\boldsymbol{\sigma}^{-1}$, we have $\mathrm{tr}(\boldsymbol{\sigma} - \ln\boldsymbol{\sigma} - \boldsymbol{I}) \leq \mathrm{tr}(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{-1} - 2\boldsymbol{I})$. So, using the Poincaré inequality which states that there exists a constant C_P depending only on $\mathcal{D}$ such that, for all $\boldsymbol{u} \in H_0^1(\mathcal{D})$,

$$\int_{\mathcal{D}} |\boldsymbol{u}|^2 \leq C_P \int_{\mathcal{D}} |\boldsymbol{\nabla}\boldsymbol{u}|^2,$$

we finally obtain that $F(\boldsymbol{u},\boldsymbol{\sigma})$ goes exponentially fast to 0. Indeed, we have from (2-II.4) :

$$\frac{d}{dt}F(\boldsymbol{u},\boldsymbol{\sigma}) \leq -\frac{1-\varepsilon}{C_P}\int_{\mathcal{D}} |\boldsymbol{u}|^2 - \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}} \mathrm{tr}\left(\boldsymbol{\sigma} + \boldsymbol{\sigma}^{-1} - 2\boldsymbol{I}\right) \leq -\min\left(\frac{2(1-\varepsilon)}{\mathrm{Re}\ C_P}, \frac{1}{\mathrm{Wi}} \right) F(\boldsymbol{u},\boldsymbol{\sigma}),$$

so that, by a direct application of Gronwall's lemma, we get :

$$F(\boldsymbol{u},\boldsymbol{\sigma}) \leq F(\boldsymbol{u}(t=0),\boldsymbol{\sigma}(t=0)) \exp\left(-\min\left(\frac{2(1-\varepsilon)}{\mathrm{Re}\ C_P}, \frac{1}{\mathrm{Wi}} \right) t \right). \qquad \square$$

2-II-B Free energy estimate for the log-formulation of the Oldroyd-B system

2-II-B-a Log-formulation of the Oldroyd-B system

Let us now introduce the log-formulation proposed in [FK04]. We want to map solutions of the system (2-II.2) with solutions of another system of equations where a partial differential equation for the logarithm of the conformation tensor is substituted to the Oldroyd-B partial differential equation for the conformation tensor $\boldsymbol{\sigma}$.

In order to obtain a constitutive equation in terms of $\boldsymbol{\psi} = \ln\boldsymbol{\sigma}$, following [FK04], we make use of the following decomposition of the deformation tensor $\boldsymbol{\nabla}\boldsymbol{u} \in \mathbb{R}^{d\times d}$ (see Appendix C for a proof) :

Lemma 4. *For any matrix $\boldsymbol{\nabla}\boldsymbol{u}$ and any symmetric positive definite matrix $\boldsymbol{\sigma}$ in $\mathbb{R}^{d\times d}$, there exist in $\mathbb{R}^{d\times d}$ two antisymmetric matrices $\boldsymbol{\Omega}$, $\mathcal{N}$ and a symmetric matrix $\boldsymbol{B}$ that commutes with $\boldsymbol{\sigma}$, such that :*

$$\boldsymbol{\nabla}\boldsymbol{u} = \boldsymbol{\Omega} + \boldsymbol{B} + \mathcal{N}\boldsymbol{\sigma}^{-1}. \tag{2-II.9}$$

Moreover, we have $\mathrm{tr}\boldsymbol{\nabla}\boldsymbol{u} = \mathrm{tr}\boldsymbol{B}$.

We now proceed to the change of variable $\boldsymbol{\psi} = \ln\boldsymbol{\sigma}$. The system (2-II.2) then rewrites (see [FK04] for a proof) :

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial\boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u} \right) = -\boldsymbol{\nabla}p + (1-\varepsilon)\Delta\boldsymbol{u} + \dfrac{\varepsilon}{\mathrm{Wi}}\mathrm{div}\, e^{\boldsymbol{\psi}}, \\ \mathrm{div}\,\boldsymbol{u} = 0, \\ \dfrac{\partial\boldsymbol{\psi}}{\partial t} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\psi} = \boldsymbol{\Omega}\boldsymbol{\psi} - \boldsymbol{\psi}\boldsymbol{\Omega} + 2\boldsymbol{B} + \dfrac{1}{\mathrm{Wi}}\left(\mathrm{e}^{-\boldsymbol{\psi}} - \boldsymbol{I}\right). \end{cases} \tag{2-II.10}$$

It is supplied with unchanged initial and boundary conditions for $\boldsymbol{u}$, plus the initial condition $\boldsymbol{\psi}(t=0) = \ln\boldsymbol{\sigma}(t=0)$ for the log-conformation tensor.

2-II-B-b Reformulation of the free energy estimate

A result similar to Theorem 1 can be obtained for system (2-II.10), where the free energy is written in terms of $\boldsymbol{\psi}$ as :

$$F(\boldsymbol{u},\mathrm{e}^{\boldsymbol{\psi}}) = \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} |\boldsymbol{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(e^{\boldsymbol{\psi}} - \boldsymbol{\psi} - \boldsymbol{I}). \tag{2-II.11}$$

The following theorem then holds :

Theorem 2. *Let $(\boldsymbol{u},p,\boldsymbol{\psi})$ be a smooth solution to system* (2-II.10) *supplied with homogeneous Dirichlet boundary conditions for $\boldsymbol{u}$. The free energy satisfies :*

$$\frac{d}{dt}F(\boldsymbol{u},\mathrm{e}^{\boldsymbol{\psi}}) + (1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla u}|^2 + \frac{\varepsilon}{2\,Wi^2}\int_{\mathcal{D}} \mathrm{tr}(e^{\boldsymbol{\psi}} + e^{-\boldsymbol{\psi}} - 2\boldsymbol{I}) = 0. \tag{2-II.12}$$

From this estimate, we get that $F(\boldsymbol{u},\mathrm{e}^{\boldsymbol{\psi}})$ decreases exponentially fast in time to zero.

Proof of Theorem (2). The proof of this theorem mimics the proof of Theorem 1. We go over the steps of the proof, and point out the differences with the previous case. Let $(\boldsymbol{u},p,\boldsymbol{\psi})$ be a smooth solution to (2-II.10).

From the inner product of the momentum conservation equation in (2-II.10) with the velocity $\boldsymbol{u}$, we obtain :

$$\frac{\mathrm{Re}}{2}\frac{d}{dt}\int_{\mathcal{D}} |\boldsymbol{u}|^2 = -(1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla u}|^2 - \frac{\varepsilon}{\mathrm{Wi}}\int_{\mathcal{D}} \boldsymbol{\nabla u} : e^{\boldsymbol{\psi}}, \tag{2-II.13}$$

which is equivalent to (2-II.5). Taking the trace of the evolution equation for the conformation tensor, we get :

$$\frac{d}{dt}\int_{\mathcal{D}} \mathrm{tr}\boldsymbol{\psi} = \frac{1}{\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(e^{-\boldsymbol{\psi}} - \boldsymbol{I}), \tag{2-II.14}$$

which is equivalent to (2-II.8). Contracting the evolution equation for $\boldsymbol{\psi}$ with $e^{\boldsymbol{\psi}}$ and using (2-I.13) with $\boldsymbol{\psi} = \ln\boldsymbol{\sigma}$, we rewrite the first term of this inner product :

$$\left(\frac{\partial\boldsymbol{\psi}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla\psi}\right) : e^{\boldsymbol{\psi}} = \left(\frac{\partial}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\right)\mathrm{tr}e^{\boldsymbol{\psi}}.$$

Recall that the decomposition (2-II.9) of $\boldsymbol{\nabla u}$ allows to rewrite the second term :

$$\boldsymbol{\nabla u} : e^{\boldsymbol{\psi}} = \boldsymbol{\Omega} : e^{\boldsymbol{\psi}} + \boldsymbol{B} : e^{\boldsymbol{\psi}} + (\mathcal{N}e^{-\boldsymbol{\psi}}) : e^{\boldsymbol{\psi}} = \boldsymbol{B} : e^{\boldsymbol{\psi}}, \tag{2-II.15}$$

where we have used the symmetry of $e^{\boldsymbol{\psi}}$ and the antisymmetry of $\boldsymbol{\Omega}$ and $\mathcal{N}$. Then, notice that, since $\boldsymbol{\psi}$ and $e^{\boldsymbol{\psi}}$ commute, we have :

$$(\boldsymbol{\Omega\psi} - \boldsymbol{\psi\Omega}) : e^{\boldsymbol{\psi}} = \mathrm{tr}(\boldsymbol{\Omega\psi}e^{\boldsymbol{\psi}}) - \mathrm{tr}(\boldsymbol{\psi\Omega}e^{\boldsymbol{\psi}}) = \mathrm{tr}(\boldsymbol{\Omega\psi}e^{\boldsymbol{\psi}}) - \mathrm{tr}(\boldsymbol{\Omega\psi}e^{\boldsymbol{\psi}}) = 0, \tag{2-II.16}$$

we finally obtain an equation equivalent to (2-II.6) :

$$\frac{d}{dt}\int_{\mathcal{D}} \mathrm{tr}e^{\boldsymbol{\psi}} = 2\int_{\mathcal{D}} \boldsymbol{\nabla u} : e^{\boldsymbol{\psi}} - \frac{1}{\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(e^{\boldsymbol{\psi}} - \boldsymbol{I}). \tag{2-II.17}$$

It is noticeable that in this proof, we made no use of the positivity of $\boldsymbol{\sigma} = e^{\boldsymbol{\psi}}$, in contrast to the proof of Theorem 1.

The combination (2-II.13) $-\frac{\varepsilon}{2\mathrm{Wi}}\times$ (2-II.14) $+\frac{\varepsilon}{2\mathrm{Wi}}\times$ (2-II.17) gives (2-II.12) :

$$\frac{d}{dt}\left[\frac{\mathrm{Re}}{2}\int_{\mathcal{D}} |\boldsymbol{u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}\left(e^{\boldsymbol{\psi}} - \boldsymbol{\psi} - \boldsymbol{I}\right)\right] + (1-\varepsilon)\int_{\mathcal{D}} |\boldsymbol{\nabla u}|^2 + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}} \mathrm{tr}\left(e^{\boldsymbol{\psi}} + e^{-\boldsymbol{\psi}} - 2\boldsymbol{I}\right) = 0. \tag{2-II.18}$$

This is exactly equivalent to (2-II.4). As in the proof of Theorem 1, we then obtain that $F(\boldsymbol{u},\mathrm{e}^{\boldsymbol{\psi}})$ decreases exponentially fast in time to zero. □

2-III Construction of numerical schemes with Scott-Vogelius finite elements for the velocity-pressure field $(\boldsymbol{u}_h, p_h)$

We would now like to build numerical schemes for both systems of equations (2-II.2) and (2-II.10) that respectively preserve the dissipation properties of Theorems 1 and 2 for discrete free energies similar to (2-II.3) and (2-II.11). We first present discretizations that allow for a simple and complete exposition of our reasoning in order to derive discrete free energy estimates. Possible extensions will be discussed in Section 2-IV-C (other discretizations for the velocity-pressure field) and in Appendix D (higher-order discretizations for the stress field).

2-III-A Variational formulations of the problems

To discretize (2-II.2) and (2-II.10) in space using a finite element method, we first write variational formulations for (2-II.2) and (2-II.10) that are satisfied by smooth solutions of the previous systems. Smooth solutions $(\boldsymbol{u},p,\boldsymbol{\sigma})$ and $(\boldsymbol{u},p,\boldsymbol{\psi})$ to system (2-II.2) and (2-II.10) respectively satisfy the variational formulations :

$$\begin{aligned} 0 = \int_{\mathcal{D}} \Bigg(& \mathrm{Re}\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)\cdot\boldsymbol{v} + (1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u} : \boldsymbol{\nabla}\boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma} : \boldsymbol{\nabla}\boldsymbol{v} - p\,\mathrm{div}\,\boldsymbol{v} + q\,\mathrm{div}\,\boldsymbol{u} \\ & + \left(\frac{\partial \boldsymbol{\sigma}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}\right) : \boldsymbol{\phi} - \left((\boldsymbol{\nabla}\boldsymbol{u})\boldsymbol{\sigma} + \boldsymbol{\sigma}(\boldsymbol{\nabla}\boldsymbol{u})^T\right) : \boldsymbol{\phi} + \frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}-\boldsymbol{I}) : \boldsymbol{\phi}\Bigg), \end{aligned} \tag{2-III.1}$$

and

$$\begin{aligned} 0 = \int_{\mathcal{D}} \Bigg(& \mathrm{Re}\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)\cdot\boldsymbol{v} + (1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u} : \boldsymbol{\nabla}\boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}}e^{\boldsymbol{\psi}} : \boldsymbol{\nabla}\boldsymbol{v} - p\,\mathrm{div}\,\boldsymbol{v} + q\,\mathrm{div}\,\boldsymbol{u} \\ & + \left(\frac{\partial \boldsymbol{\psi}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\psi}\right) : \boldsymbol{\phi} - (\boldsymbol{\Omega}\boldsymbol{\psi} - \boldsymbol{\psi}\boldsymbol{\Omega}) : \boldsymbol{\phi} - 2\boldsymbol{B} : \boldsymbol{\phi} - \frac{1}{\mathrm{Wi}}(e^{-\boldsymbol{\psi}} - \boldsymbol{I}) : \boldsymbol{\phi}\Bigg), \end{aligned} \tag{2-III.2}$$

for all sufficiently regular test functions $(\boldsymbol{v},q,\boldsymbol{\phi})$.

In this variational framework, we recover the free energy estimates (2-II.4) (respectively (2-II.12)) using the test functions $\left(\boldsymbol{u},p,\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I}-\boldsymbol{\sigma}^{-1})\right)$ (respectively $\left(\boldsymbol{u},p,\frac{\varepsilon}{2\mathrm{Wi}}(e^{\boldsymbol{\psi}}-\boldsymbol{I})\right)$) in (2-III.1) (respectively (2-III.2)).

2-III-B Numerical schemes with Scott-Vogelius finite elements for $(\boldsymbol{u}_h, p_h)$

Using the Galerkin discretization method, we now want to build variational numerical integration schemes that are based on the variational formulations (2-III.1) and (2-III.2) using finite-dimensional approximations of the solution/test spaces. We will then show in the next Section 2-IV that solutions to these schemes satisfy discrete free energy estimates which are equivalent to those in Theorems 1 and 2.

First, the time interval $[0,T)$ is split into N_T intervals $[t^n,t^{n+1})$ of constant size $\Delta t = \frac{T}{N_T}$, with $t^n = n\Delta t$ for $n=0,\dots,N_T$. For all $n=0,\dots,N_T-1$, we denote by $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\sigma}_h^n)$ (resp. $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\psi}_h^n)$), the value at time t_n of the discrete solutions $(\boldsymbol{u}_h, p_h, \boldsymbol{\sigma}_h)$ (resp. $(\boldsymbol{u}_h, p_h, \boldsymbol{\psi}_h)$) in finite element spaces.

In all the following sections, we will assume that the domain $\mathcal{D}$ is polyhedral. We define a conformal mesh $\mathcal{T}_h$ built from a tessellation of the domain $\mathcal{D}$,

$$\mathcal{T}_h = \bigcup_{k=1}^{N_K} K_k,$$

made of N_K simplicial elements K_k and N_D nodes at the internal vertices. We denote by h_{K_k} the diameter of the element K_k and assume that the mesh is uniformly regular, with maximal diameter $h \geq \max_{1\leq k\leq N_K} h_{K_k}$. For each element K_k of the mesh $\mathcal{T}_h$, we denote by $\boldsymbol{n}_{K_k}$ the outward unitary normal vector to element K_k, defined on its boundary ∂K_k. We also denote by $\{E_j | j=1,\dots,N_E\}$ the internal edges of the mesh $\mathcal{T}_h$ when $d=2$, or the faces of volume elements when $d=3$ (also termed as "edges" for the sake of simplicity in the following).

For the velocity-pressure field $(\boldsymbol{u}_h, p_h)$, we choose the mixed finite element space $(\mathbb{P}_2)^d \times \mathbb{P}_{1,disc}$ of Scott-Vogelius [SV85], where :

- by $\boldsymbol{u}_h \in (\mathbb{P}_2)^d$ we mean that $\boldsymbol{u}_h$ is a vector field with entries over $\mathcal{D}$ that are continuous polynomials of maximal degree 2 ;

– and by $p_h \in \mathbb{P}_{1,disc}$ we mean that p_h is a scalar field with entries over $\mathcal{T}_h$ that are piecewise continuous polynomials of maximal degree 1 (thus discontinuous over $\mathcal{D}$).

This choice is very convenient to establish the free-energy estimates at the discrete level. As mentioned earlier, other choices will be discussed in Section 2-IV-C. For general meshes, this finite element does not satisfy the Babuška-Brezzi inf-sup condition. However, for meshes built using a particular process based on a first mesh of macro-elements, this mixed finite element space is known to satisfy the Babuška-Brezzi inf-sup condition (this is detailed in [AQ92] for instance). The interest of this finite element is that the velocity field is divergence-free :

$$\operatorname{div} \boldsymbol{u}_h(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \mathcal{D}, \tag{2-III.3}$$

because $\operatorname{div} \boldsymbol{u}_h \in \mathbb{P}_{1,disc}$ can be used as a test function for the pressure field in the weak formulation of the incompressibility constraint $\int_{\mathcal{D}} (\operatorname{div} \boldsymbol{u}_h) q_h = 0$.

For the approximation of $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$, we use *discontinuous* finite elements to derive the free energy estimates. For simplicity, we first consider piecewise constant approximations of $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$ in Sections 2-III and 2-IV. In Appendix D, we will come back to this assumption and discuss the use of higher-order finite element spaces for $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$. All along this work, we denote by $\boldsymbol{\sigma}_h \in (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ the fact that the symmetric-tensor field $\boldsymbol{\sigma}_h$ is discretized using a $\frac{d(d+1)}{2}$-dimensional so-called *stress* field, which stands for the entries in $\mathbb{P}_0$ of a symmetric $(d \times d)$-dimensional tensor field, thus enforcing the symmetry in the discretization.

The advection terms $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\sigma}$ and $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\psi}$ will be discretized either through a *characteristic* method in the spirit of [Pir82, BM97, WKL00], or with the *discontinuous Galerkin* (DG) method in the spirit of [HFK05]. Notice already that the characteristic method requires the velocity field to be more regular than the discontinuous Galerkin method in order to define the flow associated with the vector field $\boldsymbol{u}_h$.

For the discontinuous Galerkin method, we will need the following notation. Let E_j be some internal edge in the mesh $\mathcal{T}_h$. To each edge E_j, we associate a unitary orthogonal vector $\boldsymbol{n} \equiv \boldsymbol{n}_{E_j}$, whose orientation will not matter in the following. Then, for a given velocity field $\boldsymbol{u}_h$ in $\mathcal{D}$ that is well defined on the edges, for any variable $\boldsymbol{\phi}$ in $\mathcal{D}$ and any interior point $\boldsymbol{x}$ to the edge E_j, we respectively define the downstream and upstream values of $\boldsymbol{\phi}$ by :

$$\boldsymbol{\phi}^+(\boldsymbol{x}) = \lim_{\delta \to 0^+} \boldsymbol{\phi}(\boldsymbol{x} + \delta \boldsymbol{u}_h(\boldsymbol{x})) \text{ and } \boldsymbol{\phi}^-(\boldsymbol{x}) = \lim_{\delta \to 0^-} \boldsymbol{\phi}(\boldsymbol{x} + \delta \boldsymbol{u}_h(\boldsymbol{x})). \tag{2-III.4}$$

We denote by $[\![\boldsymbol{\phi}]\!](\boldsymbol{x}) = \boldsymbol{\phi}^+(\boldsymbol{x}) - \boldsymbol{\phi}^-(\boldsymbol{x})$ the jump of $\boldsymbol{\phi}$ over the edge E_j and by $\{\boldsymbol{\phi}\}(\boldsymbol{x}) = \frac{\boldsymbol{\phi}^+(\boldsymbol{x}) + \boldsymbol{\phi}^-(\boldsymbol{x})}{2}$ the mean value over the edge. Then, one can easily check the following formula for any function ϕ :

$$\sum_{E_j} \int_{E_j} |\boldsymbol{u}_h \cdot \boldsymbol{n}| [\![\phi]\!] = -\sum_{K_k} \int_{\partial K_k} (\boldsymbol{u}_h \cdot \boldsymbol{n}_{K_k}) \phi. \tag{2-III.5}$$

Let us now present in the next section the discrete variational formulations we will consider.

Remark 1. *In what follows, we do not consider the possible instabilities occurring when advection dominates diffusion in the Navier-Stokes equation for the velocity field $\boldsymbol{u}_h$. Indeed, in practice, one typically considers small Reynolds number flows for polymeric fluids, so that we are in a regime where such instabilities are not observed. Moreover, we also assume that $0 \leq \varepsilon < 1$ so that there is no problem of compatibilities between the discretization space for the velocity and for the stress (see [BPS01] for more details).*

2-III-C Numerical schemes with $\boldsymbol{\sigma}_h$ piecewise constant

Variational formulations of the discrete problem write, for all $n = 0, \ldots, N_T - 1$, as follows :

With the *characteristic* **method :** For a given $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\sigma}_h^n)$, find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}, \boldsymbol{\sigma}_h^{n+1}) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for any test function $(\boldsymbol{v}, q, \boldsymbol{\phi}) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$\begin{aligned} 0 = \int_{\mathcal{D}} \operatorname{Re} \left(\frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{\Delta t} + \boldsymbol{u}_h^n \cdot \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \cdot \boldsymbol{v} - p_h^{n+1} \operatorname{div} \boldsymbol{v} + q \operatorname{div} \boldsymbol{u}_h^{n+1} + (1 - \varepsilon) \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}} \boldsymbol{\sigma}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} \\ + \left(\frac{\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{\sigma}_h^n \circ X^n(t^n)}{\Delta t} \right) : \boldsymbol{\phi} - \left(\left(\boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \boldsymbol{\sigma}_h^{n+1} + \boldsymbol{\sigma}_h^{n+1} \left(\boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right)^T \right) : \boldsymbol{\phi} + \frac{1}{\mathrm{Wi}} \left(\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{I} \right) : \boldsymbol{\phi}. \end{aligned} \tag{2-III.6}$$

This problem is supplied with an initial condition $(\boldsymbol{u}_h^0, p_h^0, \boldsymbol{\sigma}_h^0) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$.

The function $X^n(t): x\in\mathcal{D}\mapsto X^n(t,x)\in\mathcal{D}$ is the "backward" flow associated with the velocity field $\boldsymbol{u}_h^n$ and satisfies, for all $x\in\mathcal{D}$:

$$\begin{cases} \frac{d}{dt}X^n(t,x)=\boldsymbol{u}_h^n(X^n(t,x)), \quad \forall t\in[t^n,t^{n+1}], \\ X^n(t^{n+1},x)=x. \end{cases} \tag{2-III.7}$$

With the *discontinuous Galerkin* **method :** For a given $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\sigma}_h^n)$, find $(\boldsymbol{u}_h^{n+1},p_h^{n+1},\boldsymbol{\sigma}_h^{n+1})\in(\mathbb{P}_2)^d\times\mathbb{P}_{1,disc}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for any test function $(\boldsymbol{v},q,\boldsymbol{\phi})\in(\mathbb{P}_2)^d\times\mathbb{P}_{1,disc}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$\begin{aligned} 0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v} \\ +\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n}{\Delta t}\right):\boldsymbol{\phi}-\left(\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\boldsymbol{\sigma}_h^{n+1}+\boldsymbol{\sigma}_h^{n+1}\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)^T\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\boldsymbol{\phi} \\ +\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\,[\![\boldsymbol{\sigma}_h^{n+1}]\!]:\boldsymbol{\phi}^+. \end{aligned} \tag{2-III.8}$$

Since $\boldsymbol{\sigma}_h\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ is discontinuous, we have discretized the advection term for $\boldsymbol{\sigma}_h$ with a sum of jumps similar to the usual upwind technique, where $\boldsymbol{\phi}^+=\left(\frac{1}{2}[\![\boldsymbol{\phi}]\!]+\{\boldsymbol{\phi}\}\right)$ (see [EG04, HFK05]).

Remark 2. *In all the following, we assume that, when using the characteristic method :*
- *the characteristics are exactly integrated ;*
- *and the integrals involving the backward flow X^n are exactly computed.*

We are aware of the fact that these assumptions are strong, and that numerical instabilities may be induced by bad integration schemes [MPS88, S̈88]. Hence, considering the lack for an analysis of those integration schemes for the characteristics in the present study, our analysis of discontinuous Galerkin discretizations of the advection terms may seem closer to the real implementation than that of the discretizations using the characteristic method.

2-III-D Numerical schemes with ψ_h piecewise constant

We now show how to discretize the variational log-formulation similarly as above. For this, we will need the following elementwise decomposition of the velocity gradient (see Lem. 4 above) :

$$\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}=\boldsymbol{\Omega}_h^{n+1}+\boldsymbol{B}_h^{n+1}+\mathcal{N}_h^{n+1}e^{-\boldsymbol{\psi}_h^{n+1}}. \tag{2-III.9}$$

Moreover, for the decomposition (2-III.9) with $\boldsymbol{u}\in(\mathbb{P}_2)^d$, we will need the following Lemma 5 for $k=1$, which is proved in Appendix C :

Lemma 5. *Let $\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\in(\mathbb{P}_{k,\mathrm{disc}})^{d\times d}$ for some $k\in\mathbb{N}$. Then, for any symmetric positive definite matrix $e^{\boldsymbol{\psi}_h^{n+1}}\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, there exist two antisymmetric matrices $\boldsymbol{\Omega}_h^{n+1},\mathcal{N}_h^{n+1}\in(\mathbb{P}_{k,\mathrm{disc}})^{\frac{d(d-1)}{2}}$ and a symmetric matrix $\boldsymbol{B}_h^{n+1}\in(\mathbb{P}_{k,\mathrm{disc}})^{\frac{d(d+1)}{2}}$ that commutes with $e^{\boldsymbol{\psi}_h^{n+1}}$, such that the matrix-valued function $\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}$ can be decomposed pointwise as : $\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}=\boldsymbol{\Omega}_h^{n+1}+\boldsymbol{B}_h^{n+1}+\mathcal{N}_h^{n+1}e^{-\boldsymbol{\psi}_h^{n+1}}$.*

Variational formulations of the discrete problem write, for all $n=0,\dots,N_T-1$, as follows :

With the *characteristic* method : For a given $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\psi}_h^n)$, find $(\boldsymbol{u}_h^{n+1},p_h^{n+1},\boldsymbol{\psi}_h^{n+1})\in(\mathbb{P}_2)^d\times\mathbb{P}_{1,disc}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for any test function $(\boldsymbol{v},q,\boldsymbol{\phi})\in(\mathbb{P}_2)^d\times\mathbb{P}_{1,disc}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$\begin{aligned} 0=\int_{\mathcal{D}}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}e^{\boldsymbol{\psi}_h^{n+1}}:\boldsymbol{\nabla}\boldsymbol{v} \\ +\left(\frac{\boldsymbol{\psi}_h^{n+1}-\boldsymbol{\psi}_h^n\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-\left(\boldsymbol{\Omega}_h^{n+1}\boldsymbol{\psi}_h^{n+1}-\boldsymbol{\psi}_h^{n+1}\boldsymbol{\Omega}_h^{n+1}\right):\boldsymbol{\phi}-2\boldsymbol{B}_h^{n+1}:\boldsymbol{\phi}-\frac{1}{\mathrm{Wi}}\left(e^{-\boldsymbol{\psi}_h^{n+1}}-\boldsymbol{I}\right):\boldsymbol{\phi}, \end{aligned} \tag{2-III.10}$$

where the initial condition $(\boldsymbol{u}_h^0,p_h^0,\boldsymbol{\psi}_h^0)\in(\mathbb{P}_2)^d\times\mathbb{P}_{1,disc}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ is given and where $X^n(t)$ is again defined by (2-III.7).

With the *discontinuous Galerkin* method : For a given $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\psi}_h^n)$, find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}, \boldsymbol{\psi}_h^{n+1}) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for any test function $(\boldsymbol{v}, q, \boldsymbol{\phi}) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$
\begin{aligned}
0 = \sum_{k=1}^{N_K} \int_{K_k} \mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{\Delta t} + \boldsymbol{u}_h^n \cdot \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \cdot \boldsymbol{v} - p_h^{n+1} \operatorname{div} \boldsymbol{v} + q \operatorname{div} \boldsymbol{u}_h^{n+1} + (1-\varepsilon) \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}} \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} : \boldsymbol{\nabla} \boldsymbol{v} \\
+ \left(\frac{\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n}{\Delta t} \right) : \boldsymbol{\phi} - \left(\boldsymbol{\Omega}_h^{n+1} \boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^{n+1} \boldsymbol{\Omega}_h^{n+1} \right) : \boldsymbol{\phi} - 2 \boldsymbol{B}_h^{n+1} : \boldsymbol{\phi} - \frac{1}{\mathrm{Wi}} \left(e^{-\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I} \right) : \boldsymbol{\phi} \\
+ \sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}| \, [\![\boldsymbol{\psi}_h^{n+1}]\!] : \boldsymbol{\phi}^+ . \quad \text{(2-III.11)}
\end{aligned}
$$

Remark 3. *Notice that the numerical schemes we propose are nonlinear due to the implicit terms corresponding to the discretization of the upper-convective derivative $(\boldsymbol{\nabla} \boldsymbol{u}) \boldsymbol{\sigma} + \boldsymbol{\sigma} (\boldsymbol{\nabla} \boldsymbol{u})^T$ (resp. $\boldsymbol{\Omega} \boldsymbol{\psi} - \boldsymbol{\psi} \boldsymbol{\Omega}$). In practice, this nonlinear system can be solved by fixed point procedures, either using the values at the previous time step as an initial guess, or using a predictor obtained by solving another scheme where the nonlinear terms are explicited.*

2-III-E Local existence and uniqueness of the discrete solutions

Before we show how to recover free energy estimates at the discrete level, let us now deal with the local-in-time existence and uniqueness of solutions to the discrete problems presented above.

First, since the mixed finite element space of Scott-Vogelius chosen in the systems above for the velocity-pressure field satisfies the Babuška-Brezzi inf-sup condition, notice that the system (2-III.6) is equivalent to the following for all $n = 0, \dots, N_T - 1$: For a given $(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n)$, find $(\boldsymbol{u}_h^{n+1}, \boldsymbol{\sigma}_h^{n+1}) \in (\mathbb{P}_2)^d_{\mathrm{div}=0} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for any test function $(\boldsymbol{v}, \boldsymbol{\phi}) \in (\mathbb{P}_2)^d_{\mathrm{div}=0} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$
\begin{aligned}
0 = \int_{\mathcal{D}} \mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{\Delta t} + \boldsymbol{u}_h^n \cdot \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \cdot \boldsymbol{v} + (1-\varepsilon) \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}} \boldsymbol{\sigma}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} \\
+ \left(\frac{\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{\sigma}_h^n \circ X^n(t^n)}{\Delta t} \right) : \boldsymbol{\phi} - \left(\left(\boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \boldsymbol{\sigma}_h^{n+1} + \boldsymbol{\sigma}_h^{n+1} \left(\boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right)^T \right) : \boldsymbol{\phi} + \frac{1}{\mathrm{Wi}} \left(\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{I} \right) : \boldsymbol{\phi}, \quad \text{(2-III.12)}
\end{aligned}
$$

where the flow $X^n(t)$ is defined by (2-III.7) and where we have used the following notation :

$$
(\mathbb{P}_2)^d_{\mathrm{div}=0} = \left\{ \boldsymbol{v} \in (\mathbb{P}_2)^d, \int_{\mathcal{D}} q \operatorname{div} \boldsymbol{v} = 0, \ \forall q \in \mathbb{P}_{1,disc} \right\}. \quad \text{(2-III.13)}
$$

Notice that it is also straightforward to rewrite the systems (2-III.8), (2-III.10) and (2-III.11) using $\boldsymbol{u}_h \in (\mathbb{P}_2)^d_{\mathrm{div}=0}$ instead of $(\boldsymbol{u}_h, p_h) \in (\mathbb{P}_2)^d \times \mathbb{P}_{1,disc}$. For instance, the system (2-III.10) is equivalent to : For a given $(\boldsymbol{u}_h^n, \boldsymbol{\psi}_h^n)$, find $(\boldsymbol{u}_h^{n+1}, \boldsymbol{\psi}_h^{n+1}) \in (\mathbb{P}_2)^d_{\mathrm{div}=0} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ such that, for all $(\boldsymbol{v}, \boldsymbol{\phi}) \in (\mathbb{P}_2)^d_{\mathrm{div}=0} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$
\begin{aligned}
0 = \int_{\mathcal{D}} \mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{\Delta t} + \boldsymbol{u}_h^n \cdot \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} \right) \cdot \boldsymbol{v} + (1-\varepsilon) \boldsymbol{\nabla} \boldsymbol{u}_h^{n+1} : \boldsymbol{\nabla} \boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}} e^{\boldsymbol{\psi}_h^{n+1}} : \boldsymbol{\nabla} \boldsymbol{v} \\
+ \left(\frac{\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)}{\Delta t} \right) : \boldsymbol{\phi} - \left(\boldsymbol{\Omega}_h^{n+1} \boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^{n+1} \boldsymbol{\Omega}_h^{n+1} \right) : \boldsymbol{\phi} - 2 \boldsymbol{B}_h^{n+1} : \boldsymbol{\phi} - \frac{1}{\mathrm{Wi}} \left(e^{-\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I} \right) : \boldsymbol{\phi}. \quad \text{(2-III.14)}
\end{aligned}
$$

Then, we have the :

Proposition 1. *Assume Scott-Vogelius finite elements are used for velocity-pressure, and piecewise constant discretization for the stress. For any couple $(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n)$ with $\boldsymbol{\sigma}_h^n$ symmetric positive definite, there exists $c_0 \equiv c_0(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n) > 0$ such that, for all $0 \leq \Delta t < c_0$, there exists a unique solution $(\boldsymbol{u}_h^{n+1}, \boldsymbol{\sigma}_h^{n+1})$ to the system* (2-III.6) *(resp.* (2-III.8)*) with $\boldsymbol{\sigma}_h^{n+1}$ symmetric positive definite.*

Proof of Proposition 1. The proofs for systems (2-III.6) and (2-III.8) are similar, so we will proceed with the proof for system (2-III.6) only, using its restatement as system (2-III.12).

For a given mesh $\mathcal{T}_h$, let us denote by $Y^{n+1} \in \mathbb{R}^{2N_D + 3N_K}$ the vector whose entries are respectively the nodal and elementwise values of $(\boldsymbol{u}_h^{n+1}, \boldsymbol{\sigma}_h^{n+1})$, solution to the system (2-III.12). The system of equations (2-III.12)

rewrites in terms of the vector $Y^{n+1}\in\mathbb{R}^{2N_D+3N_K}$ as : for a given Y^n and Δt, find a zero Y^{n+1} of the function Q defined by

$$Q\Big(\Delta t, Y^{n+1}\Big) = \Delta t A\Big(Y^{n+1}\Big)Y^{n+1} + \Delta t B(Y^n)Y^{n+1} + Y^{n+1} - C(Y^n,\Delta t), \tag{2-III.15}$$

where A and B are linear continuous matrix-valued functions in $\mathbb{R}^{(2N_D+3N_K)\times(2N_D+3N_K)}$, and where C is a vector-valued function in $\mathbb{R}^{2N_D+3N_K}$ (notice that the dependence of the function C on Δt is only related to the computation of the backward flow during a time step Δt, so that $C(Y^n,0)=Y^n$, and with the DG method it simplifies as $C(Y^n,\Delta t)=Y^n$). The functions A, B and C also implicitly depend on $\mathcal{T}_h$, as well as on the parameters $\mathrm{Re},\mathrm{Wi},\varepsilon$.

Now, $Q(\Delta t,Y)$ is continuously differentiable with respect to $(\Delta t,Y)$ and we have, with I the identity matrix in $\mathbb{R}^{(2N_D+3N_K)\times(2N_D+3N_K)}$:

$$\nabla_Y Q(\Delta t,Y) = I + \Delta t\big(B(Y^n) + A(Y) + (\nabla_Y A)Y\big). \tag{2-III.16}$$

Then, for given vectors Y^n and Y, the matrix $\nabla_Y Q(\Delta t,Y)$ is invertible for all Δt such that :

$$0\le\Delta t\le\|B(Y^n)+A(Y)+(\nabla_Y A)Y\|^{-1}$$

(with convention $\|B(Y^n)+A(Y)+(\nabla_Y A)Y\|^{-1}=\infty$ if $B(Y^n)+A(Y)+(\nabla_Y A)Y=0$), and then defines an isomorphism in $\mathbb{R}^{2N_D+3N_K}$.

Let us denote by S_+^* the subset of $\mathbb{R}^{2N_D+3N_K}$ that only contains vectors corresponding to elementwise values of *positive definite* matrix-valued functions $\boldsymbol{\sigma}_h$ in $\mathcal{D}$. Since $\mathcal{S}_+^*(\mathbb{R}^{d\times d})$ is an open (convex) domain of $\mathbb{R}^{d\times d}$, S_+^* is clearly an open (convex) domain of $\mathbb{R}^{2N_D+3N_K}$.

Since $Q(0,Y^n)=0$ and $\nabla_Y Q(0,Y^n)$ is invertible, by virtue of the implicit function theorem, there exist a neighborhood $[0,c_0)\times V(Y^n)$ of $(0,Y^n)$ in $\mathbb{R}_+\cap S_+^*$ and a continuously differentiable function $R:[0,c_0)\to V(Y^n)$, such that, for all $0\le\Delta t<c_0$:

$$Y = R(\Delta t) \Longleftrightarrow Q(\Delta t,Y)=0.$$

For a given time step $\Delta t\in[0,c_0)$ and a given symmetric positive definite tensor field $\boldsymbol{\sigma}_h^n$, $R(\Delta t)\in V(Y^n)$ is the vector of values Y^{n+1} for a solution $(\boldsymbol{u}_h^{n+1},\boldsymbol{\sigma}_h^{n+1})$ to the system (2-III.12) with a symmetric positive definite matrix $\boldsymbol{\sigma}_h^{n+1}$. Notice that, up to this point, $c_0=c_0(Y^n)$ is function of Y^n, as well as $\mathrm{Re},\mathrm{Wi},\varepsilon$ and $\mathcal{T}_h$. □

For solutions $(\boldsymbol{u}_h^n,\boldsymbol{\sigma}_h^n)$ to the systems (2-III.10) and (2-III.11), we similarly have :

Proposition 2. *Assume Scott-Vogelius finite elements are used for velocity-pressure, and piecewise constant discretization for the stress. Then, for any couple $(\boldsymbol{u}_h^n,\boldsymbol{\psi}_h^n)$, there exists a constant $c_0\equiv c_0(\boldsymbol{u}_h^n,\boldsymbol{\psi}_h^n)>0$ such that, for all $0\le\Delta t<c_0$, there exists a unique solution $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ to the system* (2-III.10) *(resp.* (2-III.11)*).*

The proof of Proposition 2 is similar to that of the Proposition 1, but for the expressions of $Q(\Delta t,Y)$ with respect to Y. An additional term $\Delta t D(Y)$ appears in Q due to $e^{\boldsymbol{\psi}_h^{n+1}}$. This term is continuously differentiable with respect to Y, and the derivative $\nabla_Y Q(0,Y^n)$ is still invertible. Thus, the proof can be completed using similar arguments.

Anticipating the results of Section 2-V, we would like to mention that the above results will be extended in two directions, using the discrete free energy estimates which will be proved in the following.

- We will show that the constant c_0 in Proposition 1 (resp. Prop. 2) can be chosen independently of $(\boldsymbol{u}_h^n,\boldsymbol{\sigma}_h^n)$ (resp. $(\boldsymbol{u}_h^n,\boldsymbol{\psi}_h^n)$), which yields a long-time existence and uniqueness result for the solutions to the discrete problems (see Props. 7 and 8 below). Of course, the limiting timestep will still depend on the parameters $\mathrm{Re},\mathrm{Wi},\varepsilon$ and on the mesh $\mathcal{T}_h$.
- We will also show, but for the log-formulation only, that it is possible to prove a long-time existence result without any restriction on the time step Δt (see Prop. 9 below).

2-IV Discrete free energy estimates with piecewise constant discretization of the stress fields $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$

In this section, we prove that various numerical schemes with piecewise constant $\boldsymbol{\sigma}_h$ or $\boldsymbol{\psi}_h$ satisfy a discrete free energy estimate. We first concentrate on Scott-Vogelius finite element spaces for $(\boldsymbol{u}_h,p_h)$ (introduced in Sect. 2-III) and then address the case of other mixed finite element spaces in Section 2-IV-C.

2-IV-A Free energy estimates with piecewise constant discretization of $\boldsymbol{\sigma}_h$

2-IV-A-a The characteristic method

Proposition 3. *Let $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\sigma}_h^n)_{0\le n\le N_T}$ be a solution to* (2-III.6)*, such that $\boldsymbol{\sigma}_h^n$ is positive definite. Then, the free energy of the solution $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\sigma}_h^n)$:*

$$F_h^n = F(\boldsymbol{u}_h^n,\boldsymbol{\sigma}_h^n) = \frac{Re}{2}\int_{\mathcal{D}}\left|\boldsymbol{u}_h^n\right|^2 + \frac{\varepsilon}{2\,Wi}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\sigma}_h^n - \ln\boldsymbol{\sigma}_h^n - \boldsymbol{I}), \tag{2-IV.1}$$

satisfies :

$$F_h^{n+1} - F_h^n + \int_{\mathcal{D}}\frac{Re}{2}\left|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\right|^2 + \Delta t\int_{\mathcal{D}}(1-\varepsilon)\left|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right|^2 + \frac{\varepsilon}{2\,Wi^2}\mathrm{tr}\left(\boldsymbol{\sigma}_h^{n+1} + \left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1} - 2I\right) \le 0. \tag{2-IV.2}$$

In particular, the sequence $(F_h^n)_{0\le n\le N_T}$ is non-increasing.

Proof of Proposition 3. Let $(\boldsymbol{u}_h^{n+1},p_h^{n+1},\boldsymbol{\sigma}_h^{n+1})$ be a solution to system (2-III.6). Notice that $(\boldsymbol{\sigma}_h^{n+1})^{-1}$ is still in $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$. We can thus use $(\boldsymbol{u}_h^{n+1},p_h^{n+1},\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-(\boldsymbol{\sigma}_h^{n+1})^{-1}\right))$ as a test function in the system (2-III.6), which yields :

$$\begin{aligned}
0 = \int_{\mathcal{D}}\mathrm{Re}\Bigg(&\frac{|\boldsymbol{u}_h^{n+1}|^2-|\boldsymbol{u}_h^n|^2}{2\Delta t} + \frac{|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n|^2}{2\Delta t} + \boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\frac{|\boldsymbol{u}_h^{n+1}|^2}{2}\Bigg)\\
&+(1-\varepsilon)\left|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right|^2 + \frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1} + \frac{\varepsilon}{2\mathrm{Wi}}\Bigg[\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n\circ X^n(t^n)}{\Delta t}\right):\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right)\\
&\qquad -2\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\boldsymbol{\sigma}_h^{n+1}:\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right) + \frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right)\Bigg].
\end{aligned}$$

We first examine the terms associated with momentum conservation and incompressibility. We recall that $\boldsymbol{u}_h^{n+1}$ satisfies (2-III.3) since we use Scott-Vogelius finite elements. By the Stokes theorem (using the no-slip boundary condition), we immediately obtain :

$$\int_{\mathcal{D}}\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\left|\boldsymbol{u}_h^{n+1}\right|^2 = -\int_{\mathcal{D}}(\mathrm{div}\,\boldsymbol{u}_h^n)\left|\boldsymbol{u}_h^{n+1}\right|^2 = 0.$$

The terms involving p_h^{n+1} also cancel. We now consider the terms involving $\boldsymbol{\sigma}_h^{n+1}$. The upper-convective term in the tensor derivative rewrites :

$$\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\boldsymbol{\sigma}_h^{n+1}:\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right) = \boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1} - \mathrm{div}\,\boldsymbol{u}_h^{n+1},$$

which vanishes after combination with the extra-stress term $\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}$ in the momentum conservation equation, and using the incompressibility property. The last term rewrites :

$$\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right) = \mathrm{tr}\left(\boldsymbol{\sigma}_h^{n+1} + \left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1} - 2\boldsymbol{I}\right).$$

The remaining term writes :

$$\begin{aligned}
\int_{\mathcal{D}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right):\left(\boldsymbol{I}-\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right) &= \int_{\mathcal{D}}\mathrm{tr}\left(\boldsymbol{\sigma}_h^{n+1}\right) - \mathrm{tr}\left(\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right)\\
&\quad + \mathrm{tr}\left(\left[\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right]\left[\boldsymbol{\sigma}_h^{n+1}\right]^{-1} - \boldsymbol{I}\right).
\end{aligned}$$

We first make use of (2-I.10) with $\boldsymbol{\sigma} = \boldsymbol{\sigma}_h^n\circ X^n(t^n)$ and $\boldsymbol{\tau} = \boldsymbol{\sigma}_h^{n+1}$:

$$\mathrm{tr}\left([\boldsymbol{\sigma}_h^n\circ X^n(t^n)][\boldsymbol{\sigma}_h^{n+1}]^{-1} - \boldsymbol{I}\right) \ge \mathrm{tr}\ln\left(\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right) - \mathrm{tr}\ln\left(\boldsymbol{\sigma}_h^{n+1}\right).$$

Then, we have :

$$\int_{\mathcal{D}} -\mathrm{tr}(\boldsymbol{\sigma}_h^n\circ X^n(t^n) + \ln(\boldsymbol{\sigma}_h^n\circ X^n(t^n))) = \int_{\mathcal{D}} -\mathrm{tr}\,(\boldsymbol{\sigma}_h^n + \ln\boldsymbol{\sigma}_h^n),$$

since the strong incompressibility property ($\mathrm{div}\,\boldsymbol{u}_h^n = 0$) implies that the flow $X^n(t)$ defines a mapping with constant Jacobian equal to 1 for all $t \in [t^n, t^{n+1}]$. Finally, we get the following lower bound :

$$\int_{\mathcal{D}} \Big(\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{\sigma}_h^n \circ X^n(t^n)\Big) : \Big(\boldsymbol{I} - \Big(\boldsymbol{\sigma}_h^{n+1}\Big)^{-1}\Big) \geq \int_{\mathcal{D}} \mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big) - \mathrm{tr}\Big(\boldsymbol{\sigma}_h^n - \ln \boldsymbol{\sigma}_h^n\Big),$$

hence the result (2-IV.2).

Notice that $\mathrm{tr}\,\big(\boldsymbol{\sigma}_h^{n+1} + (\boldsymbol{\sigma}_h^{n+1})^{-1} - 2\boldsymbol{I}\big) \geq 0$ by virtue of the equation (2-I.8), which shows that the sequence $(F_h^n)_{0 \leq n \leq N_T}$ is non-increasing. □

2-IV-A-b The discontinuous Galerkin method

Proposition 4. *Let $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\sigma}_h^n)_{0 \leq n \leq N_T}$ be a solution to* (2-III.8)*, such that $\boldsymbol{\sigma}_h^n$ is positive definite. Then, the free energy F_h^n defined by* (2-IV.1) *satisfies the free energy estimate* (2-IV.2)*. In particular, the sequence $(F_h^n)_{0 \leq n \leq N_T}$ is non-increasing.*

Proof of Proposition 4. We only point out the differences with the proof of Proposition 3. They consist in the treatment of the discretization of the advection terms for $\boldsymbol{\sigma}_h$. We recall that the test function in stress is $\boldsymbol{\phi} = \frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I} - (\boldsymbol{\sigma}_h^{n+1})^{-1})$, so that we have :

$$\sum_{j=1}^{N_E} \int_{E_j} \Big|\boldsymbol{u}_h^n \cdot \boldsymbol{n}\Big| [\![\boldsymbol{\sigma}_h^{n+1}]\!] : \Big(\boldsymbol{I} - \Big(\boldsymbol{\sigma}_h^{n+1}\Big)^{-1}\Big)^+ = \sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}|\, [\![\mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1}\Big)]\!] \\ + |\boldsymbol{u}_h^n \cdot \boldsymbol{n}|\, \mathrm{tr}\left(\boldsymbol{\sigma}_h^{n+1,-}\Big(\boldsymbol{\sigma}_h^{n+1,+}\Big)^{-1} - \boldsymbol{I}\right).$$

Again, we make use of (2-I.10), with $\boldsymbol{\sigma} = \boldsymbol{\sigma}_h^{n+1,-}$ and $\boldsymbol{\tau} = \boldsymbol{\sigma}_h^{n+1,+}$:

$$\mathrm{tr}\left(\boldsymbol{\sigma}_h^{n+1,-}\Big(\boldsymbol{\sigma}_h^{n+1,+}\Big)^{-1} - \boldsymbol{I}\right) \geq \mathrm{tr}\Big(\ln \boldsymbol{\sigma}_h^{n+1,-} - \ln \boldsymbol{\sigma}_h^{n+1,+}\Big).$$

We get, by formula (2-III.5), the fact that $\boldsymbol{\sigma}_h^{n+1} \in (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, the Stokes theorem and the incompressibility property (2-III.3) :

$$\begin{aligned}
\sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}|\, [\![\boldsymbol{\sigma}_h^{n+1}]\!] : \Big(\boldsymbol{I} - \Big(\boldsymbol{\sigma}_h^{n+1}\Big)^{-1}\Big)^+ &\geq \sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}|\, [\![\mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big)]\!] \\
&= -\sum_{k=1}^{N_K} \int_{\partial K_k} (\boldsymbol{u}_h^n \cdot \boldsymbol{n}_{K_k})\, \mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big) \\
&= -\sum_{k=1}^{N_K} \Big(\mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big)\Big)\Big|_{K_k} \int_{\partial K_k} \boldsymbol{u}_h^n \cdot \boldsymbol{n}_{K_k} \\
&= -\sum_{k=1}^{N_K} \Big(\mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big)\Big)\Big|_{K_k} \int_{K_k} \mathrm{div}(\boldsymbol{u}_h^n) \\
&= 0. \qquad \text{(2-IV.3)}
\end{aligned}$$

Moreover, it is easy to prove the following, using the same technique as in the proof of Proposition 3 :

$$\int_{\mathcal{D}} \Big(\boldsymbol{\sigma}_h^{n+1} - \boldsymbol{\sigma}_h^n\Big) : \Big(\boldsymbol{I} - \Big(\boldsymbol{\sigma}_h^{n+1}\Big)^{-1}\Big) \geq \int_{\mathcal{D}} \mathrm{tr}\Big(\boldsymbol{\sigma}_h^{n+1} - \ln \boldsymbol{\sigma}_h^{n+1}\Big) - \mathrm{tr}\Big(\boldsymbol{\sigma}_h^n - \ln \boldsymbol{\sigma}_h^n\Big).$$

This concludes the proof. □

2-IV-B Free energy estimates with piecewise constant discretization of ψ_h

This section is the equivalent of the previous section for the log-formulation.

2-IV-B-a The characteristic method

Proposition 5. *Let* $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\psi}_h^n)_{0\le n\le N_T}$ *be a solution to* (2-III.10)*. Then, the free energy of the solution* $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\psi}_h^n)$ *:*

$$F_h^n = F\Big(\boldsymbol{u}_h^n, e^{\boldsymbol{\psi}_h^n}\Big) = \frac{Re}{2}\int_{\mathcal{D}} |\boldsymbol{u}_h^n|^2 + \frac{\varepsilon}{2\,Wi}\int_{\mathcal{D}} \mathrm{tr}\Big(e^{\boldsymbol{\psi}_h^n} - \boldsymbol{\psi}_h^n - \boldsymbol{I}\Big), \tag{2-IV.4}$$

satisfies :

$$F_h^{n+1} - F_h^n + \int_{\mathcal{D}} \frac{Re}{2}\Big|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\Big|^2 + \Delta t\int_{\mathcal{D}} (1-\varepsilon)\Big|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big|^2 + \frac{\varepsilon}{2\,Wi^2}\mathrm{tr}\Big(e^{\boldsymbol{\psi}_h^{n+1}} + e^{-\boldsymbol{\psi}_h^{n+1}} - 2I\Big) \le 0. \tag{2-IV.5}$$

In particular, the sequence $(F_h^n)_{0\le n\le N_T}$ *is non-increasing.*

Proof of Proposition 5. We shall use as test functions $\Big(\boldsymbol{u}_h^{n+1}, p_h^{n+1}, \frac{\varepsilon}{2\mathrm{Wi}}(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I})\Big)$ in (2-III.10). We emphasize that, as long as the solution $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}, \boldsymbol{\psi}_h^{n+1})$ exists (see Prop. 2), $\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}}$ is well-defined, symmetric positive definite and piecewise constant.

The terms are treated similarly as in the proof of Proposition 3. For the material derivative of $\boldsymbol{\psi}_h$, we have :

$$\int_{\mathcal{D}} \Big(\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)\Big) : \Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I}\Big) = \int_{\mathcal{D}} \Big(\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)\Big) : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \mathrm{tr}\Big(\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)\Big).$$

Using the equation (2-I.11) with $\boldsymbol{\sigma} = \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}}$ and $\boldsymbol{\tau} = \mathrm{e}^{\boldsymbol{\psi}_h^n \circ X^n(t^n)}$, we obtain :

$$\Big(\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)\Big) : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} \ge \mathrm{tr}\Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \mathrm{e}^{\boldsymbol{\psi}_h^n \circ X^n(t^n)}\Big),$$

and thus :

$$\begin{aligned}\int_{\mathcal{D}} \Big(\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^n \circ X^n(t^n)\Big) : \Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I}\Big) &\ge \int_{\mathcal{D}} \mathrm{tr}\Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{\psi}_h^{n+1}\Big) - \int_{\mathcal{D}} \mathrm{tr}\Big(\mathrm{e}^{\boldsymbol{\psi}_h^n} - \boldsymbol{\psi}_h^n\Big) \circ X^n(t^n) \\ &= \int_{\mathcal{D}} \mathrm{tr}\Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{\psi}_h^{n+1}\Big) - \int_{\mathcal{D}} \mathrm{tr}(\mathrm{e}^{\boldsymbol{\psi}_h^n} - \boldsymbol{\psi}_h^n),\end{aligned}$$

where the fact that the Jacobian of the flow X^n is constant equal to one (because $\boldsymbol{u}_h^n$ is divergence-free) has been used in the change of variable in the last equality.

Besides, using the equation (2-II.16), we have :

$$\int_{\mathcal{D}} \Big(\boldsymbol{\Omega}_h^{n+1}\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^{n+1}\boldsymbol{\Omega}_h^{n+1}\Big) : \Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I}\Big) = \int_{\mathcal{D}} \Big(\boldsymbol{\Omega}_h^{n+1}\boldsymbol{\psi}_h^{n+1} - \boldsymbol{\psi}_h^{n+1}\boldsymbol{\Omega}_h^{n+1}\Big) : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} = 0.$$

Last, using (2-II.15) :

$$\begin{aligned}\int_{\mathcal{D}} \boldsymbol{B}_h^{n+1} : \Big(\mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \boldsymbol{I}\Big) &= \int_{\mathcal{D}} \boldsymbol{B}_h^{n+1} : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \int_{\mathcal{D}} \mathrm{tr}\Big(\boldsymbol{B}_h^{n+1}\Big) \\ &= \int_{\mathcal{D}} \boldsymbol{\nabla}\boldsymbol{u}_h^{n+1} : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} - \int_{\mathcal{D}} \mathrm{div}\Big(\boldsymbol{u}_h^{n+1}\Big) = \int_{\mathcal{D}} \boldsymbol{\nabla}\boldsymbol{u}_h^{n+1} : \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}},\end{aligned}$$

which cancels out with the same term $\int_{\mathcal{D}} \mathrm{e}^{\boldsymbol{\psi}_h^{n+1}} : \boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}$ in the momentum equation. □

2-IV-B-b The discontinuous Galerkin method

Proposition 6. *Let* $(\boldsymbol{u}_h^n, p_h^n, \boldsymbol{\psi}_h^n)_{0\le n\le N_T}$ *be a solution to* (2-III.11)*. Then, the free energy* F_h^n *defined by* (2-IV.4) *satisfies the free energy estimate* (2-IV.5)*. In particular, the sequence* $(F_h^n)_{0\le n\le N_T}$ *is non-increasing.*

The proof is straightforward using elements of the proofs of Propositions 5 and 4.

TAB. 2.1 – Summary of the arguments with $(\boldsymbol{u}_h, p_h, \boldsymbol{\sigma}_h)$ or $(\boldsymbol{u}_h, p_h, \boldsymbol{\psi}_h)$ in $(\mathbb{P}_2)^d \times \mathbb{P}_{1,disc} \times (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$.

Advection discretized by	Characteristics	DG
Requirements for $\boldsymbol{u}_h$	$\operatorname{div} \boldsymbol{u}_h = 0$ $(\Rightarrow \det(\nabla_{\boldsymbol{x}} X^n) \equiv 1)$ $(\Rightarrow (\boldsymbol{u}_h \cdot \boldsymbol{n})\vert_{E_j}$ well defined)	$\int_{\mathcal{D}} q \operatorname{div} \boldsymbol{u}_h = 0$, $\forall q \in \mathbb{P}_0$ and $(\boldsymbol{u}_h \cdot \boldsymbol{n})\vert_{E_j}$ well defined

2-IV-C Other finite elements for $(\boldsymbol{u}_h, p_h)$

In this section, we review some finite element spaces for $(\boldsymbol{u}_h, p_h)$ other than Scott-Vogelius for which the results of the last two sections still hold.

First, let us stress the key arguments we used in the proofs above. If the advection terms $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\sigma}$ and $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\psi}$ are discretized by the *characteristic* method, we need the velocity field $\boldsymbol{u}_h^n$ to be divergence-free :

$$\operatorname{div} \boldsymbol{u}_h^n = 0, \tag{2-IV.6}$$

in order for the flow X^n satisfying (2-III.7) to be with Jacobian one. When $\boldsymbol{u}_h^n$ is only piecewise smooth (consider below the case of $\mathbb{P}_{1,disc}$ velocity fields), the divergence in the left-hand side of (2-IV.6) should be understood in the sense of distributions. By the way, the equation (2-IV.6) ensures that the trace of the normal component $\boldsymbol{u}_h^n \cdot \boldsymbol{n}$ on the edges of the mesh is uniquely defined, which is a sufficient condition to define the flow associated with an elementwise-Lipschitz-continuous vector field $\boldsymbol{u}_h^n$ through (2-III.7), and which is necessary to treat the advection term in the Navier-Stokes equation (see [Pir82]).

If the advection terms are discretized by the *discontinuous Galerkin* method, it is necessary that the trace of the normal component of $\boldsymbol{u}_h$ be uniquely defined on the edges of the mesh since it appears in the jump terms $\sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}| [\![\boldsymbol{\sigma}_h^{n+1}]\!] : \boldsymbol{\phi}^+$ or $\sum_{j=1}^{N_E} \int_{E_j} |\boldsymbol{u}_h^n \cdot \boldsymbol{n}| [\![\boldsymbol{\psi}_h^{n+1}]\!] : \boldsymbol{\phi}^+$ in the variational formulations. But to obtain (2-IV.3), and contrary to the characteristic method, only the following *weak* incompressibility property is needed :

$$\forall k = 1, \dots, N_K, \quad \int_{K_k} \operatorname{div} \boldsymbol{u}_h^n = 0,$$

which is equivalent to writing :

$$\forall q \in \mathbb{P}_0, \quad \int_{\mathcal{D}} \operatorname{div}(\boldsymbol{u}_h^n) q = 0. \tag{2-IV.7}$$

The properties needed to obtain the discrete free energy estimates are summarized in Table 2.1.

Below, we consider the following alternative choices of the finite elements space for $(\boldsymbol{u}_h, p_h)$:

– the Taylor-Hood finite element space : $(\boldsymbol{u}_h, p_h) \in (\mathbb{P}_2)^d \times \mathbb{P}_1$, which satisfies the Babuška-Brezzi inf-sup condition, whatever the mesh ;

– the mixed Crouzeix-Raviart finite element space (see [CR73]) : $(\boldsymbol{u}_h, p_h) \in (\mathbb{P}_1^{CR})^d \times \mathbb{P}_0$, where

$$\mathbb{P}_1^{CR} = \left\{ v \in \mathbb{P}_{1,disc} \,\middle|\, \forall E_j, \int_{E_j} [\![v]\!] = 0 \right\}, \tag{2-IV.8}$$

which also satisfies the Babuška-Brezzi inf-sup condition, whatever the mesh ;

– stabilized formulations for $(\boldsymbol{u}_h, p_h) \in (\mathbb{P}_1)^d \times \mathbb{P}_1$ or $(\boldsymbol{u}_h, p_h) \in (\mathbb{P}_1)^d \times \mathbb{P}_0$.

This is not exhaustive, but it is sufficient to highlight which modifications are needed in the variational formulations, compared to the Scott-Vogelius mixed finite element, for the discrete free energy estimates to hold. In particular, some projection of the velocity field is needed in the discretization of the advection terms $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\sigma}$ and $\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{\psi}$ in order to satisfy the requirements of Table 2.1. These projection operators are introduced in the next Section 2-IV-C-a. The results of Section 2-IV-C are summarized in Table 2.2.

Remark 4. *For all the finite element spaces introduced here, similar existence results as those stated in Section 2-III-E hold. For the sake of conciseness, we do not restate these results, but rather concentrate on establishing the free energy estimates.*

2-IV-C-a Some useful projection operators for the velocity field

Let us introduce three projection operators for the velocity field.

Following [Pir82], we first define the orthogonal projection P_h^{rot} onto the piecewise constant solenoidal vector fields built from affine continuous scalar fields :

$$\{\nabla\times\zeta_h | \zeta_h\in(\mathbb{P}_1)^d, \zeta_h\times\boldsymbol{n}=0 \text{ on } \partial\mathcal{D}\}.$$

We suppose here that $d=3$, but the extension to the case $d=2$ is straightforward. We set $P_h^{rot}(\boldsymbol{u}_h)=\nabla\times\psi_h$ where $\psi_h\in(\mathbb{P}_1)^d$, such that $\psi_h\times\boldsymbol{n}|_{\partial\mathcal{D}}=0$, satisfies :

$$\int_{\mathcal{D}}(\nabla\psi_h):(\nabla\zeta_h)=\int_{\mathcal{D}}\boldsymbol{u}_h\cdot(\nabla\times\zeta_h),\ \forall\zeta_h\in(\mathbb{P}_1)^d, \zeta_h\times\boldsymbol{n}|_{\partial\mathcal{D}}=0.$$

Because $P_h^{rot}(\boldsymbol{u}_h)$ is solenoidal, we always have the strong incompressibility property (2-IV.6) :

$$\operatorname{div} P_h^{rot}(\boldsymbol{u}_h)=0,$$

for any velocity field $\boldsymbol{u}_h$. Of course, this operator is consistent only for divergence-free velocity fields $\boldsymbol{u}_h$ (or velocity field $\boldsymbol{u}_h$ with vanishing divergence when h goes to zero). See [Pir82] for consistency results.

Second, following [EG04], we define the Raviart-Thomas interpolator $P_h^{RT_0}$ onto the vector subspace of $(\mathbb{P}_{1,disc})^d$ made of the vector fields in $(\mathbb{P}_0)^d+\boldsymbol{x}\mathbb{P}_0$ with continuous normal component across the edges E_j (whose trace on E_j is then uniquely defined). The projection $P_h^{RT_0}(\boldsymbol{u}_h^n)$ clearly satisfies, for any element K_k :

$$\int_{K_k}\operatorname{div}\boldsymbol{u}_h=\int_{\partial K_k}\boldsymbol{u}_h^n\cdot\boldsymbol{n}_{K_k}=\int_{\partial K_k}P_h^{RT_0}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}_{K_k}=\int_{K_k}\operatorname{div}P_h^{RT_0}(\boldsymbol{u}_h^n). \tag{2-IV.9}$$

Thus, it satisfies the weak incompressibility property (2-IV.7) :

$$\forall q\in\mathbb{P}_0,\ \int_{\mathcal{D}}\operatorname{div}\Big(P_h^{RT_0}(\boldsymbol{u}_h^n)\Big)q=0,$$

if, and only if, the velocity field $\boldsymbol{u}_h^n$ also satisfies it.

Third, we define P_h^{BDM} as the Brezzi-Douglas-Marini projection operator [BJDM85, BJDM86]. It is with value in $(\mathbb{P}_1)^d$. This projection operator satisfies the same divergence preservation property (2-IV.9) than $P_h^{RT_0}$, but is of better accuracy.

Note that P_h^{BDM} and $P_h^{RT_0}$ are local interpolating operators in the sense that all the computations can be made elementwise. This is not the case for P_h^{rot}. In addition, we will need the following lemma :

Lemma 6. *For any velocity field $\boldsymbol{u}_h^n$ such that the previously defined interpolating operators are well defined, the normal components of the interpolated vector field, $P_h^{rot}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}$, $P_h^{RT_0}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}$ and $P_h^{BDM}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}$ are also well defined on any internal edges E_j. Moreover, if $\boldsymbol{u}_h^n\in(\mathbb{P}_{1,disc})^d$ is a velocity field such that, for all $k=1,\dots,N_K$:*

$$\int_{K_k}\operatorname{div}(\boldsymbol{u}_h^n)=0,$$

then $\operatorname{div}(P_h^{RT_0}(\boldsymbol{u}_h^n))=\operatorname{div}(P_h^{BDM}(\boldsymbol{u}_h^n))=0$ *(in the sense of distributions).*

Proof. By construction, $P_h^{RT_0}$ and P_h^{BDM} take their values in the set of velocity fields whose normal components are continuous across the edges. This is also the case for P_h^{rot} since P_h^{rot} takes its value in the set of divergence-free velocity fields. Then, from the equation (2-IV.9), we have $\int_{K_k}\operatorname{div}(P_h^{RT_0}(\boldsymbol{u}_h^n))=0$. Since $\operatorname{div}(P_h^{RT_0}(\boldsymbol{u}_h^n))$ is in $(\mathbb{P}_0)^d$, this shows that $\operatorname{div}(P_h^{RT_0}(\boldsymbol{u}_h^n))$ is zero in any element K_k. Finally, $P_h^{RT_0}(\boldsymbol{u}_h^n)$ has continuous normal components across the edges of the mesh. This shows that $\operatorname{div}(P_h^{RT_0}(\boldsymbol{u}_h^n))=0$ in the sense of distributions. The same proof holds for the projection operator P_h^{BDM}. □

2-IV-C-b Alternative mixed finite element space for $(\boldsymbol{u}_h,p_h)$ with inf-sup condition

In this section, we show how to derive discrete free energy estimates with mixed finite element spaces for the velocity and pressure fields which satisfy the inf-sup condition, but which are not the Scott-Vogelius finite elements.

Let us first consider the *Taylor-Hood* element for $(\boldsymbol{u}_h,p_h)$, that is $(\mathbb{P}_2)^d\times\mathbb{P}_1$. In this case, since the velocity field $\boldsymbol{u}_h$ is not divergence-free either in the weak form (2-IV.7), or in the strong form (2-IV.6), a projection of the velocity field is required in the discretization of the advection terms $\boldsymbol{u}\cdot\boldsymbol{\nabla\sigma}$ and $\boldsymbol{u}\cdot\boldsymbol{\nabla\psi}$. More precisely, we

need to use the projection velocity $P_h^{rot}\boldsymbol{u}_h^n$ (and, among the three projection operators we introduced above, this is the only one which is such that the strong or weak incompressibility is satisfied). For the *characteristic* method, one uses the flow $X^n(t)$ satisfying :

$$\begin{cases} \frac{d}{dt}X^n(t,x) = P_h^{rot}\boldsymbol{u}_h^n(X^n(t,x)), \quad \forall t\in[t^n,t^{n+1}], \\ X^n(t^{n+1},x) = x. \end{cases} \tag{2-IV.10}$$

For the *discontinuous Galerkin* method, the advection term in the conformation-tensor formulations writes (see the last line in (2-III.8)) :

$$+\sum_{j=1}^{N_E}\int_{E_j}\left|P_h^{rot}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}\right|[\![\boldsymbol{\sigma}_h^{n+1}]\!]:\boldsymbol{\phi}^+.$$

Notice that in the terms $[\![\boldsymbol{\sigma}_h^{n+1}]\!]:\boldsymbol{\phi}^+$, the projected velocity $P_h^{rot}\boldsymbol{u}_h^n$ is used to define the upstream and downstream values following (2-III.4). Another modification, which is specific to the Navier-Stokes equation, is needed to treat the advection term on the velocity. Namely, one needs to add to the weak formulation the so-called Temam correction term (see [Tem66]) :

$$+\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\mathrm{div}\,(\boldsymbol{u}_h^n)\left(\boldsymbol{v}\cdot\boldsymbol{u}_h^{n+1}\right) \tag{2-IV.11}$$

in such a way that, when $\boldsymbol{u}_h^{n+1}$ is used as a test function :

$$\mathrm{Re}\int_{\mathcal{D}}\left(\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{u}_h^{n+1}+\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\mathrm{div}\,(\boldsymbol{u}_h^n)\,|\boldsymbol{u}_h^{n+1}|^2=0.$$

With these modifications (projection of the velocity field in the advection terms, and Temam correction term), the free energy estimate (2-IV.2) is satisfied by the scheme. Similar results (discrete free energy estimates for $(\boldsymbol{u}_h,p_h,\boldsymbol{\psi}_h)$ in $(\mathbb{P}_2)^d\times\mathbb{P}_1\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$) can be proved on the log-formulation.

Let us now discuss the use of *Crouzeix-Raviart* finite elements for velocity : $(\boldsymbol{u}_h,p_h,\boldsymbol{\sigma}_h)$ in $(\mathbb{P}_1^{CR})^d\times\mathbb{P}_0\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ (see (2-IV.8)). In this case, the Navier-Stokes equations can be discretized using a characteristic method :

$$0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\circ X^n(t_n)}{\Delta t}\right)\cdot\boldsymbol{v}-p_h^{n+1}\,\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}, \tag{2-IV.12}$$

where X^n is obtained from the projected velocity field $P_h\boldsymbol{u}_h^n$ as :

$$\begin{cases} \frac{d}{dt}X^n(t) = P_h\boldsymbol{u}_h^n(X^n(t)), \quad \forall t\in[t^n,t^{n+1}], \\ X^n(t^{n+1}) = x. \end{cases} \tag{2-IV.13}$$

The projected velocity $P_h\boldsymbol{u}_h^n$ is defined using any of the three projectors presented above, that is $P_h^{rot}\boldsymbol{u}_h^n$, $P_h^{RT_0}\boldsymbol{u}_h^n$ or $P_h^{BDM}\boldsymbol{u}_h^n$. The Navier-Stokes equations can also be discretized using a *discontinuous Galerkin* formulation :

$$\begin{aligned} 0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}+\mathrm{Re}\sum_{j=1}^{N_E}\int_{E_j}|P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|\,[\![\boldsymbol{u}_h^{n+1}]\!]\cdot\{\boldsymbol{v}\} \\ +\sum_{k=1}^{N_K}\int_{K_k}-p_h^{n+1}\,\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}. \end{aligned} \tag{2-IV.14}$$

Here again, $P_h\boldsymbol{u}_h^n$ is any of the three projectors presented above. We would like to mention that we are not aware that the projector P_h^{rot} has ever been used with discontinuous Galerkin methods, so that the consistency of the discontinuous Galerkin approach combined with this projector still needs to be investigated. Likewise, the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ in the equation on the stress can be treated by the characteristic method or the discontinuous Galerkin method, as above for the advection term in the Navier-Stokes equations. Notice that whatever the projecting operator used, $\mathrm{div}(P_h\boldsymbol{u}_h^n)=0$ holds (see Lem. 6 above). With this property, it is easy

to check that Propositions 3 and 4 still hold for this finite element. For example, the advection term in the Navier-Stokes equations is treated as follows (using the fact that $\mathrm{div}(P_h(\boldsymbol{u}_h^n))=0$ and (2-III.5)) :

$$\begin{aligned}
&\sum_{k=1}^{N_K}\int_{K_k}\Big(P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)\cdot\boldsymbol{u}_h^{n+1}+\sum_{j=1}^{N_E}\int_{E_j}|P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|\,[\![\boldsymbol{u}_h^{n+1}]\!]\cdot\{\boldsymbol{u}_h^{n+1}\}\\
&\qquad=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{div}\left(P_h(\boldsymbol{u}_h^n)\frac{\left|\boldsymbol{u}_h^{n+1}\right|^2}{2}\right)+\sum_{j=1}^{N_E}\int_{E_j}|P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|\,\frac{1}{2}[\![\left|\boldsymbol{u}_h^{n+1}\right|^2]\!]\\
&\qquad=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{div}\left(P_h(\boldsymbol{u}_h^n)\frac{\left|\boldsymbol{u}_h^{n+1}\right|^2}{2}\right)-\sum_{k=1}^{N_K}\int_{\partial K_k}(P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}_{K_k})\frac{\left|\boldsymbol{u}_h^{n+1}\right|^2}{2}\quad=0.
\end{aligned}$$

Discrete free energy estimates for $(\boldsymbol{u}_h,p_h,\boldsymbol{\psi}_h)$ in $(\mathbb{P}_1^{CR})^d\times\mathbb{P}_0\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ can be similarly proven on the log-formulation.

2-IV-C-c Alternative mixed finite element space for $(\boldsymbol{u}_h,p_h)$ without inf-sup

It is also possible to choose a mixed finite elements space for $(\boldsymbol{u}_h,p_h)$ that does not satisfy the Babuška-Brezzi inf-sup condition, like $(\mathbb{P}_1)^d\times\mathbb{P}_0$ or $(\mathbb{P}_1)^d\times\mathbb{P}_1$. The loss of stability due to the incompatibility of the spaces can then be alleviated through a stabilization procedure, like Streamline Upwind Petrov Galerkin, Galerkin Least Square or Subgrid Scale Method (see [HF87, Cod98, Gue99]). In the following, we consider very simple stabilization procedures, for which only one simple quadratic term is added to the variational finite element formulation in order to restore stability of the discrete numerical scheme.

Let us first consider *the mixed finite element space* $(\mathbb{P}_1)^d\times\mathbb{P}_0$ for $(\boldsymbol{u}_h,p_h)$. If the term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized with the *characteristic* method, the system then writes :

$$\begin{aligned}
0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\Big(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\mathrm{div}\,\boldsymbol{u}_h^n(\boldsymbol{v}\cdot\boldsymbol{u}_h^{n+1})+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}\\
-p_h^{n+1}\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-\Big(\Big(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)\boldsymbol{\sigma}_h^{n+1}+\boldsymbol{\sigma}_h^{n+1}\Big(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)^T\Big):\boldsymbol{\phi}\\
+\frac{1}{\mathrm{Wi}}\Big(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\Big):\boldsymbol{\phi}+\sum_{j=1}^{N_E}|E_j|\int_{E_j}[\![p_h]\!][\![q]\!],\quad\text{(2-IV.15)}
\end{aligned}$$

with a flow X^n computed with the projected field $P_h^{rot}(\boldsymbol{u}_h^n)$ through (2-IV.10). The projection operator P_h^{rot} is the only one we can use among the three projectors we introduced in Section 2-IV-C-a because the weak incompressibility property (2-IV.7) is not satisfied by $\boldsymbol{u}_h^n$.

The stabilization procedure used in (2-IV.15) has been studied in [KS92]. Proposition 3 holds for system (2-IV.15), its proof being similar to the case of Taylor-Hood finite element (see Sect. 2-IV-C-b), since the additional term $\sum_{j=1}^{N_E}|E_j|\int_{E_j}[\![p_h]\!][\![q]\!]$ is non negative with the test function used in the proof. All this also holds *mutatis mutandis* for discretization of the advection terms by a *discontinuous Galerkin* method, and for the log-formulation.

Let us finally consider *the mixed finite elements space* $(\mathbb{P}_1)^d\times\mathbb{P}_1$ for $(\boldsymbol{u}_h,p_h)$. If the term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized with the characteristic method, the system then writes :

$$\begin{aligned}
0=\int_{\mathcal{D}}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\mathrm{div}\,\boldsymbol{u}_h^n\Big(\boldsymbol{v}\cdot\boldsymbol{u}_h^{n+1}\Big)+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}\\
-p_h^{n+1}\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-\Big(\Big(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)\boldsymbol{\sigma}_h^{n+1}+\boldsymbol{\sigma}_h^{n+1}\Big(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\Big)^T\Big):\boldsymbol{\phi}\\
+\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}):\boldsymbol{\phi}+\sum_{k=1}^{N_K}h_{K_k}^2\int_{K_k}\nabla p_h\cdot\nabla q,\quad\text{(2-IV.16)}
\end{aligned}$$

TAB. 2.2 – Summary of some possible finite elements for $(\boldsymbol{u}_h, p_h, \boldsymbol{\sigma}_h/\boldsymbol{\psi}_h)$ when $\boldsymbol{\sigma}_h/\boldsymbol{\psi}_h \in (\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, with some possible projections for the velocity field (see Sect. 2-IV-C).

• Advection discretized by • $(\boldsymbol{u}_h, p_h)$ in …	Characteristics or DG $\Rightarrow$ equations modified
Scott-Vogelius $(\mathbb{P}_2)^d \times \mathbb{P}_{1,disc}$	(nothing)
Taylor-Hood $(\mathbb{P}_2)^d \times \mathbb{P}_1$	+ P_h^{rot} + Temam term
Crouzeix-Raviart $(\mathbb{P}_1^{CR})^d \times \mathbb{P}_0$	+ P_h^{BDM}, $P_h^{RT_0}$ or P_h^{rot} + $P_h(\boldsymbol{u}_h^n)$ for Navier term
stabilized $(\mathbb{P}_1)^d \times \mathbb{P}_1$	+ P_h^{rot} + Temam term
stabilized $(\mathbb{P}_1)^d \times \mathbb{P}_0$	+ P_h^{rot} + Temam term

with a flow X^n again computed with the projected field $P_h^{rot}(\boldsymbol{u}_h^n)$ through (2-IV.10). Again, we are led to choose the projection operator P_h^{rot} because the weak incompressibility property (2-IV.7) is not satisfied by $\boldsymbol{u}_h^n$. The stabilization procedure used in (2-IV.16) has been studied in [BP84]. Proposition 3 holds for system (2-IV.16), its proof being similar to the case of Taylor-Hood finite element (see Sect. 2-IV-C-b), since the additional term $\sum_{k=1}^{N_K} h_{K_k}^2 \int_{K_k} \nabla p_h \cdot \nabla q$ is non negative with the test function used in the proof. Again, all this also holds *mutadis mutandis* for discretization of the advection terms by a *discontinuous Galerkin* method, and for the log-formulation.

2-V Positivity, free energy estimate and the long-time issue

Notice that both Propositions 1 and 2 impose a limitation on the time step which depends on the time iteration : $0 < \Delta t < c_0$, where $c_0 \equiv c_0(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n)$ is function of a time-dependent data. Thus, these existence results are weak insofar as the long-time existence of the discrete solutions is not ensured, *i.e.* if $\sum_{n=0}^{\infty} c_0(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n) < \infty$.

Yet, for the discretizations introduced above, we have also shown that at each time step, the solutions of those discretizations satisfy free energy estimates. This will now allow us to prove the long-time existence of the discrete solutions.

Remark 5. *In this section, we concentrate for simplicity on the discretization using Scott-Vogelius finite elements for velocity-pressure, and piecewise constant approximations for the stress. However, similar results can be proven for the other discretization methods introduced in Section 2-IV-C and Appendix D, since the solutions satisfy a free energy estimate.*

Proposition 7. *For any initial condition $(\boldsymbol{u}_h^0, \boldsymbol{\sigma}_h^0)$ with $\boldsymbol{\sigma}_h^0$ symmetric positive definite, there exists a constant $c_1 \equiv c_1(\boldsymbol{u}_h^0, \boldsymbol{\sigma}_h^0) > 0$ such that, for any time step $0 \leq \Delta t < c_1$, there exists, for all iterations $n \in \mathbb{N}$, $(\boldsymbol{u}_h^{n+1}, \boldsymbol{\sigma}_h^{n+1})$ which is the unique solution to the system* (2-III.6) *(resp.* (2-III.8)*) with $\boldsymbol{\sigma}_h^{n+1}$ symmetric positive definite.*

Proof of Proposition 7. Like in the proof of Proposition 1, we will proceed with the proof for system (2-III.6) only, using its restatement as system (2-III.12).

The proof is by induction on the time index n. With the notation of the proof of Proposition 1, for a fixed $n = 0, \ldots, N_T - 1$ and for a fixed vector Y^n of values in the subset S_+^* of $\mathbb{R}^{2N_D+3N_K}$ (standing for the nodal and elementwise values of a field $(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n)$ with $\boldsymbol{\sigma}_h^n$ symmetric positive definite), we define like in the proof of Proposition 1 (using the implicit function theorem) a function $R : \Delta t \in [0, c_0) \to R(\Delta t) \in \mathbb{R}^{2N_D+3N_K}$ (where $c_0 = c_0(\boldsymbol{u}_h^n, \boldsymbol{\sigma}_h^n)$) such that :

$$\forall \Delta t \in [0, c_0),\ Q(\Delta t, R(\Delta t)) = 0,$$

where Q is defined by (2-III.15). For any $\Delta t \in [0, c_0)$, $R(\Delta t) \in \mathbb{R}^{2N_D+3N_K}$ stands for the nodal and elementwise values of a field $(\boldsymbol{u}_h(\Delta t), \boldsymbol{\sigma}_h(\Delta t))$ (with $\boldsymbol{\sigma}_h(\Delta t)$ symmetric positive definite) that is solution to the system (2-III.12).

Then, by Proposition 3, the solution $(\boldsymbol{u}_h(\Delta t),\boldsymbol{\sigma}_h(\Delta t))$ satisfies a free energy estimate :

$$F(\boldsymbol{u}_h(\Delta t),\boldsymbol{\sigma}_h(\Delta t)) \leq F(\boldsymbol{u}_h^n,\boldsymbol{\sigma}_h^n). \tag{2-V.1}$$

Using the fact that all norms are equivalent in the finite-dimensional vector space $\mathbb{R}^{2N_D+3N_K}$, and that, for $0<\nu\leq 1-\frac{1}{e}$, we have $\nu\ x\leq x-\ln(x), \forall x>0$, we obtain that there exists two constants $\alpha>0$ and $\beta>0$ (independent of Δt), such that :

$$\alpha\|R(\Delta t)\| \leq F(\boldsymbol{u}_h(\Delta t),\boldsymbol{\sigma}_h(\Delta t))+\beta. \tag{2-V.2}$$

Let us define the function D :

$$D : \Delta t\in[0,c_0) \longrightarrow B(Y^n)+A(R(\Delta t))+(\nabla_Y A)R(\Delta t)\in\mathbb{R}^{2N_D+3N_K}.$$

We recall that (see (2-III.16)), with the new notations : $\nabla_Y Q(\Delta t,R(\Delta t))=I+\Delta t D(\Delta t)$. Using (2-V.1), (2-V.2) and the fact that the discrete free energy is non-increasing, the function D satisfies :

$$\begin{aligned}\|D(\Delta t)\| &\leq \|B\|\|Y^n\|+(\|A\|+\|\nabla_Y A\|)\|R(\Delta t)\|\\ &\leq (\|B\|+\|A\|+\|\nabla_Y A\|)\frac{1}{\alpha}\left(F(\boldsymbol{u}_h^n,\boldsymbol{\sigma}_h^n)+\beta\right)\\ &\leq (\|B\|+\|A\|+\|\nabla_Y A\|)\frac{1}{\alpha}\left(F(\boldsymbol{u}_h^0,\boldsymbol{\sigma}_h^0)+\beta\right).\end{aligned}$$

This shows that there exists a constant $c_1\equiv c_1\left(\boldsymbol{u}_h^0,\boldsymbol{\sigma}_h^0\right)>0$ such that, for any time step $0\leq\Delta t<c_1$, the matrix $\nabla_Y Q(\Delta t,R(\Delta t))$ is invertible. Using the implicit function theorem, this implies that, for any time step $0\leq\Delta t<c_1$, the system (2-III.12) admits a solution $(\boldsymbol{u}_h^{n+1},\boldsymbol{\sigma}_h^{n+1})$ with $\boldsymbol{\sigma}_h^{n+1}$ symmetric positive definite at all iterations $n\in\mathbb{N}$. □

A similar result can be proven for the log-formulations (2-III.10) and (2-III.11) :

Proposition 8. *For any initial condition $(\boldsymbol{u}_h^0,\boldsymbol{\psi}_h^0)$, there exists a constant $c_1\equiv c_1\left(\boldsymbol{u}_h^0,\boldsymbol{\psi}_h^0\right)>0$ such that, for any time step $0\leq\Delta t<c_1$, there exists, for all iterations $n\in\mathbb{N}$, $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ which is the unique solution to the system* (2-III.10) *(resp.* (2-III.11)*).*

Proof of Proposition 8. The proof of Proposition 8 is similar to that of Proposition 7 using for $Q(\Delta t,Y)$ and $D(\Delta t)$ slightly modified expressions as explained for the proof of Proposition 2. The entropic term in the free energy still helps in bounding the norm of the vector of nodal-elementwise values for $(\boldsymbol{u}_h,\boldsymbol{\psi}_h)$ like in (2-V.2) using the following scalar inequality, which is true for any fixed $\nu\in(0,1]$: $\forall x\in\mathbb{R}$, $\mathrm{e}^x-x+1\geq\nu|x|$. □

From Propositions 7 and 8, we have the global-in-time existence of solutions to those discretizations of the Oldroyd-B system presented above which satisfy a discrete free energy estimate.

The log-formulation actually also satisfies the following long-time existence result, using the fact that the *a priori* estimates can be obtained without requiring the stress tensor field to be positive definite :

Proposition 9. *For any initial condition $(\boldsymbol{u}_h^0,\boldsymbol{\psi}_h^0)$, and for any constant time step $\Delta t>0$, there exists, for all iterations $n\in\mathbb{N}$, $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ which is a solution to the system* (2-III.10) *(resp.* (2-III.11)*).*

Proof of Proposition 9. We will proceed with the proof for system (2-III.10) only, using its restatement as system (2-III.14). Note already that, since the derivation of discrete free energy estimates for the system (2-III.10) does not require the solution $\boldsymbol{\psi}_h^{n+1}$ and the test function to be non-negative like in the derivation of discrete free energy estimates for the system (2-III.6), then the manipulations used to derive the free energy estimate (2-IV.5) can also be done *a priori* for any function in the finite element space.

Let us consider a fixed time index n and a given couple $(\boldsymbol{u}_h^n,\boldsymbol{\psi}_h^n)\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$. We equip the Hilbert space $(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ with the following inner product :

$$((\boldsymbol{v}_1,\boldsymbol{\phi}_1);(\boldsymbol{v}_2,\boldsymbol{\phi}_2))=\int_{\mathcal{D}}\boldsymbol{v}_1\cdot\boldsymbol{v}_2+\boldsymbol{\phi}_1:\boldsymbol{\phi}_2,$$

for all $(\boldsymbol{v}_1,\boldsymbol{\phi}_1),(\boldsymbol{v}_2,\boldsymbol{\phi}_2)\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, and denote by $\|\cdot\|$ the associated norm. Let us introduce the mapping $\mathcal{F}:(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}\to(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ defined by duality for all $(\boldsymbol{u},\boldsymbol{\psi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times$

$(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ through the form :

$$(\mathcal{F}(\boldsymbol{u},\boldsymbol{\psi});(\boldsymbol{v},\boldsymbol{\phi})) = \int_{\mathcal{D}} \mathrm{Re}\left(\frac{\boldsymbol{u}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)\cdot\boldsymbol{v}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}e^{\boldsymbol{\psi}}:\boldsymbol{\nabla}\boldsymbol{v}$$
$$+\left(\frac{\boldsymbol{\psi}-\boldsymbol{\psi}_h^n\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-(\boldsymbol{\Omega}\boldsymbol{\psi}-\boldsymbol{\psi}\boldsymbol{\Omega}):\boldsymbol{\phi}-2\boldsymbol{B}:\boldsymbol{\phi}-\frac{1}{\mathrm{Wi}}(e^{-\boldsymbol{\psi}}-\boldsymbol{I}):\boldsymbol{\phi},$$

for any test function $(\boldsymbol{v},\boldsymbol{\phi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, where we have used the decomposition of the velocity gradient $\boldsymbol{\nabla}\boldsymbol{u}$ as explained in Lemma 4 :

$$\boldsymbol{\nabla}\boldsymbol{u}=\boldsymbol{\Omega}+\boldsymbol{B}+\mathcal{N}e^{-\boldsymbol{\psi}},$$

with $\boldsymbol{\Omega}$ and $\boldsymbol{B}$ continuous with respect to $\boldsymbol{\nabla}\boldsymbol{u}$, so that $\mathcal{F}$ is a continuous mapping on finite balls of radius $\alpha>0$:

$$\mathcal{B}_\alpha=\left\{(\boldsymbol{v},\boldsymbol{\phi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}},\|(\boldsymbol{v},\boldsymbol{\phi})\|\le\alpha\right\}.$$

Note that if $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ is a solution to (2-III.14), then we have : for all $(\boldsymbol{v},\boldsymbol{\phi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$,

$$\left(\mathcal{F}\left(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1}\right);(\boldsymbol{v},\boldsymbol{\phi})\right)=0. \tag{2-V.3}$$

Let us now assume that the mapping $\mathcal{F}$ has no zero $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ satisfying (2-V.3) in the ball $\mathcal{B}_\alpha$. Then, we define the following continuous mapping from $\mathcal{B}_\alpha$ onto itself ($\mathcal{F}$ is continuous on the finite-dimensional compact, convex ball $\mathcal{B}_\alpha$) :

$$\mathcal{G}(\boldsymbol{v},\boldsymbol{\phi})=-\alpha\frac{\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})}{\|\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})\|},\forall(\boldsymbol{v},\boldsymbol{\phi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}.$$

By the Brouwer fixed point theorem, $\mathcal{G}$ has a fixed point in $\mathcal{B}_\alpha$. Let us still denote that fixed point $(\boldsymbol{v},\boldsymbol{\phi})$ for the sake of simplicity. By definition, it satisfies :

$$\mathcal{G}(\boldsymbol{v},\boldsymbol{\phi})=(\boldsymbol{v},\boldsymbol{\phi})\in\mathcal{B}_\alpha \text{ and } \|\mathcal{G}(\boldsymbol{v},\boldsymbol{\phi})\|=\alpha. \tag{2-V.4}$$

Considering $\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})$ and using $\left(\boldsymbol{v},\frac{\varepsilon}{2\mathrm{Wi}}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{I})\right)$ as a test function, we get the following inequality after similar manipulations to those in the proof of Proposition 5 :

$$\left(\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi});\left(\boldsymbol{v},\frac{\varepsilon}{2\mathrm{Wi}}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{I})\right)\right)\ge\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{v}|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi})-\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{u}_h^n|^2-\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(e^{\boldsymbol{\psi}_h^n}-\boldsymbol{\psi}_h^n)$$
$$+\int_{\mathcal{D}}\frac{\mathrm{Re}}{2}|\boldsymbol{v}-\boldsymbol{u}_h^n|^2+\Delta t\int_{\mathcal{D}}(1-\varepsilon)|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\mathrm{tr}\left(\mathrm{e}^{\boldsymbol{\phi}}+\mathrm{e}^{-\boldsymbol{\phi}}-2I\right). \tag{2-V.5}$$

Then, using the scalar inequality $\mathrm{e}^x-x\ge|x|,\forall x\in\mathbb{R}$, we have :

$$\int_{\mathcal{D}}\mathrm{tr}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi}+\boldsymbol{I})\ge\sum_{i=1}^d\int_{\mathcal{D}}|\lambda_i|,\forall\boldsymbol{\phi}\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}, \tag{2-V.6}$$

where $(\lambda_i)_{1\le i\le d}$ are functions depending on $\boldsymbol{\phi}$ such that, for all $\boldsymbol{x}\in\mathcal{D}$, $(\lambda_i(\boldsymbol{x}))_{1\le i\le d}$ are the d (non-necessarily distinct) real eigenvalues of the symmetric matrix $\boldsymbol{\phi}(\boldsymbol{x})$. Now, since $(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ is finite-dimensional, all norms are equivalent. So there exist $\gamma_1,\gamma_2>0$ such that, for all $(\boldsymbol{v},\boldsymbol{\phi})\in(\mathbb{P}_2)^d_{\mathrm{div}=0}\times(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$:

$$\gamma_1\|(\boldsymbol{v},\boldsymbol{\phi})\|\le\left(\int_{\mathcal{D}}|\boldsymbol{v}|^2\right)^{\frac{1}{2}}+\|\max_{1\le i\le d}|\lambda_i(\boldsymbol{x})|\|_\infty\le\gamma_2\|(\boldsymbol{v},\boldsymbol{\phi})\|, \tag{2-V.7}$$

where it is easy to prove that $\|\max_{1\le i\le d}|\lambda_i(\boldsymbol{x})|\|_\infty$ defines a norm in the vector space $L^\infty\left(\mathcal{D},\mathcal{S}(\mathbb{R}^{d\times d})\right)$. Using the equation (2-V.6) with the norm equivalence (2-V.7), we obtain :

$$\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{v}|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi}+\boldsymbol{I})$$
$$\ge\min\left(\frac{\mathrm{Re}}{2},\frac{\varepsilon}{2\mathrm{Wi}}\frac{1}{\|\max_{1\le i\le d}|\lambda_i(\boldsymbol{x})|\|_\infty}\right)\left(\int_{\mathcal{D}}|\boldsymbol{v}|^2+\|\max_{1\le i\le d}|\lambda_i(\boldsymbol{x})|\|_\infty\sum_{i=1}^d\int_{\mathcal{D}}|\lambda_i|\right)$$
$$\ge\min\left(\frac{\mathrm{Re}}{2},\frac{\varepsilon}{2\mathrm{Wi}}\frac{1}{\|\max_{1\le i\le d}|\lambda_i(\boldsymbol{x})|\|_\infty}\right)\left(\int_{\mathcal{D}}|\boldsymbol{v}|^2+\sum_{i=1}^d\int_{\mathcal{D}}|\lambda_i|^2\right).$$

Last, since the fixed-point $(\boldsymbol{v},\boldsymbol{\phi})\in\mathcal{B}_\alpha$ satisfies $\|(\boldsymbol{v},\boldsymbol{\phi})\|=\alpha$ because of (2-V.4), we can choose α large enough so that :

$$\min\left(\frac{\mathrm{Re}}{2},\frac{\varepsilon}{2\mathrm{Wi}\gamma_2\alpha}\right)\|(\boldsymbol{v},\boldsymbol{\phi})\|^2>\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{u}_h^n|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(e^{\boldsymbol{\psi}_h^n}-\boldsymbol{\psi}_h^n+\boldsymbol{I}),$$

and we get :

$$\begin{aligned}\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{v}|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi}+\boldsymbol{I})-\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}|\boldsymbol{u}_h^n|^2-\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(e^{\boldsymbol{\psi}_h^n}-\boldsymbol{\psi}_h^n+\boldsymbol{I})\\+\int_{\mathcal{D}}\frac{\mathrm{Re}}{2}|\boldsymbol{v}-\boldsymbol{u}_h^n|^2+\Delta t\int_{\mathcal{D}}(1-\varepsilon)|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\mathrm{tr}\left(\mathrm{e}^{\boldsymbol{\phi}}+\mathrm{e}^{-\boldsymbol{\phi}}-2\boldsymbol{I}\right)>0,\end{aligned}$$

that is :

$$\left(\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi});\left(\boldsymbol{v},\frac{\varepsilon}{2\mathrm{Wi}}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{I})\right)\right)>0. \tag{2-V.8}$$

Now, using the equation (2-V.4) we have :

$$\left(\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi});\left(\boldsymbol{v},\frac{\varepsilon}{2\mathrm{Wi}}(\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{I})\right)\right)=-\frac{\|\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})\|}{\alpha}\left(\int_{\mathcal{D}}|\boldsymbol{v}|^2+\frac{\varepsilon}{2\mathrm{Wi}}\mathrm{tr}\left(\boldsymbol{\phi}\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi}\right)\right)\leq 0 \tag{2-V.9}$$

which is obviously in contradiction with (2-V.8) since, for all $\boldsymbol{\phi}\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, we have $\mathrm{tr}(\boldsymbol{\phi}\mathrm{e}^{\boldsymbol{\phi}}-\boldsymbol{\phi})\geq 0$ by virtue of the scalar inequality $x(\mathrm{e}^x-1)\geq 0,\forall x\in\mathbb{R}$.

Thus, for any $\Delta t>0$, if we choose α sufficiently large, the mapping $\mathcal{F}$ has a zero $(\boldsymbol{u}_h^{n+1},\boldsymbol{\psi}_h^{n+1})$ satisfying (2-V.3) in the ball $\mathcal{B}_\alpha$, which concludes the proof. □

Notice that Proposition 9 does not ensure the uniqueness of solutions. There may be bifurcations, hence many possible solutions to the log-formulation, in the case where the CFL condition is not fulfilled. Though, all those solutions will satisfy a free energy estimate, which is not the case for the usual formulation in terms of $\boldsymbol{\tau}$. The fact that we are able to prove such a stability result without any assumption on the timestep for the log-formulation, and not for the classical formulation, may be related to the fact that discretizations of the log-formulation have been reported to yield solutions beyond the limiting Weissenberg number for standard discretizations (see [HFK05]).

Remark 6 (other positivity preserving schemes)**.** *There exist other means than using the log-formulation to preserve the non-negativity of the conformation tensor. A very natural way of preserving the non-negativity is to reformulate the constitutive equation with the deformation gradient instead of the stress or the conformation tensor, using a Lie-derivative like in [LX06]. It is also possible to build free-energy-dissipative schemes for a Lie-formulation, as shown in Appendix E. But discretizations of a Lie-formulation seem to necessitate the numerical integration of ordinary differential equations like* (2-III.7) *for the characteristic flow, which may introduce new instabilities (see Rem. 2).*

2-VI Appendix to the Chapter 2

2-VI-A Appendix A. Some properties of symmetric positive definite matrices

2-VI-A-a Proof of Lemma 1

Formula (2-I.6), (2-I.7) and (2-I.8) are simply obtained by diagonalizing the symmetric positive definite matrix $\boldsymbol{\sigma}$, and using the inequalities : $\forall x,y>0$, $\ln(xy)=\ln x+\ln y$, $x-1\geq\ln x$ and $x+1/x\geq 2$.

Let us now prove formula (2-I.9). By diagonalization, we have $\boldsymbol{\sigma}=\Omega^T D\Omega$ with Ω orthogonal and D diagonal positive, which gives :

$$\mathrm{tr}(\boldsymbol{\sigma}\boldsymbol{\tau}^{-1})=\mathrm{tr}(\Omega^T\sqrt{D}\sqrt{D}\Omega\boldsymbol{\tau}^{-1})=\mathrm{tr}(\sqrt{D}\Omega\boldsymbol{\tau}^{-1}\Omega^T\sqrt{D})\geq 0,$$

because $A=\sqrt{D}\Omega\boldsymbol{\tau}^{-1}\Omega^T\sqrt{D}$ is clearly a symmetric positive definite matrix. Likewise, we have :

$$\det(\boldsymbol{\sigma}\boldsymbol{\tau}^{-1})=\det(\Omega^T D\Omega\boldsymbol{\tau}^{-1})=\det(\sqrt{D}\Omega\boldsymbol{\tau}^{-1}\Omega^T\sqrt{D}).$$

The proof of (2-I.10) is then equivalent to show :

$$\ln(\det(A))\leq\mathrm{tr}(A-\boldsymbol{I}),$$

for any symmetric positive definite matrix A, which simply derives from (2-I.6) and (2-I.7).

It remains to prove formula (2-I.11). By diagonalization, we write $\boldsymbol{\sigma}=O^T DO$ and $\boldsymbol{\tau}=R^T\Lambda R$ with O and R orthogonal, and D and Λ diagonal positive. Let us introduce the orthogonal matrix $\Omega=OR^T$. We denote by D_i (resp. Λ_i) the (i,i)-th entry of D (resp. of Λ). We have :

$$\begin{aligned}\mathrm{tr}\left((\ln\boldsymbol{\sigma}-\ln\boldsymbol{\tau})\boldsymbol{\sigma}-(\boldsymbol{\sigma}-\boldsymbol{\tau})\right)&=\sum_i D_i\ln D_i-D_i+\Lambda_i-\sum_{i,j}(\Omega_{ij})^2 D_i\ln\Lambda_j\\&=\sum_i\left(\Lambda_i-D_i-\sum_j(\Omega_{ij})^2 D_i(\ln\Lambda_j-\ln D_i)\right),\end{aligned}$$

since Ω is an orthogonal matrix ($\sum_j(\Omega_{ij})^2=1$ for all i). Using the convexity inequality $x-y\le x(\ln x-\ln y)$ for all $x,y>0$, we thus obtain $\mathrm{tr}\left((\ln\boldsymbol{\sigma}-\ln\boldsymbol{\tau})\boldsymbol{\sigma}-(\boldsymbol{\sigma}-\boldsymbol{\tau})\right)\ge 0$ which concludes the proof of (2-I.11).

2-VI-A-b Proof of Lemma 2

First, since $\boldsymbol{\sigma}\in\left(C^1([0,T))\right)^{\frac{d(d+1)}{2}}$ is symmetric positive definite, $\det(\boldsymbol{\sigma})$ is positive and $C^1([0,T))$. So we immediately get the classical Jacobi formula (2-I.12) :

$$\frac{d}{dt}\ln(\det(\boldsymbol{\sigma}))=(1/\det(\boldsymbol{\sigma}))\frac{d}{dt}\det(\boldsymbol{\sigma})=\mathrm{tr}\Big(\boldsymbol{\sigma}^{-1}\frac{d}{dt}\boldsymbol{\sigma}\Big),$$

on noting that $\ln(\det(\boldsymbol{\sigma}))=\mathrm{tr}(\ln(\boldsymbol{\sigma}))$.

Then, for the proof of (2-I.13), first note that the matrix exponential is a C^∞-diffeomorphism from the set of symmetric matrices onto the set of symmetric positive definite matrices by virtue of the local inversion theorem (see [MT86], Cor. 3.8.5, for instance), whose inverse mapping coincides with the matrix logarithm defined in (2-I.5). Then, there exists $\boldsymbol{\tau}\in\left(C^1([0,T))\right)^{\frac{d(d+1)}{2}}$ such that $\boldsymbol{\sigma}=\mathrm{e}^{\boldsymbol{\tau}}$, and on noting that $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ commute, we immediately get (2-I.13) :

$$\mathrm{tr}\left(\boldsymbol{\sigma}\frac{\mathrm{d}\ln\boldsymbol{\sigma}}{\mathrm{d}t}\right)=\mathrm{tr}\left(\mathrm{e}^{\boldsymbol{\tau}}\frac{\mathrm{d}\boldsymbol{\tau}}{\mathrm{d}t}\right)=\mathrm{tr}\left(\frac{\mathrm{de}^{\boldsymbol{\tau}}}{\mathrm{d}t}\right)=\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}\boldsymbol{\sigma}.$$

2-VI-B Appendix B. Proof of Lemma 3

Let us introduce $t_0=\inf\{t>0|\boldsymbol{\sigma}(t)\text{ is not symmetric positive definite}\}$, with convention $t_0=\infty$ if $\{t>0,\boldsymbol{\sigma}\text{ is not symmetric positive definite}\}=\emptyset$. Since $\boldsymbol{\sigma}(t=0)$ is symmetric positive definite, it remains so at least for small times $0\le t<\Delta t$, by continuity of $\det(\boldsymbol{\sigma})$ with respect to the time variable t. Thus, $t_0\ge\Delta t>0$.

Let us assume that $t_0<\infty$. First, one can define the logarithm $\ln\boldsymbol{\sigma}$ of $\boldsymbol{\sigma}$, which satisfies the equation for $\boldsymbol{\psi}$ in system (2-II.10) for $t\in[0,t_0)$. Taking the trace of the equation for $\boldsymbol{\psi}$ in system (2-II.10), we get for $\ln\boldsymbol{\sigma}$:

$$\frac{\mathrm{D}}{\mathrm{D}t}\ln\det\boldsymbol{\sigma}=\frac{1}{\mathrm{Wi}}\mathrm{tr}(\boldsymbol{\sigma}^{-1}-\boldsymbol{I}),\tag{2-VI.1}$$

where we have introduced the convective derivative $\frac{\mathrm{D}}{\mathrm{D}t}=\left(\frac{d}{dt}+(\boldsymbol{u}\cdot\boldsymbol{\nabla})\right)$ (the next formulae (2-VI.3) and (2-VI.4) thus hold along the characteristics, which are well defined because $\boldsymbol{u}\in C^1\left([0,T),C^{0,1}(\mathcal{D})\right)$). Besides, for any positive definite matrix $\boldsymbol{\sigma}^{-1}$, we have :

$$\frac{\mathrm{tr}(\boldsymbol{\sigma}^{-1})}{d}\ge(\det\boldsymbol{\sigma}^{-1})^{1/d},\tag{2-VI.2}$$

which follows from the convex inequality between geometrical and arithmetical means. Thus, combining (2-VI.1) and (2-VI.2), we get, on the time interval $[0,t_0)$:

$$\frac{\mathrm{D}}{\mathrm{D}t}(\det\boldsymbol{\sigma})^{1/d}=\frac{1}{d}(\det\boldsymbol{\sigma})^{1/d}\frac{\mathrm{D}}{\mathrm{D}t}\ln\det\boldsymbol{\sigma}\ge\frac{1}{\mathrm{Wi}}\left(1-(\det\boldsymbol{\sigma})^{1/d}\right).\tag{2-VI.3}$$

Now, by continuity of $\det(\boldsymbol{\sigma})$ with respect to t, one eigenvalue at least converges to zero as $t\to t_0^-$, which implies $\det\boldsymbol{\sigma}\to 0^+$. Then, there exists $\eta>0$ such that, for times $t_0-\eta<t<t_0$, we have :

$$0<\det\boldsymbol{\sigma}<1,$$

and by (2-VI.3) :

$$\frac{\mathrm{D}}{\mathrm{D}t}(\det\boldsymbol{\sigma})^{1/d}>0.\tag{2-VI.4}$$

But then, t_0 cannot be the first time when $\det\boldsymbol{\sigma}=0$, otherwise one should have $\frac{\mathrm{D}}{\mathrm{D}t}(\det\boldsymbol{\sigma})^{1/d}(t_0^-)\le 0$, which contradicts (2-VI.4). Thus $t_0=\infty$ which ends the proof of Lemma 3.

2-VI-C Appendix C. Proof of Lemmas 4 and 5

Lemmas 4 and 5 are consequences of the following result, which is a slight modification of a result proved in [FK04].

Lemma 7. *Let $\boldsymbol{M}$ be a $d\times d$ matrix and $\boldsymbol{\sigma}$ be a symmetric positive definite $d\times d$ matrix. Then, there exists three $d\times d$ matrices $\boldsymbol{\Omega}$, $\boldsymbol{B}$ and $\boldsymbol{N}$ such that*

$$\boldsymbol{M}=\boldsymbol{\Omega}+\boldsymbol{B}+\boldsymbol{N}\boldsymbol{\sigma}^{-1}$$

and $\boldsymbol{B}$ is a symmetric matrix which commutes with $\boldsymbol{\sigma}$, $\boldsymbol{\Omega}$ and $\boldsymbol{N}$ are antisymmetric. Moreover, the entries of $\boldsymbol{\Omega}$, $\boldsymbol{B}$ and $\boldsymbol{N}$ are linear with respect to the entries of $\boldsymbol{M}$.

Proof. First, it is easy to check by diagonalization that it is sufficient to prove the result for a diagonal matrix $\boldsymbol{\sigma}$ (more precisely, by rewriting everything in a diagonalizing basis for $\boldsymbol{\sigma}$). In the following, we thus assume that $\boldsymbol{\sigma}=\mathrm{diag}(\Lambda_1,\dots,\Lambda_d)$, where $(\Lambda_i)_{1\le i\le d}$ are positive numbers. Moreover, we restrict ourselves to the physical case $d=3$, but the arguments can be generalized to any dimension.
Let us first consider the case $\Lambda_i\neq\Lambda_j$ for $i\neq j$. In this case, we set :

- $B_{i,i}=M_{i,i}$ and $B_{i,j}=0$ for $i\neq j$;
- $N_{i,i}=0$ and $N_{i,j}=\frac{M_{i,j}+M_{j,i}}{\Lambda_j^{-1}-\Lambda_i^{-1}}$ for $i\neq j$;
- $\Omega_{i,i}=0$ and $\Omega_{i,j}=-\frac{M_{i,j}\Lambda_i^{-1}+M_{j,i}\Lambda_j^{-1}}{\Lambda_j^{-1}-\Lambda_i^{-1}}$ for $i\neq j$.

It is easy to check that these matrices satisfy the requirements of the lemma.

Let us now consider the case $\Lambda_1=\Lambda_2=\Lambda_3$. In this case, we simply set $\boldsymbol{N}=0$, $\boldsymbol{B}=\frac{\boldsymbol{M}+\boldsymbol{M}^T}{2}$ and $\boldsymbol{\Omega}=\frac{\boldsymbol{M}-\boldsymbol{M}^T}{2}$. It is again straightforward to check that these matrices satisfy the requirements of the lemma.

Finally, let us consider the case when only two Λ_i's are equal. Without loss of generality, we can suppose $\Lambda_1=\Lambda_2\neq\Lambda_3$. In this case, we set :

- $B_{3,3}=M_{3,3}$, $B_{i,j}=\frac{M_{i,j}+M_{j,i}}{2}$ for $1\le i,j\le 2$ and $B_{i,j}=0$ otherwise ;
- $N_{i,j}=0$ for $1\le i,j\le 2$, $N_{3,3}=0$ and $N_{i,j}=\frac{M_{i,j}+M_{j,i}}{\Lambda_j^{-1}-\Lambda_i^{-1}}$ otherwise ;
- $\Omega_{i,j}=\frac{M_{i,j}-M_{j,i}}{2}$ for $1\le i,j\le 2$, $\Omega_{3,3}=0$ and $\Omega_{i,j}=-\frac{M_{i,j}\Lambda_i^{-1}+M_{j,i}\Lambda_j^{-1}}{\Lambda_j^{-1}-\Lambda_i^{-1}}$ otherwise.

This case is a combination of the two previous cases, and one can check that these matrices satisfy the requirements of the lemma. □

Notice in particular that the linear dependence of the entries of the matrices $\boldsymbol{\Omega}$, $\boldsymbol{B}$ and $\boldsymbol{N}$ with respect to the entries of $\boldsymbol{M}$ implies that if $\boldsymbol{\sigma}$ is piecewise constant (with respect to the space variable) and $\boldsymbol{M}$ is $(\mathbb{P}_{k,\mathrm{disc}})^{d\times d}$, then $\boldsymbol{\Omega}$, $\boldsymbol{B}$ and $\boldsymbol{N}$ are also $(\mathbb{P}_{k,\mathrm{disc}})^{d\times d}$ (which is the result of Lem. 5).

2-VI-D Appendix D. Higher order discretization of the stress fields $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$

We now show how to build numerical schemes with higher order discretization spaces for the stress that still satisfy a discrete free energy estimate. We typically have in mind piecewise linear spaces for $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$.

From the previous proofs establishing discrete free energy estimates at low order in $\mathbb{P}_0$, it is clear that we need to use nonlinear functionals of $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$ as test functions, namely $\boldsymbol{\sigma}_h^{-1}$ and $e^{\boldsymbol{\psi}_h}$. Finite element spaces other than $\mathbb{P}_0$ are typically not invariant under such nonlinear functionals, and this brings us to introduce projections of these nonlinear terms on $\mathbb{P}_0$, and finite element spaces to discretize the stress that contain $\mathbb{P}_0$, thus discontinuous.

We will use a $\mathbb{P}_0$-Lagrange interpolation operator π_h which is convenient because it commutes with nonlinear functionals (see Lem. 9 below). Moreover, we will need that this interpolation operator coincides with an L^2 orthogonal projection onto $\mathbb{P}_0$ (see Lem. 8 below). The need for π_h to coincide with an L^2 orthogonal projection onto $\mathbb{P}_0$ limits the maximum regularity of the discretization of the stress, essentially to piecewise $\mathbb{P}_1$ finite elements. Therefore, we consider $\boldsymbol{\sigma}_h$ and $\boldsymbol{\psi}_h$ in either of the following finite element spaces[2] :

$$(\mathbb{P}_1+\mathbb{P}_0)^{\frac{d(d+1)}{2}} \text{ or } (\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}.$$

In Section 2-VI-D-a, we introduce the interpolation operator π_h. Then we prove that, for a Scott-Vogelius discretization of the velocity-pressure field, a free energy estimate can be obtained for discretization schemes

[2]Note that, clearly, $(\mathbb{P}_1+\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ is only a subspace of $(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$.

close to those considered in Section 2-IV, when $\boldsymbol{\sigma}_h$ (respectively $\boldsymbol{\psi}_h$) is in $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$. This is the purpose of the Section 2-VI-D-b (respectively Sect. 2-VI-D-c). Finally, we show in Section 2-VI-D-d how these results can be extended to other finite element discretizations of the velocity-pressure field.

Remark 7. *In this appendix, we concentrate on establishing free energy estimates, and do not prove existence results as those stated in Sections 2-III-E and 2-V. It is easy to extend these existence results to the numerical schemes considered here.*

2-VI-D-a The interpolation operator π_h

Let us introduce the projection operator π_h as the $\mathbb{P}_0$ Lagrange interpolation at barycenter θ_{K_k} for each $K_k \in \mathcal{T}_h$.

Definition 1. *For $k=1,\dots,N_K$, we denote by θ_{K_k} the barycenter of the triangle K_k. For any $\boldsymbol{\phi}$ such that $\forall k=1,\dots,N_K$, $\boldsymbol{\phi}(\theta_{K_k})$ is well-defined (for example ϕ is a tensor-valued function, continuous at points θ_{K_k}), we define its piecewise constant interpolation by :*

$$\forall k=1,\dots,N_K,\, \pi_h(\boldsymbol{\phi})|_{K_k} = \boldsymbol{\phi}(\theta_{K_k}).$$

Notice that this definition also makes sense for the case in which $\boldsymbol{\phi}$ is matrix-valued. And this interpolation operator π_h coincides with the L^2 orthogonal projection from $(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$ onto $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$:

Lemma 8. *Let π_h be the interpolation operator introduced in Definition 1. Then, for any $\boldsymbol{\phi}_h \in (\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$, we have :*

$$\int_{\mathcal{D}} \boldsymbol{\phi}_h : \tilde{\boldsymbol{\phi}}_h = \int_{\mathcal{D}} \pi_h(\boldsymbol{\phi}_h) : \tilde{\boldsymbol{\phi}}_h,\ \forall \tilde{\boldsymbol{\phi}}_h \in (\mathbb{P}_0)^{\frac{d(d+1)}{2}}.$$

Proof. It is enough to prove Lemma 8 on each simplex $K_k \in \mathcal{T}_h$ and in the scalar case. Let $(x_i)_{1\le i\le 3}$ be the vertices of the simplex K_k and $(\psi_i)_{1\le i\le 3}$ the corresponding (linear) basis functions in $\mathbb{P}_1$. Then, the function $\boldsymbol{\phi}_h|_{K_k} \in \mathbb{P}_1$ reads $\boldsymbol{\phi}_h|_{K_k}(x) = \boldsymbol{\phi}_h(x_1)\psi_1(x) + \boldsymbol{\phi}_h(x_2)\psi_2(x) + \boldsymbol{\phi}_h(x_3)\psi_3(x)$, $\forall x \in K_k$. For every $\tilde{\boldsymbol{\phi}}_h \in \mathbb{P}_0$,

$$\int_{K_k} \boldsymbol{\phi}_h \tilde{\boldsymbol{\phi}}_h = \tilde{\boldsymbol{\phi}}_h \left(\int_{K_k} \boldsymbol{\phi}_h \right) = \tilde{\boldsymbol{\phi}}_h \frac{|K_k|}{3} \left(\boldsymbol{\phi}_h(x_1) + \boldsymbol{\phi}_h(x_2) + \boldsymbol{\phi}_h(x_3) \right)$$

because $\int_{K_k} \psi_i = \frac{|K_k|}{3}$. Moreover, $\boldsymbol{\phi}_h|_{K_k} \in \mathbb{P}_1$, hence

$$\frac{1}{3}\left(\boldsymbol{\phi}_h(x_1) + \boldsymbol{\phi}_h(x_2) + \boldsymbol{\phi}_h(x_3) \right) = \boldsymbol{\phi}_h\left(\frac{x_1+x_2+x_3}{3} \right) = \boldsymbol{\phi}_h(\theta_{K_k})$$

which means

$$\int_{K_k} \boldsymbol{\phi}_h \tilde{\boldsymbol{\phi}}_h = \int_{K_k} \tilde{\boldsymbol{\phi}}_h \boldsymbol{\phi}_h(\theta_{K_k}) = \int_{K_k} \pi_h(\boldsymbol{\phi}) \tilde{\boldsymbol{\phi}}_h.$$ □

In addition, the following property holds, which is important in the choice of this particular interpolation :

Lemma 9. *Let π_h be the interpolation operator introduced in Definition 1. The interpolation operator π_h commutes with any function f : for any functions f and $\boldsymbol{\phi}_h$ such that $\boldsymbol{\phi}_h$ and $f(\boldsymbol{\phi}_h)$ are well-defined at the barycenters θ_k,*

$$\pi_h(f(\boldsymbol{\phi}_h)) = f(\pi_h(\boldsymbol{\phi}_h)).$$

The proof of Lemma 9 is straightforward since, by Definition 1, the interpolation π_h only uses specific values at fixed points in the spatial domain $\mathcal{D}$.

2-VI-D-b Free energy estimates with discontinuous piecewise linear $\boldsymbol{\sigma}_h$

In this section, we consider the following finite element discretization : Scott-Vogelius $(\mathbb{P}_2)^d \times \mathbb{P}_{1,disc}$ for $(\boldsymbol{u}_h, p_h)$ and $(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$ or $(\mathbb{P}_1 + \mathbb{P}_0)^{\frac{d(d+1)}{2}}$ for $\boldsymbol{\sigma}_h$.

2-VI-D-b.1 The characteristic method If the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized by the *characteristic* method, the system writes :

$$0=\int_{\mathcal{D}}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}$$
$$+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{v}$$
$$+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\pi_h\left(\boldsymbol{\sigma}_h^n\right)\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)+\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1})^T):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}):\boldsymbol{\phi},\quad\text{(2-VI.5)}$$

where X^n is defined as in (2-III.7). Notice that we have used the projection operator π_h in four terms. It will become clearer from the proof of the free energy estimate below why those projections are needed.

Proposition 10. *Let $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\sigma}_h^n)_{0\le n\le N_T}$ be a solution to* (2-VI.5)*, such that $\pi_h(\boldsymbol{\sigma}_h^n)$ is positive definite. Then, the free energy of the solution $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\sigma}_h^n)$:*

$$F_h^n=F(\boldsymbol{u}_h^n,\pi_h(\boldsymbol{\sigma}_h^n))=\frac{Re}{2}\int_{\mathcal{D}}|\boldsymbol{u}_h^n|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}\left(\pi_h(\boldsymbol{\sigma}_h^n)-\ln\pi_h(\boldsymbol{\sigma}_h^n)-\boldsymbol{I}\right),\quad\text{(2-VI.6)}$$

satisfies :

$$F_h^{n+1}-F_h^n+\int_{\mathcal{D}}\frac{Re}{2}\left|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\right|^2+\Delta t\int_{\mathcal{D}}(1-\varepsilon)\left|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)+\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}-2\boldsymbol{I}\right)\le 0.\quad\text{(2-VI.7)}$$

In particular, the sequence $(F_h^n)_{0\le n\le N_T}$ is non-increasing.

Remark 8. *The ensemble of symmetric positive definite matrices is convex. This implies that a piecewise linear tensor field is symmetric positive definite as soon as it is symmetric positive definite at the nodes of the mesh. Moreover, this also implies that $\pi_h(\boldsymbol{\sigma}_h)$ is symmetric positive definite as soon as $\boldsymbol{\sigma}_h$ is a piecewise linear (possibly discontinuous) symmetric positive definite tensor field.*

Proof of Proposition 10. The test functions we choose are $\left(\boldsymbol{u}_h^{n+1},p_h^{n+1},\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I}-\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1})\right)$. Recall that by Lemma 9, $\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right)^{-1}=\pi_h\left((\boldsymbol{\sigma}_h^{n+1})^{-1}\right)$. The proof is similar to the one of Proposition 3 except in the treatment of the constitutive equation. The upper-convective term in the tensor derivative writes (using Lem. 8 and the incompressibility property (2-III.3)) :

$$\begin{aligned}\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):(\boldsymbol{I}-\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1})&=\int_{\mathcal{D}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\\&=\int_{\mathcal{D}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{I}\\&=\int_{\mathcal{D}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\operatorname{div}\boldsymbol{u}_h^{n+1}\\&=\int_{\mathcal{D}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1},\end{aligned}$$

which vanishes after combination with the extra-stress term in the momentum equation.

The last term rewrites (using again Lem. 8) :

$$\int_{\mathcal{D}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\left(\boldsymbol{I}-\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right)=\int_{\mathcal{D}}\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)+\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}-2\boldsymbol{I}\right).$$

The remaining term writes (using Lem. 8, Eq. (2-I.10) with $\boldsymbol{\sigma}=\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)$ and $\boldsymbol{\tau}=\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)$, and the fact that the Jacobian of X^n remains equal to one due to the incompressibility property (2-III.3)) :

$$\begin{aligned}\int_{\mathcal{D}}\left(\boldsymbol{\sigma}_h^{n+1}-\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)\right):\left(\boldsymbol{I}-\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right)&=\int_{\mathcal{D}}\mathrm{tr}\boldsymbol{\sigma}_h^{n+1}-\mathrm{tr}\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)\\&+\mathrm{tr}\left(\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}-\boldsymbol{I}\right)\\&\ge\int_{\mathcal{D}}\mathrm{tr}\boldsymbol{\sigma}_h^{n+1}-\mathrm{tr}\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)+\mathrm{tr}\ln\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)-\mathrm{tr}\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\\&=\int_{\mathcal{D}}\mathrm{tr}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)-\mathrm{tr}\pi_h(\boldsymbol{\sigma}_h^n)+\mathrm{tr}\ln\pi_h(\boldsymbol{\sigma}_h^n)-\mathrm{tr}\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right).\end{aligned}$$

This completes the proof. □

2-VI-D-b.2 The discontinuous Galerkin method If the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized by the *discontinuous Galerkin* method, the system writes :

$$\begin{aligned}0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}\\ +(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\nabla}\boldsymbol{v}\\ +\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n}{\Delta t}\right):\boldsymbol{\phi}-\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)+\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)^T\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}):\boldsymbol{\phi}\\ +\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\left[\!\left[\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right]\!\right]:\boldsymbol{\phi}^+. \quad \text{(2-VI.8)}\end{aligned}$$

As for the characteristic method, the projection operator π_h is used in four terms. Besides, like in the case where $\boldsymbol{\sigma}_h\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized using a jump term only. Indeed, in order to derive discrete free energy estimates, we treat the discrete advection term using the projection $\pi_h(\boldsymbol{\sigma}_h)\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ of the stress field $\boldsymbol{\sigma}_h$, the derivative of which is zero.

Proposition 10 still holds for the system (2-VI.8). The proof is straightforward using all the arguments of the previous sections, except for the treatment of the discrete advection term for $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$. Using equations (2-I.10), (2-III.5), the fact that $\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ and the weak incompressibility property (2-IV.7), we have :

$$\begin{aligned}&\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\left[\!\left[\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right]\!\right]\left(\boldsymbol{I}-\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)^{-1}\right)^+=\\ &\quad\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\left[\!\left[\mathrm{tr}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right]\!\right]+|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\,\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1,-}\right)\pi_h\left(\boldsymbol{\sigma}_h^{n+1,+}\right)^{-1}-\boldsymbol{I}\right)\\ &\quad\geq\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\left[\!\left[\mathrm{tr}\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right]\!\right]+|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\,\mathrm{tr}\left(\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1,-}\right)-\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1,+}\right)\right).\end{aligned}$$

Now, the right-hand-side vanishes since it is equal to

$$\begin{aligned}\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\left[\!\left[\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)-\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right)\right]\!\right]&=\sum_{k=1}^{N_K}-\int_{\partial K_k}(\boldsymbol{u}_h^n\cdot\boldsymbol{n}_{K_k})\,\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)-\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right)\\ &=\sum_{k=1}^{N_K}-\mathrm{tr}\left(\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)-\ln\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)\right)\Big|_{K_k}\int_{K_k}\operatorname{div}(\boldsymbol{u}_h^n)=0.\end{aligned}$$

2-VI-D-c Free energy estimates with discontinuous piecewise linear $\boldsymbol{\psi}_h$

In the following section, we write free-energy-dissipative schemes using the log-formulation with $\boldsymbol{\psi}_h$ piecewise linear. For this, we again need the projection operator π_h introduced in Definition 1. We consider the Scott-Vogelius finite element space for $(\boldsymbol{u}_h,p_h)$ and the following decomposition of the velocity gradient $\boldsymbol{\nabla}\boldsymbol{u}_h\in(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$:

$$\boldsymbol{\nabla}\boldsymbol{u}_h=\boldsymbol{\Omega}_h+\boldsymbol{B}_h+\mathcal{N}_h\pi_h\left(e^{\boldsymbol{\psi}_h}\right)^{-1}. \quad \text{(2-VI.9)}$$

Notice that since $\pi_h\left(e^{\boldsymbol{\psi}_h}\right)^{-1}=\mathrm{e}^{-\pi_h(\boldsymbol{\psi}_h)}$ is in $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$, we have $\boldsymbol{\Omega}_h,\boldsymbol{B}_h,\mathcal{N}_h\in(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$ by virtue of Lemma 5 with $k=1$.

2-VI-D-c.1 The characteristic method If the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized by the *characteristic* method, the system writes :

$$\begin{aligned}0=\int_{\mathcal{D}}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right):\boldsymbol{\nabla}\boldsymbol{v}\\+\left(\frac{\boldsymbol{\psi}_h^{n+1}-\pi_h(\boldsymbol{\psi}_h^n)\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}-\left(\boldsymbol{\Omega}_h^{n+1}\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\boldsymbol{\Omega}_h^{n+1}\right):\boldsymbol{\phi}\\-2\boldsymbol{B}_h^{n+1}:\boldsymbol{\phi}-\frac{1}{\mathrm{Wi}}\left(\pi_h\left(e^{-\boldsymbol{\psi}_h^{n+1}}\right)-\boldsymbol{I}\right):\boldsymbol{\phi}.\quad\text{(2-VI.10)}\end{aligned}$$

In the system above, we have used the projection operator π_h to treat the same terms as in the system (2-VI.5). But in addition to these, we have also used the projection operator for the exponential term $e^{-\boldsymbol{\psi}_h^{n+1}}$ in the Oldroyd-B equation.

Proposition 11. *Let $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\psi}_h^n)_{0\le n\le N_T}$ be a solution to* (2-VI.10)*. Then, the free energy of the solution $(\boldsymbol{u}_h^n,p_h^n,\boldsymbol{\psi}_h^n)$:*

$$F_h^n=F\left(\boldsymbol{u}_h^n,\mathrm{e}^{\pi_h(\boldsymbol{\psi}_h^n)}\right)=\frac{Re}{2}\int_{\mathcal{D}}|\boldsymbol{u}_h|^2+\frac{\varepsilon}{2\,Wi}\int_{\mathcal{D}}\mathrm{tr}\left(\mathrm{e}^{\pi_h(\boldsymbol{\psi}_h^n)}-\pi_h(\boldsymbol{\psi}_h^n)-\boldsymbol{I}\right),\qquad\text{(2-VI.11)}$$

satisfies :

$$F_h^{n+1}-F_h^n+\int_{\mathcal{D}}\frac{Re}{2}|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n|^2+\Delta t\int_{\mathcal{D}}(1-\varepsilon)|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}|^2+\frac{\varepsilon}{2\,Wi^2}\mathrm{tr}\left(\mathrm{e}^{\pi_h(\boldsymbol{\psi}_h^n)}+\mathrm{e}^{-\pi_h(\boldsymbol{\psi}_h^n)}-2\boldsymbol{I}\right)\le 0.\qquad\text{(2-VI.12)}$$

In particular, the sequence $(F_h^n)_{0\le n\le N_T}$ is non-increasing.

Proof of Proposition 11. The proof is similar to that of Proposition 5 except for the terms using the interpolation operator π_h. We shall use as test functions $\left(\boldsymbol{u}_h^{n+1},p_h^{n+1},\frac{\varepsilon}{2\mathrm{Wi}}\left(\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)-\boldsymbol{I}\right)\right)$ in (2-III.10). Also, we will make use of the following property throughout the proof (see Lem. 9) : $\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)=\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}$.

For the material derivative of $\boldsymbol{\psi}_h$, using Lemma 8, equation (2-I.11) with $\boldsymbol{\sigma}=e^{\boldsymbol{\psi}_h^{n+1}}$ and $\boldsymbol{\tau}=\mathrm{e}^{\boldsymbol{\psi}_h^n\circ X^n(t^n)}$, and the fact that the Jacobian of the flow X^n is one for divergence-free velocity field $\boldsymbol{u}_h^n$, we have :

$$\begin{aligned}\int_{\mathcal{D}}\left(\boldsymbol{\psi}_h^{n+1}-\pi_h(\boldsymbol{\psi}_h^n)\circ X^n(t^n)\right):\left(\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)-\boldsymbol{I}\right)&=\int_{\mathcal{D}}\left(\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h(\boldsymbol{\psi}_h^n)\circ X^n(t^n)\right):\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}\\&\quad-\mathrm{tr}\left(\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h(\boldsymbol{\psi}_h^n)\circ X^n(t^n)\right)\\&\ge\int_{\mathcal{D}}\mathrm{tr}\left(\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\right)-\int_{\mathcal{D}}\mathrm{tr}\left(\mathrm{e}^{\pi_h(\boldsymbol{\psi}_h^n)}-\pi_h(\boldsymbol{\psi}_h^n)\right)\circ X^n(t^n)\\&=\int_{\mathcal{D}}\mathrm{tr}\left(\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\right)-\int_{\mathcal{D}}\mathrm{tr}\left(\mathrm{e}^{\pi_h(\boldsymbol{\psi}_h^n)}-\pi_h(\boldsymbol{\psi}_h^n)\right).\end{aligned}$$

Besides, using equation (2-II.16), we have :

$$\begin{aligned}\int_{\mathcal{D}}\left(\boldsymbol{\Omega}_h^{n+1}\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\boldsymbol{\Omega}_h^{n+1}\right):\left(\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}-\boldsymbol{I}\right)=\\\int_{\mathcal{D}}\left(\boldsymbol{\Omega}_h^{n+1}\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\boldsymbol{\Omega}_h^{n+1}\right):\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}=0,\end{aligned}$$

and using equations (2-II.15) and (2-III.3) :

$$\int_{\mathcal{D}}\boldsymbol{B}_h^{n+1}:\left(\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)-\boldsymbol{I}\right)=\int_{\mathcal{D}}\boldsymbol{B}_h^{n+1}:\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}-\int_{\mathcal{D}}\operatorname{div}(\boldsymbol{u}_h^{n+1})=\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)},$$

which cancels out with the same term $\int_{\mathcal{D}}\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}$ in the momentum equation. □

2-VI-D-c.2 The discontinuous Galerkin method If the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized by the *discontinuous Galerkin* method, the system writes :

$$\begin{aligned}0=&\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\,\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}\\&+\frac{\varepsilon}{\mathrm{Wi}}\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right):\boldsymbol{\nabla}\boldsymbol{v}\left(\frac{\boldsymbol{\psi}_h^{n+1}-\pi_h\left(\boldsymbol{\psi}_h^n\right)}{\Delta t}\right):\boldsymbol{\phi}-\left(\boldsymbol{\Omega}_h^{n+1}\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)-\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\boldsymbol{\Omega}_h^{n+1}\right):\boldsymbol{\phi}-2\boldsymbol{B}_h^{n+1}:\boldsymbol{\phi}\\&-\frac{1}{\mathrm{Wi}}(\pi_h\left(e^{-\boldsymbol{\psi}_h^{n+1}}\right)-\boldsymbol{I}):\boldsymbol{\phi}+\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{u}_h^n\cdot\boldsymbol{n}|\,[\![\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)]\!]:\boldsymbol{\phi}^+.\end{aligned}\quad(2\text{-VI.13})$$

Proposition 11 still holds for solutions of the system (2-VI.13). The proof follows that of the previous Section 2-VI-D-c.1 except for the treatment of the jump term, which follows that of Section 2-IV-A-b (see also Sect. 2-IV-B-b), because $\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)\in(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ and $\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)=\mathrm{e}^{\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)}$ is also in $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$.

2-VI-D-d Other finite elements for $(\boldsymbol{u}_h,p_h)$

In this section, we review the modifications that apply to the systems in the two previous Sections 2-VI-D-b and 2-VI-D-c when the different mixed finite element spaces for $(\boldsymbol{u}_h,p_h)$ proposed in Section 2-IV-C are used instead of Scott-Vogelius. Notice that the conclusions of Table 2.1 about the conditions that the velocity field has to satisfy still hold for the two previous Sections 2-VI-D-b and 2-VI-D-c with piecewise linear approximations of $\boldsymbol{\sigma}_h,\boldsymbol{\psi}_h$.

Other finite elements space for $(\boldsymbol{u}_h,p_h)$ than Scott-Vogelius and adequate projections of the velocity field (see summary in Tab. 2.2) have to be combined with interpolations of the stress field $\boldsymbol{\sigma}_h,\boldsymbol{\psi}_h$ using π_h (see the two previous Sects. 2-VI-D-b and 2-VI-D-c above). We give a summary of the projections that are required in Table 2.3.

2-VI-D-d.1 Alternative mixed finite element space for $(\boldsymbol{u}_h,p_h)$ with inf-sup condition The situation is very similar to that in Section 2-IV-C-b. Among the mixed finite element space that satisfy the inf-sup condition, let us first choose the *Taylor-Hood* $(\mathbb{P}_2)^d\times\mathbb{P}_1$. Again, because the velocity is not even weakly incompressible in the sense of equation (2-IV.7), we need to use the projection of the velocity field onto the solenoidal vector fields for the treatment of some terms in the variational formulations. When the advection terms $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ and $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\psi}$ are discretized using the *characteristic* method, we define the flow with $P_h^{rot}(\boldsymbol{u}_h^n)$ like in (2-IV.10) and use the same systems (2-VI.5) and (2-VI.10) as above. When the advection terms $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ and $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\psi}$ are discretized using the *discontinuous Galerkin* method, we use systems similar to (2-VI.5) and (2-VI.10) above, where the jump term rewrites (in the conformation-tensor formulation) :

$$+\sum_{j=1}^{N_E}\int_{E_j}|P_h^{rot}(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|[\![\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)]\!]:\boldsymbol{\phi}^+.$$

Also, one still needs to add the so-called Temam correction term (2-IV.11) to the weak formulation.

We can also use the *Crouzeix-Raviart* finite elements for velocity (see (2-IV.8)) : $(\boldsymbol{u}_h,p_h,\boldsymbol{\sigma}_h)$ in $(\mathbb{P}_1^{CR})^d\times\mathbb{P}_0\times(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$. Similarly to the advection terms $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ and $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\psi}$, the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}$ in the Navier-Stokes equations should then be discretized either using a characteristic method with the flow defined in (2-IV.13) with any of the projections P_h introduced above for the velocity field, or using the *discontinuous Galerkin* method formulated in equation (2-IV.14).

It is noticeable that choosing the mixed finite elements of Crouzeix-Raviart simplifies all the variational formulations presented above in the present Section D. Indeed, since $\boldsymbol{\nabla}\boldsymbol{u}\in(\mathbb{P}_0)^{d\times d}$ and we have the Lemma 8, it is then unnecessary to project the velocity except in the advection terms. For instance, for the conformation-

tensor formulation using the discontinuous Galerkin method, the formulation writes :

$$0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}$$
$$+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\pi_h(\boldsymbol{\sigma}_h^n)}{\Delta t}\right):\boldsymbol{\phi}-\left((\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1})\boldsymbol{\sigma}_h^{n+1}+\boldsymbol{\sigma}_h^{n+1}(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1})^T\right):\boldsymbol{\phi}$$
$$+\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\boldsymbol{\phi}+\sum_{j=1}^{N_E}\int_{E_j}|P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|\,[\![\pi_h(\boldsymbol{\sigma}_h^{n+1})]\!]:\boldsymbol{\phi}^+ +\mathrm{Re}|P_h(\boldsymbol{u}_h^n)\cdot\boldsymbol{n}|\,[\![\boldsymbol{u}_h^{n+1}]\!]\cdot\{\boldsymbol{v}\}\cdot \quad \text{(2-VI.14)}$$

Note that the second term in the sum of integrals over edges E_j is due to the use of the Crouzeix-Raviart element, and is uncorrelated to the treatment of the advection by a discontinuous Galerkin method.

The discrete free energy estimate (2-VI.7) holds. Its proof combines arguments of the proofs above, except for the treatment of the upper-convective term in (2-VI.14). This term writes, on any element K_k of the mesh (using Lem. 8, the fact that $\boldsymbol{\nabla}\boldsymbol{u}\in(\mathbb{P}_0)^{d\times d}$ and the incompressibility (2-III.3)) :

$$\begin{aligned}\int_{K_k}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\boldsymbol{\sigma}_h^{n+1}:(\boldsymbol{I}-\pi_h(\boldsymbol{\sigma}_h^{n+1})^{-1}) &= \int_{K_k}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\boldsymbol{\sigma}_h^{n+1}:\pi_h(\boldsymbol{\sigma}_h^{n+1})^{-1}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\\ &= \int_{K_k}\pi_h(\boldsymbol{\sigma}_h^{n+1}):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\pi_h(\boldsymbol{\sigma}_h^{n+1}):\pi_h(\boldsymbol{\sigma}_h^{n+1})^{-1}\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\\ &= \int_{K_k}\pi_h(\boldsymbol{\sigma}_h^{n+1}):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}-\int_{\mathcal{D}}\mathrm{div}\,\boldsymbol{u}_h^{n+1}\\ &= \int_{K_k}\pi_h(\boldsymbol{\sigma}_h^{n+1}):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1},\end{aligned}$$

which vanishes after combination with the extra-stress term in the momentum equation, the latter satisfying :

$$\int_{K_k}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}=\int_{K_k}\pi_h(\boldsymbol{\sigma}_h^{n+1}):\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1},$$

because of the fact that $\boldsymbol{\nabla}\boldsymbol{u}\in(\mathbb{P}_0)^{d\times d}$ and using Lemma 8.

2-VI-D-d.2 Alternative mixed finite element space for $(\boldsymbol{u}_h,p_h)$ without inf-sup It is also possible to use finite element spaces for $(\boldsymbol{u}_h,p_h)$ that do not satisfy the inf-sup condition like in Section 2-IV-C-c, while the stress field is discretized using discontinuous piecewise linear approximations. The construction of systems of equations and the derivation of discrete free energy estimates then directly follow from the combination of results from Section 2-IV-C-c with those used above in Section D, after upgrading the degree of the polynomial approximations for the stress field.

If we consider the mixed finite element space $(\mathbb{P}_1)^d\times\mathbb{P}_0$ for $(\boldsymbol{u}_h,p_h)$, and if the term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$ is discretized with the *characteristic* method, the system then writes :

$$0=\sum_{k=1}^{N_K}\int_{K_k}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\mathrm{div}\,\boldsymbol{u}_h^n(\boldsymbol{v}\cdot\boldsymbol{u}_h^{n+1})+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}$$
$$-p_h^{n+1}\mathrm{div}\,\boldsymbol{v}+q\,\mathrm{div}\,\boldsymbol{u}_h^{n+1}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\pi_h(\boldsymbol{\sigma}_h^n)\circ X^n(t^n)}{\Delta t}\right):\boldsymbol{\phi}$$
$$-\left(\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\boldsymbol{\sigma}_h^{n+1}+\boldsymbol{\sigma}_h^{n+1}\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)^T\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}):\boldsymbol{\phi}+\sum_{j=1}^{N_E}|E_j|\int_{E_j}[\![p_h]\!][\![q]\!], \quad \text{(2-VI.15)}$$

with a flow X^n computed with the projected field $P_h^{rot}(\boldsymbol{u}_h^n)$ through (2-IV.10). It is noteworthy that, for the same reason as above in equation (2-VI.14), the projection operator π_h is needed only for the discretization of the advection term $\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}$.

If we consider the mixed finite element space $(\mathbb{P}_1)^d\times\mathbb{P}_1$ for $(\boldsymbol{u}_h,p_h)$, it is straightforward to rewrite the system (2-IV.16) where the stress field was only piecewise constant, while using the same argument as above to see that only the advection term for the stress field needs a projected velocity.

TAB. 2.3 – Summary of projected terms in the Navier-Stokes (NS) and Oldroyd-B (OB) equations for $\boldsymbol{\sigma}_h/\boldsymbol{\psi}_h$ in $(\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$.

$\boldsymbol{\nabla u}\ldots$	$\boldsymbol{\sigma}_h \in (\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$	$\boldsymbol{\psi}_h \in (\mathbb{P}_{1,disc})^{\frac{d(d+1)}{2}}$
In $(\mathbb{P}_0)^{d^2}$	$\pi_h(\boldsymbol{\sigma}_h^n)$ in time derivative (incl. flux term in DG)	$\pi_h(\boldsymbol{\psi}_h^n)$ in time derivative (incl. flux term in DG) + implicit source term $\pi_h\left(e^{-\boldsymbol{\psi}_h^{n+1}}\right)$ in OB + implicit coupling term $\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)$ in NS
Not in $(\mathbb{P}_0)^{d^2}$	$\pi_h(\boldsymbol{\sigma}_h^n)$ in time derivative (incl. flux term in DG) + implicit coupling terms ($\pi_h\left(\boldsymbol{\sigma}_h^{n+1}\right)$ in NS, OB)	$\pi_h(\boldsymbol{\psi}_h^n)$ in time derivative (incl. flux term in DG) + implicit source term $\pi_h\left(e^{-\boldsymbol{\psi}_h^{n+1}}\right)$ in OB + implicit coupling terms ($\pi_h\left(e^{\boldsymbol{\psi}_h^{n+1}}\right)$ in NS, $\pi_h\left(\boldsymbol{\psi}_h^{n+1}\right)$ in OB)

Remark 9. *We were not able to establish discrete free energy estimates without interpolating some terms in the formulations above thanks to the operator π_h. This operator projects the stress $\boldsymbol{\sigma}_h$ (or $\boldsymbol{\psi}_h$) onto $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$. Thus, for the formulations we have considered in this section, the interest of using larger dimensional spaces for $\boldsymbol{\sigma}_h$ (or $\boldsymbol{\psi}_h$) than $(\mathbb{P}_0)^{\frac{d(d+1)}{2}}$ is not clear. Our aim in this section is simply to exhibit discrete formulations with piecewise linear approximations of the stress, for which we are able to derive a free energy estimate.*

2-VI-E Appendix E. Free-energy-dissipative discretization of a Lie-formulation

We discuss here some discretization of the Oldroyd-B system where the equation for the stress tensor is reformulated using a Lie derivative along the deformation gradient (see Rem. 6 and [LX06]). We want to show that some discretizations of the Lie-formulation could also satisfy a discrete free energy inequality.

Using Scott-Vogelius elements for $(\boldsymbol{u}_h,p_h)$ and piecewise constant approximations for $\boldsymbol{\sigma}_h$, one possible (low-order) discretization of a Lie-formulation from [LX06] writes :

$$\begin{aligned}0=&\int_{\mathcal{D}}\mathrm{Re}\left(\frac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\cdot\boldsymbol{v}-p_h^{n+1}\operatorname{div}\boldsymbol{v}+q\operatorname{div}\boldsymbol{u}_h^{n+1}+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}_h^{n+1}:\boldsymbol{\nabla}\boldsymbol{v}\\&+\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right)^{-1}\left(\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right)\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right)^{-T}}{\Delta t}\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{I}\right):\boldsymbol{\phi},\end{aligned}\tag{2-VI.16}$$

where the characteristic flow $X^n(t)$ is defined like in (2-III.7). The system (2-VI.16) admits a solution such that $(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right))^{-1}$ is well-defined, provided Δt is sufficiently small (but possible very small when $\left\|\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right\|$ is large). Besides, taking $\boldsymbol{\phi}$ as the characteristic function of some element K_k, we have the following equality inside K_k :

$$\left(1+\frac{\Delta t}{\mathrm{Wi}}\right)\boldsymbol{\sigma}_h^{n+1}=\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right)^{-1}\left(\boldsymbol{\sigma}_h^n\circ X^n(t^n)\right)\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right)^{-T}+\frac{\Delta t}{\mathrm{Wi}}\boldsymbol{I}.\tag{2-VI.17}$$

Then it is clear that the system (2-VI.16) preserves the non-negativity of $\boldsymbol{\sigma}_h^n$. Moreover, it is possible to derive the free energy estimate (2-IV.2) for the system (2-VI.16). It suffices to take as a test function for the stress :

$$\boldsymbol{\phi}=\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right)^T\left(\boldsymbol{I}-(\boldsymbol{\sigma}_h^{n+1})^{-1}\right)\left(\boldsymbol{I}-\Delta t\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\right),$$

and to proceed to the derivation of a free energy estimate using both ideas of the present work and of the work [LX06], after noting that :

$$\mathrm{tr}\left(\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)^T\left(\boldsymbol{I}-(\boldsymbol{\sigma}_h^{n+1})^{-1}\right)\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)\left(\left(1+\frac{\Delta t}{\mathrm{Wi}}\right)\boldsymbol{\sigma}_h^{n+1}-\frac{\Delta t}{\mathrm{Wi}}\boldsymbol{I}\right)\right)\geq 0,\tag{2-VI.18}$$

the proof of which is completely similar to the proof of (2-I.9), using the fact that $\left(1+\frac{\Delta t}{\mathrm{Wi}}\right)\boldsymbol{\sigma}_h^{n+1}-\frac{\Delta t}{\mathrm{Wi}}\boldsymbol{I}$ is symmetric positive definite (provided Δt is sufficiently small) and $\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)^T\left(\boldsymbol{I}-(\boldsymbol{\sigma}_h^{n+1})^{-1}\right)\pi_h\left(\boldsymbol{\nabla}\boldsymbol{u}_h^{n+1}\right)$ is symmetric positive semi-definite.

Acknowledgement 1. *Thank to A. Ern and J.W. Barrett for fruitful discussions.*

2-VII Addendum to Chapter 2

In this section, we show yet unpublished materials : numerical results obtained with a stabilised $\mathbb{P}_1\times\mathbb{P}_1\times\mathbb{P}_1$ discretization, implemented in the MISTRAL FE code initially developped by J.F. Gerbeau et T. Lelièvre, and with $\mathbb{P}_2\times\mathbb{P}_1\times\mathbb{P}_0$ or $\mathbb{P}_2\times\mathbb{P}_0\times\mathbb{P}_0$ discretizations, using the FreeFem++ code of O. Pironneau and F. Hecht.

2-VII-A Numerical Results with MISTRAL : $\mathbb{P}_1\times\mathbb{P}_1\times\mathbb{P}_1$ FE code

We simulate the Oldroyd-B flow (2-I.1) in a lid-driven cavity $\mathcal{D}=(0,1)\times(0,1)$ ($d=2$). We take $\mathrm{Re}=10^{-7}$, which makes our simulations close to the creeping flow results in [FK04, HP07a]. We choose $\varepsilon=.5$ to avoid instabilities due to incompatibility of the velocity–stress (in fact, conformation here) approximation spaces [BS92b].

For different meshes with triangular elements of size upper-bounded by $h=.1$, .05 and .03, we compute for $N_T=1000$ time steps with $\Delta t=.01$ ($T=10$) the solution to a system for $(\boldsymbol{u}_h,p_h,\boldsymbol{\sigma}_h)$ in $(\mathbb{P}_1)^2\times\mathbb{P}_1\times(\mathbb{P}_1)^3$ close to (2-IV.16), where we rewrite the material derivative for $\boldsymbol{\sigma}_h$ as :

$$\int_{\mathcal{D}}\left(\frac{\boldsymbol{\sigma}_h^{n+1}-\boldsymbol{\sigma}_h^n}{\Delta t}+\boldsymbol{u}_h^n\cdot\boldsymbol{\nabla}\boldsymbol{\sigma}_h^{n+1}\right):\boldsymbol{\phi},$$

which is possible since $\boldsymbol{\sigma}_h\in\mathbb{P}_1$ is continuous. The velocity-pressure incompatibility is stabilized using SUPG [BH82], and the Oldroyd-B equation is stabilized with a GaLS method like in [BFT93, BPS01] (even though this may seem unnecessary because of the choice $\varepsilon=.5$).

Note that we are not able to show that such a discretization with a *continuous* stress field satisfies a discrete free energy estimate. Though, it seems quite close to our discretizations and to many discretizations used by practitionners.

We apply the following Dirichlet boundary condition at the top $\Gamma=[0,1]\times\{1\}$ of the cavity $\mathcal{D}$:

$$\boldsymbol{u}_h(x,1,t)=\begin{cases}\left(8x^2(1-x)^2(1+\tanh\left(8(t-.5)\right)\right), & \forall(x,t)\in[0,1]\times[0,t_{max})\\ 0, & \forall(x,t)\in[0,1]\times(t_{max},T)\end{cases}. \tag{2-VII.1}$$

The initial condition at $t=0$ is taken as the stationary state $\left(\boldsymbol{u}_h^0\equiv 0,\boldsymbol{\sigma}_h^0\equiv\boldsymbol{I}\right)$. When, $t_{max}\in(0,T)$, provided Δt is small enough (depending on the initial condition), we could expect the flow to converge back to the stationary solution $\left(\boldsymbol{u}_h^0\equiv 0,\boldsymbol{\sigma}_h^0\equiv\boldsymbol{I}\right)$ for $t>t_{max}$ (recall Proposition 7 for homogeneous Dirichlet boundary conditions).

The non-linearity will be solved using Picard iterations (and GMRES for the linear steps at each Picard iteration), which may not be optimal (and may then induce inaccuracies responsible for subsequent numerical instabilities). Yet, quite often in practice, the numerical results barely change using more than one Picard iteration. So we conjecture a fast convergence of the non-linear iterations in this simple testcase and will indeed show results obtained with only one Picard iteration (temporarily assuming this is not responsible for subsequent numerical instabilities).

First fixing $t_{\max}=T=10$ like in [FK04, HP07a], we show a numerical evidence of a High-Weissenberg-Number problem, assuming that the numerical solutions should converge to a stationary value (we have no proof for the existence of such a stationary state). It seems the higher Wi, the more difficult it is to converge to a stationary value, see Fig. 2.1. This manifestation of a High-Weissenberg-Number problem seems commonly observed in the literature.

We also performed numerical simulations for the log-formulation using the same discretization with a SUPG-stabilized mixed finite element space $(\boldsymbol{u}_h,p_h)\in(\mathbb{P}_1)^2\times\mathbb{P}_1$ and a continuous stress field $\boldsymbol{\psi}_h\in\mathbb{P}_1$. Then, in this first test, the log-formulation seems more "stable".

But this testcase does not coincide with our Proposition 7. Indeed, it is not obvious to compute the free-energy of a system, the stationary solution of which we do not know (even if we assume it). Let us now use $t_{\max}=2$.

With our *a priori* not necessarily free-energy-dissipative numerical scheme for $(\boldsymbol{u},p,\boldsymbol{\sigma})$, the solutions obtained with $\Delta t=10^{-2}$ still seem to converge to the stationary state after $t>t_{\max}$ (when the boundary condition (2-VII.1) vanishes) while $\mathrm{Wi}\leq 1$. But for higher Wi (and always with $\Delta t=10^{-2}$), the solution blows up all the faster after $t>t_{\max}$ as Wi grows, see Fig. 2.2. This appears to be a manifestation of the High-Weissenberg-Number problem in our new testcase. But is the numerical instability due to a lack of free-energy-dissipation here ? (And if yes, could we kill that instability by improving our scheme toward a better free-energy dissipation ?)

First, we can hope for our (reasonable) numerical scheme to be close to free-energy-dissipative, even if we cannot show this. Then, from the existence results of Propositions 1 and 7, one could then think of taking a

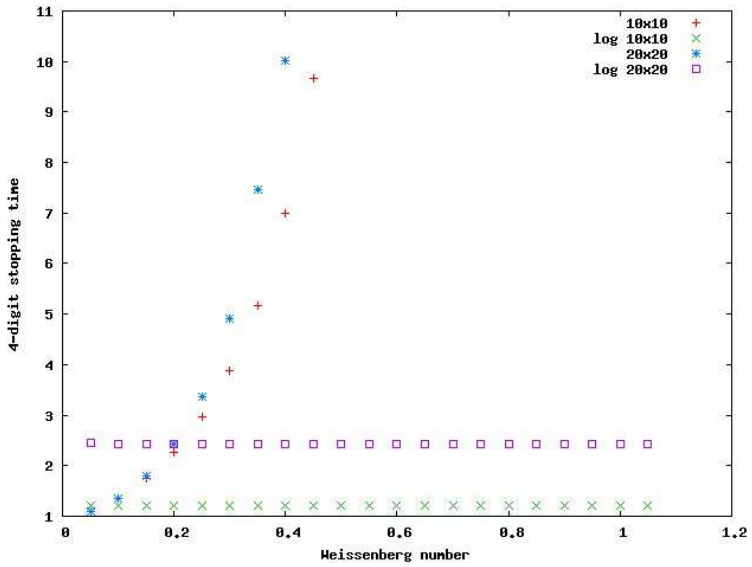


FIG. 2.1 – First time when the four digits in the L^∞ norm of the three fields $(\boldsymbol{u},p,\boldsymbol{\sigma})$ seem stabilized (that is, do not evolve after three consecutive steps), fwith respect to the Weissenberg numbers, for two meshes : $h=.1$ (10×10) and .05 (20×20). Also for $(\boldsymbol{u},p,\boldsymbol{\psi})$, using the log-formulation (log 10×10 and log 20×20).

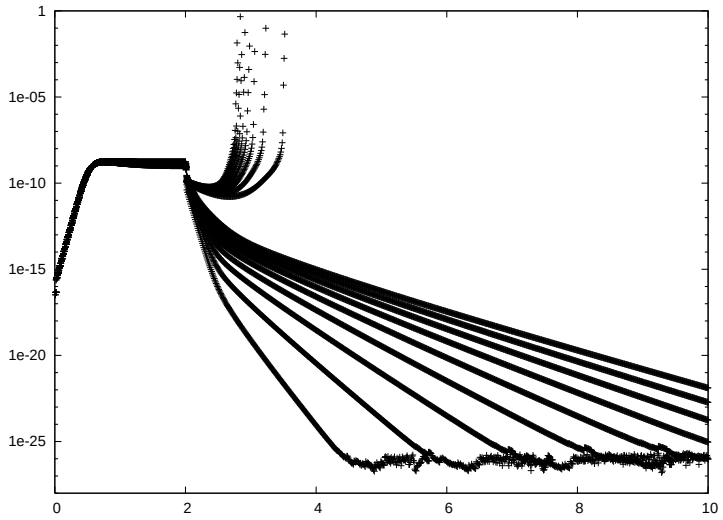


FIG. 2.2 – Kinetic energy (in logarithmic scale) versus time for $\mathrm{Wi}\in\{.1,.2\ldots 1.9,2\}$. Blow up happens with $\Delta t=10^{-2}$ for $\mathrm{Wi}>1$, all the faster after $t_{max}=2$ as Wi is large.

smaller time step $\Delta t=10^{-3}$. And indeed, using the smaller time step $\Delta t=10^{-3}$, the limit Weissenberg number until which convergence to a stationary state apparently occurs is a bit higher : $\mathrm{Wi}\leq 1.2$ (see Figure 2.3), which is coherent with the fact that more stringent initial data (for instance, in the case of higher Wi) may ask for smaller time steps. Yet, the improvement in Weissenberg number seems small compared to the much more computationally expensive simulations with $\Delta t=10^{-3}$.

Moreover, like many practitioners, we noticed that the blow-up starts at approximatively the same time than a loss of positivity for $\boldsymbol{\sigma}$ (close to the top-right corner, where the stress becomes singular for $t_{\max}=T$). So a question naturally arises : is in fact the loss of positivity the cause or the consequence of blow-up ? Indeed, if we use the log-formulation, then nothing happens (no blow-up) ! Then, from this viewpoint, it seems the log-formulation is better, apparently because it ensures positivity of $\boldsymbol{\sigma}$.

Yet, this numerical fact is not a definitive evidence for the superiority of the log-formulation. At this step, one should still check whether a numerical scheme that really satisfies a free energy estimate does better. Besides,

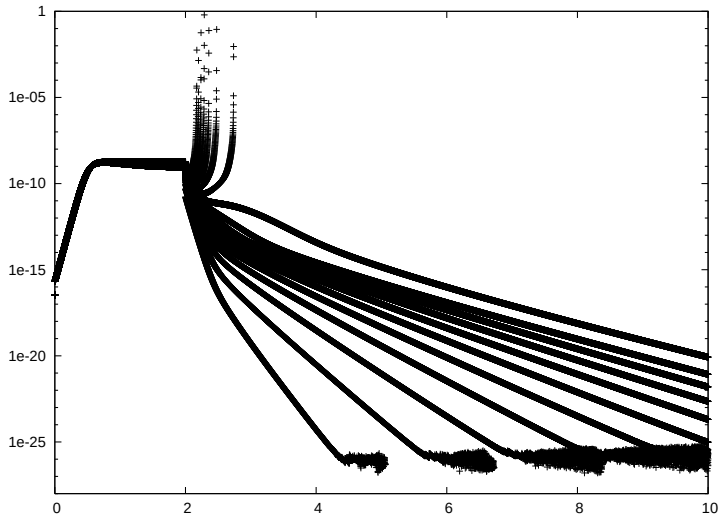


FIG. 2.3 – Kinetic energy (in logarithmic scale) versus time for $\mathrm{Wi} \in \{.1,.2 \ldots 1.9,2\}$. Blow up happens with $\Delta t = 10^{-3}$ for $\mathrm{Wi} > 1.2$, all the faster after $t_{max} = 2$ as Wi is large.

even if our (nonlinear) scheme was theoretically free-energy-dissipative, the numerical instabilities we observed might also be due to its numerical solution. That is, Picard iterations may converge more slowly as Wi grows, and one may for instance have to take very small Δt in order to actually see that convergence happen (with large Wi). By the way, since the scheme for the log-formulation is even more nonlinear, why is it not confronted with such problems ?

Now, we observed in the numerical experiments that it is essential for the stability of the discrete log-formulation to project $e^{\boldsymbol{\psi}_h}$ and $e^{-\boldsymbol{\psi}_h}$ on the discontinuous fields $\pi_h\left(e^{\boldsymbol{\psi}_h}\right)$ and $\pi_h\left(e^{-\boldsymbol{\psi}_h}\right)$. For instance, if we use discrete fields for the stress that are interpolated in $\mathbb{P}_1$, with the same values as $e^{\boldsymbol{\psi}_h}$ and $e^{-\boldsymbol{\psi}_h}$ at each node of the mesh, rather than the $\mathbb{P}_0$ discontinuous projected stress fields $\pi_h\left(e^{\boldsymbol{\psi}_h}\right)$ and $\pi_h\left(e^{-\boldsymbol{\psi}_h}\right)$ that we introduced in our previous theoretical investigation, then the log-formulation does not correctly dissipate the free-energy. Coming back to the test with $t_{\max} = 10$, we indeed observe that the free-energy does not converge to a stationary value in the badly-discretized log-formulation, see Fig. 2.4.

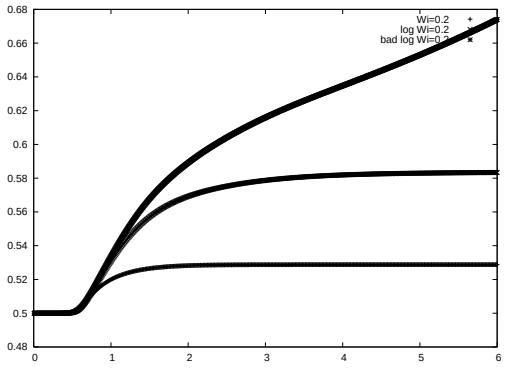


FIG. 2.4 – Blow-up of the free energy for the $\boldsymbol{\psi}_h$ formulation at $\mathrm{Wi} = .2$ when $\mathbb{P}_1$ interpolations at the nodes are used for $e^{\boldsymbol{\psi}_h}$,$e^{-\boldsymbol{\psi}_h}$ instead of $\mathbb{P}_0$ projections.

Note also in Fig. 2.4 the discrepancy between the stationary value of the free energy for the usual formulation with $\boldsymbol{\sigma}_h$ and that for the log-formulation. This is probably explained by the relative large numerical error in solutions obtained with the log-formulation, which is probably of poor consistency after taking the exponential of the field $\boldsymbol{\psi}_h$.

We show in Fig. 2.5 the consistency of our different spatial discretizations (log and not log) by comparing the relative errors in the stationary value of the free-energy at $t_{\max} = T = 10$ for three different meshes, with respect to a finer one (at a Wi number where convergence happens, of course). This corresponds to the time-evolutions

shown in Fig. 2.6.

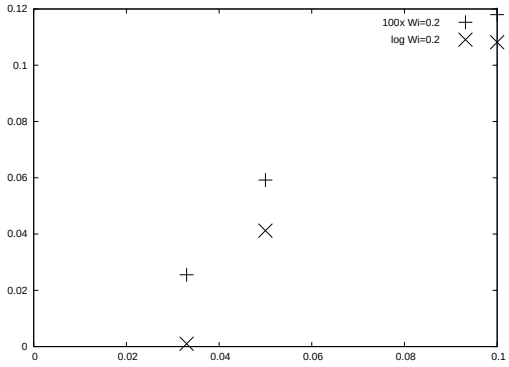


FIG. 2.5 – Relative errors for the stationary free energy at $\mathrm{Wi}=.2$ with numerical schemes for $\boldsymbol{\psi}_h$ and $\boldsymbol{\sigma}_h$ (then multiplied by 100 in the latter case), with respect to the size h of the mesh.

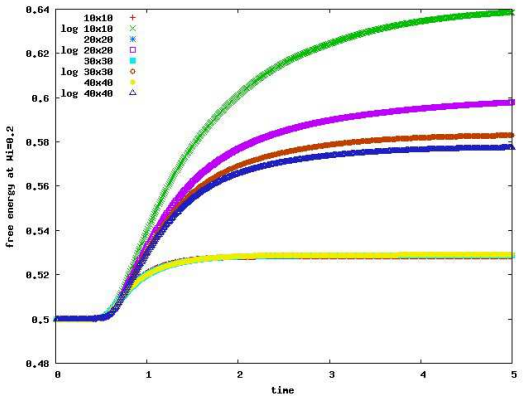


FIG. 2.6 – Free energies at $\mathrm{Wi}=.2$ for refining meshes. From top to bottom, $h=.1,.05,.0333,.25$ for the equations with $\boldsymbol{\psi}_h$ forumlation, and for the equations with $\boldsymbol{\sigma}_h$ forumlation (the relative error is then much smaller).

2-VII-B Numerical Results with FreeFem++ : $\mathbb{P}_2\times\mathbb{P}_1\times\mathbb{P}_0$ FE code

We first tried to use a $\mathbb{P}_2\times\mathbb{P}_1\times\mathbb{P}_0$ discretization of the three-field problem, but without using the projection operator which we needed in the previous sections in order for a $\mathbb{P}_2\times\mathbb{P}_1\times\mathbb{P}_0$-based numerical scheme to satisfy a free-energy inequality (in the case of no-flow boundary conditions).

We tried both the characteristic method (as implemented in FreeFem++ v2.240002) and the DG method to compute the advection term in the Oldroyd-B equation. To test these discretizations, we simulate the two same problems as in the above section. Like in the previous section above, the linear systems (corresponding to one Picard iteration) are solved by the GMRES method.

First, we wanted to check whether our discretizations actually dissipate the free-energy after $t_{\max}=2$ (under no flow boundary conditions), for a constant time step $\Delta t=10^{-2}$, three different meshes (satisfying respectively $h=.1$, .05 and .03) and $\mathrm{Wi}=.8,1,1.2,\ldots,2$.

All the computations have indeed been observed to dissipate the free-energy after $t_{\max}=2$, until $t_{\max}=T$, see Fig. 2.7. Note yet that this could not be checked for the characteristic method when $h=.03$, since then the positivity of the conformation tensor is lost before $t=2$ (followed by blow-up a few iterations later), whatever the Weissenberg number tested here. (With the DG approach, the positivity of the conformation tensor is also lost before $t=2$, whatever the Weissenberg number tested here, if we use a finer mesh $h=.025$; but then blow-up does not follow : the computations run until final time T.)

Second, we tried to capture a stationary state using $t_{\max}=T$ in our different discrete problems.

For the characteristic method, the convergence to a stationary state after $t=2$ could only be observed for the coarser mesh $h=.1$, see Fig. 2.8. Then, the convergence rate seems all the slower as Wi increases, in the sense

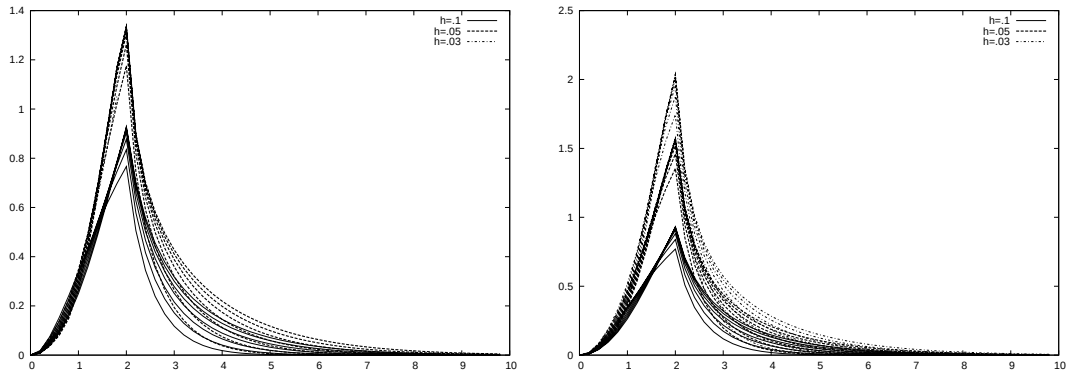


FIG. 2.7 – Free energy with respect to time using the characteristic (left) / DG method (right) with $\mathbb{P}_1$ approximations of the pressure. We use $t_{\max}=2$, three meshes and $\mathrm{Wi}\in[.8,2]$.

that the difference between two successives steps decreases more slowly when Wi increases. For finer meshes, the positivity is lost soon after $t=2$, all the earlier as Wi increases.

The DG method always converged to a stationary state. Though, the positivity of the conformation tensor is also lost for meshes finer than $h=.1$, soon after $t=2$ (and all the sooner as Wi increases and h decreases), although always later than the characteristic method. (It is never followed by a blow-up a few iterations later, contrary to the characteristic method).

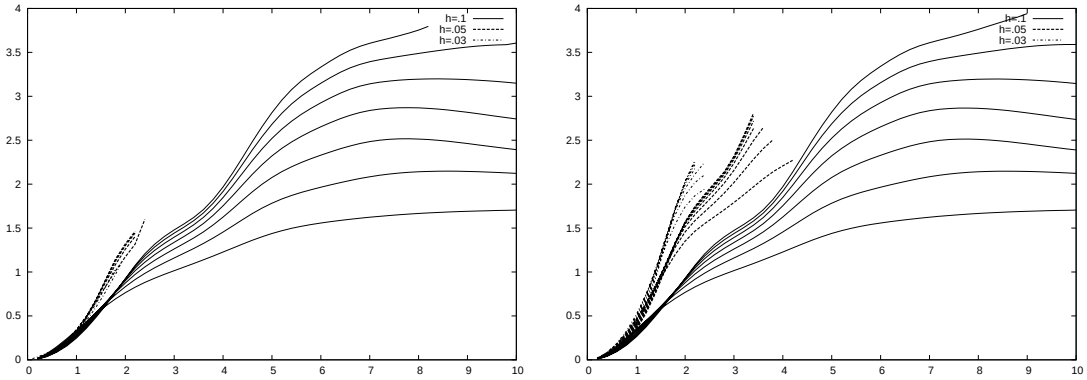


FIG. 2.8 – Free energy with respect to time when $t_{\max}=T$, using the characteristic (left) / DG method (right) with $\mathbb{P}_1$ approximations of the pressure. When the curve stops, positivity has been lost (left : the computations then blow up). We use three different meshes $h=.1,.05,.03$ and $\mathrm{Wi}\in[.8,2]$.

2-VII-C Numerical Results with FreeFem++ : $\mathbb{P}_2\times\mathbb{P}_0\times\mathbb{P}_0$ FE code

Compared to the previous section, we use here a lower-order finite-element space for the pressure, which includes discontinuous elements (piecewise constant). Then, the mixed formulation for velocity and pressure spaces is still inf-sup stable, and we can even show that the DG approach is free-energy-dissipative (under no flow boundary conditions, see the next chapter for details).

First, using $t_{\max}=2$, we again observe that all our schemes dissipate the free-energy after $t_{\max}$ (under no flow boundary conditions), again for a constant time step $\Delta t=10^{-2}$, three different meshes ($h=.1$, .05 and .03) and $\mathrm{Wi}=.8,1,1.2,\ldots,2$, provided they remain computable. (No loss of positivity happened before $t_{\max}=2$.)

Second, we try to capture a stationary state using $t_{\max}=T$.

The characteristic method reaches a stationary state only for the coarser mesh $h=.1$, all the slower as Wi increases, in the sense the convergence rate of the difference between two succesive steps goes slower to zero. For finer mesh, no convergence to a stationary state could be observed because the positivity of the conformation tensor is lost soon after $t=2$, all the sooner as Wi increases, and then the computations stop (there is blow up).

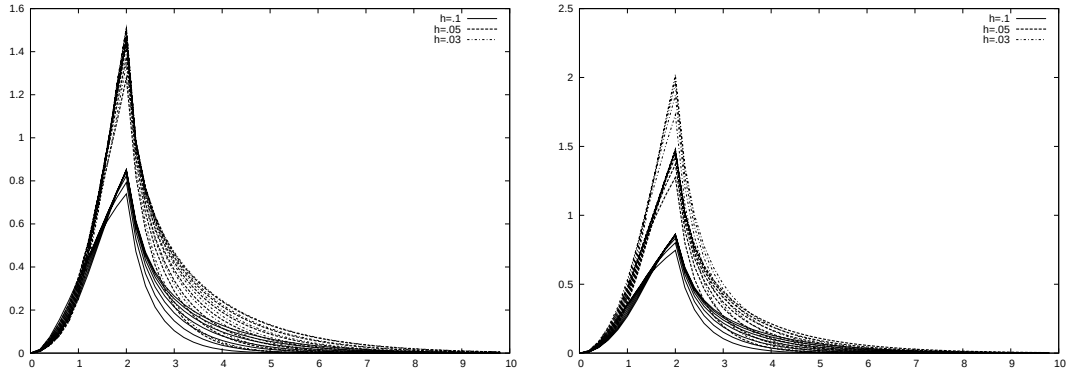


FIG. 2.9 – Free energy with respect to time using the characteristic (left) / DG method (right) with $\mathbb{P}_0$ approximations of the pressure. We use $t_{\max}=2$, three meshes and $\mathrm{Wi}\in[.8,2]$.

On the contrary, the DG method seems to converge to a stationary state for all meshes (though again all the slower as Wi increases, and h decreases). The conformation tensor also loses positivity after $t_{\max}=2$ (all the sooner as $\mathrm{Wi}\in[1,2]$ increases and h decreases) for the two finer meshes, again a bit later than the characteristic method, see Fig. 2.8. Note that for $\mathrm{Wi}=.8$ and $h=.05$ or $h=.03$, the positivity was not even lost ! (Also, for $\mathrm{Wi}=.8$, the convergence of the free energy and the kinetic energy is not monotone.)

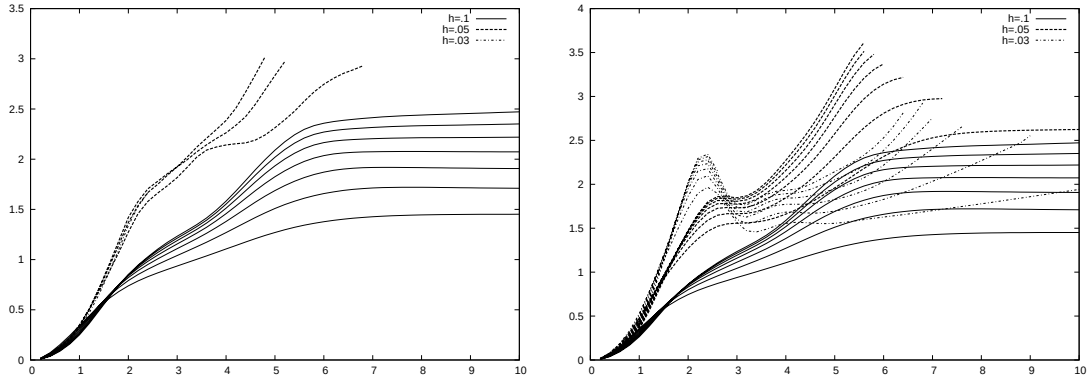


FIG. 2.10 – Free energy with respect to time when $t_{\max}=T$, using the characteristic (left) / DG method (right) with $\mathbb{P}_0$ approximations of the pressure. When the curve stops, positivity has been lost (left : the computations then blow up). We use three different meshes $h=.1,.05,.03$ and $\mathrm{Wi}\in[.8,2]$.

So it seems that, the scheme for which we can actually show the stability property of free-energy-dissipation (under no flow boundary conditions, here using the DG approach with piecewise constant approximation of the pressure) behaves a bit better than the other ones. But this scheme is not a definitive remedy to all the High-Weissenberg-Number problems, as shown by the loss of positivity (probably produced by a bad linear approximation of the nonlinear terms), it is at best one solution to a numerical problem among (probably) many different numerical problems mixed together.

To finish this section, we show in Fig. 2.11 the value of the diagonal component of the conformation tensor field, as it is computed at $t=5$ (when $t_{\max}=T=10$) by the DG approach above. For the lid-driven test-case, this is an indication about the origin of problems linked to a lack of regularity for the solutions to the continuous problem. Indeed, one can expect a numerical instability in the lid-driven cavity to arise from the singularity at the right-top corner. And as a matter of fact, when positivity is lost during the computation, it is also in this singular area that a pointwise value of the conformation tensor field is the first non-positive one encountered (then, the loss of positivity quickly propagates in the spatial domain and the computation blows up).

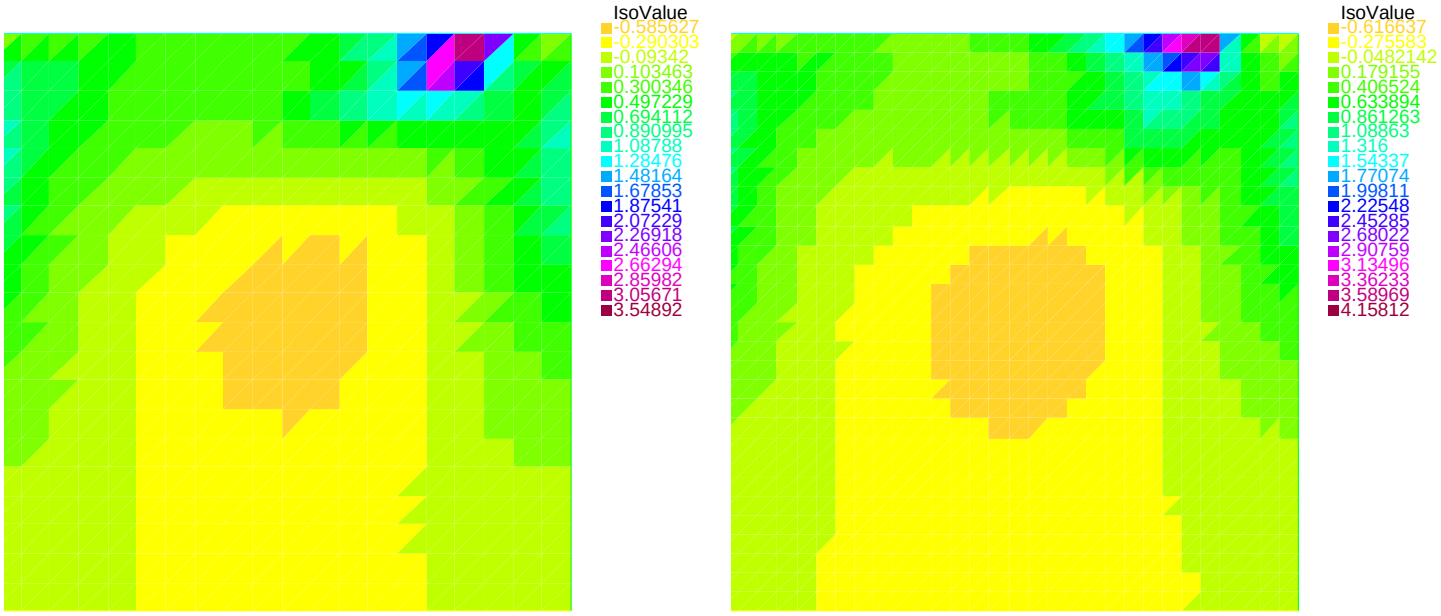


FIG. 2.11 – Value of the diagonal component of the conformation tensor field at $t=5$, for $t_{\max}=T$ and two different meshes (left : $h=.05$; and right : $h=.03$).

Chapitre 3

Existence et approximation d'un modèle Oldroyd-B (régularisé)

Dans ce chapitre, on présente des résultats obtenus en collaboration avec J.W. Barrett, et qui sont soumis pour publication dans une revue à comité de lecture.

On y considère des approximations par la méthode des éléments finis du système d'équations Oldroyd-B, qui modélise un fluide polymérique avec conditions de collement dans un domaine borné $\mathcal{D} \subset \mathbb{R}^d$ ($d = 2,3$). Nos schémas sont basés sur des approximations $\mathbb{P}_2$ (continues) du champ de vitesse et soit (a) $\mathbb{P}_0$ (discontinues), soit (b) $\mathbb{P}_1$ (continues), pour les champs de pression et conformation (un champ tensoriel symétrique).

D'abord, on montre qu'une solution (au moins) de ces schémas satisfait une inégalité d'*énergie libre*, s'exprimant avec le logarithme du tenseur des conformations, sans contrainte sur le pas de la discrétisation en temps – de type Euler rétrograde –. C'est une extension des résultats obtenus dans le chapitre précédent pour la même estimation d'énergie libre (voir aussi [BLM09]). Nous avions alors besoin d'une discrétisation $\mathbb{P}_0$ (discontinue et constante par morceaux) du champ de conformation pour traiter le terme d'advection dans l'équation sur la contrainte, et d'une restriction sur le pas de temps, formulée en fonction des conditions initiales, pour s'assurer que le tenseur de conformation restait bien symétrique défini positif. Toutefois, dans cette extension, nous n'avons plus unicité de la solution (on utilise en effet le théorème abstrait du point fixe de Brouwer), et il se pourrait que d'autres solutions du système d'équations discrétisées existassent et ne satisfassent pas l'estimation d'énergie libre (en particulier, ne fussent pas définies positives en le champs de conformation). Par ailleurs, avec des éléments finis continus, on utilise une discrétisation consistante mais *ad hoc* du terme de convection dans l'équation d'Oldroyd-B, afin qu'une procédure de test permette de vérifier simplement l'inégalité d'énergie libre attendue. En pratique, on pourrait donc encore tester numériquement ces schémas (leur facilité d'implémentation, leur ordre de convergence,...). De plus, on observerait alors si l'introduction de troncatures en le champs de conformation dans les termes de la dérivée sur-convective (équation d'Oldroyd-B) et de l'extra-contrainte polymérique (équation de Navier-Stokes) a un effet sur les instabilités numériques de type HWNP (mentionnées précédemment dans le Chapitre 2). En particulier, si on impose au champs de conformation une borne inférieure $\delta > 0$, on préserverait sa positivité, et on pourrait alors vraiment s'intéresser à la dissipation d'énergie libre sans se soucier du maintien de la positivité. (Numériquement, on finirait par la limite $\delta \to 0^+$: si pour chaque δ, les solutions numériques du système non-linéaire avec troncature à δ sont sur une même branche solution continue en δ, cette approche pourrait être une manière de capter, à la limite $\delta \to 0^+$, une solution du système de départ sans troncature qui dissipe bien l'énergie libre).

Ensuite, avec un terme dissipatif additionnel dans l'équation sur la contrainte et une troncature du tenseur de conformation (dans certains termes du système), troncature similaire à ce qui a été introduit dans [BS08] pour le modèle micro-macro des haltères dites *FENE dumbbells*, on montre la convergence (à extraction près) de l'approximation (b) vers une solution (faible) globale en temps du système d'équations Oldroyd-B *régularisé* (quand les paramètres de discrétisation en espace et en temps tendent vers 0). Dans le cas où $d = 2$, on peut de plus obtenir cette convergence sans la troncature, quoiqu'avec une restriction sur le pas de temps fonction du pas d'espace. On a ainsi montré l'existence de solutions faibles globales à un système d'équations Oldroyd-B régularisé directement à l'échelle macroscopique, sans faire appel à un modèle micro-macro sous-jacent (celui des haltères dites *Hookean dumbbells*), contrairement à [BSS05, BS07, BS08], où de fait, la régularisation était donc aussi quelque peu différente. On pourrait aussi étudier numériquement ce nouveau schéma – régularisé – dans des écoulements très instables (de type HWNP), car le terme dissipatif additionnel dans l'équation sur

la contrainte a une origine physique [BS07, Sch06] et a par le passé déjà montré une bonne aptitude dans des simulations [SB95]. De plus, le schéma numérique en question, dissipant une énergie libre, utilise des éléments finis continus $\mathbb{P}_1$, ce qui pourrait être encore relativement facile à implémenter (en comparaison d'une méthode de Galerkin discontinue) et efficace numériquement (quoique l'ordre global du schéma reste certainement peu élevé à cause des régularisations introduites).

Remarque 6 (Sur l'utilisation d'éléments finis $\mathbb{P}_1$ *continus* pour $\boldsymbol{\sigma}$)**.** *On a déjà tenté au Chap. 2 d'utiliser une méthode par éléments finis de plus grand ordre que $\mathbb{P}_0$ pour discrétiser $\boldsymbol{\sigma}$, avec des fonctions $\mathbb{P}_1$. On avait alors besoin d'utiliser un espace d'approximation de grande dimension pour $\boldsymbol{\sigma}$, qui contienne au moins les fonctions $\mathbb{P}_0$ constantes par morceaux sur chaque élément fini (et donc tel que les fonctions soient discontinues d'un élément fini à l'autre).*

En effet, la procédure de test de la formulation variationnelle qui mène à l'inégalité d'énergie libre fait appel à une fonction test non-linéaire en l'inconnue $\boldsymbol{\sigma}$ pour l'équation d'Oldroyd-B. Pour imiter cette procédure après discrétisation de l'équation d'Oldroyd-B, nous avions donc proposé de projeter la fonction test (cette fois non-linéaire en l'inconnue discrète $\boldsymbol{\sigma}_h$) dans un sous-espace de l'espace d'approximation pour $\boldsymbol{\sigma}$ où nous pouvions mener les calculs conduisant à une inégalité d'énergie libre discrète. Un sous-espace naturel pour effectuer cette projection était alors $\mathbb{P}_0$, puisqu'on venait de montrer qu'une méthode par éléments finis de bas ordre, avec $\boldsymbol{\sigma}$ discrétisé dans $\mathbb{P}_0$, permettait justement d'obtenir une inégalité d'énergie libre discrète. Mais, comme nous venons de le dire, discrétiser $\boldsymbol{\sigma}$ dans $\mathbb{P}_1+\mathbb{P}_0$ ou dans $\mathbb{P}_{1,disc}$ nécessite un grand nombre de degrés de liberté (donc de grandes matrices à inverser), et on pourrait souhaiter utiliser plutôt des fonctions $\mathbb{P}_1$ continues.

Dans le chapitre qui suit, nous utilisons bien des fonctions $\mathbb{P}_1$ continues pour discrétiser $\boldsymbol{\sigma}$, qui permet entre autre d'utiliser simplement la dérivée spatiale de l'inconnue $\boldsymbol{\sigma}_h$ afin d'introduire de la dissipation dans l'équation d'Oldroyd-B et d'effectuer ensuite une preuve de convergence (avec la bonne compacité). Il n'y a toutefois pas de miracle dans cette nouvelle discrétisation, qui permettrait d'obtenir simplement une inégalité d'énergie libre discrète avec des fonctions $\mathbb{P}_1$ continues. En effet, on doit ici aussi trouver un moyen de revenir dans l'espace d'approximation pour $\boldsymbol{\sigma}$ en vue d'effectuer les manipulations menant à une inégalité d'énergie libre discrète. On utilise ici pour cela l'interpolation de la fonction test, plutôt que la projection dans un sous-espace. Puis, comme précédemment avec la méthode de projection dans un sous-espace, il faut discrétiser chacun des termes de l'équation d'Oldroyd-B de manière ad hoc (mais consistante) pour pouvoir effectuer les manipulations menant à une inégalité d'énergie libre discrète.

D'abord, on fait appel à des formules de quadrature pour calculer certaines intégrales dans la formulation variationnelle uniquement avec les valeurs aux points d'interpolation. Ensuite, on se limite à un domaine $\mathcal{D}$ convexe qui peut être recouvert par une famille de maillages quasi-uniforme par rapport au paramètre de discrétisation spatiale, avec des simplexes dont tous les angles entre deux faces adjacentes sont aigus (en fait non-obtus[1]*, c'est-à-dire $\leq \frac{\pi}{2}$). Enfin, on discrétise le terme d'advection dans l'équation d'Oldroyd-B par une formule ad hoc (consistante), qui imite au niveau discret la formule de dérivation composée que nous invoquons au niveau continu pour annuler ce terme dans la procédure de test (voir* (3-V.16) *ci-après).*

Noter également que, comme dans le cas $\mathbb{P}_0$, on a besoin ici d'une première régularisation (qu'on lève ensuite par passage à la limite) pour montrer l'existence d'une solution (non-unique et dissipant l'énergie libre) à notre schéma discret, quelque soit le pas de temps dans une discrétisation temporelle de type Euler rétrograde (à cause de la contrainte de positivité sur $\boldsymbol{\sigma}$).

[1] Cette limitation permet de prolonger sur chaque élément fini à des dérivées de fonctions $\mathbb{P}_1$ des inégalités qui sont vraies pour les "dérivées discrètes" (différences finies d'ordre 1 construites avec les sommets des simplexes), voir le Lemma 11 ci-après.

Existence and approximation of a (regularized) Oldroyd-B model

John W. Barrett[a], Sébastien Boyaval[b,c]

[a]Imperial College London, Department of Mathematics (London SW7 2AZ, UK).
[b]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).
[c]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

We consider the finite element approximation of the Oldroyd-B system of equations, which models a dilute polymeric fluid, in a bounded domain $\mathcal{D}\subset\mathbb{R}^d$, $d=2$ or 3, subject to no flow boundary conditions. Our schemes are based on approximating the velocity field with continuous piecewise quadratics and either (a) piecewise constants or (b) continuous piecewise linears for the pressure and the symmetric conformation tensor. We show that both of these schemes satisfy a *free energy* bound, which involves the logarithm of the conformation tensor, without any constraint on the time step for the backward Euler type time discretization. This extends the results of [Boyaval *et al. M2AN* **43** *(2009) 523–561*] on this free energy bound. There a piecewise constant approximation of the conformation tensor was necessary to treat the advection term in the stress equation, and a restriction on the time step, based on the initial data, was required to ensure that the approximation to the conformation tensor remained positive definite. Furthermore , for our approximation (b) in the presence of an additional dissipative term in the stress equation and a cut-off on the conformation tensor on certain terms in the system, similar to those introduced in [Barrett *et al. M3AS* **18** *(2008) 935–971*] for the microscopic-macroscopic FENE model of a dilute polymeric fluid, we show (subsequence) *convergence*, as the spatial and temporal discretization parameters tend to zero, towards global-in-time weak solutions of this *regularized* Oldroyd-B system. Hence, we prove existence of global-in-time weak solutions to this regularized model. Moreover, in the case $d=2$ we carry out this convergence in the absence of cut-offs, but with a time step restriction dependent on the spatial discretization parameter, and hence show existence of a global-in-time weak solution to the Oldroyd-B system with an additional dissipative term in the stress equation.

3-I Introduction

3-I-A The standard Oldroyd-B model

We consider the Oldroyd-B model for a dilute polymeric fluid. The fluid, confined to an open bounded domain $\mathcal{D}\subset\mathbb{R}^d$ ($d=2$ or 3) with a Lipschitz boundary $\partial\mathcal{D}$, is governed by the following non-dimensionalized system :

(P) Find $\boldsymbol{u}:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\mapsto\boldsymbol{u}(t,\boldsymbol{x})\in\mathbb{R}^d$, $p:(t,\boldsymbol{x})\in(0,T)\times\mathcal{D}\mapsto p(t,\boldsymbol{x})\in\mathbb{R}$ and $\boldsymbol{\sigma}:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\mapsto\boldsymbol{\sigma}(t,\boldsymbol{x})\in\mathbb{R}^{d\times d}_S$ such that

$$\mathrm{Re}\left(\frac{\partial\boldsymbol{u}}{\partial t}+(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{u}\right)=-\boldsymbol{\nabla}p+(1-\varepsilon)\Delta\boldsymbol{u}+\frac{\varepsilon}{\mathrm{Wi}}\,\mathrm{div}\,\boldsymbol{\sigma}+\boldsymbol{f}\qquad\text{on }\mathcal{D}_T:=(0,T)\times\mathcal{D},\tag{3-I.1a}$$

$$\mathrm{div}\,\boldsymbol{u}=0\qquad\text{on }\mathcal{D}_T,\tag{3-I.1b}$$

$$\frac{\partial\boldsymbol{\sigma}}{\partial t}+(\boldsymbol{u}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}=(\boldsymbol{\nabla}\boldsymbol{u})\boldsymbol{\sigma}+\boldsymbol{\sigma}(\boldsymbol{\nabla}\boldsymbol{u})^T-\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}-\boldsymbol{I})\qquad\text{on }\mathcal{D}_T,\tag{3-I.1c}$$

$$\boldsymbol{u}(0,\boldsymbol{x})=\boldsymbol{u}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D},\tag{3-I.1d}$$

$$\boldsymbol{\sigma}(0,\boldsymbol{x})=\boldsymbol{\sigma}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D},\tag{3-I.1e}$$

$$\boldsymbol{u}=\boldsymbol{0}\qquad\text{on }(0,T)\times\partial\mathcal{D}.\tag{3-I.1f}$$

Here $\boldsymbol{u}$ is the velocity of the fluid, p is the hydrostatic pressure, and $\boldsymbol{\sigma}$ is the symmetric conformation tensor of the polymer molecules linked to the symmetric polymeric extra-stress tensor $\boldsymbol{\tau}$ through the relation $\boldsymbol{\sigma}=\boldsymbol{I}+\frac{\mathrm{Wi}}{\varepsilon}\boldsymbol{\tau}$, where $\boldsymbol{I}$ is the d-dimensional identity tensor and $\mathbb{R}^{d\times d}_S$ denotes symmetric real $d\times d$ matrices. In addition, $\boldsymbol{f}:(t,\boldsymbol{x})\in\mathcal{D}_T\mapsto\boldsymbol{f}(t,\boldsymbol{x})\in\mathbb{R}^d$ is the given density of body forces acting on the fluid ; and the following given parameters are dimensionless : the Reynolds number $\mathrm{Re}\in\mathbb{R}_{>0}$, the Weissenberg number $\mathrm{Wi}\in\mathbb{R}_{>0}$, and the elastic-to-viscous viscosity fraction $\varepsilon\in(0,1)$. For the sake of simplicity, we will limit ourselves to the no flow boundary condition (3-I.1f). Finally, $\boldsymbol{\nabla}\boldsymbol{u}(t,\boldsymbol{x})\in\mathbb{R}^{d\times d}$ with $[\boldsymbol{\nabla}\boldsymbol{u}]_{ij}=\frac{\partial\boldsymbol{u}_i}{\partial\boldsymbol{x}_j}$, and $(\mathrm{div}\,\boldsymbol{\sigma})(t,\boldsymbol{x})\in\mathbb{R}^d$ with $[\mathrm{div}\,\boldsymbol{\sigma}]_i=\sum_{j=1}^d\frac{\partial\boldsymbol{\sigma}_{ij}}{\partial\boldsymbol{x}_j}$.

Unfortunately, at present there is no proof of existence of global-in-time weak solutions to (P) available in the literature. Local-in-time existence results for (P) for sufficiently smooth initial data, and global-in-time existence results for sufficiently small initial data can be found in [GS90]. Global-in-time existence results for the corotational version of (P); that is, where $\boldsymbol{\nabla u}$ in (3-I.1c) is replaced by its anti-symmetric part $\frac{1}{2}(\boldsymbol{\nabla u} - (\boldsymbol{\nabla u})^T)$ can be found in [LM00]. We note that such a simple change to the model leads to a vast simplification mathematically, but, of course, it is not justified on physical grounds. Finally, global-in-time existence results for (P) in the case $\boldsymbol{f} \equiv \boldsymbol{0}$ and for initial data close to equilibrium can be found in [LLZ08].

This paper considers some finite element approximations of the Oldroyd-B system, possibly with some regularization. In the regularized case, we show (subsequence) convergence of the approximation, as the spatial and temporal discretization parameters tend to zero, and so establish the existence of global-in-time weak solutions of these regularized versions of the Oldroyd-B system. The first of these regularized problems is (P_α) obtained by adding the dissipative term $\alpha\Delta\boldsymbol{\sigma}$ for a given $\alpha \in \mathbb{R}_{>0}$ to the right-hand side of (3-I.1c), as considered computationally in [SB95], with an additional no flux boundary condition for $\boldsymbol{\sigma}$ on $\partial\mathcal{D}$. The second is (P_α^L) where, in addition to the regularization in (P_α), the conformation tensor $\boldsymbol{\sigma}$ is replaced by the cut-off $\beta^L(\boldsymbol{\sigma})$ on the right-hand side of (3-I.1a) and in the terms involving $\boldsymbol{u}$ in (3-I.1c), where $\beta^L(s) := \min\{s, L\}$ for a given $L \gg 1$. Similar regularizations have been introduced for the microscopic-macroscopic dumbbell model of dilute polymers with a finitely extensible nonlinear elastic (FENE) spring law, see [BS08], and for the convergence of the finite element approximation of such models, see [BS09]. In fact, it is argued in [BS07] and [Sch06] that the dissipative term $\alpha\Delta\boldsymbol{\sigma}$ is not a regularization, but is present in the original model with a positive $\alpha \ll 1$. Here we recall that the Oldroyd-B system is the the macroscopic closure of the microscopic-macroscopic dumbbell model with a Hookean spring law, see e.g. [BS07].

Overall the aims of this paper are threefold. First, we extend previous results in [BLM09] for a a finite element approximation of (P) using essentially the backward Euler scheme in time and based on approximating the velocity field with continuous piecewise quadratics, and the pressure and the symmetric conformation tensor by piecewise constants. We show that solutions of this numerical scheme satisfy a discrete free energy bound, which involves the logarithm of the conformation tensor, *without* any constraint on the time step, whereas a time constraint based on the initial data was required in [BLM09] in order to ensure that the approximation to the conformation tensor $\boldsymbol{\sigma}$ remained positive definite. See also [LX06], where the difficulties of maintaining the positive definiteness of approximations to $\boldsymbol{\sigma}$ are also discussed. We achieve our result by first introducing problem (P_δ), based on a regularization parameter $\delta \in \mathbb{R}_{>0}$. ($\mathrm{P}_\delta$) satisfies a regularized free energy estimate based on a regularization of ln and is valid without the positive definiteness constraint on the deformation tensor.

Second, we show that it is possible to approximate (P) with a *continuous* (piecewise linear) approximation of the conformation tensor, such that a discrete free energy bound still holds. We note that a piecewise constant approximation of the conformation tensor was necessary in [BLM09] in order to treat the advection term in (3-I.1c) and still obtain a discrete free energy bound.

Third, we show (subsequence) convergence, as the spatial and temporal discretization parameters tend to zero, of this latter approximation in the presence of the regularization terms stated above to global-in-time weak solutions of the corresponding regularized form of (P).

The outline of this paper is as follows. We end Section 3-I by introducing our notation and auxiliary results. In Section 3-II we introduce our regularizations of ln based on the parameter $\delta \in (0, \frac{1}{2}]$ and the cut-off $L \geq 2$. We introduce our regularized problem (P_δ), and show a formal free energy estimate for it. In Section 3-III, on assuming that $\mathcal{D}$ is a polytope for ease of exposition, we introduce our finite element approximation of (P_δ), ($\mathrm{P}_{\delta,h}^{\Delta t}$), based on approximating the velocity field with continuous piecewise quadratics, and the pressure and the symmetric conformation tensor by piecewise constants. Using the Brouwer fixed point theorem, we prove existence of a solution to ($\mathrm{P}_{\delta,h}^{\Delta t}$) and show that it satisfies a discrete regularized free energy estimate for any choice of time step; see Theorem 3. We conclude by showing that, in the limit $\delta \to 0_+$, these solutions of ($\mathrm{P}_{\delta,h}^{\Delta t}$) converge to a solution of ($\mathrm{P}_h^{\Delta t}$) with the approximation of the conformation tensor being positive definite. Moreover, this solution of ($\mathrm{P}_h^{\Delta t}$) satisfies a discrete free energy estimate; see Theorem 4.

In Section 3-IV we introduce our regularizations ($\mathrm{P}_\alpha^{(L)}$) of (P) involving the dissipative term $\alpha\Delta\boldsymbol{\sigma}$ on the right-hand side of (3-I.1c), and possibly the cut-off $\beta^L(\boldsymbol{\sigma})$ on certain terms involving $\boldsymbol{\sigma}$ in (3-I.1a,c). We then introduce the corresponding regularized version ($\mathrm{P}_{\alpha,\delta}^{(L)}$), and show a formal free energy estimate for it. In Section 3-V we introduce our finite element approximation of ($\mathrm{P}_{\alpha,\delta}^{(L)}$), ($\mathrm{P}_{\alpha,\delta,h}^{(L,)\Delta t}$), based on approximating the velocity field with continuous piecewise quadratics, and the pressure and the symmetric conformation tensor by continuous piecewise linears. Here we assume that $\mathcal{D}$ is a convex polytope and that the finite element mesh consists of quasi-uniform non-obtuse simplices. Using the Brouwer fixed point theorem, we prove existence of a solution to

$(\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h})$ and show that is satisfies a discrete regularized free energy estimate for any choice of time step; see Theorem 5. We conclude by showing that, in the limit $\delta \to 0_+$, these solutions of $(\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h})$ converge to a solution of $(\mathrm{P}^{(L,)\Delta t}_{\alpha,h})$ with the approximation of the conformation tensor being positive definite. Moreover, this solution of $(\mathrm{P}^{(L,)\Delta t}_{\alpha,h})$ satisfies a discrete free energy estimate; see Theorem 6.

In Section 3-VI we prove (subsequence) convergence of the solutions of $(\mathrm{P}^{L,\Delta t}_{\alpha,h})$, as the spatial and temporal discretization parameters tend to zero, to global-in-time weak solutions of $(\mathrm{P}^{L}_{\alpha})$; see Theorem 8. Finally in Section 3-VII, on further assuming that $d=2$ and a time step restriction dependent on the spatial discretization parameter, we prove (subsequence) convergence of the solutions of $(\mathrm{P}^{\Delta t}_{\alpha,h})$, as the spatial and temporal discretization parameters tend to zero, to global-in-time weak solutions of (P_{α}); see Theorem 10. We note that these existence results for $(\mathrm{P}^{(L)}_{\alpha})$ are new to the literature. In addition, the corresponding $L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H^1(\Omega))$ norms of the velocity solution $\boldsymbol{u}^{(L)}_{\alpha}$ of $(\mathrm{P}^{(L)}_{\alpha})$ are independent of the regularization parameters α (and L).

In a forthcoming paper, [BB09a], we will extend the ideas in this paper to a related macroscopic model, the FENE-P model; see [HL07], where a free energy estimate is developed for such a model, as well as Oldroyd-B. In addition, we will report in the near future on numerical computations based on the finite element approximations in this paper and [BB09a].

3-I-B Notation and auxiliary results

The absolute value and the negative part of a real number $s\in\mathbb{R}$ are denoted by $|s|:=\max\{s,-s\}$ and $[s]_-=\min\{s,0\}$, respectively. In addition to $\mathbb{R}^{d\times d}_S$, the set of symmetric $\mathbb{R}^{d\times d}$ matrices, we let $\mathbb{R}^{d\times d}_{SPD}$ be the set of symmetric positive definite $\mathbb{R}^{d\times d}$ matrices. We adopt the following notation for inner products :

$$\boldsymbol{v}\cdot\boldsymbol{w}:=\sum_{i=1}^{d}\boldsymbol{v}_i\boldsymbol{w}_i\equiv\boldsymbol{v}^T\boldsymbol{w}=\boldsymbol{w}^T\boldsymbol{v} \qquad \forall \boldsymbol{v},\boldsymbol{w}\in\mathbb{R}^d, \tag{3-I.2a}$$

$$\boldsymbol{\phi}:\boldsymbol{\psi}:=\sum_{i=1}^{d}\sum_{j=1}^{d}\boldsymbol{\phi}_{ij}\boldsymbol{\psi}_{ij}\equiv\mathrm{tr}\left(\boldsymbol{\phi}^T\boldsymbol{\psi}\right)=\mathrm{tr}\left(\boldsymbol{\psi}^T\boldsymbol{\phi}\right) \qquad \forall \boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}, \tag{3-I.2b}$$

$$\boldsymbol{\nabla\phi}::\boldsymbol{\nabla\psi}:=\sum_{i=1}^{d}\sum_{j=1}^{d}\boldsymbol{\nabla\phi}_{ij}\cdot\boldsymbol{\nabla\psi}_{ij} \qquad \forall \boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}; \tag{3-I.2c}$$

where $\cdot^T$ and $\mathrm{tr}(\cdot)$ denote transposition and trace, respectively. The corresponding norms are

$$\|\boldsymbol{v}\|:=(\boldsymbol{v}\cdot\boldsymbol{v})^{\frac{1}{2}}, \qquad \|\boldsymbol{\nabla v}\|:=(\boldsymbol{\nabla v}:\boldsymbol{\nabla v})^{\frac{1}{2}} \qquad \forall \boldsymbol{v}\in\mathbb{R}^d; \tag{3-I.3a}$$

$$\|\boldsymbol{\phi}\|:=(\boldsymbol{\phi}:\boldsymbol{\phi})^{\frac{1}{2}}, \qquad \|\boldsymbol{\nabla\phi}\|:=(\boldsymbol{\nabla\phi}::\boldsymbol{\nabla\phi})^{\frac{1}{2}}, \qquad \forall \boldsymbol{\phi}\in\mathbb{R}^{d\times d}. \tag{3-I.3b}$$

We will use on several occasions that $\mathrm{tr}(\boldsymbol{\phi})=\mathrm{tr}(\boldsymbol{\phi}^T)$ and $\mathrm{tr}(\boldsymbol{\phi\psi})=\mathrm{tr}(\boldsymbol{\psi\phi})$ for all $\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}$. In particular, we note that :

$$\boldsymbol{\phi\chi}^T:\boldsymbol{\psi}=\boldsymbol{\chi\phi}:\boldsymbol{\psi}=\boldsymbol{\chi}:\boldsymbol{\psi\phi} \qquad \forall \boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}_S,\ \boldsymbol{\chi}\in\mathbb{R}^{d\times d}, \tag{3-I.4a}$$

$$\|\boldsymbol{\psi\phi}\|\le\|\boldsymbol{\psi}\|\,\|\boldsymbol{\phi}\| \qquad \forall \boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}. \tag{3-I.4b}$$

In addition, for any $\boldsymbol{\phi}\in\mathbb{R}^{d\times d}_S$, there exists a diagonal decomposition

$$\boldsymbol{\phi}=\mathbf{O}^T\mathbf{D}\mathbf{O} \qquad \Rightarrow \qquad \mathrm{tr}\,(\boldsymbol{\phi})=\mathrm{tr}\,(\mathbf{D})\,, \tag{3-I.5}$$

where $\mathbf{O}\in\mathbb{R}^{d\times d}$ is an orthogonal matrix and $\mathbf{D}\in\mathbb{R}^{d\times d}$ a diagonal matrix. Hence, for any $g:\mathbb{R}\to\mathbb{R}$, one can define $g(\boldsymbol{\phi})\in\mathbb{R}^{d\times d}_S$ as

$$g(\boldsymbol{\phi}):=\mathbf{O}^T g(\mathbf{D})\mathbf{O} \qquad \Rightarrow \qquad \mathrm{tr}\,(g(\boldsymbol{\phi}))=\mathrm{tr}\,(g(\mathbf{D}))\,, \tag{3-I.6}$$

where $g(\mathbf{D})\in\mathbb{R}^{d\times d}_S$ is the diagonal matrix with entries $[g(\mathbf{D})]_{ii}=g(\mathbf{D}_{ii})$, $i=1\to d$. Although the diagonal decomposition (3-I.5) is not unique, (3-I.6) uniquely defines $g(\boldsymbol{\phi})$. Similarly, one can define $g(\boldsymbol{\phi})\in\mathbb{R}^{d\times d}_{SPD}$, when $\boldsymbol{\phi}\in\mathbb{R}^{d\times d}_{SPD}$ and $g:\mathbb{R}_{>0}\to\mathbb{R}$. We note for later purposes that

$$d^{-1}(\mathrm{tr}(|\boldsymbol{\phi}|))^2\le\|\boldsymbol{\phi}\|^2\le(\mathrm{tr}(|\boldsymbol{\phi}|))^2 \qquad \forall \boldsymbol{\phi}\in\mathbb{R}^{d\times d}_S. \tag{3-I.7}$$

We adopt the standard notation for Sobolev spaces, e.g. $H^1(\mathcal{D}) := \{\eta : \mathcal{D} \mapsto \mathbb{R} : \int_{\mathcal{D}} [|\eta|^2 + \|\nabla \eta\|^2] < \infty\}$ with $H^1_0(\mathcal{D})$ being the closure of $C^\infty_0(\mathcal{D})$ for the corresponding norm $\|\cdot\|_{H^1(\mathcal{D})}$. We denote the associated semi-norm as $|\cdot|_{H^1(\mathcal{D})}$. The topological dual of the Hilbert space $H^1_0(\mathcal{D})$, with pivot space $L^2(\mathcal{D})$, will be denoted by $H^{-1}(\mathcal{D})$. We denote the duality pairing between $H^{-1}(\mathcal{D})$ and $H^1_0(\mathcal{D})$ as $\langle\cdot,\cdot\rangle_{H^1_0(\mathcal{D})}$. Such function spaces are naturally extended when the range $\mathbb{R}$ is replaced by $\mathbb{R}^d$, $\mathbb{R}^{d\times d}$ and $\mathbb{R}^{d\times d}_S$; e.g. $H^1(\mathcal{D})$ becomes $[H^1(\mathcal{D})]^d$, $[H^1(\mathcal{D})]^{d\times d}$ and $[H^1(\mathcal{D})]^{d\times d}_S$, respectively. For ease of notation, we write the corresponding norms and semi-norms as $\|\cdot\|_{H^1(\mathcal{D})}$ and $|\cdot|_{H^1(\mathcal{D})}$, respectively, as opposed to e.g. $\|\cdot\|_{[H^1(\mathcal{D})]^d}$ and $|\cdot|_{[H^1(\mathcal{D})]^d}$, respectively. Similarly, we write $\langle\cdot,\cdot\rangle_{H^1_0(\mathcal{D})}$ for the duality pairing between e.g. $[H^{-1}(\mathcal{D})]^d$ and $[H^1_0(\mathcal{D})]^d$. We recall the Poincaré inequality

$$\int_{\mathcal{D}} \|\boldsymbol{v}\|^2 \le C_P \int_{\mathcal{D}} \|\boldsymbol{\nabla v}\|^2 \qquad \forall \boldsymbol{v} \in [H^1_0(D)]^d, \tag{3-I.8}$$

where $C_P \in \mathbb{R}_{>0}$ depends only on $\mathcal{D}$. For notational convenience, we introduce also convex sets such as $[H^1(\mathcal{D})]^{d\times d}_{SPD} := \{\boldsymbol{\phi} \in [H^1(\mathcal{D})]^{d\times d}_S : \boldsymbol{\phi} \in \mathbb{R}^{d\times d}_{SPD}$ a.e. in $\mathcal{D}\}$. Moreover, in order to analyse (P), we adopt the notation

$$\begin{aligned} &\mathrm{W} := [H^1_0(\mathcal{D})]^d, \qquad \mathrm{Q} := L^2(\mathcal{D}), \qquad \mathrm{V} := \left\{ \boldsymbol{v} \in \mathrm{W} : \int_{\mathcal{D}} q \operatorname{div} \boldsymbol{v} = 0 \quad \forall q \in \mathrm{Q} \right\}, \\ &\mathrm{S} := [L^1(\mathcal{D})]^{d\times d}_S \qquad \text{and} \qquad \mathrm{S}_{PD} := [L^1(\mathcal{D})]^{d\times d}_{SPD}. \end{aligned} \tag{3-I.9}$$

Finally, throughout the paper C will denote a generic positive constant independent of the regularization parameters δ, L and α; and the mesh parameters h and Δt.

3-II Formal free energy estimates for a regularized problem (P_δ)

3-II-A Some regularizations

Let $G : s \in \mathbb{R}_{>0} \mapsto \ln s \in \mathbb{R}$ be the logarithm function, whose domain of definition can be straightforwardly extended to the set of symmetric positive definite matrices using (3-I.5). We define the following two concave $C^1(\mathbb{R})$ regularizations of G based on given parameters $L > 1 > \delta > 0$:

$$\begin{aligned} &G_\delta : s \in \mathbb{R} \mapsto \begin{cases} G(s), & \forall s \ge \delta \\ \frac{s}{\delta} + G(\delta) - 1, & \forall s \le \delta \end{cases} \qquad \text{and} \qquad G^L_\delta : s \in \mathbb{R} \mapsto \begin{cases} G^L(s), & \forall s \ge \delta \\ G_\delta(s), & \forall s \le \delta \end{cases}, \\ &\text{where} \qquad G^L : s \in \mathbb{R}_{>0} \mapsto \begin{cases} \frac{s}{L} + G(L) - 1, & \forall s \ge L \\ G(s), & \forall s \in (0, L] \end{cases}. \end{aligned} \tag{3-II.1}$$

We define also the following scalar functions

$$\beta^{(L)}_\delta(s) := \left(G^{(L)\prime}_\delta(s) \right)^{-1} \qquad \forall s \in \mathbb{R} \qquad \text{and} \qquad \beta^{(L)}(s) := \left(G^{(L)\prime}(s) \right)^{-1} \qquad \forall s \in \mathbb{R}_{>0}; \tag{3-II.2}$$

where, here and throughout this paper, $\cdot^{(\star)}$ denotes an expression with or without the superscript $\star$, and a similar convention with subscripts. Hence we have that

$$\begin{aligned} &\beta_\delta : s \in \mathbb{R} \mapsto \max\{s, \delta\}, \qquad \beta^L_\delta : s \in \mathbb{R} \mapsto \min\{\beta_\delta(s), L\}, \\ &\beta : s \in \mathbb{R}_{>0} \mapsto s \qquad \text{and} \qquad \beta^L : s \in \mathbb{R}_{>0} \mapsto \min\{\beta(s), L\}. \end{aligned} \tag{3-II.3}$$

We note for example that

$$\left\| \beta^L_\delta(\boldsymbol{\phi}) \right\|^2 \le dL^2 \quad \forall \boldsymbol{\phi} \in \mathbb{R}^{d\times d}_S \quad \text{and} \quad \left\| \beta^L(\boldsymbol{\phi}) \right\|^2 \le dL^2 \quad \forall \boldsymbol{\phi} \in \mathbb{R}^{d\times d}_{SPD}. \tag{3-II.4}$$

Introducing the concave $C^1(\mathbb{R})$ functions

$$H^L_\delta(s) := G^{\delta^{-1}}_{L^{-1}}(s) \qquad \forall s \in \mathbb{R} \qquad \text{and} \qquad H_\delta(s) := G^{\delta^{-1}}(s) \qquad \forall s \in \mathbb{R}_{>0}, \tag{3-II.5}$$

it follows from (3-II.1) and (3-II.3) that

$$H^{(L)\prime}_\delta(G^{(L)\prime}_\delta(s)) = \beta^{(L)}_\delta(s) \qquad \forall s \in \mathbb{R}. \tag{3-II.6}$$

For later purposes, we prove the following results concerning these functions.

Lemma 10. *The following hold for any $\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}_S$ and for any $L>1>\delta>0$ that*

$$[\beta^{(L)}_\delta(\boldsymbol{\phi})][G^{(L)'}_\delta(\boldsymbol{\phi})]=[G^{(L)'}_\delta(\boldsymbol{\phi})][\beta^{(L)}_\delta(\boldsymbol{\phi})]=\boldsymbol{I}, \tag{3-II.7a}$$

$$\operatorname{tr}\left(\beta^{(L)}_\delta(\boldsymbol{\phi})+[\beta^{(L)}_\delta(\boldsymbol{\phi})]^{-1}-2\boldsymbol{I}\right)\geq 0, \tag{3-II.7b}$$

$$\operatorname{tr}\left(\boldsymbol{\phi}-G^{(L)}_\delta(\boldsymbol{\phi})-\boldsymbol{I}\right)\geq 0, \tag{3-II.7c}$$

$$\left(\boldsymbol{\phi}-\beta^{(L)}_\delta(\boldsymbol{\phi})\right):\left(\boldsymbol{I}-G^{(L)'}_\delta(\boldsymbol{\phi})\right)\geq 0, \tag{3-II.7d}$$

$$(\boldsymbol{\phi}-\boldsymbol{\psi}):\left(G^{(L)'}_\delta(\boldsymbol{\psi})\right)\geq\operatorname{tr}\left(G^{(L)}_\delta(\boldsymbol{\phi})-G^{(L)}_\delta(\boldsymbol{\psi})\right), \tag{3-II.7e}$$

$$-(\boldsymbol{\phi}-\boldsymbol{\psi}):\left(G^{(L)'}_\delta(\boldsymbol{\phi})-G^{(L)'}_\delta(\boldsymbol{\psi})\right)\geq\delta^2\left\|G^{(L)'}_\delta(\boldsymbol{\phi})-G^{(L)'}_\delta(\boldsymbol{\psi})\right\|^2. \tag{3-II.7f}$$

In addition, if $\delta\in(0,\frac{1}{2}]$ and $L\geq 2$ we have that

$$\operatorname{tr}\left(\boldsymbol{\phi}-G^{(L)}_\delta(\boldsymbol{\phi})\right)\geq\begin{cases}\frac{1}{2}\|\boldsymbol{\phi}\|\\ \frac{1}{2\delta}\|[\boldsymbol{\phi}]_-\|\end{cases}\quad\text{and}\quad\boldsymbol{\phi}:\left(\boldsymbol{I}-G^{(L)'}_\delta(\boldsymbol{\phi})\right)\geq\tfrac{1}{2}\|\boldsymbol{\phi}\|-d. \tag{3-II.8}$$

Proof. The result (3-II.7a) follows immediately from (3-I.6) and as $\beta^{(L)}_\delta(s)=G^{(L)'}_\delta(s)$ for all $s\in\mathbb{R}$. The desired results (3-II.7b–d) follow similarly, on noting the scalar inequalities $\beta^{(L)}_\delta(s)+[\beta^{(L)}_\delta(s)]^{-1}\geq 2$, $s-G^{(L)}_\delta(s)\geq 1$ and $(s-\beta^{(L)}_\delta(s))(1-G^{(L)'}_\delta(s))\geq 0$ for all $s\in\mathbb{R}$.

We note that $G^{(L)}_\delta$ are concave functions like G, and hence they satisfy the following inequality

$$(s_1-s_2)G^{(L)'}_\delta(s_2)\geq G^{(L)}_\delta(s_1)-G^{(L)}_\delta(s_2)\qquad\forall s_1,s_2\in\mathbb{R}; \tag{3-II.9}$$

Hence for any $\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}^{d\times d}_S$ with $\boldsymbol{\phi}=\mathbf{O}^T_\phi\mathbf{D}_\phi\mathbf{O}_\phi$ and $\mathbf{O}^T_\psi\mathbf{D}_\psi\mathbf{O}_\psi$, where $\mathbf{O}_\phi,\mathbf{O}_\psi\in\mathbb{R}^{d\times d}$ orthogonal and $\mathbf{D}_\phi,\mathbf{D}_\psi\in\mathbb{R}^{d\times d}$ diagonal, we have, on noting the properties of trace, that

$$(\boldsymbol{\phi}-\boldsymbol{\psi}):G^{(L)'}_\delta(\boldsymbol{\psi})=\operatorname{tr}\left((\boldsymbol{\phi}-\boldsymbol{\psi})G^{(L)'}_\delta(\boldsymbol{\psi})\right)=\operatorname{tr}\left((\mathbf{O}^T\mathbf{D}_\phi\mathbf{O}-\mathbf{D}_\psi)G^{(L)'}_\delta(\mathbf{D}_\psi)\right), \tag{3-II.10}$$

where $\mathbf{O}=\mathbf{O}_\phi\mathbf{O}^T_\psi\in\mathbb{R}^{d\times d}$ is orthogonal and hence $\sum_{i=1}^d[\mathbf{O}_{ij}]^2=\sum_{i=1}^d[\mathbf{O}_{ji}]^2=1$ for $j=1\to d$. Therefore we have, on noting these properties of $\mathbf{O}$, (3-II.9) and (3-I.6), that

$$\begin{aligned}\operatorname{tr}\left((\mathbf{O}^T\mathbf{D}_\phi\mathbf{O}-\mathbf{D}_\psi)G^{(L)'}_\delta(\mathbf{D}_\psi)\right)&=\sum_{i=1}^d\left(\sum_{j=1}^d[\mathbf{O}_{ji}]^2[\mathbf{D}_\phi]_{jj}-[\mathbf{D}_\psi]_{ii}\right)[G^{(L)'}_\delta(\mathbf{D}_\psi)]_{ii}\\&=\sum_{i=1}^d\sum_{j=1}^d[\mathbf{O}_{ji}]^2\left([\mathbf{D}_\phi]_{jj}-[\mathbf{D}_\psi]_{ii}\right)[G^{(L)'}_\delta(\mathbf{D}_\psi)]_{ii}\\&\geq\sum_{i=1}^d\sum_{j=1}^d[\mathbf{O}_{ji}]^2\left([G^{(L)}_\delta(\mathbf{D}_\phi)]_{jj}-[G^{(L)}_\delta(\mathbf{D}_\psi)]_{ii}\right)\\&=\operatorname{tr}\left(G^{(L)}_\delta(\mathbf{D}_\phi)\right)-\operatorname{tr}\left(G^{(L)}_\delta(\mathbf{D}_\psi)\right)\\&=\operatorname{tr}\left(G^{(L)}_\delta(\boldsymbol{\phi})-G^{(L)}_\delta(\boldsymbol{\psi})\right).\end{aligned} \tag{3-II.11}$$

Combining (3-II.10) and (3-II.11) yields the desired result (3-II.7e).

We note that $-G^{(L)'}_\delta\in C^{0,1}(\mathbb{R})$ is monotonically increasing with Lipschitz constant δ^{-2} and so

$$-(s_1-s_2)(G^{(L)'}_\delta(s_1)-G^{(L)'}_\delta(s_2))\geq\delta^2[G^{(L)'}_\delta(s_1)-G^{(L)'}_\delta(s_2)]^2\qquad\forall s_1,s_2\in\mathbb{R}. \tag{3-II.12}$$

Then, similarly to (3-II.10) and (3-II.11), we have, on noting (3-II.12), that

$$
\begin{aligned}
-(\boldsymbol{\phi}-\boldsymbol{\psi}):(G_\delta^{(L)'}(\boldsymbol{\phi})-G_\delta^{(L)'}(\boldsymbol{\psi})) \qquad & (3\text{-II.13})\\
&=-\Big[\mathrm{tr}\Big((\mathbf{D}_\phi-\mathbf{O}\mathbf{D}_\psi\mathbf{O}^T)G_\delta^{(L)'}(\mathbf{D}_\phi)\Big)-\mathrm{tr}\Big((\mathbf{O}^T\mathbf{D}_\phi\mathbf{O}-\mathbf{D}_\psi)G_\delta^{(L)'}(\mathbf{D}_\psi)\Big)\Big]\\
&=-\sum_{i=1}^{d}\sum_{j=1}^{d}[\mathbf{O}_{ji}]^2([\mathbf{D}_\phi]_{jj}-[\mathbf{D}_\psi]_{ii})([G_\delta^{(L)'}(\mathbf{D}_\phi)]_{jj}-[G_\delta^{(L)'}(\mathbf{D}_\psi)]_{ii})\\
&\geq\delta^2\sum_{i=1}^{d}\sum_{j=1}^{d}[\mathbf{O}_{ji}]^2([G_\delta^{(L)'}(\mathbf{D}_\phi)]_{jj}-[G_\delta^{(L)'}(\mathbf{D}_\psi)]_{ii})^2\\
&=\delta^2\mathrm{tr}\Big((G_\delta^{(L)'}(\boldsymbol{\phi})-G_\delta^{(L)'}(\boldsymbol{\psi}))^2\Big)=\delta^2\,\|G_\delta^{(L)'}(\boldsymbol{\phi})-G_\delta^{(L)'}(\boldsymbol{\psi})\|^2
\end{aligned}
$$

and hence the desired result (3-II.7f).

Finally the results (3-II.8) follow from (3-I.6) and (3-I.7) on noting the following scalar inequalities

$$
s-G_\delta^{(L)}(s)\geq\begin{cases}\frac{1}{2}\,|s|\\ \frac{1}{2\delta}\,|[s]_-|\end{cases}\qquad\text{and}\qquad s(1-G_\delta^{(L)'}(s))\geq\tfrac{1}{2}|s|-1\qquad\forall s\in\mathbb{R}, \tag{3-II.14}
$$

which are easily deduced if $\delta\in(0,\frac{1}{2}]$ and $L\geq 2$. □

Clearly (3-II.7e) holds for any concave function $g\in C^1(\mathbb{R})$, not just $G_\delta^{(L)}$, and this implies that

$$
(\boldsymbol{\phi}-\boldsymbol{\psi}):g'(\boldsymbol{\psi})\geq\mathrm{tr}\,(g(\boldsymbol{\phi})-g(\boldsymbol{\psi}))\geq(\boldsymbol{\phi}-\boldsymbol{\psi}):g'(\boldsymbol{\phi})\qquad\forall\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}_S^{d\times d}. \tag{3-II.15}
$$

For a convex function $g\in C^1(\mathbb{R})$, the inequalities in (3-II.15) are reversed. Hence for any concave or convex function $g\in C^1(\mathbb{R})$ and for any $\boldsymbol{\phi}\in C^1([0,T);\mathbb{R}_S^{d\times d})$ one can deduce from the above that

$$
\frac{d}{dt}\mathrm{tr}\,(g(\boldsymbol{\phi}))=\mathrm{tr}\left(\frac{d\boldsymbol{\phi}}{dt}g'(\boldsymbol{\phi})\right)=\left(\frac{d\boldsymbol{\phi}}{dt}\right):g'(\boldsymbol{\phi})\qquad\forall t\in[0,T). \tag{3-II.16}
$$

Of course, a similar result holds true for spatial derivatives. Furthermore, these results hold true if $\boldsymbol{\phi}$ is in addition positive definite, and $g\in C^1(\mathbb{R}_{>0})$ is a concave or convex function. Finally, we note that one can use the approach in (3-II.11) to show that if $g\in C^{0,1}(\mathbb{R})$ with Lipschitz constant λ, then

$$
\|g(\boldsymbol{\phi})-g(\boldsymbol{\psi})\|\leq\lambda\|\boldsymbol{\phi}-\boldsymbol{\psi}\|\qquad\forall\boldsymbol{\phi},\boldsymbol{\psi}\in\mathbb{R}_S^{d\times d}. \tag{3-II.17}
$$

3-II-B Regularized problem ($\mathbf{P}_\delta$)

Using the regularizations G_δ introduced above with parameter δ we consider the following regularization of (P) for a given $\delta\in(0,\frac{1}{2}]$:

($\mathbf{P}_\delta$) Find $\boldsymbol{u}_\delta:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\mapsto\boldsymbol{u}_\delta(t,\boldsymbol{x})\in\mathbb{R}^d$, $p_\delta:(t,\boldsymbol{x})\in(0,T)\times\mathcal{D}\mapsto p_\delta(t,\boldsymbol{x})\in\mathbb{R}$ and $\boldsymbol{\sigma}_\delta:(t,\boldsymbol{x})\in[0,T)\times\mathcal{D}\mapsto\boldsymbol{\sigma}_\delta(t,\boldsymbol{x})\in\mathbb{R}_S^{d\times d}$ such that

$$
\mathrm{Re}\left(\frac{\partial\boldsymbol{u}_\delta}{\partial t}+(\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})\boldsymbol{u}_\delta\right)=-\boldsymbol{\nabla}p_\delta+(1-\varepsilon)\Delta\boldsymbol{u}_\delta+\frac{\varepsilon}{\mathrm{Wi}}\,\mathrm{div}\,\beta_\delta(\boldsymbol{\sigma}_\delta)+\boldsymbol{f}\qquad\text{on }\mathcal{D}_T, \tag{3-II.18a}
$$

$$
\mathrm{div}\,\boldsymbol{u}_\delta=0\qquad\text{on }\mathcal{D}_T, \tag{3-II.18b}
$$

$$
\frac{\partial\boldsymbol{\sigma}_\delta}{\partial t}+(\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}_\delta=(\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)+\beta_\delta(\boldsymbol{\sigma}_\delta)(\boldsymbol{\nabla}\boldsymbol{u}_\delta)^T-\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_\delta-\boldsymbol{I})\qquad\text{on }\mathcal{D}_T, \tag{3-II.18c}
$$

$$
\boldsymbol{u}_\delta(0,\boldsymbol{x})=\boldsymbol{u}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D}, \tag{3-II.18d}
$$

$$
\boldsymbol{\sigma}_\delta(0,\boldsymbol{x})=\boldsymbol{\sigma}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D}, \tag{3-II.18e}
$$

$$
\boldsymbol{u}_\delta=\mathbf{0}\qquad\text{on }(0,T)\times\partial\mathcal{D}. \tag{3-II.18f}
$$

3-II-C Formal energy estimates for (P_δ)

In this section, we derive *formal* energy estimates, see e.g. (3-II.21) below, where we will assume that the triple $(\boldsymbol{u}_\delta, p_\delta, \boldsymbol{\sigma}_\delta)$, which is a solution to problem (P_δ), has sufficient regularity for all the subsequent manipulations.

We will assume throughout that

$$\begin{aligned} &\boldsymbol{f} \in L^2\left(0,T;[H^{-1}(\mathcal{D})]^d\right), \qquad \boldsymbol{u}^0 \in [L^2(\mathcal{D})]^d, \qquad \text{and} \qquad \boldsymbol{\sigma}^0 \in [L^\infty(\mathcal{D})]^{d\times d}_{SPD} \quad \text{with} \\ &\sigma^0_{\min}\|\boldsymbol{\xi}\|^2 \le \boldsymbol{\xi}^T\boldsymbol{\sigma}^0(\boldsymbol{x})\boldsymbol{\xi} \le \sigma^0_{\max}\|\boldsymbol{\xi}\|^2 \qquad \forall \boldsymbol{\xi}\in\mathbb{R}^d \qquad \text{for } a.e.\ \boldsymbol{x} \text{ in } \mathcal{D}; \end{aligned} \tag{3-II.19}$$

where $\sigma^0_{\min}, \sigma^0_{\max} \in \mathbb{R}_{>0}$.

Let $F_\delta(\boldsymbol{u}_\delta, \boldsymbol{\sigma}_\delta)$ denote the free energy of the solution $(\boldsymbol{u}_\delta, p_\delta, \boldsymbol{\sigma}_\delta)$ to problem (P_δ), where $F_\delta(\cdot,\cdot) : \mathrm{W}\times\mathrm{S} \mapsto \mathbb{R}$ is defined as

$$F_\delta(\boldsymbol{v},\boldsymbol{\phi}) := \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} \|\boldsymbol{v}\|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(\boldsymbol{\phi} - G_\delta(\boldsymbol{\phi}) - \boldsymbol{I}). \tag{3-II.20}$$

Here the first term $\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{v}\|^2$ corresponds to the usual kinetic energy term, and the second term $\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\phi} - G_\delta(\boldsymbol{\phi}) - \boldsymbol{I})$ is a regularized version of the relative entropy term $\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\phi} - G(\boldsymbol{\phi}) - \boldsymbol{I})$ introduced in [HL07], see also [JLBLO06].

Proposition 12. *Let $(\boldsymbol{u}_\delta, p_\delta, \boldsymbol{\sigma}_\delta)$ be a sufficiently smooth solution to problem (P_δ). Then the free energy $F_\delta(\boldsymbol{u}_\delta, \boldsymbol{\sigma}_\delta)$ satisfies for a.a. $t \in (0,T)$*

$$\frac{d}{dt}F_\delta(\boldsymbol{u}_\delta,\boldsymbol{\sigma}_\delta) + (1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_\delta\|^2 + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_\delta) + [\beta_\delta(\boldsymbol{\sigma}_\delta)]^{-1} - 2\boldsymbol{I}) \le \langle \boldsymbol{f}, \boldsymbol{u}_\delta\rangle_{H^1_0(\mathcal{D})}. \tag{3-II.21}$$

Proof. Multiplying the Navier-Stokes equation with $\boldsymbol{u}_\delta$ and the stress equation with $\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta))$, summing and integrating over $\mathcal{D}$ yields, after using integrations by parts and the incompressibility property in the standard way, that

$$\begin{aligned} &\int_{\mathcal{D}}\left[\frac{\mathrm{Re}}{2}\frac{\partial}{\partial t}\|\boldsymbol{u}_\delta\|^2 + (1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}_\delta\|^2 + \frac{\varepsilon}{\mathrm{Wi}}\beta_\delta(\boldsymbol{\sigma}_\delta) : \boldsymbol{\nabla}\boldsymbol{u}_\delta\right] \\ &\quad + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left[\left(\frac{\partial}{\partial t}\boldsymbol{\sigma}_\delta + (\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}_\delta\right) + \frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_\delta - \boldsymbol{I})\right] : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) \\ &\quad - \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta) + \beta_\delta(\boldsymbol{\sigma}_\delta)(\boldsymbol{\nabla}\boldsymbol{u}_\delta)^T\right) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) = \langle \boldsymbol{f}, \boldsymbol{u}_\delta\rangle_{H^1_0(\mathcal{D})}. \end{aligned} \tag{3-II.22}$$

Using (3-II.16) and its spatial counterpart, we first note that

$$\left(\frac{\partial}{\partial t}\boldsymbol{\sigma}_\delta + (\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}_\delta\right) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) = \left(\frac{\partial}{\partial t} + (\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})\right)\mathrm{tr}\left(\boldsymbol{\sigma}_\delta - G_\delta(\boldsymbol{\sigma}_\delta)\right). \tag{3-II.23}$$

On integrating over $\mathcal{D}$, the $(\boldsymbol{u}_\delta\cdot\boldsymbol{\nabla})$ part of this term vanishes as $\boldsymbol{u}_\delta(t,\cdot)\in\mathrm{V}$. On using trace properties, (3-II.7a) and the incompressibility property, we obtain that

$$\begin{aligned} ((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) &= \mathrm{tr}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta) - (\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)G'_\delta(\boldsymbol{\sigma}_\delta)\right), \\ &= \mathrm{tr}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta) - \boldsymbol{\nabla}\boldsymbol{u}_\delta\right), \\ &= \mathrm{tr}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)\right) - \mathrm{div}\,\boldsymbol{u}_\delta, \\ &= \mathrm{tr}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)\right). \end{aligned} \tag{3-II.24}$$

On noting (3-I.4a), we have also that

$$\left(\beta_\delta(\boldsymbol{\sigma}_\delta)(\boldsymbol{\nabla}\boldsymbol{u}_\delta)^T\right) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) = \mathrm{tr}\left((\boldsymbol{\nabla}\boldsymbol{u}_\delta)\beta_\delta(\boldsymbol{\sigma}_\delta)\right). \tag{3-II.25}$$

Therefore the terms involving the left-hand sides of (3-II.24) and (3-II.25) in (3-II.22) cancel with the term $\frac{\varepsilon}{\mathrm{Wi}}\beta_\delta(\boldsymbol{\sigma}_\delta) : \boldsymbol{\nabla}\boldsymbol{u}_\delta$ in (3-II.22) arising from the Navier-Stokes equation. Finally, for the remaining term we have on noting (3-I.2b) and (3-II.7a,d) that

$$\begin{aligned} (\boldsymbol{\sigma}_\delta - \boldsymbol{I}) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) &= [(\boldsymbol{\sigma}_\delta - \beta_\delta(\boldsymbol{\sigma}_\delta)) + (\beta_\delta(\boldsymbol{\sigma}_\delta) - \boldsymbol{I})] : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) \\ &\ge (\beta_\delta(\boldsymbol{\sigma}_\delta) - \boldsymbol{I}) : (\boldsymbol{I} - G'_\delta(\boldsymbol{\sigma}_\delta)) = \mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_\delta) + [\beta_\delta(\boldsymbol{\sigma}_\delta)]^{-1} - 2\boldsymbol{I}). \end{aligned} \tag{3-II.26}$$

Hence we obtain the desired free energy inequality (3-II.21). □

Corollary 1. *Let $(\boldsymbol{u}_\delta,p_\delta,\boldsymbol{\sigma}_\delta)$ be a sufficiently smooth solution to problem (P_δ). Then it follows that*

$$\sup_{t\in(0,T)} F_\delta(\boldsymbol{u}_\delta(t,\cdot),\boldsymbol{\sigma}_\delta(t,\cdot))+\int_{\mathcal{D}_T}\left[\frac{1}{2}(1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}_\delta\|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_\delta)+[\beta_\delta(\boldsymbol{\sigma}_\delta)]^{-1}-2\boldsymbol{I})\right] \qquad \text{(3-II.27)}$$
$$\leq 2\left(F_\delta(\boldsymbol{u}^0,\boldsymbol{\sigma}^0)+\frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}\|^2_{L^2(0,T;H^{-1}(\mathcal{D}))}\right).$$

Proof. Smooth solutions $(\boldsymbol{u}_\delta,p_\delta,\boldsymbol{\sigma}_\delta)$ of (P_δ) satisfy the free energy estimate (3-II.20). One can bound the term $\langle\boldsymbol{f},\boldsymbol{u}_\delta\rangle_{H^1_0(\mathcal{D})}$ there, using the Cauchy-Schwarz and Young inequalities for $\nu\in\mathbb{R}_{>0}$, and the Poincaré inequality (3-I.8), by

$$\langle\boldsymbol{f},\boldsymbol{u}_\delta\rangle_{H^1_0(\mathcal{D})}\leq\|\boldsymbol{f}\|_{H^{-1}(\mathcal{D})}\|\boldsymbol{u}_\delta\|_{H^1(\mathcal{D})}\leq\frac{1}{2\nu^2}\|\boldsymbol{f}\|^2_{H^{-1}(\mathcal{D})}+\frac{\nu^2}{2}\|\boldsymbol{u}_\delta\|^2_{H^1(\mathcal{D})} \qquad \text{(3-II.28)}$$
$$\leq\frac{1}{2\nu^2}\|\boldsymbol{f}\|^2_{H^{-1}(\mathcal{D})}+\frac{\nu^2}{2}(1+C_P)\|\boldsymbol{\nabla}\boldsymbol{u}_\delta\|^2_{L^2(\mathcal{D})}.$$

Combining (3-II.28) and (3-II.20) with $\nu^2=(1-\varepsilon)/(1+C_P)$, and integrating in time yields the desired result (3-II.27). □

We note that the right-hand side of (3-II.27) is independent of the regularization parameter δ if $\boldsymbol{\sigma}^0$ is positive definite.

3-III Finite element approximation of (P_δ) and (P)

3-III-A Finite element discretization

We now introduce a finite element discretization of the problem (P_δ), which satisfies a discrete analogue of (3-II.21).

The time interval $[0,T)$ is split into intervals $[t^{n-1},t^n)$ with $\Delta t_n=t^n-t^{n-1}$, $n=1,\dots,N_T$. We set $\Delta t:=\max_{n=1,\dots,N_T}\Delta t_n$. We will assume throughout, for ease of exposition, that the domain $\mathcal{D}$ is a polytope. We define a regular family of meshes $\{\mathcal{T}_h\}_{h>0}$ with discretization parameter $h>0$, which is built from partitionings of the domain $\mathcal{D}$ into regular open simplices so that

$$\mathcal{D}\subset\mathcal{T}_h:=\bigcup_{k=1}^{N_K}\overline{K_k}.$$

We denote by h_{K_k} the diameter of the simplex K_k, so that $h=\max_{k=1,\dots,N_K}h_{K_k}$. For each element K_k of the mesh $\mathcal{T}_h$, we denote by $\boldsymbol{n}_{K_k}$ the outward unit normal vector of boundary ∂K_k of K_k. We introduce also $\partial\mathcal{T}_h:=\{E_j\}_{j=1}^{N_E}$ as the set of internal edges E_j of triangles in the mesh $\mathcal{T}_h$ when $d=2$, or the set of internal faces E_j of tetrahedra when $d=3$.

We approximate the problem (P_δ) by the problem $(\mathrm{P}^{\Delta t}_{\delta,h})$ based on the finite element spaces $\mathrm{W}_h\times\mathrm{Q}^0_h\times\mathrm{S}^0_h$ satisfying the following. As is standard, the discrete velocity-pressure spaces $\mathrm{W}_h\times\mathrm{Q}^0_h\subset\mathrm{W}\times\mathrm{Q}$ satisfy the discrete Ladyshenskaya-Babuška-Brezzi (LBB) inf-sup condition

$$\inf_{q\in\mathrm{Q}^0_h}\sup_{\boldsymbol{v}\in\mathrm{W}_h}\frac{(\mathrm{div}\,\boldsymbol{v},q)}{\|q\|_{L^2(\mathcal{D})}\|\boldsymbol{v}\|_{H^1(\mathcal{D})}}\geq\mu_\star>0. \qquad \text{(3-III.1)}$$

In the following, we set

$$\mathrm{W}_h:=\{\boldsymbol{v}\in[C(\overline{\mathcal{D}})]^d\cap\mathrm{W}:\boldsymbol{v}|_{K_k}\in[\mathbb{P}_2]^d\quad k=1,\dots,N_K\}\subset\mathrm{W}, \qquad \text{(3-III.2a)}$$
$$\mathrm{Q}^0_h:=\{q\in\mathrm{Q}:q|_{K_k}\in\mathbb{P}_0\quad k=1,\dots,N_K\}\subset\mathrm{Q}, \qquad \text{(3-III.2b)}$$
$$\text{and}\qquad \mathrm{S}^0_h:=\{\boldsymbol{\phi}\in\mathrm{S}:\boldsymbol{\phi}|_{K_k}\in[\mathbb{P}_0]^{d\times d}_S\quad k=1,\dots,N_K\}\subset\mathrm{S}; \qquad \text{(3-III.2c)}$$

where $\mathbb{P}_m$ denotes polynomials of maximal degree m in $\boldsymbol{x}$. We introduce also

$$\mathrm{V}^0_h:=\left\{\boldsymbol{v}\in\mathrm{W}_h:\int_{\mathcal{D}}q\,\mathrm{div}\,\boldsymbol{v}=0\quad\forall q\in\mathrm{Q}^0_h\right\},$$

which approximates V. It is well-known the choice (3-III.2a,b) satisfies (3-III.1), see e.g. [BF91]. Moreover, this particular choice of S_h^0 has the desirable property that

$$\boldsymbol{\phi}\in\mathrm{S}_h^0 \qquad \Rightarrow \qquad \boldsymbol{I}-G_\delta'(\boldsymbol{\phi})\in\mathrm{S}_h^0 \qquad \text{and} \qquad \mathrm{tr}\,(\boldsymbol{\phi}-G_\delta(\boldsymbol{\phi})-\boldsymbol{I})\in\mathrm{Q}_h^0\,. \tag{3-III.3}$$

Since S_h^0 is discontinuous, we will use the discontinuous Galerkin method to approximate the advection term $(\boldsymbol{u}_\delta\cdot\nabla)\boldsymbol{\sigma}_\delta$ in the following. Then, for the boundary integrals, we will make use of the following definitions (see e.g. [EG04, p267]). Given $\boldsymbol{v}\in\mathrm{W}_h$, then for any $\boldsymbol{\phi}\in\mathrm{S}_h^0$ (or Q_h^0) and for any point $\boldsymbol{x}$ that is in the interior of some $E_j\in\partial\mathcal{T}_h$, we define the downstream and upstream values of $\boldsymbol{\phi}$ at $\boldsymbol{x}$ by

$$\boldsymbol{\phi}^{+\boldsymbol{v}}(\boldsymbol{x})=\lim_{\rho\to0^+}\boldsymbol{\phi}(\boldsymbol{x}+\rho\boldsymbol{v}(\boldsymbol{x})) \qquad \text{and} \qquad \boldsymbol{\phi}^{-\boldsymbol{v}}(\boldsymbol{x})=\lim_{\rho\to0^-}\boldsymbol{\phi}(\boldsymbol{x}+\rho\boldsymbol{v}(\boldsymbol{x}))\,; \tag{3-III.4}$$

respectively. In addition, we denote by

$$[\![\boldsymbol{\phi}]\!]_{\to\boldsymbol{v}}(\boldsymbol{x})=\boldsymbol{\phi}^{+\boldsymbol{v}}(\boldsymbol{x})-\boldsymbol{\phi}^{-\boldsymbol{v}}(\boldsymbol{x}) \qquad \text{and} \qquad \{\boldsymbol{\phi}\}(\boldsymbol{x})=\frac{\boldsymbol{\phi}^{+\boldsymbol{v}}(\boldsymbol{x})+\boldsymbol{\phi}^{-\boldsymbol{v}}(\boldsymbol{x})}{2}\,, \tag{3-III.5}$$

the jump and mean value, respectively, of $\boldsymbol{\phi}$ at the point $\boldsymbol{x}$ of boundary E_j. From (3-III.4), it is clear that the values of $\boldsymbol{\phi}^{+\boldsymbol{v}}|_{E_j}$ and $\boldsymbol{\phi}^{-\boldsymbol{v}}|_{E_j}$ can change along $E_j\in\partial\mathcal{T}_h$. In addition, it is easily deduced that

$$\sum_{j=1}^{N_E}\int_{E_j}|\boldsymbol{v}\cdot\boldsymbol{n}|[\![q]\!]_{\to\boldsymbol{v}}=-\sum_{k=1}^{N_K}\int_{\partial K_k}(\boldsymbol{v}\cdot\boldsymbol{n}_{K_k})\,q \qquad \forall\boldsymbol{v}\in\mathrm{W}_h,\ q\in\mathrm{Q}_h^0\,; \tag{3-III.6}$$

where $\boldsymbol{n}\equiv\boldsymbol{n}(E_j)$ is a unit normal to E_j, whose sign is of no importance.

3-III-B A free energy preserving approximation, $(\mathrm{P}_{\delta,h}^{\Delta t})$, of (P_δ)

For any source term $\boldsymbol{f}\in L^2\left(0,T;[H^{-1}(\mathcal{D})]^d\right)$, we define the following piecewise constant function with respect to the time variable

$$\boldsymbol{f}^{\Delta t,+}(t,\cdot)=\boldsymbol{f}^n(\cdot):=\frac{1}{\Delta t_n}\int_{t^{n-1}}^{t^n}\boldsymbol{f}(t,\cdot)\,dt, \qquad t\in[t^{n-1},t^n), \qquad n=1,\dots,N_T\,. \tag{3-III.7}$$

It is easily deduced that

$$\sum_{n=1}^{N_T}\Delta t_n\,\|\boldsymbol{f}^n\|_{H^{-1}(\mathcal{D})}^r\le\int_0^T\|f(t,\cdot)\|_{H^{-1}(\mathcal{D})}^r dt \qquad \text{for any } r\in[1,2]\,, \tag{3-III.8a}$$

$$\text{and} \qquad \boldsymbol{f}^{\Delta t,+}\to\boldsymbol{f} \quad \text{strongly in } L^2(0,T;[H^{-1}(\mathcal{D})]^d) \text{ as } \Delta t\to0_+\,. \tag{3-III.8b}$$

Throughout this section we choose $\boldsymbol{u}_h^0\in\mathrm{V}_h^0$ to be a suitable approximation of $\boldsymbol{u}^0$ such as the L^2 projection of $\boldsymbol{u}^0$ onto V_h^0. We will also choose $\boldsymbol{\sigma}_h^0\in\mathrm{S}_h^0$ to be the L^2 projection of $\boldsymbol{\sigma}^0$ onto S_h^0. Hence for $k=1,\dots,N_K$

$$\boldsymbol{\sigma}_h^0\,|_{K_k}=\frac{1}{|K_k|}\int_{K_k}\boldsymbol{\sigma}^0\,, \tag{3-III.9a}$$

where $|K_k|$ is the measure of K_k ; and it immediately follows from (3-II.19) that

$$\sigma_{\min}^0\,\|\boldsymbol{\xi}\|^2\le\boldsymbol{\xi}^T\boldsymbol{\sigma}_h^0\,|_{K_k}\,\boldsymbol{\xi}\le\sigma_{\max}^0\,\|\boldsymbol{\xi}\|^2 \qquad \forall\boldsymbol{\xi}\in\mathbb{R}^d\,. \tag{3-III.9b}$$

Our approximation $(\mathrm{P}_{\delta,h}^{\Delta t})$ of (P_δ) is then :

$(\mathbf{P}_{\delta,h}^{\Delta t})$ Setting $(\boldsymbol{u}_{\delta,h}^0,\boldsymbol{\sigma}_{\delta,h}^0)=(\boldsymbol{u}_h^0,\boldsymbol{\sigma}_h^0)\in\mathrm{V}_h^0\times\mathrm{S}_h^0$, then for $n=1,\dots,N_T$ find $(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n)\in\mathrm{V}_h^0\times\mathrm{S}_h^0$ such that for any test functions $(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}_h^0\times\mathrm{S}_h^0$

$$\int_{\mathcal{D}}\left[\mathrm{Re}\left(\frac{\boldsymbol{u}_{\delta,h}^n-\boldsymbol{u}_{\delta,h}^{n-1}}{\Delta t_n}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\left[\left((\boldsymbol{u}_{\delta,h}^{n-1}\cdot\nabla)\boldsymbol{u}_{\delta,h}^n\right)\cdot\boldsymbol{v}-\boldsymbol{u}_{\delta,h}^n\cdot\left((\boldsymbol{u}_{\delta,h}^{n-1}\cdot\nabla)\boldsymbol{v}\right)\right]\right. \tag{3-III.10a}$$

$$\left.+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n):\boldsymbol{\nabla}\boldsymbol{v}\right]=\langle\boldsymbol{f}^n,\boldsymbol{v}\rangle_{H_0^1(\mathcal{D})}\,,$$

$$\int_{\mathcal{D}}\left[\left(\frac{\boldsymbol{\sigma}_{\delta,h}^n-\boldsymbol{\sigma}_{\delta,h}^{n-1}}{\Delta t_n}\right):\boldsymbol{\phi}-2\left((\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n)\,\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_{\delta,h}^n-\boldsymbol{I}\right):\boldsymbol{\phi}\right] \tag{3-III.10b}$$

$$+\sum_{j=1}^{N_E}\int_{E_j}\left|\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}\right|[\![\boldsymbol{\sigma}_{\delta,h}^n]\!]_{\to\boldsymbol{u}_{\delta,h}^{n-1}}:\boldsymbol{\phi}^{+\boldsymbol{u}_{\delta,h}^{n-1}}=0\,.$$

In deriving ($\mathrm{P}_{\delta,h}^{\Delta t}$), we have noted (3-I.4a) and that

$$\int_{\mathcal{D}} \boldsymbol{v}\cdot[(\boldsymbol{z}\cdot\nabla)\boldsymbol{w}] = -\int_{\mathcal{D}} \boldsymbol{w}\cdot[(\boldsymbol{z}\cdot\nabla)\boldsymbol{v}] \qquad \forall \boldsymbol{z}\in \mathrm{V}, \qquad \forall \boldsymbol{v},\boldsymbol{w}\in[H^1(\mathcal{D})]^d. \tag{3-III.11}$$

Once again we refer to [EG04, p267] for the consistency of our stated approximation of the stress convection term, see also [BLM09].

Before proving existence of a solution to ($\mathrm{P}_{\delta,h}^{\Delta t}$), we first derive a discrete analogue of the energy estimate (3-II.21) for ($\mathrm{P}_{\delta,h}^{\Delta t}$); which uses the elementary equality

$$2s_1(s_1-s_2) = s_1^2 - s_2^2 + (s_1-s_2)^2 \qquad \forall s_1,s_2\in\mathbb{R}. \tag{3-III.12}$$

3-III-C Energy bound for ($\mathrm{P}_{\delta,h}^{\Delta t}$)

Proposition 13. *For $n=1,\ldots,N_T$, a solution $\left(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n\right)\in \mathrm{V}_h^0\times\mathrm{S}_h^0$ to (3-III.10a,b), if it exists, satisfies*

$$\begin{aligned}
&\frac{F_\delta(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n)-F_\delta(\boldsymbol{u}_{\delta,h}^{n-1},\boldsymbol{\sigma}_{\delta,h}^{n-1})}{\Delta t_n} + \frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}}\|\boldsymbol{u}_{\delta,h}^n-\boldsymbol{u}_{\delta,h}^{n-1}\|^2 + (1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n\|^2 \\
&\qquad + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)+[\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)]^{-1}-2\boldsymbol{I}) \\
&\le \langle \boldsymbol{f}^n,\boldsymbol{u}_{\delta,h}^n\rangle_{H_0^1(\mathcal{D})} \le \frac{1}{2}(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n\|^2 + \frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|^2_{H^{-1}(\mathcal{D})}.
\end{aligned} \tag{3-III.13}$$

Proof. Similarly to the proof of Proposition 12, we choose as test functions $\boldsymbol{v}=\boldsymbol{u}_{\delta,h}^n\in\mathrm{V}_h^0$ and $\boldsymbol{\phi}=\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-G'_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)\in\mathrm{S}_h^0$ in (3-III.10a,b), and obtain, on noting (3-III.12) and (3-II.7a,d), that

$$\begin{aligned}
\langle \boldsymbol{f}^n,\boldsymbol{u}_{\delta,h}^n\rangle_{H_0^1(\mathcal{D})} \ge{}& \int_{\mathcal{D}}\left[\frac{\mathrm{Re}}{2}\left(\frac{\|\boldsymbol{u}_{\delta,h}^n\|^2-\|\boldsymbol{u}_{\delta,h}^{n-1}\|^2}{\Delta t_n}+\frac{\|\boldsymbol{u}_{\delta,h}^n-\boldsymbol{u}_{\delta,h}^{n-1}\|^2}{\Delta t_n}\right)+(1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n\|^2\right] \\
&+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left(\frac{\boldsymbol{\sigma}_{\delta,h}^n-\boldsymbol{\sigma}_{\delta,h}^{n-1}}{\Delta t_n}\right):\left(\boldsymbol{I}-G'_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right) \\
&+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)+[\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)]^{-1}-2\boldsymbol{I}) \\
&+\frac{\varepsilon}{2\mathrm{Wi}}\sum_{j=1}^{N_E}\int_{E_j}\left[\left|\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}\right|[\![\boldsymbol{\sigma}_{\delta,h}^n]\!]_{\to\boldsymbol{u}_{\delta,h}^{n-1}}:\left(\boldsymbol{I}-G'_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)^{+\boldsymbol{u}_{\delta,h}^{n-1}}\right].
\end{aligned} \tag{3-III.14}$$

We consequently obtain from (3-III.14), on noting (3-I.2b) and (3-II.7e) applied to the edge terms as well as the discrete time derivative term for the stress variable, that

$$\begin{aligned}
\langle \boldsymbol{f}^n,\boldsymbol{u}_{\delta,h}^n\rangle_{H_0^1(\mathcal{D})} \ge{}& \int_{\mathcal{D}}\left[\frac{\mathrm{Re}}{2}\left(\frac{\|\boldsymbol{u}_{\delta,h}^n\|^2-\|\boldsymbol{u}_{\delta,h}^{n-1}\|^2}{\Delta t_n}+\frac{\|\boldsymbol{u}_{\delta,h}^n-\boldsymbol{u}_{\delta,h}^{n-1}\|^2}{\Delta t_n}\right)+(1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}_{\delta,h}^n\|^2\right] \\
&+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left(\frac{\mathrm{tr}\left(\boldsymbol{\sigma}_{\delta,h}^n-G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)-\mathrm{tr}\left(\boldsymbol{\sigma}_{\delta,h}^{n-1}-G_\delta(\boldsymbol{\sigma}_{\delta,h}^{n-1})\right)}{\Delta t_n}\right) \\
&+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)+[\beta_\delta(\boldsymbol{\sigma}_{\delta,h}^n)]^{-1}-2\boldsymbol{I}) \\
&+\frac{\varepsilon}{2\mathrm{Wi}}\sum_{j=1}^{N_E}\int_{E_j}\left|\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}\right|[\![\mathrm{tr}\boldsymbol{\sigma}_{\delta,h}^n-\mathrm{tr}G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)]\!]_{\to\boldsymbol{u}_{\delta,h}^{n-1}}.
\end{aligned} \tag{3-III.15}$$

Finally, we note from (3-III.6), (3-III.3) and as $\boldsymbol{u}_{\delta,h}^{n-1}\in \mathrm{V}_h^0$ that

$$\begin{aligned}\sum_{j=1}^{N_E}\int_{E_j}\left|\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}\right|[\![\mathrm{tr}\boldsymbol{\sigma}_{\delta,h}^n-\mathrm{tr}G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)]\!]_{\rightarrow\boldsymbol{u}_{\delta,h}^{n-1}} \qquad (3\text{-III.16})\\ =-\sum_{k=1}^{N_K}\int_{\partial K_k}\left(\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}_{K_k}\right)\mathrm{tr}\left(\boldsymbol{\sigma}_{\delta,h}^n-G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)\\ =-\sum_{k=1}^{N_K}\int_{K_k}\mathrm{div}\left(\boldsymbol{u}_{\delta,h}^{n-1}\mathrm{tr}\left(\boldsymbol{\sigma}_{\delta,h}^n-G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)\right)\\ =-\int_{\mathcal{D}}\mathrm{tr}\left(\boldsymbol{\sigma}_{\delta,h}^n-G_\delta(\boldsymbol{\sigma}_{\delta,h}^n)\right)\mathrm{div}\,\boldsymbol{u}_{\delta,h}^{n-1}=0.\end{aligned}$$

Combining (3-III.15) and (3-III.16) yields the first desired inequality in (3-III.13). The second inequality in (3-III.13) follows immediately from (3-II.28) with $\nu^2=(1-\varepsilon)/(1+C_P)$. □

3-III-D Existence of a solution to $(\mathrm{P}_{\delta,h}^{\Delta t})$

Proposition 14. *Given $(\boldsymbol{u}_{\delta,h}^{n-1},\boldsymbol{\sigma}_{\delta,h}^{n-1})\in \mathrm{V}_h^0\times\mathrm{S}_h^0$ and for any time step $\Delta t_n>0$, then there exists at least one solution $\left(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n\right)\in \mathrm{V}_h^0\times\mathrm{S}_h^0$ to (3-III.10a,b).*

Proof. We introduce the following inner product on the Hilbert space $\mathrm{V}_h^0\times\mathrm{S}_h^0$

$$((\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi}))_{\mathcal{D}}=\int_{\mathcal{D}}[\boldsymbol{w}\cdot\boldsymbol{v}+\boldsymbol{\psi}:\boldsymbol{\phi}]\qquad\forall(\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0. \qquad (3\text{-III.17})$$

Given $(\boldsymbol{u}_{\delta,h}^{n-1},\boldsymbol{\sigma}_{\delta,h}^{n-1})\in \mathrm{V}_h^0\times\mathrm{S}_h^0$, let $\mathcal{F}:\mathrm{V}_h^0\times\mathrm{S}_h^0\mapsto\mathrm{V}_h^0\times\mathrm{S}_h^0$ be such that for any $(\boldsymbol{w},\boldsymbol{\psi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0$

$$\begin{aligned}(\mathcal{F}(\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi}))_{\mathcal{D}}:=\int_{\mathcal{D}}\Bigg[&\mathrm{Re}\left(\frac{\boldsymbol{w}-\boldsymbol{u}_{\delta,h}^{n-1}}{\Delta t_n}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\left[\left((\boldsymbol{u}_{\delta,h}^{n-1}\cdot\nabla)\boldsymbol{w}\right)\cdot\boldsymbol{v}-\boldsymbol{w}\cdot\left((\boldsymbol{u}_{\delta,h}^{n-1}\cdot\nabla)\boldsymbol{v}\right)\right]\\ &+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{w}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\beta_\delta(\boldsymbol{\psi}):\boldsymbol{\nabla}\boldsymbol{v}+\left(\frac{\boldsymbol{\psi}-\boldsymbol{\sigma}_{\delta,h}^{n-1}}{\Delta t_n}\right):\boldsymbol{\phi}\\ &-2\left((\boldsymbol{\nabla}\boldsymbol{w})\beta_\delta(\boldsymbol{\psi})\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\psi}-\boldsymbol{I}):\boldsymbol{\phi}\Bigg]-\langle\boldsymbol{f}^n,\boldsymbol{v}\rangle_{H_0^1(\mathcal{D})}\\ &+\sum_{j=1}^{N_E}\int_{E_j}\left|\boldsymbol{u}_{\delta,h}^{n-1}\cdot\boldsymbol{n}\right|[\![\boldsymbol{\psi}]\!]_{\rightarrow\boldsymbol{u}_{\delta,h}^{n-1}}:\boldsymbol{\phi}^{+\boldsymbol{u}_{\delta,h}^{n-1}}\qquad\forall(\boldsymbol{v},\boldsymbol{\phi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0.\end{aligned} \qquad (3\text{-III.18})$$

We note that a solution $(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n)$ to (3-III.10a,b), if it exists, corresponds to a zero of $\mathcal{F}$; that is,

$$\left(\mathcal{F}(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n),(\boldsymbol{v},\boldsymbol{\phi})\right)_{\mathcal{D}}=0\qquad\forall(\boldsymbol{v},\boldsymbol{\phi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0. \qquad (3\text{-III.19})$$

In addition, it is easily deduced that the mapping $\mathcal{F}$ is continuous.

For any $(\boldsymbol{w},\boldsymbol{\psi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0$, on choosing $(\boldsymbol{v},\boldsymbol{\phi})=\left(\boldsymbol{w},\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I}-G'_\delta(\boldsymbol{\psi}))\right)$, we obtain analogously to (3-III.13) that

$$\begin{aligned}&\left(\mathcal{F}(\boldsymbol{w},\boldsymbol{\psi}),\left(\boldsymbol{w},\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I}-G'_\delta(\boldsymbol{\psi}))\right)\right)_{\mathcal{D}} \qquad (3\text{-III.20})\\ &\quad\geq\frac{F_\delta(\boldsymbol{w},\boldsymbol{\psi})-F_\delta(\boldsymbol{u}_{\delta,h}^{n-1},\boldsymbol{\sigma}_{\delta,h}^{n-1})}{\Delta t_n}+\frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}}\|\boldsymbol{w}-\boldsymbol{u}_{\delta,h}^{n-1}\|^2+\frac{1-\varepsilon}{2}\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{w}\|^2\\ &\qquad+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\psi})+[\beta_\delta(\boldsymbol{\psi})]^{-1}-2\boldsymbol{I})-\frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|_{H^{-1}(\mathcal{D})}^2.\end{aligned}$$

Let us now assume that for any $\gamma\in\mathbb{R}_{>0}$, the continuous mapping $\mathcal{F}$ has no zero $(\boldsymbol{u}_{\delta,h}^n,\boldsymbol{\sigma}_{\delta,h}^n)$ satisfying (3-III.19), which lies in the ball

$$\mathcal{B}_\gamma:=\left\{(\boldsymbol{v},\boldsymbol{\phi})\in \mathrm{V}_h^0\times\mathrm{S}_h^0:\|(\boldsymbol{v},\boldsymbol{\phi})\|_{\mathcal{D}}\leq\gamma\right\}; \qquad (3\text{-III.21})$$

where

$$\|(\boldsymbol{v},\boldsymbol{\phi})\|_D := [((\boldsymbol{v},\boldsymbol{\phi}),(\boldsymbol{v},\boldsymbol{\phi}))_D]^{\frac{1}{2}} = \left(\int_{\mathcal{D}} [\|\boldsymbol{v}\|^2 + \|\boldsymbol{\phi}\|^2]\right)^{\frac{1}{2}}. \tag{3-III.22}$$

Then for such γ, we can define the continuous mapping $\mathcal{G}_\gamma : \mathcal{B}_\gamma \mapsto \mathcal{B}_\gamma$ such that for all $(\boldsymbol{v},\boldsymbol{\phi}) \in \mathcal{B}_\gamma$

$$\mathcal{G}_\gamma(\boldsymbol{v},\boldsymbol{\phi}) := -\gamma \frac{\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})}{\|\mathcal{F}(\boldsymbol{v},\boldsymbol{\phi})\|_{\mathcal{D}}}. \tag{3-III.23}$$

By the Brouwer fixed point theorem, $\mathcal{G}_\gamma$ has at least one fixed point $(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)$ in $\mathcal{B}_\gamma$. Hence it satisfies

$$\big\|(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\big\|_{\mathcal{D}} = \big\|\mathcal{G}_\gamma(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\big\|_{\mathcal{D}} = \gamma. \tag{3-III.24}$$

It follows, on noting (3-III.2c), that

$$\|\boldsymbol{\phi}\|^2_{L^\infty(\mathcal{D})} \le \frac{1}{\min_{k\in N_K}|K_k|}\int_{\mathcal{D}} \|\boldsymbol{\phi}\|^2 \equiv \mu_h^2 \int_{\mathcal{D}} \|\boldsymbol{\phi}\|^2 \qquad \forall \boldsymbol{\phi} \in \mathrm{S}_h^0, \tag{3-III.25}$$

where $\mu_h := [1/(\min_{k\in N_K}|K_k|)]^{\frac{1}{2}}$. It follows from (3-II.20), (3-II.8), (3-III.25) and (3-III.24) that

$$\begin{aligned} F_\delta(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma) &= \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} \|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \mathrm{tr}(\boldsymbol{\psi}_\gamma - G_\delta(\boldsymbol{\psi}_\gamma) - \boldsymbol{I}) \\ &\ge \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} \|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{4\mathrm{Wi}}\left[\int_{\mathcal{D}} \|\boldsymbol{\psi}_\gamma\| - 2d|\mathcal{D}|\right] \\ &\ge \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} \|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\|\boldsymbol{\psi}_\gamma\|_{L^\infty(\mathcal{D})}\left[\int_{\mathcal{D}} \|\boldsymbol{\psi}_\gamma\|\right] - \frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}} \\ &\ge \min\left(\frac{\mathrm{Re}}{2}, \frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\left(\int_{\mathcal{D}} \left[\|\boldsymbol{w}_\gamma\|^2 + \|\boldsymbol{\psi}_\gamma\|^2\right]\right) - \frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}} \\ &= \min\left(\frac{\mathrm{Re}}{2}, \frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\gamma^2 - \frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}}. \end{aligned} \tag{3-III.26}$$

Hence for all γ sufficiently large, it follows from (3-III.20) and (3-III.26) that

$$\left(\mathcal{F}(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma), \left(\boldsymbol{w}_\gamma, \frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I} - G'_\delta(\boldsymbol{\psi}_\gamma)\right)\right)\right)_{\mathcal{D}} \ge 0. \tag{3-III.27}$$

On the other hand as $(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)$ is a fixed point of $\mathcal{G}_\gamma$, we have that

$$\begin{aligned} &\left(\mathcal{F}(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma), \left(\boldsymbol{w}_\gamma, \frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I} - G'_\delta(\boldsymbol{\psi}_\gamma)\right)\right)\right)_D \\ &\qquad = -\frac{\big\|\mathcal{F}(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\big\|_{\mathcal{D}}}{\gamma}\int_{\mathcal{D}} \left[\|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\boldsymbol{\psi}_\gamma : \left(\boldsymbol{I} - G'_\delta(\boldsymbol{\psi}_\gamma)\right)\right]. \end{aligned} \tag{3-III.28}$$

It follows from (3-II.8), and similarly to (3-III.26), on noting (3-III.25) and (3-III.24) that

$$\begin{aligned} \int_{\mathcal{D}} \left[\|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\boldsymbol{\psi}_\gamma : \left(\boldsymbol{I} - G'_\delta(\boldsymbol{\psi}_\gamma)\right)\right] &\ge \int_{\mathcal{D}} \left[\|\boldsymbol{w}_\gamma\|^2 + \frac{\varepsilon}{4\mathrm{Wi}}\left[\|\boldsymbol{\psi}_\gamma\| - 2d\right]\right] \\ &\ge \min\left(1, \frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\gamma^2 - \frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}}. \end{aligned} \tag{3-III.29}$$

Therefore on combining (3-III.28) and (3-III.29), we have for all γ sufficiently large that

$$\left(\mathcal{F}(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma), \left(\boldsymbol{w}_\gamma, \frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I} - G'_\delta(\boldsymbol{\psi}_\gamma)\right)\right)\right)_D < 0, \tag{3-III.30}$$

which obviously contradicts (3-III.27). Hence the mapping $\mathcal{F}$ has a zero in $\mathcal{B}_\gamma$ for γ sufficiently large. □

Theorem 3. *For any $\delta\in(0,\frac{1}{2}]$, $N_T\geq 1$ and any partitioning of $[0,T]$ into N_T time steps, then there exists a solution $\{(\boldsymbol{u}^n_{\delta,h},\boldsymbol{\sigma}^n_{\delta,h})\}_{n=1}^{N_T}\in[\mathrm{V}^0_h\times\mathrm{S}^0_h]^{N_T}$ to $(P^{\Delta t}_{\delta,h})$.*

In addition, it follows for $n=1,\ldots,N_T$ that

$$F_\delta(\boldsymbol{u}^n_{\delta,h},\boldsymbol{\sigma}^n_{\delta,h})+\frac{1}{2}\sum_{m=1}^{n}\int_{\mathcal{D}}\Big[\mathrm{Re}\|\boldsymbol{u}^m_{\delta,h}-\boldsymbol{u}^{m-1}_{\delta,h}\|^2+(1-\varepsilon)\Delta t_m\|\boldsymbol{\nabla}\boldsymbol{u}^m_{\delta,h}\|^2\Big] \tag{3-III.31}$$
$$+\frac{\varepsilon}{2\mathrm{Wi}^2}\sum_{m=1}^{n}\Delta t_m\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta(\boldsymbol{\sigma}^m_{\delta,h})+[\beta_\delta(\boldsymbol{\sigma}^m_{\delta,h})]^{-1}-2\boldsymbol{I})$$
$$\leq F_\delta(\boldsymbol{u}^0_h,\boldsymbol{\sigma}^0_h)+\frac{1+C_P}{2(1-\varepsilon)}\sum_{m=1}^{n}\Delta t_m\|\boldsymbol{f}^m\|^2_{H^{-1}(\mathcal{D})}\leq C.$$

Moreover, it follows that

$$\max_{n=0,\ldots,N_T}\int_{\mathcal{D}}\big[\|\boldsymbol{u}^n_{\delta,h}\|^2+\|\boldsymbol{\sigma}^n_{\delta,h}\|+\delta^{-1}\|[\boldsymbol{\sigma}^n_{\delta,h}]_-\|\big]+\sum_{n=1}^{N_T}\Delta t_n\int_{\mathcal{D}}\|[\beta_\delta(\boldsymbol{\sigma}^n_{\delta,h})]^{-1}\|\leq C. \tag{3-III.32}$$

Proof. Existence and the stability result (3-III.31) follow immediately from Propositions 14 and 13, respectively, on noting (3-II.20), (3-III.9b), (3-III.8a) and (3-II.19). The bounds (3-III.32) follow immediately from (3-III.31), on noting (3-II.7b), (3-II.8), (3-I.7) and the fact that $\beta_\delta(\boldsymbol{\phi})\in\mathbb{R}^{d\times d}_{SPD}$ for any $\boldsymbol{\phi}\in\mathbb{R}^{d\times d}_S$. □

3-III-E Convergence of $(\mathrm{P}^{\Delta t}_{\delta,h})$ to $(\mathrm{P}^{\Delta t}_h)$

We now consider the corresponding direct finite element approximation of (P), i.e. $(\mathrm{P}^{\Delta t}_{\delta,h})$ without the regularization δ :

$(\mathbf{P}^{\Delta t}_h)$ Given initial conditions $(\boldsymbol{u}^0_h,\boldsymbol{\sigma}^0_h)\in\mathrm{V}^0_h\times\mathrm{S}^0_h$ with $\boldsymbol{\sigma}^0_h$ satisfying (3-III.9a,b), then for $n=1,\ldots,N_T$ find $(\boldsymbol{u}^n_h,\boldsymbol{\sigma}^n_h)\in\mathrm{V}^0_h\times\mathrm{S}^0_h$ such that for any test functions $(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^0_h\times\mathrm{S}^0_h$

$$\int_{\mathcal{D}}\bigg[\mathrm{Re}\bigg(\frac{\boldsymbol{u}^n_h-\boldsymbol{u}^{n-1}_h}{\Delta t_n}\bigg)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\big[((\boldsymbol{u}^{n-1}_h\cdot\nabla)\boldsymbol{u}^n_h)\cdot\boldsymbol{v}-\boldsymbol{u}^n_h\cdot((\boldsymbol{u}^{n-1}_h\cdot\nabla)\boldsymbol{v})\big] \tag{3-III.33a}$$
$$+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}^n_h:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\boldsymbol{\sigma}^n_h:\boldsymbol{\nabla}\boldsymbol{v}\bigg]=\langle\boldsymbol{f}^n,\boldsymbol{v}\rangle_{H^1_0(\mathcal{D})},$$
$$\int_{\mathcal{D}}\bigg[\bigg(\frac{\boldsymbol{\sigma}^n_h-\boldsymbol{\sigma}^{n-1}_h}{\Delta t_n}\bigg):\boldsymbol{\phi}-2((\boldsymbol{\nabla}\boldsymbol{u}^n_h)\boldsymbol{\sigma}^n_h):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}^n_h-\boldsymbol{I}):\boldsymbol{\phi}\bigg] \tag{3-III.33b}$$
$$+\sum_{j=1}^{N_E}\int_{E_j}\big|\boldsymbol{u}^{n-1}_h\cdot\boldsymbol{n}\big|\,[\![\boldsymbol{\sigma}^n_h]\!]_{\rightarrow\boldsymbol{u}^{n-1}_h}:\boldsymbol{\phi}^{+\boldsymbol{u}^{n-1}_h}=0.$$

We introduce also the unregularised free energy

$$F(\boldsymbol{v},\boldsymbol{\phi}):=\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{v}\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\phi}-G(\boldsymbol{\phi})-\boldsymbol{I}), \tag{3-III.34}$$

which is well defined for $(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^0_h\times\mathrm{S}^0_h$ with $\boldsymbol{\phi}$ being positive definite on $\mathcal{D}$.

Theorem 4. *For all regular partitionings $\mathcal{T}_h$ of $\mathcal{D}$ into simplices $\{K_k\}_{k=1}^{N_K}$ and all partitionings $\{\Delta t_n\}_{n=1}^{N_T}$ of $[0,T]$, there exists a subsequence $\{\{(\boldsymbol{u}^n_{\delta,h},\boldsymbol{\sigma}^n_{\delta,h})\}_{n=1}^{N_T}\}_{\delta>0}$, where $\{(\boldsymbol{u}^n_{\delta,h},\boldsymbol{\sigma}^n_{\delta,h})\}_{n=1}^{N_T}\in[\mathrm{V}^0_h\times\mathrm{S}^0_h]^{N_T}$ solves $(P^{\Delta t}_{\delta,h})$, and $\{(\boldsymbol{u}^n_h,\boldsymbol{\sigma}^n_h)\}_{n=1}^{N_T}\in[\mathrm{V}^0_h\times\mathrm{S}^0_h]^{N_T}$ such that for the subsequence*

$$\boldsymbol{u}^n_{\delta,h}\rightarrow\boldsymbol{u}^n_h,\qquad\boldsymbol{\sigma}^n_{\delta,h}\rightarrow\boldsymbol{\sigma}^n_h\qquad\text{as }\delta\rightarrow 0_+,\qquad\text{for}\quad n=1,\ldots,N_T. \tag{3-III.35}$$

In addition, for $n=1,\ldots,N_T$, $\boldsymbol{\sigma}^n_h|_{K_k}\in\mathbb{R}^{d\times d}_{SPD}$, $k=1,\ldots,N_K$,. Moreover, $\{(\boldsymbol{u}^n_h,\boldsymbol{\sigma}^n_h)\}_{n=1}^{N_T}\in[\mathrm{V}^0_h\times\mathrm{S}^0_h]^{N_T}$ solves $(P^{\Delta t}_h)$ and for $n=1,\ldots,N_T$

$$\frac{F(\boldsymbol{u}^n_h,\boldsymbol{\sigma}^n_h)-F(\boldsymbol{u}^{n-1}_h,\boldsymbol{\sigma}^{n-1}_h)}{\Delta t_n}+\frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}}\|\boldsymbol{u}^n_h-\boldsymbol{u}^{n-1}_h\|^2+(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}^n_h\|^2 \tag{3-III.36}$$
$$+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\sigma}^n_h+[\boldsymbol{\sigma}^n_h]^{-1}-2\boldsymbol{I})$$
$$\leq\frac{1}{2}(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}^n_h\|^2+\frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|^2_{H^{-1}(\mathcal{D})}.$$

Proof. For any integer $n \in [1, N_T]$, the desired subsequence convergence result (3-III.35) follows immediately from (3-III.32), as $(\boldsymbol{u}^n_{\delta,h}, \boldsymbol{\sigma}^n_{\delta,h})$ are finite dimensional for fixed $\mathrm{V}^0_h \times S^0_h$. It also follows from (3-III.32), (3-III.35) and (3-II.17) that $[\boldsymbol{\sigma}^n_h]_-$ vanishes on $\mathcal{D}$, so that $\boldsymbol{\sigma}^n_h$ must be non-negative definite on $\mathcal{D}$. Hence on noting this, (3-II.17) and (3-III.35), we have the following subsequence convergence results

$$\beta_\delta(\boldsymbol{\sigma}^n_h) \to \boldsymbol{\sigma}^n_h \qquad \text{as} \quad \delta \to 0_+ \qquad \text{and} \qquad \beta_\delta(\boldsymbol{\sigma}^n_{\delta,h}) \to \boldsymbol{\sigma}^n_h \qquad \text{as} \quad \delta \to 0_+ . \tag{3-III.37}$$

It also follows from (3-III.32), (3-III.37) and as $[\beta_\delta(\boldsymbol{\sigma}^n_{\delta,h})]^{-1}\beta_\delta(\boldsymbol{\sigma}^n_{\delta,h}) = \boldsymbol{I}$ that the following subsequence result

$$[\beta_\delta(\boldsymbol{\sigma}^n_{\delta,h})]^{-1} \to [\boldsymbol{\sigma}^n_h]^{-1} \qquad \text{as} \quad \delta \to 0_+ \tag{3-III.38}$$

holds, and so $\boldsymbol{\sigma}^n_h$ is positive definite on $\mathcal{D}$. Therefore, we have from (3-III.35) and (3-II.1) that

$$G_\delta(\boldsymbol{\sigma}^n_{\delta,h}) \to G(\boldsymbol{\sigma}^n_h) \qquad \text{as} \quad \delta \to 0_+ . \tag{3-III.39}$$

Since $\boldsymbol{u}^{n-1}_{\delta,h}, \boldsymbol{u}^{n-1}_h \in C(\overline{\mathcal{D}})$, it follows from (3-III.35), (3-III.4) and (3-III.5) that for $j = 1, \ldots, N_E$ and for all $\boldsymbol{\phi} \in \mathrm{S}^0_h$

$$\int_{E_j} \left| \boldsymbol{u}^{n-1}_{\delta,h} \cdot \boldsymbol{n} \right| [\![\boldsymbol{\sigma}^n_{\delta,h}]\!]_{\to \boldsymbol{u}^{n-1}_{\delta,h}} : \boldsymbol{\phi}^{+\boldsymbol{u}^{n-1}_{\delta,h}} \to \int_{E_j} \left| \boldsymbol{u}^{n-1}_h \cdot \boldsymbol{n} \right| [\![\boldsymbol{\sigma}^n_h]\!]_{\to \boldsymbol{u}^{n-1}_h} : \boldsymbol{\phi}^{+\boldsymbol{u}^{n-1}_h} \quad \text{as} \quad \delta \to 0_+ . \tag{3-III.40}$$

Hence using (3-III.35), (3-III.37) and (3-III.40), we can pass to the limit $\delta \to 0_+$ in $(\mathrm{P}^{\Delta t}_{\delta,h})$, (3-III.10a,b), to show that $\{(\boldsymbol{u}^n_h, \boldsymbol{\sigma}^n_h)\}_{n=1}^{N_T} \in [\mathrm{V}^0_h \times \mathrm{S}^0_h]^{N_T}$ solves $(\mathrm{P}^{\Delta t}_h)$, (3-III.33a,b). Similarly, using (3-III.35), (3-III.37), (3-III.38) and (3-III.39), and noting (3-II.20) and (3-III.34), we can pass to the limit $\delta \to 0_+$ in (3-III.13) to obtain the desired result (3-III.36). □

Remark 10. *Most numerical approximations of (P) suffer from instabilities when* Wi *is relatively large, the so-called high Weissenberg number problem (HWNP). This problem is still not fully understood. Some reasons for these instabilities are discussed in [BLM09], e.g. poor numerical scheme or the lack of existence of a solution to (P) itself. In addition in [BLM09], finite element approximations of (P) such as $(P^{\Delta t}_h)$, approximating the primitive variables $(\boldsymbol{u}, p, \boldsymbol{\sigma})$, are compared with finite element approximations of the log-formulation of (P), introduced in [FK05], which is based on the variables $(\boldsymbol{u}, p, \boldsymbol{\psi})$, where $\boldsymbol{\psi} = \ln \boldsymbol{\sigma}$. The equivalent free energy estimate for this log-formulation is based on testing the Navier-Stokes equation with $\boldsymbol{u}$ as before, but the log-form of the stress equation with $(\exp \boldsymbol{\psi} - \boldsymbol{I})$. Whereas the free energy estimate for (P) requires $\boldsymbol{\sigma}$ to be positive definite, due to the testing with $\ln \boldsymbol{\sigma}$, the free energy estimate for the log-formulation requires no such constraint. In [BLM09] a constraint, based on the initial data, was required on the time step in order to ensure that the approximation to $\boldsymbol{\sigma}$ remained positive definite for schemes such as $(P^{\Delta t}_h)$ approximating (P); whereas existence of a solution to finite element approximations of the log-formulation, and satisfying a discrete log-form of the free energy estimate, were shown for any choice of time step. It was suggested in [BLM09] that this may be the reason why the approximations of the log-formulation are reported to be more stable than those based on (P). However, Theorem 4 above shows that there does exist (at least) one solution to $(P^{\Delta t}_h)$, which satisfies the free energy estimate (3-III.36), whatever the time step. Of course, we do not have a uniqueness proof for $(P^{\Delta t}_h)$.*

3-IV Regularized problems with stress diffusion and possibly the cut-off β^L

3-IV-A Regularizations $(\mathrm{P}^{(L)}_\alpha)$ of (P) with stress diffusion and possibly the cut-off

In this section, we consider the following modified versions of (P) for given constants $\alpha \in \mathbb{R}_{>0}$ and $L \geq 2$:

$(\mathbf{P}^{(L)}_\alpha)$ Find $\boldsymbol{u}^{(L)}_\alpha : (t, \boldsymbol{x}) \in [0,T) \times \mathcal{D} \mapsto \boldsymbol{u}^{(L)}_\alpha(t, \boldsymbol{x}) \in \mathbb{R}^d$, $p^{(L)}_\alpha : (t, \boldsymbol{x}) \in (0,T) \times \mathcal{D} \mapsto p^{(L)}_\alpha(t, \boldsymbol{x}) \in \mathbb{R}$ and $\boldsymbol{\sigma}^{(L)}_\alpha : (t, \boldsymbol{x}) \in$

$[0,T)\times\mathcal{D}\mapsto\boldsymbol{\sigma}_\alpha^{(L)}(t,\boldsymbol{x})\in\mathbb{R}^{d\times d}_S$ such that

$$\mathrm{Re}\left(\frac{\partial\boldsymbol{u}_\alpha^{(L)}}{\partial t}+(\boldsymbol{u}_\alpha^{(L)}\cdot\boldsymbol{\nabla})\boldsymbol{u}_\alpha^{(L)}\right)=-\boldsymbol{\nabla}p_\alpha^{(L)}+(1-\varepsilon)\Delta\boldsymbol{u}_\alpha^{(L)}+\frac{\varepsilon}{\mathrm{Wi}}\operatorname{div}\beta^{(L)}(\boldsymbol{\sigma}_\alpha^{(L)})+\boldsymbol{f}\qquad\text{on }\mathcal{D}_T,\tag{3-IV.1a}$$

$$\operatorname{div}\boldsymbol{u}_\alpha^{(L)}=0\qquad\text{on }\mathcal{D}_T,\tag{3-IV.1b}$$

$$\frac{\partial\boldsymbol{\sigma}_\alpha^{(L)}}{\partial t}+(\boldsymbol{u}_\alpha^{(L)}\cdot\boldsymbol{\nabla})\beta^{(L)}(\boldsymbol{\sigma}_\alpha^{(L)})=(\boldsymbol{\nabla}\boldsymbol{u}_\alpha^{(L)})\beta^{(L)}(\boldsymbol{\sigma}_\alpha^{(L)})+\beta^{(L)}(\boldsymbol{\sigma}_\alpha^{(L)})(\boldsymbol{\nabla}\boldsymbol{u}_\alpha^{(L)})^T\tag{3-IV.1c}$$
$$-\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_\alpha^{(L)}-\boldsymbol{I}\right)+\alpha\Delta\boldsymbol{\sigma}_\alpha^{(L)}\qquad\text{on }\mathcal{D}_T,$$

$$\boldsymbol{u}_\alpha^{(L)}(0,\boldsymbol{x})=\boldsymbol{u}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D},\tag{3-IV.1d}$$

$$\boldsymbol{\sigma}_\alpha^{(L)}(0,\boldsymbol{x})=\boldsymbol{\sigma}^0(\boldsymbol{x})\qquad\forall\boldsymbol{x}\in\mathcal{D},\tag{3-IV.1e}$$

$$\boldsymbol{u}_\alpha^{(L)}=\boldsymbol{0}\qquad\text{on }(0,T)\times\partial\mathcal{D},\tag{3-IV.1f}$$

$$(\boldsymbol{n}_{\partial\mathcal{D}}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}_\alpha^{(L)}=\boldsymbol{0}\qquad\text{on }(0,T)\times\partial\mathcal{D};\tag{3-IV.1g}$$

where $\boldsymbol{n}_{\mathcal{D}}$ is normal to the boundary $\partial\mathcal{D}$.

Hence problem $(\mathrm{P}_\alpha^{(L)})$ is the same as (P), but with the added diffusion term $\alpha\Delta\boldsymbol{\sigma}_\alpha^{(L)}$ for the stress equation (3-IV.1c), and the associated Neumann boundary condition (3-IV.1g); and in the case of (P_α^{L}) with certain terms in (3-IV.1a,c) involving $\boldsymbol{\sigma}_\alpha^{L}$ replaced by $\beta^{L}(\boldsymbol{\sigma}_\alpha^{L})$, recall (3-II.3). Of course, it is naturally assumed in $(\mathrm{P}_\alpha^{(L)})$ that $\boldsymbol{\sigma}_\alpha^{(L)}$ is positive definite on $\mathcal{D}_T$ in order for $\beta^{(L)}(\boldsymbol{\sigma}_\alpha^{(L)})$ to be well defined.

We will also be interested in the corresponding regularization $(\mathrm{P}_{\alpha,\delta}^{(L)})$ of $(\mathrm{P}_\alpha^{(L)})$ with solution $(\boldsymbol{u}_{\alpha,\delta}^{(L)},p_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})$; where $\beta^{(L)}(\cdot)$ in (3-IV.1a–g) is replaced by $\beta_\delta^{(L)}(\cdot)$, and so that $\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}$ is not required to be positive definite.

3-IV-B Formal energy estimates for $(\mathrm{P}_{\alpha,\delta}^{(L)})$

Let $F_\delta^{(L)}(\boldsymbol{u}_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})$ denote the free energy of the solution $(\boldsymbol{u}_{\alpha,\delta}^{(L)},\, p_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})$ to problem $(\mathrm{P}_{\alpha,\delta}^{(\mathrm{L})})$, where $F_\delta^{(L)}:\mathrm{W}\times\mathrm{S}\mapsto\mathbb{R}$ is defined as

$$F_\delta^{(L)}(\boldsymbol{v},\boldsymbol{\phi}):=\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{v}\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\mathrm{tr}(\boldsymbol{\phi}-G_\delta^{(L)}(\boldsymbol{\phi})-\boldsymbol{I}).\tag{3-IV.2}$$

We have the following analogue of Proposition 12.

Proposition 15. *Let $(\boldsymbol{u}_{\alpha,\delta}^{(L)},p_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})$ be a sufficiently smooth solution to problem $(\mathrm{P}_{\alpha,\delta}^{(\mathrm{L})})$. Then the free energy $F_\delta^{(L)}(\boldsymbol{u}_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})$ satisfies for a.a. $t\in(0,T)$*

$$\frac{d}{dt}F_\delta^{(L)}(\boldsymbol{u}_{\alpha,\delta}^{(L)},\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})+(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,\delta}^{(L)}\|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\mathrm{tr}(\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})+[\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})]^{-1}-2\boldsymbol{I})\tag{3-IV.3}$$
$$+\frac{\alpha\varepsilon\delta^2}{2\mathrm{Wi}}\int_{\mathcal{D}}\|\boldsymbol{\nabla}{G_\delta^{(L)}}'(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})\|^2\le\langle\boldsymbol{f},\boldsymbol{u}_{\alpha,\delta}^{(L)}\rangle_{H_0^1(\mathcal{D})}.$$

Proof. Multiplying the Navier-Stokes equation (3-IV.1a) with $\boldsymbol{u}_{\alpha,\delta}^{(L)}$ and the stress equation (3-IV.1c) with $\frac{\varepsilon}{2\mathrm{Wi}}(\boldsymbol{I}-{G_\delta^{(L)}}'(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}))$, summing and integrating over $\mathcal{D}$ yields, after using integrations by parts and the incompressibility property in the standard way, that

$$\int_{\mathcal{D}}\left[\frac{\mathrm{Re}}{2}\frac{\partial}{\partial t}\|\boldsymbol{u}_{\alpha,\delta}^{(L)}\|^2+(1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,\delta}^{(L)}\|^2+\frac{\varepsilon}{\mathrm{Wi}}\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}):\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,\delta}^{(L)}-\frac{\varepsilon\alpha}{2\mathrm{Wi}}\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}::\boldsymbol{\nabla}{G_\delta^{(L)}}'(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})\right]\tag{3-IV.4}$$
$$+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left[\left(\frac{\partial}{\partial t}\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}+(\boldsymbol{u}_{\alpha,\delta}^{(L)}\cdot\boldsymbol{\nabla})\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}\right)+\frac{1}{\mathrm{Wi}}\left(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}-\boldsymbol{I}\right)\right]:\left(\boldsymbol{I}-{G_\delta^{(L)}}'(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})\right)$$
$$-\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\left(\left(\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,\delta}^{(L)}\right)\beta_\delta^{(L)}\left(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}\right)+\beta_\delta^{(L)}\left(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}\right)\left(\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,\delta}^{(L)}\right)^T\right):\left(\boldsymbol{I}-{G_\delta^{(L)}}'(\boldsymbol{\sigma}_{\alpha,\delta}^{(L)})\right)$$
$$=\langle\boldsymbol{f},\boldsymbol{u}_{\alpha,\delta}^{(L)}\rangle_{H_0^1(\mathcal{D})}.$$

Similarly to (3-II.12), we have that

$$-\boldsymbol{\nabla}\boldsymbol{\sigma}^{(L)}_{\alpha,\delta} :: \boldsymbol{\nabla} G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) \geq \delta^2 \|\boldsymbol{\nabla} G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\|^2 \qquad a.e.\ \text{in } \mathcal{D}_T. \tag{3-IV.5}$$

Using (3-II.16), we have that

$$\frac{\partial}{\partial t}\boldsymbol{\sigma}^{(L)}_{\alpha,\delta} : \left(\boldsymbol{I} - G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\right) = \frac{\partial}{\partial t}\mathrm{tr}\left(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta} - G^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\right). \tag{3-IV.6}$$

We will deal with the convection term differently to the approach used in (3-II.23), as that cannot be mimicked at a discrete level using continuous piecewise linear elements to approximate $\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}$. Note that we cannot use S^0_h with the desirable property (3-III.3) to approximate $\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}$, as we now have the added diffusion term. Instead, as $\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}$ has been replaced by $\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) \equiv H^{(L)'}_{\delta}(G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}))$, on recalling (3-II.6), in this convective term and as $\boldsymbol{u}^{(L)}_{\alpha,\delta} \in \mathrm{V}$, we have that

$$\begin{aligned}\int_{\mathcal{D}} (\boldsymbol{u}^{(L)}_{\alpha,\delta}\cdot\boldsymbol{\nabla})\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) : \left(\boldsymbol{I} - G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\right) &= \int_{\mathcal{D}} \beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) : (\boldsymbol{u}^{(L)}_{\alpha,\delta}\cdot\boldsymbol{\nabla}) G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) \\ &= \int_{\mathcal{D}} (\boldsymbol{u}^{(L)}_{\alpha,\delta}\cdot\boldsymbol{\nabla})\mathrm{tr}\left(H^{(L)}_{\delta}(G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}))\right) = 0,\end{aligned} \tag{3-IV.7}$$

where we have noted the spatial counterpart of (3-II.16). Similarly to (3-II.24) and (3-II.25) we obtain that

$$\begin{aligned}\left(\left(\boldsymbol{\nabla}\boldsymbol{u}^{(L)}_{\alpha,\delta}\right)\beta^{(L)}_{\delta}\left(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}\right) + \beta^{(L)}_{\delta}\left(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}\right)\left(\boldsymbol{\nabla}\boldsymbol{u}^{(L)}_{\alpha,\delta}\right)^T\right) &: \left(\boldsymbol{I} - G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\right) \\ &= 2\mathrm{tr}\left(\left(\boldsymbol{\nabla}\boldsymbol{u}^{(L)}_{\alpha,\delta}\right)\beta^{(L)}_{\delta}\left(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}\right)\right),\end{aligned} \tag{3-IV.8}$$

and once again the terms involving the left-hand side of (3-IV.8) in (3-IV.4) cancel with the term $\frac{\varepsilon}{\mathrm{Wi}}\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) : \boldsymbol{\nabla}\boldsymbol{u}^{(L)}_{\alpha,\delta}$ in (3-IV.4) arising from the Navier-Stokes equation. Finally, the treatment of the remaining term $\left(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta} - \boldsymbol{I}\right) : \left(\boldsymbol{I} - G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\right)$ follows similarly to (3-II.26); and so we obtain the desired free energy inequality (3-IV.3). □

The following Corollary follows from (3-IV.3) on noting the proof of Corollary 1.

Corollary 2. *Let $(\boldsymbol{u}^{(L)}_{\alpha,\delta}, p^{(L)}_{\alpha,\delta}, \boldsymbol{\sigma}^{(L)}_{\alpha,\delta})$ be a sufficiently smooth solution to problem $(\mathrm{P}^{(L)}_{\alpha,\delta})$. Then it follows that*

$$\begin{aligned}&\sup_{t\in(0,T)} F^{(L)}_{\delta}(\boldsymbol{u}^{(L)}_{\alpha,\delta}(t,\cdot), \boldsymbol{\sigma}^{(L)}_{\alpha,\delta}(t,\cdot)) \\ &+ \frac{1}{2}\int_{\mathcal{D}_T}\left[(1-\varepsilon)\|\boldsymbol{\nabla}\boldsymbol{u}^{(L)}_{\alpha,\delta}\|^2 + \frac{\varepsilon}{\mathrm{Wi}^2}\mathrm{tr}(\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}) + [\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})]^{-1} - 2\boldsymbol{I}) + \frac{\alpha\varepsilon\delta^2}{\mathrm{Wi}}\|\boldsymbol{\nabla} G^{(L)'}_{\delta}(\boldsymbol{\sigma}^{(L)}_{\alpha,\delta})\|^2\right] \\ &\qquad \leq 2\left(F^{(L)}_{\delta}(\boldsymbol{u}^0, \boldsymbol{\sigma}^0) + \frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}\|^2_{L^2(0,T;H^{-1}(\mathcal{D}))}\right).\end{aligned} \tag{3-IV.9}$$

3-V Finite element approximation of $(\mathrm{P}^{(L)}_{\alpha,\delta})$ and $(\mathrm{P}^{(L)}_{\alpha})$

3-V-A Finite element discretization

We now introduce a conforming finite element discretization of $(\mathrm{P}^{(L)}_{\alpha,\delta})$, which satisfies a discrete analogue of (3-IV.3). As noted in the proof of Proposition 15 above, we cannot use S^0_h with the desirable property (3-III.3) to approximate $\boldsymbol{\sigma}^{(L)}_{\alpha,\delta}$, as we now have the added diffusion term. In the following, we choose

$$\mathrm{Q}^1_h = \{q \in C(\overline{\mathcal{D}}) : q\,|_{K_k} \in \mathbb{P}_1 \quad k = 1,\ldots,N_K\} \subset \mathrm{Q}, \tag{3-V.1a}$$

$$\mathrm{S}^1_h = \{\boldsymbol{\phi} \in [C(\overline{\mathcal{D}})]^{d\times d}_S : \boldsymbol{\phi}\,|_{K_k} \in [\mathbb{P}_1]^{d\times d}_S \quad k = 1,\ldots,N_K\} \subset \mathrm{S}. \tag{3-V.1b}$$

$$\text{and} \qquad \mathrm{V}^1_h = \left\{\boldsymbol{v} \in \mathrm{W}_h : \int_{\mathcal{D}} q\,\mathrm{div}\,\boldsymbol{v} = 0 \quad \forall q \in \mathrm{Q}^1_h\right\}. \tag{3-V.1c}$$

This velocity-pressure choice, $\mathrm{W}_h \times \mathrm{Q}_h^1$, is the lowest order Taylor-Hood element; and it is well-known that it satisfies (3-III.1) with Q_h^0 replaced by Q_h^1, see e.g. [BF91]. Hence for all $\boldsymbol{v} \in \mathrm{V}$, there exists a sequence $\{\boldsymbol{v}_h\}_{h>0}$, with $\boldsymbol{v}_h \in \mathrm{V}_h^1$, such that

$$\lim_{h \to 0_+} \|\boldsymbol{v} - \boldsymbol{v}_h\|_{H^1(\mathcal{D})} = 0. \tag{3-V.2}$$

We introduce the interpolation operator $\pi_h : C(\overline{\mathcal{D}}) \mapsto \mathrm{Q}_h^1$, and extended naturally to $\pi_h : [C(\overline{\mathcal{D}})]_S^{d\times d} \mapsto \mathrm{S}_h^1$, such that for all $\eta \in C(\overline{\mathcal{D}})$ and $\boldsymbol{\phi} \in [C(\overline{\mathcal{D}})]_S^{d\times d}$

$$\pi_h \eta(P_p) = \eta(P_p) \qquad \text{and} \qquad \pi_h \boldsymbol{\phi}(P_p) = \boldsymbol{\phi}(P_p) \qquad p = 1, \ldots, N_P, \tag{3-V.3}$$

where $\{P_p\}_{p=1}^{N_P}$ are the vertices of $\mathcal{T}_h$. We require also the L^2 projector $\mathcal{R}_h : \mathrm{V} \mapsto \mathrm{V}_h^1$ defined by

$$\int_{\mathcal{D}} (\boldsymbol{v} - \mathcal{R}_h \boldsymbol{v}) \boldsymbol{w} = 0 \qquad \forall \boldsymbol{w} \in \mathrm{V}_h^1. \tag{3-V.4}$$

In addition, we require $\mathcal{P}_h : \mathrm{S} \mapsto \mathrm{S}_h^1$ defined by

$$\int_{\mathcal{D}} \pi_h[\mathcal{P}_h \boldsymbol{\chi} : \boldsymbol{\phi}] = \int_{\mathcal{D}} \boldsymbol{\chi} : \boldsymbol{\phi} \qquad \forall \boldsymbol{\phi} \in \mathrm{S}_h^1. \tag{3-V.5}$$

As $\boldsymbol{\phi} \in \mathrm{S}_h^1$ does not imply that $G_\delta^{(L)'}(\boldsymbol{\phi}) \in \mathrm{S}_h^1$, we have to test the finite element approximation of the $(\mathrm{P}_{\alpha,\delta}^{(L)})$ version of (3-IV.1c) with $\boldsymbol{I} - \pi_h[G_\delta^{L'}(\boldsymbol{\sigma}_{\alpha,\delta,h}^{(L,)n})] \in \mathrm{S}_h^1$, where $\boldsymbol{\sigma}_{\alpha,\delta,h}^{(L,)n} \in \mathrm{S}_h^1$ is our finite element approximation to $\boldsymbol{\sigma}_{\alpha,\delta}^{(L)}$ at time level t_n. This approximation of the $(\mathrm{P}_{\alpha,\delta}^{(L)})$ version of (3-IV.1c) has to be constructed to mimic the results (3-IV.5)–(3-IV.8), when tested with $\boldsymbol{I} - \pi_h[G_\delta^{(L)'}(\boldsymbol{\sigma}_{\alpha,\delta,h}^{(L,)n})] \in \mathrm{S}_h^1$.

In order to mimic (3-IV.5), see (3-V.24) below, we shall assume from now on that the mesh $\mathcal{T}_h$ for the polytope $\mathcal{D}$ consists of non-obtuse simplices only, i.e. all dihedral angles of any simplex in $\mathcal{T}_h$ are less than or equal to $\frac{\pi}{2}$. We shall also assume that $\mathcal{D}$ is convex and the family $\{\mathcal{T}_h\}_{h>0}$ is quasi-uniform, in order to guarantee the H^1 stability of the L^2 projections $\mathcal{R}_h$ and $\mathcal{P}_h$, see (3-VI.6) and (3-VI.7) below.

In order to mimic (3-IV.6) and (3-IV.8), we need to use numerical integration (vertex sampling).

In order to mimic (3-IV.7), we have to carefully construct our finite element approximation of the convective term in the $(\mathrm{P}_{\alpha,\delta}^{(L)})$ version of (3-IV.1c) This we now do. Let $\{\boldsymbol{e}_i\}_{i=1}^d$ be the orthonormal vectors in $\mathbb{R}^d$, such that the j^{th} component of $\boldsymbol{e}_i$ is δ_{ij}, $i,j = 1, \ldots, d$, on using the Kronecker notation. Let $\widehat{K}$ be the standard open reference simplex in $\mathbb{R}^d$ with vertices $\{\widehat{P}_i\}_{i=0}^d$, where $\widehat{P}_0$ is the origin and $\widehat{P}_i = \boldsymbol{e}_i$, $i = 1, \ldots, d$. Given a simplex $K_k \in \mathcal{T}_h$ with vertices $\{P_i^k\}_{i=0}^d$, then there exists a non-singular matrix B_{K_k} such that the linear mapping

$$\mathcal{B}_{K_k} : \widehat{\boldsymbol{x}} \in \mathbb{R}^d \to P_0^k + B_{K_k} \widehat{\boldsymbol{x}} \in \mathbb{R}^d \tag{3-V.6}$$

maps vertex $\widehat{P}_i$ to vertex P_i^k, $i = 0, \ldots, d$. Hence $\mathcal{B}_{K_k}$ maps $\widehat{K}$ to K_k. For all $\eta \in \mathrm{Q}_h^1$ and $K_k \in \mathcal{T}_h$, we define

$$\widehat{\eta}(\widehat{\boldsymbol{x}}) := \eta(\mathcal{B}_{K_k}(\widehat{\boldsymbol{x}})) \qquad \forall \widehat{\boldsymbol{x}} \in \widehat{K} \qquad \Rightarrow \qquad \boldsymbol{\nabla} \eta(\mathcal{B}_{K_k}(\widehat{\boldsymbol{x}})) = (B_{K_k}^T)^{-1} \widehat{\boldsymbol{\nabla}} \widehat{\eta}(\widehat{\boldsymbol{x}}) \qquad \forall \widehat{\boldsymbol{x}} \in \widehat{K}, \tag{3-V.7}$$

where for all $\widehat{\boldsymbol{x}} \in \widehat{K}$

$$[\widehat{\boldsymbol{\nabla}} \widehat{\eta}(\widehat{\boldsymbol{x}})]_j = \frac{\partial}{\partial \widehat{x}_j} \widehat{\eta}(\widehat{\boldsymbol{x}}) = \widehat{\eta}(\widehat{P}_j) - \widehat{\eta}(\widehat{P}_0) = \eta(P_j^k) - \eta(P_0^k) \qquad j = 1, \ldots, d. \tag{3-V.8}$$

Such notation is easily extended to $\boldsymbol{\phi} \in \mathrm{S}_h^1$.

Given $\boldsymbol{\phi} \in \mathrm{S}_h^1$ and $K_k \in \mathcal{T}_h$, then firstly, for $j = 1, \ldots, d$, we find $\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}}) \in \mathbb{R}_S^{d\times d}$ such that

$$\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}}) : \frac{\partial}{\partial \widehat{x}_j} \widehat{\pi}_h[G_\delta^{(L)'}(\widehat{\boldsymbol{\phi}})] = \frac{\partial}{\partial \widehat{x}_j} \widehat{\pi}_h[\mathrm{tr}(H_\delta^{(L)}(G_\delta^{(L)'}(\widehat{\boldsymbol{\phi}})))] \qquad \text{on } \widehat{K}, \tag{3-V.9}$$

where $(\widehat{\pi}_h \widehat{\eta})(\widehat{\boldsymbol{x}}) \equiv (\pi_h \eta)(\mathcal{B}_{K_k} \widehat{\boldsymbol{x}})$ for all $\widehat{\boldsymbol{x}} \in \widehat{K}$ and $\eta \in C(\overline{K_k})$. To construct $\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}})$ we note the following. We have from (3-II.5), (3-II.6) and (3-II.15) that

$$\begin{aligned} \beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k)) : (G_\delta^{(L)'}(\boldsymbol{\phi}(P_j^k)) - G_\delta^{(L)'}(\boldsymbol{\phi}(P_0^k))) &\le \mathrm{tr}(H_\delta^{(L)}(G_\delta^{(L)'}(\boldsymbol{\phi}(P_j^k)) - H_\delta^{(L)}(G_\delta^{(L)'}(\boldsymbol{\phi}(P_0^k))) \\ &\le \beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k)) : (G_\delta^{(L)'}(\boldsymbol{\phi}(P_j^k)) - G_\delta^{(L)'}(\boldsymbol{\phi}(P_0^k))). \end{aligned} \tag{3-V.10}$$

Next we note from that (3-II.5), (3-II.6), (3-II.7f) and (3-I.2b) that

$$-(\beta_\delta^L(\boldsymbol{\phi}(P_j^k))-\beta_\delta^L(\boldsymbol{\phi}(P_0^k))):(G_\delta^{L\prime}(\boldsymbol{\phi}(P_j^k))-G_\delta^{L\prime}(\boldsymbol{\phi}(P_0^k)))\geq L^{-2}\,\|\beta_\delta^L(\boldsymbol{\phi}(P_j^k))-\beta_\delta^L(\boldsymbol{\phi}(P_0^k))\|^2; \tag{3-V.11}$$

and so the left-hand side is zero if and only if $\beta_\delta^L(\boldsymbol{\phi}(P_j^k))=\beta_\delta^L(\boldsymbol{\phi}(P_0^k))$. Similarly, we see from (3-II.5), (3-II.6) and the proof of (3-II.7f) ; that is, (3-II.13) ; that

$$\begin{aligned}&-(\beta_\delta(\boldsymbol{\phi}(P_j^k))-\beta_\delta(\boldsymbol{\phi}(P_0^k))):(G_\delta'(\boldsymbol{\phi}(P_j^k))-G_\delta'(\boldsymbol{\phi}(P_0^k)))=0\\ &\qquad\Leftrightarrow\qquad \|\beta_\delta(\boldsymbol{\phi}(P_j^k))-\beta_\delta(\boldsymbol{\phi}(P_0^k))\|^2=0\qquad\Leftrightarrow\qquad \beta_\delta(\boldsymbol{\phi}(P_j^k))=\beta_\delta(\boldsymbol{\phi}(P_0^k)).\end{aligned} \tag{3-V.12}$$

Hence, on noting (3-V.8), (3-V.10), (3-V.11), (3-V.12) and (3-I.2b), we have that

$$\begin{aligned}&\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}}):=(1-\lambda_{\delta,j}^{(L)})\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k))+\lambda_{\delta,j}^{(L)}\beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k))\\ &\qquad\qquad\text{if}\quad(\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k))-\beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k))):(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_j^k))-G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_0^k)))\neq 0,\\ \text{and}\qquad&\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}}):=\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k))=\beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k))\\ &\qquad\qquad\text{if}\quad(\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k))-\beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k))):(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_j^k))-G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_0^k)))=0\end{aligned} \tag{3-V.13}$$

satisfies (3-V.9) for $j=1,\dots,d$; where $\lambda_{\delta,j}^{(L)}\in[0,1]$ is defined as

$$\lambda_{\delta,j}^{(L)}:=\frac{\mathrm{tr}(H_\delta^{(L)}(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_j^k))-H_\delta^{(L)}(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_0^k)))-\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k)):(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_j^k))-G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_0^k)))}{(\beta_\delta^{(L)}(\boldsymbol{\phi}(P_0^k))-\beta_\delta^{(L)}(\boldsymbol{\phi}(P_j^k))):(G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_j^k))-G_\delta^{(L)\prime}(\boldsymbol{\phi}(P_0^k)))}.$$

Furthermore, $\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}})\in\mathbb{R}_S^{d\times d}$, $j=1,\dots,d$, depends continuously on $\boldsymbol{\phi}\,|_{K_k}$.

Therefore given $\boldsymbol{\phi}\in\mathrm{S}_h^1$, we introduce, for $m,p=1,\dots,d$,

$$\Lambda_{\delta,m,p}^{(L)}(\boldsymbol{\phi})=\sum_{j=1}^d[(B_{K_k}^T)^{-1}]_{mj}\,\widehat{\Lambda}_{\delta,j}^{(L)}(\widehat{\boldsymbol{\phi}})\,[B_{K_k}^T]_{jp}\in\mathbb{R}_S^{d\times d}\qquad\text{on } K_k,\qquad k=1,\dots,N_K. \tag{3-V.14}$$

It follows from (3-V.14), (3-V.9) and (3-V.7) that

$$\Lambda_{\delta,m,p}^{(L)}(\boldsymbol{\phi})\approx\beta_\delta^{(L)}(\boldsymbol{\phi})[\boldsymbol{I}]_{mp}\qquad m,p=1,\dots,d; \tag{3-V.15}$$

and for $m=1,\dots,d$

$$\sum_{p=1}^d\Lambda_{\delta,m,p}^{(L)}(\boldsymbol{\phi}):\frac{\partial}{\partial x_p}\pi_h[G_\delta^{(L)\prime}(\boldsymbol{\phi})]=\frac{\partial}{\partial x_m}\pi_h[\mathrm{tr}(H_\delta^{(L)}(G_\delta^{(L)\prime}(\boldsymbol{\phi})))]\qquad\text{on } K_k,\qquad k=1,\dots,N_K. \tag{3-V.16}$$

For a more precise version of (3-V.15), see Lemma 12 below. Finally, as the partitioning $\mathcal{T}_h$ consists of regular simplices, we have that

$$\|(B_{K_k}^T)^{-1}\|\,\|B_{K_k}^T\|\leq C,\qquad k=1,\dots,N_K\,. \tag{3-V.17}$$

Hence, it follows from (3-V.14), (3-V.17), (3-V.13) and (3-II.4) that

$$\|\Lambda_{\delta,m,p}^L(\boldsymbol{\phi})\|_{L^\infty(\mathcal{D})}\leq C\,L\qquad\forall\boldsymbol{\phi}\in\mathrm{S}_h^1. \tag{3-V.18}$$

3-V-B A free energy preserving approximation, $(\mathbf{P}_{\alpha,\delta,h}^{(L,)\Delta t})$, of $(\mathbf{P}_{\alpha,\delta}^{(L)})$

In addition to the assumptions on the finite element discretization stated in subsection 3-V-A, and our definition of Δt in subsection 3-III-A, we shall assume that there exists a $C\in\mathbb{R}_{>0}$ such that

$$\Delta t_n\leq C\,\Delta t_{n-1},\qquad n=2,\dots,N,\qquad\text{as}\quad\Delta t\to 0_+. \tag{3-V.19}$$

With Δt_1 and C as above, let $\Delta t_0 \in \mathbb{R}_{>0}$ be such that $\Delta t_1 \le C\Delta t_0$. Given initial data satisfying (3-II.19), we choose $\boldsymbol{u}_h^0 \in \mathrm{V}_h^1$ and $\boldsymbol{\sigma}_h^0 \in \mathrm{S}_h^1$ throughout the rest of this paper such that

$$\int_{\mathcal{D}} \left[\boldsymbol{u}_h^0 \cdot \boldsymbol{v} + \Delta t_0 \boldsymbol{\nabla} \boldsymbol{u}_h^0 : \boldsymbol{\nabla} \boldsymbol{v} \right] = \int_{\mathcal{D}} \boldsymbol{u}^0 \cdot \boldsymbol{v} \qquad \forall \boldsymbol{v} \in \mathrm{V}_h^1, \tag{3-V.20a}$$

$$\boldsymbol{\sigma}_h^0 = \mathcal{P}_h \boldsymbol{\sigma}^0 . \tag{3-V.20b}$$

It follows from (3-V.20b) and (3-V.5) for $p = 1, \ldots, N_P$ and $i, j = 1, \ldots, d$ that

$$[\boldsymbol{\sigma}_h^0]_{ij}(P_p) = \frac{1}{\int_{\mathcal{D}} \eta_p} \int_{\mathcal{D}} [\boldsymbol{\sigma}^0]_{ij} \, \eta_p , \tag{3-V.21}$$

where $\eta_p \in \mathrm{Q}_h^1$ is such that $\eta_p(P_r) = \delta_{pr}$ for $p, r = 1, \ldots, N_P$. Hence it follows from (3-V.20a,b), (3-V.21) and (3-II.19) that

$$\int_{\mathcal{D}} \left[\|\boldsymbol{u}_h^0\|^2 + \Delta t_0 \|\boldsymbol{\nabla} \boldsymbol{u}_h^0\|^2 \right] \le C \tag{3-V.22}$$

$$\text{and for } p = 1, \ldots, N_P \qquad \sigma^0_{\min} \|\boldsymbol{\xi}\|^2 \le \boldsymbol{\xi}^T \boldsymbol{\sigma}_h^0(P_p) \boldsymbol{\xi} \le \sigma^0_{\max} \|\boldsymbol{\xi}\|^2 \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d .$$

On recalling (3-III.7), our approximation $(\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h})$ of $(\mathrm{P}^{(L)}_{\alpha,\delta})$ is then :

$(\mathbf{P}^{(L,)\Delta t}_{\alpha,\delta,h})$ Setting $(\boldsymbol{u}^{(L),0}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L),0}_{\alpha,\delta,L,h}) = (\boldsymbol{u}_h^0, \boldsymbol{\sigma}_h^0) \in \mathrm{V}_h^1 \times \mathrm{S}_h^1$, then for $n = 1, \ldots, N_T$ find $(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) \in \mathrm{V}_h^1 \times \mathrm{S}_h^1$ such that for any test functions $(\boldsymbol{v}, \boldsymbol{\phi}) \in \mathrm{V}_h^1 \times \mathrm{S}_h^1$

$$\int_{\mathcal{D}} \Bigg[\mathrm{Re} \left(\frac{\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} - \boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}}{\Delta t_n} \right) \cdot \boldsymbol{v} + \frac{\mathrm{Re}}{2} \left[\left((\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h} \cdot \nabla) \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} \right) \cdot \boldsymbol{v} - \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} \cdot \left((\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h} \cdot \nabla) \boldsymbol{v} \right) \right] \tag{3-V.23a}$$

$$+ (1-\varepsilon) \boldsymbol{\nabla} \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} : \boldsymbol{\nabla} \boldsymbol{v} + \frac{\varepsilon}{\mathrm{Wi}} \pi_h [\beta^{(L)}_\delta (\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})] : \boldsymbol{\nabla} \boldsymbol{v} \Bigg] = \langle \boldsymbol{f}^n, \boldsymbol{v} \rangle_{H^1_0(\mathcal{D})} ,$$

$$\int_{\mathcal{D}} \pi_h \left[\left(\frac{\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h} - \boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h}}{\Delta t_n} \right) : \boldsymbol{\phi} + \frac{1}{\mathrm{Wi}} \left(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h} - \boldsymbol{I} \right) : \boldsymbol{\phi} \right] + \alpha \int_D \boldsymbol{\nabla} \boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h} :: \boldsymbol{\nabla} \boldsymbol{\phi} \tag{3-V.23b}$$

$$- 2 \int_{\mathcal{D}} \boldsymbol{\nabla} \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} : \pi_h [\boldsymbol{\phi} \beta^{(L)}_\delta (\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})] - \int_{\mathcal{D}} \sum_{m=1}^{d} \sum_{p=1}^{d} [\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}]_m \Lambda^{(L)}_{\delta,m,p} (\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) : \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{x}_p} = 0 .$$

In deriving $(\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h})$, we have noted (3-III.11), (3-I.4a) and (3-V.14).

We now prove the discrete analogue of (3-IV.5).

Lemma 11. *If $\mathcal{T}_h$ consists of only non-obtuse simplices, then we have for all $\boldsymbol{\phi} \in \mathrm{S}_h^1$ that*

$$-\boldsymbol{\nabla} \pi_h [G^{(L)'}_\delta(\boldsymbol{\phi})] :: \boldsymbol{\nabla} \boldsymbol{\phi} \ge \delta^2 \|\boldsymbol{\nabla} \pi_h [G^{(L)'}_\delta(\boldsymbol{\phi})]\|^2 \qquad \text{on } K_k, \quad k = 1, \ldots, N_K . \tag{3-V.24}$$

Proof. Let K_k have vertices $\{P_j^k\}_{j=0}^d$, and let $\eta_j^k(\boldsymbol{x})$ be the basis functions on K_k associated with $\mathrm{Q}^1_{\mathrm{h}}$ and S_h^1, i.e. $\eta_j^k|_{K_k} \in \mathbb{P}_1$ and $\eta_j^k(P_i^k) = \delta_{ij}$, $i, j = 0, \ldots, d$. As K_k is non-obtuse it follows that

$$\boldsymbol{\nabla} \eta_i^k \cdot \boldsymbol{\nabla} \eta_j^k \le 0 \qquad \text{on } K_k, \qquad i, j = 0, \ldots, d, \ \text{ with } \ i \ne j . \tag{3-V.25}$$

We note that

$$\sum_{j=0}^{d} \eta_j^k \equiv 1 \quad \text{on} \quad K_k \quad \Rightarrow \quad \|\boldsymbol{\nabla} \eta_i^k\|^2 = - \sum_{j=0,\ j \ne i}^{d} \boldsymbol{\nabla} \eta_i^k \cdot \boldsymbol{\nabla} \eta_j^k \quad \text{on} \quad K_k, \quad i = 0, \ldots, d. \tag{3-V.26}$$

Hence for $a_i, b_i \in \mathbb{R}$, $i = 0, \ldots d$, we have that

$$\boldsymbol{\nabla} (\sum_{i=0}^{d} a_i \eta_i^k) \cdot \boldsymbol{\nabla} (\sum_{j=0}^{d} b_j \eta_j^k) = \sum_{i=0}^{d} \left[a_i b_i \|\boldsymbol{\nabla} \eta_i^k\|^2 + \sum_{j=0,\ j \ne i}^{d} a_i b_j \boldsymbol{\nabla} \eta_i^k \cdot \boldsymbol{\nabla} \eta_j^k \right] \tag{3-V.27}$$

$$= - \sum_{i=0}^{d} \sum_{j=0,\ j \ne i}^{d} a_i (b_i - b_j) \boldsymbol{\nabla} \eta_i^k \cdot \boldsymbol{\nabla} \eta_j^k$$

$$= - \sum_{i=0}^{d} \sum_{j>i}^{d} (a_i - a_j)(b_i - b_j) \boldsymbol{\nabla} \eta_i^k \cdot \boldsymbol{\nabla} \eta_j^k .$$

Similarly for $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^{d\times d}_S$, $i=0,\ldots,d$, we have that

$$\boldsymbol{\nabla}(\sum_{i=0}^{d} \boldsymbol{a}_i \eta_i^k) :: \boldsymbol{\nabla}(\sum_{j=0}^{d} \boldsymbol{b}_j \eta_j^k) = -\sum_{i=0}^{d}\sum_{j>i}^{d} [(\boldsymbol{a}_i - \boldsymbol{a}_j) : (\boldsymbol{b}_i - \boldsymbol{b}_j)]\, \boldsymbol{\nabla}\eta_i^k \cdot \boldsymbol{\nabla}\eta_j^k . \tag{3-V.28}$$

The desired result (3-V.24) then follows immediately from (3-V.28), (3-V.25) and (3-II.7f). □

Before proving existence of a solution to ($\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h}$), we first derive a discrete analogue of the energy estimate (3-IV.3) for ($\mathrm{P}^{(L)}_{\alpha,\delta}$).

3-V-C Energy estimate

On setting

$$F^{(L)}_{\delta,h}(\boldsymbol{v},\boldsymbol{\phi}) := \frac{\mathrm{Re}}{2}\int_{\mathcal{D}} \|\boldsymbol{v}\|^2 + \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \pi_h[\mathrm{tr}\left(\boldsymbol{\phi} - G^{(L)}_{\delta}(\boldsymbol{\phi}) - \boldsymbol{I}\right)] \qquad \forall (\boldsymbol{v},\boldsymbol{\phi}) \in \mathrm{V}^1_h \times \mathrm{S}^1_h , \tag{3-V.29}$$

we have the following discrete analogue of Proposition 15.

Proposition 16. *For $n=1,\ldots,N_T$, a solution $\left(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}\right) \in \mathrm{V}^1_h \times \mathrm{S}^1_h$ to (3-V.23a,b), if it exists, satisfies*

$$\begin{aligned}
&\frac{F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) - F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h})}{\Delta t_n} + \frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}} \|\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} - \boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\|^2 \\
&\quad + (1-\varepsilon)\int_{\mathcal{D}} \|\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\|^2 + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}} \pi_h[\mathrm{tr}(\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) + [\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]^{-1} - 2\boldsymbol{I})] \\
&\quad + \frac{\alpha\varepsilon\delta^2}{2\mathrm{Wi}}\int_{\mathcal{D}} \|\boldsymbol{\nabla}\pi_h[{G^{(L)}_{\delta}}'(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]\|^2 \\
&\le \langle \boldsymbol{f}^n, \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\rangle_{H^1_0(\mathcal{D})} \le \frac{1}{2}(1-\varepsilon)\int_{\mathcal{D}} \|\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\|^2 + \frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|^2_{H^{-1}(\mathcal{D})} .
\end{aligned} \tag{3-V.30}$$

Proof. The proof is similar to that of Proposition 13, we choose as test functions $\boldsymbol{v} = \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} \in \mathrm{V}^1_h$ and $\boldsymbol{\phi} = \frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I} - \pi_h[{G^{(L)}_{\delta}}'(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]\right) \in \mathrm{S}^1_h$ in (3-V.23a,b), and obtain, on noting (3-III.12), (3-II.7a,d,e), (3-V.24), (3-V.16) and (3-V.29) that

$$\begin{aligned}
&\langle \boldsymbol{f}^n, \boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\rangle_{H^1_0(\mathcal{D})} \\
&\ge \frac{F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) - F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}, \boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h})}{\Delta t_n} + \frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}} \|\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h} - \boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\|^2 \\
&\quad + (1-\varepsilon)\int_{\mathcal{D}} \|\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\|^2 + \frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}} \pi_h[\mathrm{tr}(\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}) + [\beta^{(L)}_{\delta}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]^{-1} - 2\boldsymbol{I})] \\
&\quad + \frac{\alpha\varepsilon\delta^2}{2\mathrm{Wi}}\int_{\mathcal{D}} \|\boldsymbol{\nabla}\pi_h[{G^{(L)}_{\delta}}'(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]\|^2 + \int_{\mathcal{D}} \boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h} \cdot \boldsymbol{\nabla}\pi_h[\mathrm{tr}(H^{(L)}_{\delta}({G^{(L)}_{\delta}}'(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})))] .
\end{aligned} \tag{3-V.31}$$

The first desired inequality in (3-V.30) follows immediately from (3-V.31) on noting (3-V.1c), (3-III.2a), (3-I.9) and that $\pi_h : C(\overline{\mathcal{D}}) \mapsto \mathrm{Q}^1_h$. The second inequality in (3-V.30) follows immediately from (3-II.28) with $\nu^2 = (1-\varepsilon)/(1+C_P)$. □

3-V-D Existence of discrete solutions

We recall the well-known local inverse inequality for Q^1_h

$$\begin{aligned}
&\|q\|_{L^\infty(K_k)} \le C\,|K_k|^{-1}\int_{K_k} |q| \qquad \forall q \in \mathrm{Q}^1_h, \qquad k=1,\ldots,N_K \\
\Rightarrow \qquad &\|\boldsymbol{\chi}\|_{L^\infty(K_k)} \le C\,|K_k|^{-1}\int_{K_k} \|\boldsymbol{\chi}\| \qquad \forall \boldsymbol{\chi} \in \mathrm{S}^1_h, \qquad k=1,\ldots,N_K .
\end{aligned} \tag{3-V.32}$$

We recall a similar well-known local inverse inequality for V^1_h

$$\|\boldsymbol{\nabla}\boldsymbol{v}\|_{L^2(K_k)} \le C\,h^{-1}\|\boldsymbol{v}\|_{L^2(K_k)} \qquad \forall \boldsymbol{v} \in \mathrm{V}^1_h, \qquad k=1,\ldots,N_K . \tag{3-V.33}$$

Proposition 17. *Given* $(\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h})\in \mathrm{V}^1_h\times\mathrm{S}^1_h$ *and for any time step* $\Delta t_n>0$*, then there exists at least one solution* $\left(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}\right)\in \mathrm{V}^1_h\times\mathrm{S}^1_h$ *to (3-V.23a,b).*

Proof. The proof is similar to that of Proposition 14. We introduce the following inner product on the Hilbert space $\mathrm{V}^1_h\times\mathrm{S}^1_h$

$$((\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi}))^h_{\mathcal{D}}=\int_{\mathcal{D}}[\boldsymbol{w}\cdot\boldsymbol{v}+\pi_h[\boldsymbol{\psi}:\boldsymbol{\phi}]]\qquad\forall(\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h.$$

Given $(\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h})\in\mathrm{V}^1_h\times\mathrm{S}^1_h$, let $\mathcal{F}^h:\mathrm{V}^1_h\times\mathrm{S}^1_h\mapsto\mathrm{V}^1_h\times\mathrm{S}^1_h$ be such that for any $(\boldsymbol{w},\boldsymbol{\psi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h$

$$\begin{aligned}&\left(\mathcal{F}^h(\boldsymbol{w},\boldsymbol{\psi}),(\boldsymbol{v},\boldsymbol{\phi})\right)^h_{\mathcal{D}}\\&:=\int_{\mathcal{D}}\left[\mathrm{Re}\left(\frac{\boldsymbol{w}-\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}}{\Delta t_n}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\left[\left((\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\cdot\nabla)\boldsymbol{w}\right)\cdot\boldsymbol{v}-\boldsymbol{w}\cdot\left((\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\cdot\nabla)\boldsymbol{v}\right)\right]\right.\\&\quad+(1-\varepsilon)\boldsymbol{\nabla w}:\boldsymbol{\nabla v}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h[\beta^{(L)}_\delta(\boldsymbol{\psi})]:\boldsymbol{\nabla v}+\alpha\boldsymbol{\nabla\psi}::\boldsymbol{\nabla\phi}-2\boldsymbol{\nabla w}:\pi_h[\boldsymbol{\phi}\beta^{(L)}_\delta(\boldsymbol{\psi})]\\&\quad\left.+\pi_h\left[\left(\frac{\boldsymbol{\psi}-\boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h}}{\Delta t_n}\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}(\boldsymbol{\psi}-\boldsymbol{I}):\boldsymbol{\phi}\right]\right]-\langle\boldsymbol{f}^n,\boldsymbol{v}\rangle_{H^1_0(\mathcal{D})}\\&\quad-\int_{\mathcal{D}}\sum_{m=1}^{d}\sum_{p=1}^{d}[\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}]_m\Lambda^{(L)}_{\delta,m,p}(\boldsymbol{\psi}):\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{x}_p}\qquad\forall(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h.\end{aligned}\tag{3-V.34}$$

A solution $(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})$ to (3-III.10a,b), if it exists, corresponds to a zero of $\mathcal{F}^h$. On recalling (3-V.14) and (3-V.13), it is easily deduced that the mapping $\mathcal{F}^h$ is continuous.

For any $(\boldsymbol{w},\boldsymbol{\psi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h$, on choosing $(\boldsymbol{v},\boldsymbol{\phi})=\left(\boldsymbol{w},\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\pi_h[{G^{(L)}_\delta}'(\boldsymbol{\psi})]\right)\right)$, we obtain analogously to (3-V.30) that

$$\begin{aligned}&\left(\mathcal{F}^h(\boldsymbol{w},\boldsymbol{\psi}),\left(\boldsymbol{w},\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\pi_h[{G^{(L)}_\delta}'(\boldsymbol{\psi})]\right)\right)\right)^h_{\mathcal{D}}\\&\geq\frac{F^{(L)}_{\delta,h}(\boldsymbol{w},\boldsymbol{\psi})-F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n-1}_{\alpha,\delta,L,h})}{\Delta t_n}+\frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}}\|\boldsymbol{w}-\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\|^2+\frac{1-\varepsilon}{2}\int_{\mathcal{D}}\|\boldsymbol{\nabla w}\|^2\\&\quad+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\pi_h[\mathrm{tr}(\beta^{(L)}_\delta(\boldsymbol{\psi})+[\beta^{(L)}_\delta(\boldsymbol{\psi})]^{-1}-2\boldsymbol{I})]+\frac{\alpha\varepsilon\delta^2}{2\mathrm{Wi}}\int_{\mathcal{D}}\|\boldsymbol{\nabla}\pi_h[{G^{(L)}_\delta}'(\boldsymbol{\psi})]\|^2\\&\quad-\frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|^2_{H^{-1}(\mathcal{D})}.\end{aligned}\tag{3-V.35}$$

If for any $\gamma\in\mathbb{R}_{>0}$, the continuous mapping $\mathcal{F}^h$ has no zero $(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})$, which lies in the ball

$$\mathcal{B}^h_\gamma:=\left\{(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h:\|(\boldsymbol{v},\boldsymbol{\phi})\|^h_{\mathcal{D}}\leq\gamma\right\};$$

where

$$\|(\boldsymbol{v},\boldsymbol{\phi})\|^h_D:=\left[((\boldsymbol{v},\boldsymbol{\phi}),(\boldsymbol{v},\boldsymbol{\phi}))^h_D\right]^{\frac12}=\left(\int_{\mathcal{D}}\left[\|\boldsymbol{v}\|^2+\pi_h[\|\boldsymbol{\phi}\|^2]\right]\right)^{\frac12}.$$

Then for such γ, we can define the continuous mapping $\mathcal{G}^h_\gamma:\mathcal{B}^h_\gamma\mapsto\mathcal{B}^h_\gamma$ such that for all $(\boldsymbol{v},\boldsymbol{\phi})\in\mathcal{B}^h_\gamma$

$$\mathcal{G}^h_\gamma(\boldsymbol{v},\boldsymbol{\phi}):=-\gamma\frac{\mathcal{F}^h(\boldsymbol{v},\boldsymbol{\phi})}{\|\mathcal{F}^h(\boldsymbol{v},\boldsymbol{\phi})\|^h_{\mathcal{D}}}.$$

By the Brouwer fixed point theorem, $\mathcal{G}^h_\gamma$ has at least one fixed point $(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)$ in $\mathcal{B}^h_\gamma$. Hence it satisfies

$$\left\|(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\right\|^h_{\mathcal{D}}=\left\|\mathcal{G}^h_\gamma(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\right\|^h_{\mathcal{D}}=\gamma.\tag{3-V.36}$$

On noting (3-V.32), we have that there exists a $\mu_h\in\mathbb{R}_{>0}$ such that for all $\boldsymbol{\phi}\in\mathrm{S}^1_h$,

$$\|\pi_h[\|\boldsymbol{\phi}\|]\|^2_{L^\infty(\mathcal{D})}\leq\|\pi_h[\|\boldsymbol{\phi}\|^2]\|_{L^\infty(\mathcal{D})}\leq\mu_h^2\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\phi}\|^2].\tag{3-V.37}$$

It follows from (3-V.29), (3-II.8), (3-V.37) and (3-V.36) that

$$
\begin{aligned}
F^{(L)}_{\delta,h}(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma) &= \frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\mathrm{tr}(\boldsymbol{\psi}_\gamma-G^{(L)}_\delta(\boldsymbol{\psi}_\gamma)-\boldsymbol{I})] \\
&\geq \frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{4\mathrm{Wi}}\left[\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\psi}_\gamma\|]-2d|\mathcal{D}|\right] \\
&\geq \frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\|\pi_h[\|\boldsymbol{\psi}_\gamma\|]\|_{L^\infty(\mathcal{D})}\left[\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\psi}_\gamma\|]\right]-\frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}} \\
&\geq \min\left(\frac{\mathrm{Re}}{2},\frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\left(\int_{\mathcal{D}}\left[\|\boldsymbol{w}_\gamma\|^2+\pi_h[\|\boldsymbol{\psi}_\gamma\|^2]\right]\right)-\frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}} \\
&= \min\left(\frac{\mathrm{Re}}{2},\frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\gamma^2-\frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}}\,.
\end{aligned}
\tag{3-V.38}
$$

Hence for all γ sufficiently large, it follows from (3-V.35) and (3-V.38) that

$$
\left(\mathcal{F}^h(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma),\left(\boldsymbol{w}_\gamma,\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\pi_h[G^{(L)'}_\delta(\boldsymbol{\psi}_\gamma)]\right)\right)\right)^h_{\mathcal{D}}\geq 0. \tag{3-V.39}
$$

On the other hand as $(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)$ is a fixed point of $\mathcal{G}^h_\gamma$, we have that

$$
\begin{aligned}
&\left(\mathcal{F}^h(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma),\left(\boldsymbol{w}_\gamma,\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\pi_h[G^{(L)'}_\delta(\boldsymbol{\psi}_\gamma)]\right)\right)\right)^h_D \\
&\qquad = -\frac{\left\|\mathcal{F}^h(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma)\right\|^h_{\mathcal{D}}}{\gamma}\int_{\mathcal{D}}\left[\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\pi_h[\boldsymbol{\psi}_\gamma:\left(\boldsymbol{I}-G^{(L)'}_\delta(\boldsymbol{\psi}_\gamma)\right)]\right].
\end{aligned}
\tag{3-V.40}
$$

It follows from (3-II.8), and similarly to (3-V.38), on noting (3-V.37) and (3-V.36) that

$$
\begin{aligned}
\int_{\mathcal{D}}\left[\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\pi_h[\boldsymbol{\psi}_\gamma:\left(\boldsymbol{I}-G^{(L)'}_\delta(\boldsymbol{\psi}_\gamma)\right)]\right] &\geq \int_{\mathcal{D}}\left[\|\boldsymbol{w}_\gamma\|^2+\frac{\varepsilon}{4\mathrm{Wi}}\left[\pi_h[\|\boldsymbol{\psi}_\gamma\|]-2d\right]\right] \\
&\geq \min\left(1,\frac{\varepsilon}{4\mathrm{Wi}\mu_h\gamma}\right)\gamma^2-\frac{\varepsilon d|\mathcal{D}|}{2\mathrm{Wi}}.
\end{aligned}
\tag{3-V.41}
$$

Therefore on combining (3-V.40) and (3-V.41), we have for all γ sufficiently large that

$$
\left(\mathcal{F}^h(\boldsymbol{w}_\gamma,\boldsymbol{\psi}_\gamma),\left(\boldsymbol{w}_\gamma,\frac{\varepsilon}{2\mathrm{Wi}}\left(\boldsymbol{I}-\pi_h[G^{(L)'}_\delta(\boldsymbol{\psi}_\gamma)]\right)\right)\right)^h_D<0, \tag{3-V.42}
$$

which obviously contradicts (3-V.39). Hence the mapping $\mathcal{F}^h$ has a zero in $\mathcal{B}^h_\gamma$ for γ sufficiently large. □

We now have the analogue of stability Theorem 3.

Theorem 5. *For any $\delta\in(0,\frac{1}{2}]$, $L\geq 2$, $N_T\geq 1$ and any partitioning of $[0,T]$ into N_T time steps, there exists a solution $\{(\boldsymbol{u}^{(L),n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L),n}_{\alpha,\delta,L,h})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_h]^{N_T}$ to $(P^{(L,)\Delta t}_{\alpha,\delta,h})$.*

In addition, it follows for $n=1,\ldots,N_T$ that

$$
\begin{aligned}
&F^{(L)}_{\delta,h}(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})+\frac{1}{2}\sum_{m=1}^{n}\int_{\mathcal{D}}\left[\mathrm{Re}\|\boldsymbol{u}^{(L,)m}_{\alpha,\delta,L,h}-\boldsymbol{u}^{(L,)m-1}_{\alpha,\delta,L,h}\|^2+(1-\varepsilon)\Delta t_m\|\boldsymbol{\nabla}\boldsymbol{u}^{(L,)m}_{\alpha,\delta,L,h}\|^2\right] \\
&\qquad+\frac{\varepsilon}{2\mathrm{Wi}^2}\sum_{m=1}^{n}\Delta t_m\int_{\mathcal{D}}\pi_h[\mathrm{tr}(\beta^{(L)}_\delta(\boldsymbol{\sigma}^{(L,)m}_{\alpha,\delta,L,h})+[\beta^{(L)}_\delta(\boldsymbol{\sigma}^{(L,)m}_{\alpha,\delta,L,h})]^{-1}-2\boldsymbol{I})] \\
&\qquad+\frac{\alpha\varepsilon\delta^2}{2\mathrm{Wi}}\sum_{m=1}^{n}\Delta t_m\int_{\mathcal{D}}\|\boldsymbol{\nabla}\pi_h[G^{(L)'}_\delta(\boldsymbol{\sigma}^{(L,)m}_{\alpha,\delta,L,h})]\|^2 \\
&\leq F^{(L)}_{\delta,h}(\boldsymbol{u}^0_h,\boldsymbol{\sigma}^0_h)+\frac{1+C_P}{2(1-\varepsilon)}\sum_{m=1}^{n}\Delta t_m\|\boldsymbol{f}^m\|^2_{H^{-1}(\mathcal{D})}\leq C.
\end{aligned}
\tag{3-V.43}
$$

Moreover, it follows that

$$\max_{n=0,\dots,N_T}\int_{\mathcal{D}}\Big[\|\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\|^2+\pi_h[\|\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}\|]+\delta^{-1}\pi_h[\|[\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}]_-\|]\Big] \tag{3-V.44}$$
$$+\sum_{n=1}^{N_T}\int_{\mathcal{D}}\Big[\Delta t_n\|\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\|^2+\Delta t_n\pi_h[\|[\beta^{(L)}_\delta(\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})]^{-1}\|]+\|\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}-\boldsymbol{u}^{(L,)n-1}_{\alpha,\delta,L,h}\|^2\Big]\le C\,.$$

Proof. Existence and the stability result (3-V.43) follow immediately from Propositions 17 and 16, respectively, on noting (3-V.29), (3-V.22), (3-III.8a) and (3-II.19). The bounds (3-V.44) follow immediately from (3-V.43), on noting (3-II.7b), (3-II.8), (3-I.7) and the fact that $\beta^{(L)}_\delta(\boldsymbol{\phi})\in\mathbb{R}^{d\times d}_{SPD}$ for any $\boldsymbol{\phi}\in\mathbb{R}^{d\times d}_S$. □

3-V-E Convergence of $(\mathrm{P}^{(L,)\Delta t}_{\alpha,\delta,h})$ to $(\mathrm{P}^{(L,)\Delta t}_{\alpha,h})$

We now consider the corresponding direct finite element approximation of $(\mathrm{P}^{(L)}_\alpha)$, i.e. $(\mathrm{P}^{(L,)\Delta t}_{\alpha,h})$ without the regularization δ :

We introduce

$$\mathrm{S}^1_{h,PD}=\{\boldsymbol{\phi}\in\mathrm{S}^1_h:\boldsymbol{\phi}(P_p)\in\mathbb{R}^{d\times d}_{SPD}\quad\text{for}\quad p=1,\dots,N_P\}\subset\mathrm{S}_{PD}\,. \tag{3-V.45}$$

It follows from (3-V.22) that $\boldsymbol{\sigma}^0_h\in\mathrm{S}^1_{h,PD}$.

$(\mathbf{P}^{(L,)\Delta t}_{\alpha,h})$ Setting $(\boldsymbol{u}^{(L),0}_{\alpha,h},\boldsymbol{\sigma}^{(L),0}_{\alpha,h})=(\boldsymbol{u}^0_h,\boldsymbol{\sigma}^0_h)\in\mathrm{V}^1_h\times\mathrm{S}^1_{h,PD}$, then for $n=1,\dots,N_T$ find $(\boldsymbol{u}^{(L,)n}_{\alpha,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,h})\in\mathrm{V}^1_h\times\mathrm{S}^1_h$ such that for any test functions $(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^1_h\times\mathrm{S}^1_h$

$$\int_{\mathcal{D}}\Big[\mathrm{Re}\left(\frac{\boldsymbol{u}^{(L,)n}_{\alpha,h}-\boldsymbol{u}^{(L,)n-1}_{\alpha,h}}{\Delta t_n}\right)\cdot\boldsymbol{v}+\frac{\mathrm{Re}}{2}\Big[\Big((\boldsymbol{u}^{(L,)n-1}_{\alpha,h}\cdot\nabla)\boldsymbol{u}^{(L,)n}_{\alpha,h}\Big)\cdot\boldsymbol{v}-\boldsymbol{u}^{(L,)n}_{\alpha,h}\cdot\Big((\boldsymbol{u}^{(L,)n-1}_{\alpha,h}\cdot\nabla)\boldsymbol{v}\Big)\Big] \tag{3-V.46a}$$
$$+(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,h}:\boldsymbol{\nabla}\boldsymbol{v}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h[\beta^{(L)}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,h})]:\boldsymbol{\nabla}\boldsymbol{v}\Big]=\langle\boldsymbol{f}^n,\boldsymbol{v}\rangle_{H^1_0(\mathcal{D})}\,,$$
$$\int_{\mathcal{D}}\pi_h\left[\left(\frac{\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}-\boldsymbol{\sigma}^{(L,)n-1}_{\alpha,h}}{\Delta t_n}\right):\boldsymbol{\phi}+\frac{1}{\mathrm{Wi}}\Big(\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}-\boldsymbol{I}\Big):\boldsymbol{\phi}\right]+\alpha\int_D\boldsymbol{\nabla}\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}::\boldsymbol{\nabla}\boldsymbol{\phi} \tag{3-V.46b}$$
$$-2\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}^{(L,)n}_{\alpha,h}:\pi_h[\boldsymbol{\phi}\beta^{(L)}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,h})]-\int_{\mathcal{D}}\sum_{m=1}^d\sum_{p=1}^d[\boldsymbol{u}^{(L,)n-1}_{\alpha,h}]_m\,\Lambda^{(L)}_{m,p}(\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}):\frac{\partial\boldsymbol{\phi}}{\partial\boldsymbol{x}_p}=0\,.$$

Remark 11. *Due to the presence of $\beta^{(L)}$ in (3-V.46a,b), it is implicitly assumed that $\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}\in\mathrm{S}^1_{h,PD}$, $n=1,\dots,N_T$; recall (3-II.2). In addition, $\Lambda^{(L)}_{m,p}(\boldsymbol{\phi})$ for $\boldsymbol{\phi}\in\mathrm{S}^1_{h,PD}$ is defined similarly to (3-V.14) with $\widehat{\Lambda}^{(L)}_{\delta,j}(\widehat{\boldsymbol{\phi}})$ replaced by $\widehat{\Lambda}^{(L)}_j(\widehat{\boldsymbol{\phi}})$, which is defined similarly to (3-V.13) with $\lambda^{(L)}_{\delta,j},\beta^{(L)}_\delta$ and $G^{(L)}_\delta$ replaced by $\lambda^{(L)}_j,\beta^{(L)}$ and $G^{(L)}$, with $\lambda^{(L)}_j$ defined similarly to $\lambda^{(L)}_{\delta,j}$ with $\beta^{(L)}_\delta,G^{(L)}_\delta$ and $H^{(L)}_\delta$ replaced by $\beta^{(L)},G^{(L)}$ and $H^{(L)}$. Hence, similarly to (3-V.18), we have that*

$$\|\Lambda^L_{m,p}(\boldsymbol{\phi})\|_{L^\infty(\mathcal{D})}\le C\,L\qquad\forall\boldsymbol{\phi}\in\mathrm{S}^1_{h,PD}. \tag{3-V.47}$$

We introduce also the unregularised free energy

$$F^{(L)}_h(\boldsymbol{v},\boldsymbol{\phi}):=\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\|\boldsymbol{v}\|^2+\frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\mathrm{tr}(\boldsymbol{\phi}-G^{(L)}(\boldsymbol{\phi})-\boldsymbol{I})]\,, \tag{3-V.48}$$

which is well defined for all $(\boldsymbol{v},\boldsymbol{\phi})\in\mathrm{V}^1_h\times\mathrm{S}^1_{h,PD}$.

Theorem 6. *For all regular partitionings $\mathcal{T}_h$ of $\mathcal{D}$ into simplices $\{K_k\}_{k=1}^{N_K}$ and all partitionings $\{\Delta t_n\}_{n=1}^{N_T}$ of $[0,T]$, there exists a subsequence $\{\{(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})\}_{n=1}^{N_T}\}_{\delta>0}$, where $\{(\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_h]^{N_T}$ solves $(P^{(L,)\Delta t}_{\alpha,\delta,h})$, and $\{(\boldsymbol{u}^{(L,)n}_{\alpha,h},\boldsymbol{\sigma}^{(L,)n}_{\alpha,h})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_h]^{N_T}$ such that for the subsequence*

$$\boldsymbol{u}^{(L,)n}_{\alpha,\delta,L,h}\to\boldsymbol{u}^{(L,)n}_{\alpha,h},\qquad\boldsymbol{\sigma}^{(L,)n}_{\alpha,\delta,L,h}\to\boldsymbol{\sigma}^{(L,)n}_{\alpha,h}\qquad\text{as }\delta\to0_+,\qquad\text{for}\quad n=1,\dots,N_T\,. \tag{3-V.49}$$

In addition, for $n=1,\dots,N_T$, $\boldsymbol{\sigma}_{\alpha,h}^{(L,)n}\in \mathrm{S}^1_{h,PD}$, *and* $\{(\boldsymbol{u}_{\alpha,h}^{(L,)n},\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_{h,PD}]^{N_T}$ *solves* ($P_{\alpha,h}^{(L,)\Delta t}$).

Moreover, we have for $n=1,\dots,N_T$ *that*

$$\begin{aligned}\frac{F_h^{(L)}(\boldsymbol{u}_{\alpha,h}^{(L,)n},\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})-F_h^{(L)}(\boldsymbol{u}_{\alpha,h}^{(L,)n-1},\boldsymbol{\sigma}_{\alpha,h}^{(L,)n-1})}{\Delta t_n}+\frac{\mathrm{Re}}{2\Delta t_n}\int_{\mathcal{D}}\|\boldsymbol{u}_{\alpha,h}^{(L,)n}-\boldsymbol{u}_{\alpha,h}^{(L,)n-1}\|^2 \\ +(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{(L,)n}\|^2+\frac{\varepsilon}{2\mathrm{Wi}^2}\int_{\mathcal{D}}\pi_h[\mathrm{tr}(\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})+[\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})]^{-1}-2\boldsymbol{I})] \\ \le\frac{1}{2}(1-\varepsilon)\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{(L,)n}\|^2+\frac{1+C_P}{2(1-\varepsilon)}\|\boldsymbol{f}^n\|^2_{H^{-1}(\mathcal{D})},\end{aligned}\tag{3-V.50}$$

and

$$\begin{aligned}&\max_{n=0,\dots,N_T}\int_{\mathcal{D}}\Big[\|\boldsymbol{u}_{\alpha,h}^{(L,)n}\|^2+\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{(L,)n}\|]\Big] \\ &+\sum_{n=1}^{N_T}\int_{\mathcal{D}}\Big[\Delta t_n\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{(L,)n}\|^2+\Delta t_n\pi_h[\|[\beta^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})]^{-1}\|]+\|\boldsymbol{u}_{\alpha,h}^{(L,)n}-\boldsymbol{u}_{\alpha,h}^{(L,)n-1}\|^2\Big]\le C.\end{aligned}\tag{3-V.51}$$

Proof. For any integer $n\in[1,N_T]$, the desired subsequence convergence result (3-V.49) follows immediately from (3-V.44), as ($\boldsymbol{u}_{\alpha,\delta,L,h}^{(L,)n}$, $\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n}$) are finite dimensional for fixed $\mathrm{V}^1_h\times S^1_h$. It also follows from (3-V.44), (3-V.49) and (3-II.17) that $\pi_h[[\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n}]_-]$ vanishes on $\mathcal{D}$, so that $\boldsymbol{\sigma}_{\alpha,h}^{(L,)n}$ must be non-negative definite on $\mathcal{D}$. Hence on noting this, (3-II.3), (3-II.17) and (3-V.49), we have the following subsequence convergence results

$$\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\to\beta^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\quad\text{and}\quad\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})\to\beta^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\quad\text{as}\quad\delta\to0_+.\tag{3-V.52}$$

It also follows from (3-V.44), (3-V.52) and as $[\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})]^{-1}\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})=\boldsymbol{I}$ that the following subsequence result

$$\pi_h[[\beta_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})]^{-1}]\to\pi_h[[\beta^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})]^{-1}]\qquad\text{as}\quad\delta\to0_+\tag{3-V.53}$$

holds, and so $\boldsymbol{\sigma}_{\alpha,h}^{(L,)n}\in\mathrm{S}^1_{h,PD}$. Therefore, we have from (3-V.49) and (3-II.1) that

$$G_\delta^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})\to G^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\qquad\text{as}\quad\delta\to0_+.\tag{3-V.54}$$

Similarly to (3-V.54), it follows from (3-V.49), (3-V.52), (3-V.14) and (3-V.13) as $\boldsymbol{\sigma}_{\alpha,h}^{(L,),n}\in\mathrm{S}^1_{h,PD}$ that for $m,p=1,\dots,d$

$$\Lambda_{\delta,m,p}^{(L)}(\boldsymbol{\sigma}_{\alpha,\delta,L,h}^{(L,)n})\to\Lambda_{m,p}^{(L)}(\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\quad\text{as}\quad\delta\to0_+.\tag{3-V.55}$$

Hence using (3-V.49), (3-V.52) and (3-V.55), we can pass to the limit $\delta\to0_+$ in ($\mathrm{P}_{\alpha,\delta,h}^{(L,)\Delta t}$), (3-V.23a,b), to show that $\{(\boldsymbol{u}_{\alpha,h}^{(L,)n},\boldsymbol{\sigma}_{\alpha,h}^{(L,)n})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_{h,PD}]^{N_T}$ solves ($\mathrm{P}_{\alpha,h}^{(L,)\Delta t}$), (3-V.46a,b). Similarly, using (3-V.49), (3-V.52), (3-V.53) and (3-V.54), and noting (3-V.29) and (3-V.48), we can pass to the limit $\delta\to0_+$ in (3-V.30) and (3-V.44) to obtain the desired results (3-V.50) and (3-V.51). □

For later purposes, we introduce the following notation in line with (3-III.7). Let $\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t}\in C([0,T];\mathrm{V}^1_h)$ and $\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t,\pm}\in L^\infty(0,T;\mathrm{V}^1_h)$ be such that

$$\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t}(t,\cdot):=\frac{t-t^{n-1}}{\Delta t_n}\boldsymbol{u}_{\alpha,h}^{(L,)n}(\cdot)+\frac{t^n-t}{\Delta t_n}\boldsymbol{u}_{\alpha,h}^{(L,)n-1}(\cdot)\qquad t\in[t^{n-1},t^n],\quad n=1,\dots,N_T,\tag{3-V.56a}$$

$$\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t,+}(t,\cdot):=\boldsymbol{u}_{\alpha,h}^{(L,)n}(\cdot),\qquad\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t,-}(t,\cdot):=\boldsymbol{u}_{\alpha,h}^{(L,)n-1}(\cdot)\qquad t\in[t^{n-1},t^n),\quad n=1,\dots,N_T,\tag{3-V.56b}$$

$$\text{and}\quad\Delta(t):=\Delta t_n\qquad t\in[t^{n-1},t^n),\quad n=1,\dots,N_T.\tag{3-V.56c}$$

We note that

$$\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t}-\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t,\pm}=(t-t_\pm^n)\frac{\partial\boldsymbol{u}_{\alpha,h}^{(L,)\Delta t}}{\partial t}\qquad t\in(t^{n-1},t^n),\quad n=1,\dots,N_T,\tag{3-V.57}$$

where $t^n_+ := t^n$ and $t^n_- := t^{n-1}$. We define $\boldsymbol{\sigma}^{(L,)\Delta t}_{\alpha,h} \in C([0,T];\mathrm{S}^1_{h,PD})$ and $\boldsymbol{\sigma}^{(L,)\Delta t,\pm}_{\alpha,h} \in L^\infty(0,T;\,\mathrm{S}^1_{h,PD})$ similarly to (3-V.56a,b).

Using the notation (3-V.56a,b), (3-V.46a) multiplied by Δt_n and summed for $n,\dots,N_T$ can be restated as :

$$\begin{aligned}
&\int_0^T \int_{\mathcal{D}} \left[\mathrm{Re} \frac{\partial \boldsymbol{u}^{(L,)\Delta t}_{\alpha,h}}{\partial t} \cdot \boldsymbol{v} + (1-\varepsilon) \boldsymbol{\nabla} \boldsymbol{u}^{(L,)\Delta t,+}_{\alpha,h} : \boldsymbol{\nabla} \boldsymbol{v} \right] dt \qquad \text{(3-V.58)} \\
&+ \frac{\mathrm{Re}}{2} \int_0^T \int_{\mathcal{D}} \left[\left[(\boldsymbol{u}^{(L,)\Delta t,-}_{\alpha,h} \cdot \boldsymbol{\nabla}) \boldsymbol{u}^{(L,)\Delta t,+}_{\alpha,h} \right] \cdot \boldsymbol{v} - \left[(\boldsymbol{u}^{(L,)\Delta t,-}_{\alpha,h} \cdot \boldsymbol{\nabla}) \boldsymbol{v} \right] \cdot \boldsymbol{u}^{(L,)\Delta t,+}_{\alpha,h} \right] dt \\
&= \int_0^T \left[\langle \boldsymbol{f}^+, \boldsymbol{v} \rangle_{H^1_0(\mathcal{D})} - \frac{\varepsilon}{\mathrm{Wi}} \int_{\mathcal{D}} \pi_h [\beta^{(L)}(\boldsymbol{\sigma}^{(L,)\Delta t,+}_{\alpha,h})] : \boldsymbol{\nabla} \boldsymbol{v} \right] dt \qquad \forall \boldsymbol{v} \in L^2(0,T;\mathrm{V}^1_h).
\end{aligned}$$

Similarly, (3-V.46b) multiplied by Δt_n and summed for $n=1,\dots,N_T$ can be restated as :

$$\begin{aligned}
&\int_0^T \int_{\mathcal{D}} \pi_h \left[\frac{\partial \boldsymbol{\sigma}^{(L,)\Delta t}_{\alpha,h}}{\partial t} : \boldsymbol{\chi} + \frac{1}{\mathrm{Wi}} (\boldsymbol{\sigma}^{(L,)\Delta t,+}_{\alpha,h} - \boldsymbol{I}) : \boldsymbol{\chi} \right] dt + \alpha \int_0^T \int_{\mathcal{D}} \boldsymbol{\nabla} \boldsymbol{\sigma}^{(L,)\Delta t,+}_{\alpha,h} :: \boldsymbol{\nabla} \boldsymbol{\chi} \, dt \qquad \text{(3-V.59)} \\
&- 2 \int_0^T \int_{\mathcal{D}} \boldsymbol{\nabla} \boldsymbol{u}^{(L,)\Delta t,+}_{\alpha,h} : \pi_h [\boldsymbol{\chi} \beta^{(L)}(\boldsymbol{\sigma}^{(L,)\Delta t,+}_{\alpha,h})] \, dt \\
&- \int_0^T \int_{\mathcal{D}} \sum_{m=1}^d \sum_{p=1}^d [\boldsymbol{u}^{(L,)\Delta t,-}_{\alpha,h}]_m \Lambda^{(L)}_{m,p}(\boldsymbol{\sigma}^{(L,)\Delta t,+}_{\alpha,h}) : \frac{\partial \boldsymbol{\chi}}{\partial \boldsymbol{x}_p} \, dt = 0 \qquad \forall \boldsymbol{\chi} \in L^2(0,T;\mathrm{S}^1_h).
\end{aligned}$$

We note also the following Lemma for later purposes.

Lemma 12. *For all $K_k \in \mathcal{T}_h$, and for all $\boldsymbol{\phi} \in \mathrm{S}^1_{h,PD}$ we have that*

$$\int_{K_k} \|\pi_h[\beta^{(L)}(\boldsymbol{\phi})] - \beta^{(L)}(\boldsymbol{\phi})\|^2 + \max_{m,p=1,\dots,d} \int_{K_k} \|\Lambda^{(L)}_{m,p}(\boldsymbol{\phi}) - \beta^{(L)}(\boldsymbol{\phi})[\boldsymbol{I}]_{mp}\|^2 \le C h^2 \int_{K_k} \|\boldsymbol{\nabla} \boldsymbol{\phi}\|^2 . \qquad \text{(3-V.60)}$$

Proof. First, we have from (3-II.17) that for all $\boldsymbol{\phi} \in \mathrm{S}^1_{h,PD}$

$$\begin{aligned}
\int_{K_k} \|\pi_h[\beta^{(L)}(\boldsymbol{\phi})] - \beta^{(L)}(\boldsymbol{\phi})\|^2 &\le C |K_k| \sum_{j=0}^d \|\beta^{(L)}(\boldsymbol{\phi}(P^k_j)) - \beta^{(L)}(\boldsymbol{\phi})\|^2_{L^\infty(K_k)} \qquad \text{(3-V.61)} \\
&\le C |K_k| \sum_{j=0}^d \|\boldsymbol{\phi}(P^k_j) - \boldsymbol{\phi}\|^2_{L^\infty(K_k)} \\
&\le C h^2 |K_k| \, \|\boldsymbol{\nabla} \boldsymbol{\phi}\|^2_{L^\infty(K_k)} \le C h^2 \int_{K_k} \|\boldsymbol{\nabla} \boldsymbol{\phi}\|^2 .
\end{aligned}$$

where $\{P^k_j\}_{j=0}^d$ are the vertices of K_k. Hence we have the desired first bound in (3-V.60).

It follows from the δ independent versions of (3-V.14) and (3-V.13), recall Remark 11, (3-V.17) and (3-II.17) that for all $\boldsymbol{\phi} \in \mathrm{S}^1_{h,PD}$

$$\begin{aligned}
\int_{K_k} \|\Lambda^{(L)}_{m,p}(\boldsymbol{\phi}) - \pi_h[\beta^{(L)}(\boldsymbol{\phi})][\boldsymbol{I}]_{mp}\|^2 &\le C \int_{K_k} \sum_{j=1}^d \|\widehat{\Lambda}^{(L)}_j(\widehat{\boldsymbol{\phi}}) - \pi_h[\beta^{(L)}(\boldsymbol{\phi})]\|^2 \qquad \text{(3-V.62)} \\
&\le C |K_k| \max_{i,j=0,\dots,d} \|\beta^{(L)}(\boldsymbol{\phi}(P^k_j)) - \beta^{(L)}(\boldsymbol{\phi}(P^k_i))\|^2 \\
&\le C |K_k| \max_{i,j=0,\dots,d} \|\boldsymbol{\phi}(P^k_j) - \boldsymbol{\phi}(P^k_i)\|^2 \\
&\le C h^2 \int_{K_k} \|\boldsymbol{\nabla} \boldsymbol{\phi}\|^2 .
\end{aligned}$$

Combining (3-V.62) and the first bound in (3-V.60) yields the second bound in (3-V.60). □

3-VI Convergence of $(\mathrm{P}_{\alpha,h}^{L,\Delta t})$ to $(\mathrm{P}_{\alpha}^{L})$

Before proving our convergence result, we first deduce some simple inequalities that will be required. We recall the following well-known results concerning the interpolant π_h :

$$\|(\boldsymbol{I}-\pi_h)\boldsymbol{\phi}\|_{W^{1,\infty}(K_k)} \le C\,h\,|\boldsymbol{\phi}|_{W^{2,\infty}(K_k)} \qquad \forall \boldsymbol{\phi}\in[W^{2,\infty}(K_k)]_S^{d\times d},\quad k=1,\ldots,N_K; \tag{3-VI.1a}$$

$$\begin{aligned}\|(\boldsymbol{I}-\pi_h)[\boldsymbol{\chi}:\boldsymbol{\phi}]\|_{L^2(\mathcal{D})} &\le C\,h^2\,\|\nabla\boldsymbol{\chi}\|_{L^2(\mathcal{D})}\,\|\nabla\boldsymbol{\phi}\|_{L^\infty(\mathcal{D})} \\ &\le C\,h\,\|\boldsymbol{\chi}\|_{L^2(\mathcal{D})}\,\|\nabla\boldsymbol{\phi}\|_{L^\infty(\mathcal{D})} \qquad \forall \boldsymbol{\chi},\boldsymbol{\phi}\in \mathrm{S}_h^1.\end{aligned} \tag{3-VI.1b}$$

We note for any $\zeta\in\mathbb{R}_{>0}$ that

$$\begin{aligned}[\pi_h[\boldsymbol{\chi}:\boldsymbol{\phi}]](\boldsymbol{x}) \le \frac{1}{2}\left[\pi_h\left[\zeta\,\|\boldsymbol{\chi}\|^2+\zeta^{-1}\|\boldsymbol{\phi}\|^2\right]\right](\boldsymbol{x}) \\ \forall \boldsymbol{x}\in K_k,\quad \forall \boldsymbol{\chi},\boldsymbol{\phi}\in[C(\overline{K_k})]^{d\times d},\quad k=1,\ldots,N_K\,.\end{aligned} \tag{3-VI.2}$$

Moreover, as the basis functions associated with Q_h^1 and S_h^1 are nonnegative and sum to unity everywhere, we have for $k=1,\ldots,N_K$ that

$$\|[\pi_h\boldsymbol{\chi}](\boldsymbol{x})\|^2 \le (\pi_h[\|\boldsymbol{\chi}\|^2])(\boldsymbol{x}) \qquad \forall \boldsymbol{x}\in K_k, \qquad \forall \boldsymbol{\chi}\in[C(\overline{K_k})]^{d\times d}. \tag{3-VI.3}$$

Combining (3-VI.3), (3-I.4b) and (3-II.4), we have for all $\boldsymbol{\phi}\in\mathrm{S}_{h,PD}^1$ and for all $\boldsymbol{\psi}\in\mathrm{S}_h^1$ that

$$\int_D \|\pi_h[\boldsymbol{\psi}\,\beta^L(\boldsymbol{\phi})]\|^2 \le \int_D \pi_h[\|\boldsymbol{\psi}\,\beta^L(\boldsymbol{\phi})\|^2] \le \int_D \pi_h[\|\boldsymbol{\psi}\|^2\,\|\beta^L(\boldsymbol{\phi})\|^2] \le dL^2\int_D \pi_h[\|\boldsymbol{\psi}\|^2]. \tag{3-VI.4}$$

We note the following results concerning the projections $\mathcal{R}_h$ and $\mathcal{P}_h$, recall (3-V.4) and (3-V.5). It follows from (3-VI.3) and (3-V.5) that

$$\int_{\mathcal{D}} \|\mathcal{P}_h\boldsymbol{\chi}\|^2 \le \int_{\mathcal{D}} \pi_h[\|\mathcal{P}_h\boldsymbol{\chi}\|^2] \le \int_{\mathcal{D}} \|\boldsymbol{\chi}\|^2 \qquad \forall \boldsymbol{\chi}\in[L^2(\mathcal{D})]_S^{d\times d}. \tag{3-VI.5}$$

We note from the convexity of the polytope $\mathcal{D}$ and the quasi-uniformity of $\{\mathcal{T}_h\}_{h>0}$ that $\mathcal{R}_h$ is uniformly H^1 stable; that is,

$$\|\mathcal{R}_h\boldsymbol{v}\|_{H^1(\mathcal{D})} \le C\|\boldsymbol{v}\|_{H^1(\mathcal{D})} \qquad \forall \boldsymbol{v}\in \mathrm{V}, \tag{3-VI.6}$$

see [HR82]. Similarly, it is easily established that

$$\|\mathcal{P}_h\boldsymbol{\chi}\|_{H^1(\mathcal{D})} \le C\|\boldsymbol{\chi}\|_{H^1(\mathcal{D})} \qquad \forall \boldsymbol{\chi}\in[H^1(\mathcal{D})]_S^{d\times d}. \tag{3-VI.7}$$

Let $([H^1(\mathcal{D})]_S^{d\times d})'$ be the topological dual of $[H^1(\mathcal{D})]_S^{d\times d}$ with $[L^2(\mathcal{D})]_S^{d\times d}$ being the pivot space. Let $\mathcal{E}: ([H^1(\mathcal{D})]_S^{d\times d})' \mapsto [H^1(\mathcal{D})]_S^{d\times d}$ be such that $\mathcal{E}\boldsymbol{\chi}$ is the unique solution of the Helmholtz problem

$$\int_{\mathcal{D}} [\boldsymbol{\nabla}(\mathcal{E}\boldsymbol{\chi}) :: \boldsymbol{\nabla}\boldsymbol{\phi} + (\mathcal{E}\boldsymbol{\chi}):\boldsymbol{\phi}] = \langle \boldsymbol{\chi},\boldsymbol{\phi}\rangle_{H^1(\mathcal{D})} \qquad \forall \boldsymbol{\phi}\in[H^1(\mathcal{D})]_S^{d\times d}, \tag{3-VI.8}$$

where $\langle\cdot,\cdot\rangle_{H^1(\mathcal{D})}$ denotes the duality pairing between $([H^1(\mathcal{D})]_S^{d\times d})'$ and $[H^1(\mathcal{D})]_S^{d\times d}$. We note that

$$\langle \boldsymbol{\chi},\mathcal{E}\boldsymbol{\chi}\rangle_{H^1(\mathcal{D})} = \|\mathcal{E}\boldsymbol{\chi}\|^2_{H^1(\mathcal{D})} \qquad \forall \boldsymbol{\chi}\in([H^1(\mathcal{D})]_S^{d\times d})', \tag{3-VI.9}$$

and $\|\mathcal{E}\cdot\|_{H^1(\mathcal{D})}$ is a norm on $([H^1(\mathcal{D})]_S^{d\times d})'$.

Let V' be the topological dual of V with the space of weakly divergent free functions in $[L^2(\mathcal{D})]^d$ being the pivot space. Let $\mathcal{S}:\mathrm{V}'\mapsto\mathrm{V}$ be such that $\mathcal{S}\boldsymbol{w}$ is the unique solution to the Helmholtz-Stokes problem

$$\int_{\mathcal{D}} [\boldsymbol{\nabla}(\mathcal{S}\boldsymbol{w}) : \boldsymbol{\nabla}\boldsymbol{v} + (\mathcal{S}\boldsymbol{w})\cdot\boldsymbol{v}] = \langle \boldsymbol{w},\boldsymbol{v}\rangle_{\mathrm{V}} \qquad \forall \boldsymbol{v}\in\mathrm{V}, \tag{3-VI.10}$$

where $\langle\cdot,\cdot\rangle_{\mathrm{V}}$ denotes the duality pairing between V' and V. We note that

$$\langle \boldsymbol{w},\mathcal{S}\boldsymbol{w}\rangle_{\mathrm{V}} = \|\mathcal{S}\boldsymbol{w}\|^2_{H^1(\mathcal{D})} \qquad \forall \boldsymbol{w}\in\mathrm{V}'\supset[H^{-1}(\mathcal{D})]^d, \tag{3-VI.11}$$

and $\|\mathcal{S}\cdot\|_{H^1(\mathcal{D})}$ is a norm on V'.

We recall the following well-known Gagliardo-Nirenberg inequality. Let $r\in[2,\infty)$ if $d=2$, and $r\in[2,6]$ if $d=3$ and $\theta=d(\frac{1}{2}-\frac{1}{r})$. Then, there exists a positive constant $C(\mathcal{D},r,d)$ such that

$$\|\eta\|_{L^r(\mathcal{D})}\le C(\mathcal{D},r,d)\|\eta\|_{L^2(\mathcal{D})}^{1-\theta}\|\eta\|_{H^1(\mathcal{D})}^{\theta}\qquad\forall\eta\in H^1(\mathcal{D})\,. \tag{3-VI.12}$$

We recall also the following compactness result, see e.g. Temam [Tem84, Theorem 2.1, p184] and Simon [Sim87]. Let $\mathcal{Y}_0$, $\mathcal{Y}$ and $\mathcal{Y}_1$ be Banach spaces, $\mathcal{Y}_i$, $i=0,1$, reflexive, with a compact embedding $\mathcal{Y}_0\hookrightarrow\mathcal{Y}$ and a continuous embedding $\mathcal{Y}\hookrightarrow\mathcal{Y}_1$. Then, for $\mu_i>1$, $i=0,1$, the following embedding is compact :

$$\{\eta\in L^{\mu_0}(0,T;\mathcal{Y}_0):\frac{\partial\eta}{\partial t}\in L^{\mu_1}(0,T;\mathcal{Y}_1)\}\hookrightarrow L^{\mu_0}(0,T;\mathcal{Y})\ . \tag{3-VI.13}$$

Theorem 7. *Under the assumptions of Theorem 6, there exists a solution* $\{(\boldsymbol{u}_{\alpha,h}^{L,n},\boldsymbol{\sigma}_{\alpha,h}^{L,n})\}_{n=1}^{N_T}\in[\mathrm{V}_h^1\times\mathrm{S}_{h,PD}^1]^{N_T}$ *of* $(P_{\alpha,h}^{L,\Delta t})$ *such that, in addition to the bounds (3-V.50) and (3-V.51), the following bounds hold :*

$$\max_{n=0,\dots,N_T}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2]+\sum_{n=1}^{N_T}\int_{\mathcal{D}}\Big[\Delta t_n\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2+\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}\|^2]\Big]\le C(L)\,, \tag{3-VI.14a}$$

$$\sum_{n=1}^{N_T}\Delta t_n\left\|\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^{\frac{4}{\vartheta}}+\sum_{n=1}^{N_T}\Delta t_n\left\|\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^{2}\le C(L,T); \tag{3-VI.14b}$$

where

$$\vartheta\in(2,4)\quad\textit{if } d=2\qquad\textit{and}\qquad\vartheta=3\quad\textit{if } d=3. \tag{3-VI.15}$$

Proof. Existence and the bounds (3-V.50) and (3-V.51) were proved in Theorem 6. On choosing $\boldsymbol{\phi}\equiv\boldsymbol{\sigma}_{\alpha,h}^{L,n}$ in the version of (3-V.46b) dependent on L, it follows from (3-III.12), (3-VI.2), (3-VI.4), (3-V.51) and (3-V.47) on applying a Youngs' inequality that

$$\begin{aligned}
&\frac{1}{2}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2+\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}\|^2]+\Delta t_n\alpha\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2+\frac{\Delta t_n}{2\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2] \\
&\quad\le\frac{1}{2}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}\|^2]+2\Delta t_n\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|\pi_h[\boldsymbol{\sigma}_{\alpha,h}^{L,n}\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n})]\| \\
&\qquad+\Delta t_n\int_{\mathcal{D}}\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|\left(\sum_{m=1}^{d}\sum_{p=1}^{d}\|\Lambda_{m,p}^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n})\|^2\right)^{\frac{1}{2}}+\frac{\Delta t_n d|D|}{2\mathrm{Wi}} \\
&\quad\le\frac{1}{2}\left[\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}\|^2]+\Delta t_n\alpha\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2\right]+\frac{\Delta t_n}{4\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2] \\
&\qquad+\Delta t_n C(L)\left[1+\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|^2\right].
\end{aligned} \tag{3-VI.16}$$

Hence, summing (3-VI.16) from $n=1,\dots,m$ for $m=1,\dots,N_T$ yields, on noting (3-V.51), the desired result (3-VI.14a).

On choosing $\boldsymbol{w}=\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\in\mathrm{V}_h^1$ in the version of (3-V.46a) dependent on L yields, on noting

(3-V.4), (3-VI.11), (3-VI.6) and Sobolev embedding, that

$$\begin{aligned}
\mathrm{Re}\left\|\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^2 &= \mathrm{Re}\int_{\mathcal{D}}\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\cdot\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right] \qquad \text{(3-VI.17)}\\
&= -\int_{\mathcal{D}}\left[(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}+\frac{\varepsilon}{\mathrm{Wi}}\pi_h[\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n})]\right]:\boldsymbol{\nabla}\left[\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right]\\
&\quad -\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\left((\boldsymbol{u}_{\alpha,h}^{L,n-1}\cdot\boldsymbol{\nabla})\boldsymbol{u}_{\alpha,h}^{L,n}\right)\cdot\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\\
&\quad +\frac{\mathrm{Re}}{2}\int_{\mathcal{D}}\boldsymbol{u}_{\alpha,h}^{L,n}\cdot\left((\boldsymbol{u}_{\alpha,h}^{L,n-1}\cdot\boldsymbol{\nabla})\left[\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right]\right)\\
&\quad +\left\langle\boldsymbol{f}^n,\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right\rangle_{H_0^1(\mathcal{D})}\\
&\le C\big[\|\pi_h[\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n})]\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2+\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|_{L^2(\mathcal{D})}^2\\
&\quad +\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|_{L^{1+\theta}(\mathcal{D})}^2+\|\boldsymbol{f}^n\|_{H^{-1}(\mathcal{D})}^2\big],
\end{aligned}$$

for any $\theta>0$ if $d=2$ and for $\theta=\frac{1}{5}$ if $d=3$. Applying the Cauchy–Schwarz and the algebraic-geometric mean inequalities, in conjunction with (3-VI.12) and the Poincaré inequality (3-I.8) yields that

$$\begin{aligned}
\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|_{L^2(\mathcal{D})}^2 &\le \|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^4(\mathcal{D})}^2\|\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^4(\mathcal{D})}^2\le\tfrac{1}{2}\sum_{m=n-1}^{n}\|\boldsymbol{u}_{\alpha,h}^{L,m}\|_{L^4(\mathcal{D})}^4 \qquad \text{(3-VI.18)}\\
&\le C\sum_{m=n-1}^{n}\left[\|\boldsymbol{u}_{\alpha,h}^{L,m}\|_{L^2(\mathcal{D})}^{4-d}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,m}\|_{L^2(\mathcal{D})}^{d}\right].
\end{aligned}$$

Similarly, we have for any $\theta\in(0,1)$, if $d=2$, that

$$\begin{aligned}
\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|_{L^{1+\theta}(\mathcal{D})}^2 &\le \|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^{\frac{2(1+\theta)}{1-\theta}}(\mathcal{D})}^2\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2 \qquad \text{(3-VI.19a)}\\
&\le C\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^2(\mathcal{D})}^{\frac{2(1-\theta)}{1+\theta}}\sum_{m=n-1}^{n}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,m}\|_{L^2(\mathcal{D})}^{\frac{2(1+3\theta)}{1+\theta}};
\end{aligned}$$

and if $d=3$, ($\theta=\frac{1}{5}$), that

$$\begin{aligned}
\|\,\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|\,\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|\,\|_{L^{\frac{6}{5}}(\mathcal{D})}^2 &\le \|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^3(\mathcal{D})}^2\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2 \qquad \text{(3-VI.19b)}\\
&\le C\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^2(\mathcal{D})}\sum_{m=n-1}^{n}\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,m}\|_{L^2(\mathcal{D})}^3.
\end{aligned}$$

On taking the $\frac{2}{\vartheta}$ power of both sides of (3-VI.17), recall (3-VI.15), multiplying by Δt_n, summing from $n=1,\ldots,N_T$ and noting (3-VI.18), (3-VI.19a) with $\theta=\frac{\vartheta-2}{6-\vartheta}\Leftrightarrow\vartheta=\frac{2(1+3\theta)}{(1+\theta)}$, (3-VI.19b), (3-V.19), (3-III.8a), (3-V.51), (3-V.22) and (3-II.4) yields that

$$\begin{aligned}
&\sum_{n=1}^{N_T}\Delta t_n\left\|\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{L,n}-\boldsymbol{u}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^{\frac{4}{\vartheta}} \qquad \text{(3-VI.20)}\\
&\qquad\le CL^2+C(T)\left[\sum_{n=1}^{N_T}\Delta t_n\left[\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{f}^n\|_{H^{-1}(\mathcal{D})}^2\right]\right]^{\frac{2}{\vartheta}}\\
&\qquad\quad +C\left[1+\max_{n=0,\ldots,N_T}\left(\|\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2\right)\right]\left[\sum_{n=0}^{N_T}\Delta t_n\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2\right]\\
&\qquad\le C(L,T);
\end{aligned}$$

and hence the first bound in (3-VI.14b).

On choosing $\boldsymbol{\phi}=\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\in \mathrm{S}_h^1$ in the version of (3-V.46b) dependent on L yields, on noting (3-V.5), (3-VI.8), (3-VI.2), (3-VI.5), (3-VI.7), (3-VI.4) and (3-V.47), that

$$\begin{aligned}
\left\|\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^2 &= \int_{\mathcal{D}}\pi_h\left[\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right):\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right] \qquad \text{(3-VI.21)}\\
&=\frac{1}{\mathrm{Wi}}\int_{\mathcal{D}}\pi_h\left[(\boldsymbol{I}-\boldsymbol{\sigma}_{\alpha,h}^{L,n}):\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right]\\
&\quad-\alpha\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}::\boldsymbol{\nabla}\left[\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right]\\
&\quad+2\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}:\pi_h\left[\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n})\right]\\
&\quad+\int_{\mathcal{D}}\sum_{m=1}^{d}\sum_{p=1}^{d}[\boldsymbol{u}_{\alpha,h}^{L,n-1}]_m\,\Lambda_{m,p}^L(\boldsymbol{\sigma}_{\alpha,h}^{L,n}):\frac{\partial}{\partial\boldsymbol{x}_p}\left[\mathcal{P}_h\left[\mathcal{E}\left(\frac{\boldsymbol{\sigma}_{\alpha,h}^{L,n}-\boldsymbol{\sigma}_{\alpha,h}^{L,n-1}}{\Delta t_n}\right)\right]\right]\\
&\le C\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|^2]+C(L)\left[1+\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{u}_{\alpha,h}^{L,n-1}\|_{L^2(\mathcal{D})}^2\right].
\end{aligned}$$

Multiplying (3-VI.21) by Δt_n, summing from $n=1,...,N_T$ and noting (3-V.51) and (3-VI.14a) yields the second bound in (3-VI.14b). □

3-VI-A Convergence of the discrete solutions

First, we note from (3-I.2b), (3-I.6) and (3-II.3) that

$$\|\boldsymbol{\phi}^{-1}\|\le\|[\beta^L(\boldsymbol{\phi})]^{-1}\|\qquad\forall\boldsymbol{\phi}\in\mathbb{R}^{d\times d}_{SPD}. \tag{3-VI.22}$$

Second, we recall the well-known result about equivalence of norms

$$\frac{1}{d^{\frac{1}{2}}}\|\boldsymbol{\phi}\|\le\|\boldsymbol{\phi}\|_2:=\sup_{\boldsymbol{v}\in\mathbb{R}^d,\ \boldsymbol{v}\neq\boldsymbol{0}}\frac{\|\boldsymbol{\phi}\boldsymbol{v}\|}{\|\boldsymbol{v}\|}\le\|\boldsymbol{\phi}\|\qquad\forall\boldsymbol{\phi}\in\mathbb{R}^{d\times d}. \tag{3-VI.23}$$

Then using the Rayleigh quotient, it is easy to establish that

$$\|\boldsymbol{\chi}^{-1}(\boldsymbol{x})\|_2\le\left[[\pi_h[\|\boldsymbol{\chi}^{-1}\|_2^{-1}]](\boldsymbol{x})\right]^{-1}\qquad\forall\boldsymbol{x}\in K_k,\qquad k=1,\dots,N_K,\qquad\forall\boldsymbol{\chi}\in\mathrm{S}^1_{h,PD}. \tag{3-VI.24}$$

Hence it follows from (3-VI.23), (3-VI.24) and (3-V.32) that

$$\begin{aligned}
\frac{1}{d^{\frac{1}{2}}}\int_{K_k}\|\boldsymbol{\chi}^{-1}\|&\le\int_{K_k}\left[\pi_h[\|\boldsymbol{\chi}^{-1}\|_2^{-1}]\right]^{-1}\le|K_k|\,\|\pi_h[\|\boldsymbol{\chi}^{-1}\|]\|_{L^\infty(K_k)} \qquad \text{(3-VI.25)}\\
&\le C\int_{K_k}\pi_h[\|\boldsymbol{\chi}^{-1}\|]\qquad k=1,\dots,N_K,\qquad\forall\boldsymbol{\chi}\in\mathrm{S}^1_{h,PD}.
\end{aligned}$$

Therefore (3-V.51), (3-VI.14a,b), (3-VI.3), (3-VI.22), (3-VI.25) and (3-V.56a–c) yield that

$$\begin{aligned}
&\sup_{t\in(0,T)}\|\boldsymbol{u}_{\alpha,h}^{L,\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2+\int_0^T\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{L,\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2\,dt \qquad \text{(3-VI.26a)}\\
&\qquad+\int_0^T\left[\|[\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}]^{-1}\|_{L^1(\mathcal{D})}+\frac{\|\boldsymbol{u}_{\alpha,h}^{L,\Delta t,+}-\boldsymbol{u}_{\alpha,h}^{L,\Delta t,-}\|_{L^2(\mathcal{D})}^2}{\Delta(t)}\right]dt\le C,\\
&\sup_{t\in(0,T)}\|\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2+\int_0^T\left[\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}\|_{L^2(\mathcal{D})}^2+\frac{\|\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}-\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,-}\|_{L^2(\mathcal{D})}^2}{\Delta(t)}\right]dt\le C(L), \qquad \text{(3-VI.26b)}\\
&\int_0^T\left[\left\|\mathcal{S}\frac{\partial\boldsymbol{u}_{\alpha,h}^{L,\Delta t}}{\partial t}\right\|_{H^1(\mathcal{D})}^{\frac{4}{\vartheta}}+\left\|\mathcal{E}\frac{\partial\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}}{\partial t}\right\|_{H^1(\mathcal{D})}^2\right]dt\le C(L,T); \qquad \text{(3-VI.26c)}
\end{aligned}$$

where ϑ is as defined in (3-VI.15).

We are now in a position to prove the following convergence result for $(\mathrm{P}_{\alpha,h}^{L,\Delta t})$.

Theorem 8. *There exists a subsequence of* $\{(\boldsymbol{u}_{\alpha,h}^{L,\Delta t},\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t})\}_{h>0,\Delta t>0}$, *and functions* $\boldsymbol{u}_\alpha^L \in L^\infty(0,T;[L^2(\mathcal{D})]^d)\cap L^2(0,T;\mathrm{V})\cap W^{1,\frac{4}{\vartheta}}(0,T;\mathrm{V}')$ *and* $\boldsymbol{\sigma}_\alpha^L \in L^\infty(0,T;[L^2(\mathcal{D})]^{d\times d}_{SPD})\cap L^2(0,T;[H^1(\mathcal{D})]^{d\times d}_{SPD}) \cap H^1(0,T;([H^1(\mathcal{D})]^{d\times d}_S)')$ *such that, as* $h,\Delta t\to 0_+$,

$$\boldsymbol{u}_{\alpha,h}^{L,\Delta t(,\pm)} \to \boldsymbol{u}_\alpha^L \qquad \text{weak* in } L^\infty(0,T;[L^2(\mathcal{D})]^d), \tag{3-VI.27a}$$

$$\boldsymbol{u}_{\alpha,h}^{L,\Delta t(,\pm)} \to \boldsymbol{u}_\alpha^L \qquad \text{weakly in } L^2(0,T;[H^1(\mathcal{D})]^d), \tag{3-VI.27b}$$

$$\mathcal{S}\frac{\partial \boldsymbol{u}_{\alpha,h}^{L,\Delta t}}{\partial t} \to \mathcal{S}\frac{\partial \boldsymbol{u}_\alpha^L}{\partial t} \qquad \text{weakly in } L^{\frac{4}{\vartheta}}(0,T;\mathrm{V}), \tag{3-VI.27c}$$

$$\boldsymbol{u}_{\alpha,h}^{L,\Delta t(,\pm)} \to \boldsymbol{u}_\alpha^L \qquad \text{strongly in } L^2(0,T;[L^r(\mathcal{D})]^d), \tag{3-VI.27d}$$

and

$$\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)} \to \boldsymbol{\sigma}_\alpha^L \qquad \text{weak* in } L^\infty(0,T;[L^2(\mathcal{D})]^{d\times d}), \tag{3-VI.28a}$$

$$\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+} \to \boldsymbol{\sigma}_\alpha^L \qquad \text{weakly in } L^2(0,T;[H^1(\mathcal{D})]^{d\times d}), \tag{3-VI.28b}$$

$$\mathcal{E}\frac{\partial \boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}}{\partial t} \to \mathcal{E}\frac{\partial \boldsymbol{\sigma}_\alpha^L}{\partial t} \qquad \text{weakly in } L^2(0,T;[H^1(\mathcal{D})]^{d\times d}), \tag{3-VI.28c}$$

$$\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)} \to \boldsymbol{\sigma}_\alpha^L \qquad \text{strongly in } L^2(0,T;[L^r(\mathcal{D})]^{d\times d}), \tag{3-VI.28d}$$

$$\pi_h[\beta^L(\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)})] \to \beta^L(\boldsymbol{\sigma}_\alpha^L) \qquad \text{strongly in } L^2(0,T;[L^2(\mathcal{D})]^{d\times d}), \tag{3-VI.28e}$$

$$\Lambda_{m,p}^L(\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)}) \to \beta^L(\boldsymbol{\sigma}_\alpha^L)\boldsymbol{I}_{mp} \qquad \text{strongly in } L^2(0,T;[L^2(\mathcal{D})]^{d\times d}), \quad m,p=1,\dots,d, \tag{3-VI.28f}$$

where ϑ *is defined by (3-VI.15) and* $r\in[1,\infty)$ *if* $d=2$ *and* $r\in[1,6)$ *if* $d=3$.

Furthermore, $(\boldsymbol{u}_\alpha^L,\boldsymbol{\sigma}_\alpha^L)$ *solve the following problem :*

$(\mathbf{P}_\alpha^L)$ *Find* $\boldsymbol{u}_\alpha^L \in L^\infty(0,T;[L^2(\mathcal{D})]^d)\cap L^2(0,T;\mathrm{V})\cap W^{1,\frac{4}{\vartheta}}(0,T;\mathrm{V}')$ *and* $\boldsymbol{\sigma}_\alpha^L \in L^\infty(0,T;[L^2(\mathcal{D})]^{d\times d}_{SPD})\cap L^2(0,T;[H^1(\mathcal{D})]^{d\times d}_{SPD})\cap H^1(0,T;([H^1(\mathcal{D})]^{d\times d}_S)')$ *such that*

$$\int_0^T \mathrm{Re}\left\langle \frac{\partial \boldsymbol{u}_\alpha^L}{\partial t},\boldsymbol{v}\right\rangle_{\mathrm{V}} dt + \int_0^T\int_{\mathcal{D}} \left[(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_\alpha^L : \boldsymbol{\nabla}\boldsymbol{v} + \mathrm{Re}\left[(\boldsymbol{u}_\alpha^L\cdot\boldsymbol{\nabla})\boldsymbol{u}_\alpha^L\right]\cdot\boldsymbol{v}\right] dt \tag{3-VI.29a}$$

$$= \int_0^T \langle \boldsymbol{f},\boldsymbol{v}\rangle_{H_0^1(\mathcal{D})}\, dt - \frac{\varepsilon}{\mathrm{Wi}}\int_0^T\int_{\mathcal{D}} \beta^L(\boldsymbol{\sigma}_\alpha^L) : \boldsymbol{\nabla}\boldsymbol{v}\, dt \qquad \forall \boldsymbol{v}\in L^{\frac{4}{4-\vartheta}}(0,T;\mathrm{V}),$$

$$\int_0^T \left\langle \frac{\partial \boldsymbol{\sigma}_\alpha^L}{\partial t},\boldsymbol{\phi}\right\rangle_{H^1(\mathcal{D})} dt + \int_0^T\int_{\mathcal{D}} \left[(\boldsymbol{u}_\alpha^L\cdot\boldsymbol{\nabla})[\beta^L(\boldsymbol{\sigma}_\alpha^L)] : \boldsymbol{\phi} + \alpha\boldsymbol{\nabla}\boldsymbol{\sigma}_\alpha^L :: \boldsymbol{\nabla}\boldsymbol{\phi}\right] dt \tag{3-VI.29b}$$

$$= \int_0^T\int_{\mathcal{D}} \left[2(\boldsymbol{\nabla}\boldsymbol{u}_\alpha^L)\beta^L(\boldsymbol{\sigma}_\alpha^L) - \frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_\alpha^L - \boldsymbol{I})\right] : \boldsymbol{\phi}\, dt \qquad \forall \boldsymbol{\phi}\in L^2(0,T;[H^1(\mathcal{D})]^{d\times d}_S);$$

and

$$\lim_{t\to 0_+}\int_{\mathcal{D}} (\boldsymbol{u}_\alpha^L(t,\boldsymbol{x}) - \boldsymbol{u}^0(\boldsymbol{x}))\cdot\boldsymbol{v} = 0 \qquad \forall \boldsymbol{v}\in \mathrm{H} := \{\boldsymbol{w}\in[L^2(\mathcal{D})]^d : \mathrm{div}\,\boldsymbol{w} = 0 \text{ in } \mathcal{D}\}, \tag{3-VI.29c}$$

$$\lim_{t\to 0_+}\int_{\mathcal{D}} (\boldsymbol{\sigma}_\alpha^L(t,\boldsymbol{x}) - \boldsymbol{\sigma}^0(\boldsymbol{x})) : \boldsymbol{\chi} = 0 \qquad \forall \boldsymbol{\chi}\in[L^2(\mathcal{D})]^{d\times d}_{SPD}.$$

Proof. The results (3-VI.27a–c) follow immediately from the bounds (3-VI.26a,c) on noting the notation (3-V.56a–c). The denseness of $\bigcup_{h>0} Q_h^1$ in $L^2(\mathcal{D})$ and (3-V.1c) yield that $\boldsymbol{u}_\alpha^L\in L^2(0,T;\mathrm{V})$. The strong convergence result (3-VI.27d) for $\boldsymbol{u}_{\alpha,h}^{L,\Delta t}$ follows immediately from (3-VI.27a–c), (3-VI.11) and (3-VI.13), on noting that $\mathrm{V}\subset[H_0^1(\mathcal{D})]^d$ is compactly embedded in $L^r(\mathcal{D})$ for the stated values of r. We now prove (3-VI.27d) for $\boldsymbol{u}_{\alpha,h}^{L,\Delta t,\pm}$. First we obtain from the bound on the last term on the left-hand side of (3-VI.26a) and (3-V.57) that

$$\|\boldsymbol{u}_{\alpha,h}^{L,\Delta t} - \boldsymbol{u}_{\alpha,h}^{L,\Delta t,\pm}\|^2_{L^2(0,T,L^2(\mathcal{D}))} \le C\,\Delta t. \tag{3-VI.30}$$

Second, we note from Sobolev embedding that, for all $\eta\in L^2(0,T;H^1(\mathcal{D}))$,

$$\|\eta\|_{L^2(0,T;L^r(\mathcal{D}))} \le \|\eta\|^\theta_{L^2(0,T;L^2(\mathcal{D}))}\,\|\eta\|^{1-\theta}_{L^2(0,T;L^s(\mathcal{D}))} \le C\,\|\eta\|^\theta_{L^2(0,T;L^2(\mathcal{D}))}\,\|\eta\|^{1-\theta}_{L^2(0,T;H^1(\mathcal{D}))} \tag{3-VI.31}$$

for all $r\in[2,s)$, with any $s\in(2,\infty)$ if $d=2$ or any $s\in(2,6]$ if $d=3$, and $\theta=[2(s-r)]/[r(s-2)]\in(0,1]$. Hence, combining (3-VI.30), (3-VI.31) and (3-VI.27d) for $\boldsymbol{u}_{\alpha,h}^{L,\Delta t}$ yields (3-VI.27d) for $\boldsymbol{u}_{\alpha,h}^{L,\Delta t,\pm}$.

Similarly, the results (3-VI.28a–c) follow immediately from (3-VI.26b,c). The strong convergence result (3-VI.28d) for $\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}$ follows immediately from (3-VI.28a–c), (3-VI.9) and (3-VI.13). Similarly to (3-VI.30), the second bound in (3-VI.26b) then yields that (3-VI.28d) holds for $\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)}$.

Since $\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t(,\pm)}\in L^2(0,T;\mathrm{S}^1_{h,PD})$, it follows that $\boldsymbol{\sigma}_\alpha^L$ is symmetric non-negative definite a.e. in $\mathcal{D}_T$. We now establish that $\boldsymbol{\sigma}_\alpha^L$ is symmetric positive definite a.e. in $\mathcal{D}_T$. Assume that $\boldsymbol{\sigma}_\alpha^L$ is not symmetric positive definite a.e. in $\mathcal{D}_T^0\subset\mathcal{D}_T$. Let $\boldsymbol{v}\in L^2(0,T;[L^2(\mathcal{D})]^d)$ be such that $\boldsymbol{\sigma}_\alpha^L\boldsymbol{v}=\boldsymbol{0}$ with $\|\boldsymbol{v}\|=1$ a.e. in $\mathcal{D}_T^0$ and $\boldsymbol{v}=\boldsymbol{0}$ a.e. in $\mathcal{D}\setminus\mathcal{D}_T^0$. We then have from (3-VI.26a) that

$$
\begin{aligned}
|\mathcal{D}_T^0| &= \int_0^T\int_{\mathcal{D}}\|\boldsymbol{v}\|^2\,dt=\int_0^T\int_{\mathcal{D}}\Big([\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}]^{-\frac12}\boldsymbol{v}\Big):\Big([\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}]^{\frac12}\boldsymbol{v}\Big)\,dt \qquad\text{(3-VI.32)}\\
&\le C\left(\int_0^T\int_{\mathcal{D}}\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}::(\boldsymbol{v}\boldsymbol{v}^T)\,dt\right)^{\frac12}.
\end{aligned}
$$

Hence it follows from (3-VI.32) and (3-VI.28d) that $|\mathcal{D}_T^0|=0$.

Finally, the desired results (3-VI.28e,f) follow immediately from (3-V.60) the second bound in (3-VI.26b), (3-II.17), (3-VI.28d) and the fact that $\boldsymbol{\sigma}_\alpha^L\in L^\infty(0,T;[L^2(\mathcal{D})]^{d\times d}_{SPD})$.

It remains to prove that $(\boldsymbol{u}_\alpha^L,\boldsymbol{\sigma}_\alpha^L)$ solves (P_α^L). It follows from (3-V.2), (3-VI.26a–c), (3-VI.27a–d), (3-VI.28e), (3-III.8b), (3-VI.10) and (3-III.11) that we may pass to the limit, $h,\Delta t\to0_+$, in the L-dependent version of (3-V.58) to obtain that $(\boldsymbol{u}_\alpha^L,\boldsymbol{\sigma}_\alpha^L)$ satisfy (3-VI.29a). It also follows from (3-V.20a), (3-V.2), (3-VI.27c,d) and as V is dense in H that $\boldsymbol{u}_\alpha^L(0,\cdot)=\boldsymbol{u}^0(\cdot)$ in the required sense; see (3-VI.29c) and [Tem84, Lemma 1.4, p179].

It follows from (3-VI.28a–f), (3-VI.27b,d), (3-VI.8), (3-VI.26a–c), (3-VI.1a,b), (3-I.4a) and as $\boldsymbol{u}_\alpha^L\in L^2(0,T;\mathrm{V})$ that we may pass to the limit $h,\Delta t\to0_+$ in the L-dependent version of (3-V.59) with $\boldsymbol{\chi}=\pi_h\boldsymbol{\phi}$ to obtain (3-VI.29b) for any $\boldsymbol{\phi}\in C_0^\infty(0,T;[C^\infty(\overline{\mathcal{D}})]_S^{d\times d})$. For example, in order to pass to the limit on the first term in the L-dependent version of (3-V.59), we note that

$$
\begin{aligned}
&\int_0^T\int_{\mathcal{D}}\pi_h\left[\left(\frac{\partial\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}}{\partial t}+\frac{1}{\mathrm{Wi}}\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}\right):\pi_h\boldsymbol{\phi}\right]dt \qquad\text{(3-VI.33)}\\
&\quad=\int_0^T\int_{\mathcal{D}}\left\{\left(\frac{\partial\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}}{\partial t}+\frac{1}{\mathrm{Wi}}\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}\right):\pi_h\boldsymbol{\phi}+(I-\pi_h)\left[\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t}:\pi_h\left[\frac{\partial\boldsymbol{\phi}}{\partial t}\right]\right]\right\}dt.\\
&\qquad\qquad-\frac{1}{\mathrm{Wi}}\int_0^T\int_{\mathcal{D}}(I-\pi_h)\left[\boldsymbol{\sigma}_{\alpha,h}^{L,\Delta t,+}:\pi_h\boldsymbol{\phi}\right]dt.
\end{aligned}
$$

The desired result (3-VI.29b) then follows from noting that $C_0^\infty(0,T;[C^\infty(\overline{\mathcal{D}})]_S^{d\times d})$ is dense in $L^2(0,T;[H^1(\mathcal{D})]_S^{d\times d})$. Finally, it follows from (3-V.20b), (3-VI.28c,d), (3-VI.1a,b) and (3-VI.3) that $\boldsymbol{\sigma}_\alpha^L(0,\cdot)=\boldsymbol{\sigma}^0(\cdot)$ in the required sense; see (3-VI.29c) and [Tem84, Lemma 1.4, p179]. □

Remark 12. *It follows from (3-VI.26a,b), (3-VI.27a,b) and (3-VI.28a,b) that*

$$
\sup_{t\in(0,T)}\|\boldsymbol{u}_\alpha^L\|^2_{L^2(\mathcal{D})}+\int_0^T\|\boldsymbol{\nabla}\boldsymbol{u}_\alpha^L\|^2_{L^2(\mathcal{D})}dt\le C, \qquad\text{(3-VI.34a)}
$$

$$
\sup_{t\in(0,T)}\|\boldsymbol{\sigma}_\alpha^L\|^2_{L^2(\mathcal{D})}+\alpha\int_0^T\|\boldsymbol{\nabla}\boldsymbol{\sigma}_\alpha^L\|^2_{L^2(\mathcal{D})}dt\le C(L). \qquad\text{(3-VI.34b)}
$$

Hence, although we have introduced a cut-off $L\gg1$ to certain terms, and added diffusion with a positive coefficient α in the stress equation compared to the standard Oldroyd-B model; the bound (3-VI.34a) on the velocity $\boldsymbol{u}_\alpha^L$ is independent of the parameters L and α, where $(\boldsymbol{u}_\alpha^L,\boldsymbol{\sigma}_\alpha^L)$ solves (P_α^L), (3-VI.29a–c).

3-VII Convergence of $(\mathrm{P}^{\Delta t}_{\alpha,h})$ to (P_α) in the case $d=2$

First, we recall the discrete Gronwall inequality :

$$
\begin{aligned}
&(r^0)^2+(s^0)^2\le (q^0)^2, && \text{(3-VII.1)}\\
&(r^m)^2+(s^m)^2\le \sum_{n=0}^{m-1}(\eta^n)^2(r^n)^2+\sum_{n=0}^{m}(q^n)^2 \qquad m\ge 1\\
\Rightarrow\qquad &(r^m)^2+(s^m)^2\le \exp(\sum_{n=0}^{m-1}(\eta^n)^2)\sum_{n=0}^{m}(q^n)^2 \qquad m\ge 1.
\end{aligned}
$$

Theorem 9. *Under the assumptions of Theorem 6, there exists a solution* $\{(\boldsymbol{u}^n_{\alpha,h},\boldsymbol{\sigma}^n_{\alpha,h})\}_{n=1}^{N_T}\in[\mathrm{V}^1_h\times\mathrm{S}^1_{h,PD}]^{N_T}$ *of* $(P^{\Delta t}_{\alpha,h})$ *such that the bounds (3-V.50) and (3-V.51) hold.*

If $d=2$, $\alpha\le\frac{1}{2Wi}$ *and* $\Delta t\le C_\star(\zeta^{-1})\alpha^{1+\zeta}h^2$, *for a* $\zeta>0$, *then the following bounds hold :*

$$
\begin{aligned}
\max_{n=0,\dots,N_T}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2]+\sum_{n=1}^{N_T}\int_{\mathcal{D}}\Big[\Delta t_n\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}^n_{\alpha,h}\|^2+\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}-\boldsymbol{\sigma}^{n-1}_{\alpha,h}\|^2]\Big] \qquad \text{(3-VII.2)}\\
+\sum_{n=1}^{N_T}\Delta t_n\left\|\mathcal{S}\left(\frac{\boldsymbol{u}^n_{\alpha,h}-\boldsymbol{u}^{n-1}_{\alpha,h}}{\Delta t_n}\right)\right\|^{\frac{4}{\vartheta}}_{H^1(\mathcal{D})}\le C(\alpha^{-1},T);
\end{aligned}
$$

where $\vartheta\in(2,4)$.

Proof. Existence and the bounds (3-V.50) and (3-V.51) were proved in Theorem 6.

On choosing $\boldsymbol{\phi}\equiv\boldsymbol{\sigma}^n_{\alpha,h}$ in the L-independent version of (3-V.46b), it follows from (3-III.12) and on applying a Youngs' inequality for any $\zeta>0$ that

$$
\begin{aligned}
&\frac12\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2+\|\boldsymbol{\sigma}^n_{\alpha,h}-\boldsymbol{\sigma}^{n-1}_{\alpha,h}\|^2]+\Delta t_n\alpha\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{\sigma}^n_{\alpha,h}\|^2+\frac{\Delta t_n}{2\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2] \qquad \text{(3-VII.3)}\\
&\le\frac12\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^{n-1}_{\alpha,h}\|^2]+\frac{\Delta t_n d|\mathcal{D}|}{2\mathrm{Wi}}+2\Delta t_n\int_{\mathcal{D}}\boldsymbol{\nabla}\boldsymbol{u}^n_{\alpha,h}:\pi_h[(\boldsymbol{\sigma}^n_{\alpha,h})^2]\\
&\quad+\Delta t_n\int_{\mathcal{D}}\sum_{m=1}^{d}\sum_{p=1}^{d}[\boldsymbol{u}^{n-1}_{\alpha,h}]_m\Lambda_{m,p}(\boldsymbol{\sigma}^n_{\alpha,h}):\frac{\partial\boldsymbol{\sigma}^n_{\alpha,h}}{\partial\boldsymbol{x}_p}\\
&\le\frac12\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^{n-1}_{\alpha,h}\|^2]+C\Delta t_n\left[1+\|\boldsymbol{\nabla}\boldsymbol{u}^n_{\alpha,h}\|_{L^2(\mathcal{D})}\|\pi_h[(\boldsymbol{\sigma}^n_{\alpha,h})^2]\|_{L^2(\mathcal{D})}\right]\\
&\quad+C\Delta t_n\|\boldsymbol{u}^{n-1}_{\alpha,h}\|_{L^{\frac{2(2+\zeta)}{\zeta}}(\mathcal{D})}\|\Lambda_{m,p}(\boldsymbol{\sigma}^n_{\alpha,h})\|_{L^{2+\zeta}(\mathcal{D})}\|\boldsymbol{\nabla}\boldsymbol{\sigma}^n_{\alpha,h}\|_{L^2(\mathcal{D})}.
\end{aligned}
$$

It follows from (3-VI.3), (3-I.4b), (3-V.32) and (3-VI.12), as $d=2$, that

$$
\begin{aligned}
\|\pi_h[(\boldsymbol{\sigma}^n_{\alpha,h})^2]\|^2_{L^2(\mathcal{D})}&=\int_{\mathcal{D}}\|\pi_h[(\boldsymbol{\sigma}^n_{\alpha,h})^2]\|^2\le\int_{\mathcal{D}}\pi_h[\|(\boldsymbol{\sigma}^n_{\alpha,h})^2\|^2]\le\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}\|^4] \qquad \text{(3-VII.4)}\\
&=\sum_{k=1}^{N_K}\int_{K_k}\pi_h[\|\boldsymbol{\sigma}^n_{\alpha,h}\|^4]\le\sum_{k=1}^{N_K}|K_k|\|\boldsymbol{\sigma}^n_{\alpha,h}\|^4_{L^\infty(K_k)}\\
&\le C\sum_{k=1}^{N_K}|K_k|\left(|K_k|^{-1}\|\boldsymbol{\sigma}^n_{\alpha,h}\|_{L^1(K_k)}\right)^4\le C\sum_{k=1}^{N_K}\|\boldsymbol{\sigma}^n_{\alpha,h}\|^4_{L^4(K_k)}\\
&=C\|\boldsymbol{\sigma}^n_{\alpha,h}\|^4_{L^4(\mathcal{D})}\le C\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2_{L^2(\mathcal{D})}\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2_{H^1(\mathcal{D})}.
\end{aligned}
$$

Similarly, it follows from the δ-independent versions of (3-V.14), (3-V.13), recall Remark 11, (3-V.17), (3-V.32) and (3-VI.12) that for all $\zeta>0$

$$
\begin{aligned}
\|\Lambda_{m,p}(\boldsymbol{\sigma}^n_{\alpha,h})\|^{2+\zeta}_{L^{2+\zeta}(\mathcal{D})}&\le\sum_{k=1}^{N_K}|K_k|\|\Lambda_{m,p}(\boldsymbol{\sigma}^n_{\alpha,h})\|^{2+\zeta}_{L^\infty(K_k)}\le C\sum_{k=1}^{N_K}|K_k|\|\boldsymbol{\sigma}^n_{\alpha,h}\|^{2+\zeta}_{L^\infty(K_k)} \qquad \text{(3-VII.5)}\\
&=C\|\boldsymbol{\sigma}^n_{\alpha,h}\|^{2+\zeta}_{L^{2+\zeta}(\mathcal{D})}\le C(\zeta)\|\boldsymbol{\sigma}^n_{\alpha,h}\|^2_{L^2(\mathcal{D})}\|\boldsymbol{\sigma}^n_{\alpha,h}\|^\zeta_{H^1(\mathcal{D})}.
\end{aligned}
$$

In addition, we note from (3-VI.12), (3-I.8) and (3-V.51) that for all $\zeta>0$

$$\|\boldsymbol{u}_{\alpha,h}^{n-1}\|_{L^{\frac{2(2+\zeta)}{\zeta}}(\mathcal{D})}\leq C(\zeta^{-1})\,\|\boldsymbol{u}_{\alpha,h}^{n-1}\|_{L^2(\mathcal{D})}^{\frac{\zeta}{2+\zeta}}\,\|\boldsymbol{u}_{\alpha,h}^{n-1}\|_{H^1(\mathcal{D})}^{\frac{2}{2+\zeta}}\leq C(\zeta^{-1})\,\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n-1}\|_{L^2(\mathcal{D})}^{\frac{2}{2+\zeta}}. \tag{3-VII.6}$$

Combining (3-VII.3), (3-VII.4), (3-VII.5) and (3-VII.6), and on noting (3-VI.3) and that $\alpha\leq\frac{1}{2\mathrm{Wi}}$, yields on applying a Young's inequality that for all $\zeta>0$

$$\begin{aligned}&\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2+\|\boldsymbol{\sigma}_{\alpha,h}^{n}-\boldsymbol{\sigma}_{\alpha,h}^{n-1}\|^2]+\Delta t_n\alpha\int_{\mathcal{D}}\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2+\frac{\Delta t_n}{2\mathrm{Wi}}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2]\\ &\quad\leq\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n-1}\|^2]+C\,\Delta t_n\\ &\qquad+C(\zeta^{-1})\,\Delta t_n\,\alpha^{-(1+\zeta)}\left[\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n-1}\|_{L^2(\mathcal{D})}^2\right]\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2].\end{aligned} \tag{3-VII.7}$$

Hence, summing (3-VII.7) from $n=1,\ldots,m$ for $m=1,\ldots,N_T$ yields, for any $\zeta>0$ that

$$\begin{aligned}&\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{m}\|^2]+\sum_{n=1}^{m}\Delta t_n\int_{\mathcal{D}}\left[\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2+\frac{1}{2\mathrm{Wi}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2]\right]\\ &\quad+\sum_{n=1}^{m}\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}-\boldsymbol{\sigma}_{\alpha,h}^{n-1}\|^2]\\ &\quad\leq\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{0}\|^2]+C\\ &\qquad+C(\zeta^{-1})\,\alpha^{-(1+\zeta)}\sum_{n=1}^{m}\Delta t_n\left[\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n-1}\|_{L^2(\mathcal{D})}^2\right]\int_{\mathcal{D}}\pi_h[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|^2].\end{aligned} \tag{3-VII.8}$$

Applying the discrete Gronwall inequality (3-VII.1) to (3-VII.8), and noting (3-V.19), (3-V.22), (3-V.51), (3-V.33) and that $\Delta t\leq C_\star(\zeta^{-1})\,\alpha^{1+\zeta}\,h^2$, for a $\zeta>0$ where $C_\star(\zeta^{-1})$ is sufficiently small, yields the first three bounds in (3-VII.2).

Similarly to (3-VI.17), on choosing $\boldsymbol{w}=\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{n}-\boldsymbol{u}_{\alpha,h}^{n-1}}{\Delta t_n}\right)\right]\in\mathrm{V}_h^1$ in the L-independent version of (3-V.46a) yields, on noting (3-V.4), (3-VI.11), (3-VI.6) and Sobolev embedding, that

$$\begin{aligned}\mathrm{Re}\left\|\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{n}-\boldsymbol{u}_{\alpha,h}^{n-1}}{\Delta t_n}\right)\right\|_{H^1(\mathcal{D})}^2&=\mathrm{Re}\int_{\mathcal{D}}\frac{\boldsymbol{u}_{\alpha,h}^{n}-\boldsymbol{u}_{\alpha,h}^{n-1}}{\Delta t_n}\cdot\mathcal{R}_h\left[\mathcal{S}\left(\frac{\boldsymbol{u}_{\alpha,h}^{n}-\boldsymbol{u}_{\alpha,h}^{n-1}}{\Delta t_n}\right)\right]\\ &\leq C\big[\|\boldsymbol{\sigma}_{\alpha,h}^{n}\|_{L^2(\mathcal{D})}^2+\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n}\|_{L^2(\mathcal{D})}^2+\|\,\|\boldsymbol{u}_{\alpha,h}^{n-1}\|\,\|\boldsymbol{u}_{\alpha,h}^{n}\|\,\|_{L^2(\mathcal{D})}^2\\ &\qquad+\|\,\|\boldsymbol{u}_{\alpha,h}^{n-1}\|\,\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{n}\|\,\|_{L^{1+\theta}(\mathcal{D})}^2+\|\boldsymbol{f}^n\|_{H^{-1}(\mathcal{D})}^2\big]\end{aligned} \tag{3-VII.9}$$

for any $\theta>0$ as $d=2$. On taking the $\frac{2}{\vartheta}$ power of both sides of (3-VII.9), multiplying by Δt_n, summing from $n=1,\ldots,N_T$ and noting the L-independent versions of (3-VI.18) and (3-VI.19a) with $\theta=(\vartheta-2)/(6-\vartheta)$, (3-V.19), (3-III.8a), (3-V.51), (3-V.22), (3-VI.3) and the first bound in (3-VII.2) yields the last bound in (3-VII.2). □

It follows from (3-V.51), (3-VII.2), (3-VI.3), (3-VI.25) and (3-V.56a–c) that

$$\begin{aligned}&\sup_{t\in(0,T)}\|\boldsymbol{u}_{\alpha,h}^{\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2+\int_0^T\|\boldsymbol{\nabla}\boldsymbol{u}_{\alpha,h}^{\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2\,dt\\ &\qquad+\int_0^T\left[\|[\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+}]^{-1}\|_{L^1(\mathcal{D})}+\frac{\|\boldsymbol{u}_{\alpha,h}^{\Delta t,+}-\boldsymbol{u}_{\alpha,h}^{\Delta t,-}\|_{L^2(\mathcal{D})}^2}{\Delta(t)}\right]dt\leq C\end{aligned} \tag{3-VII.10a}$$

and

$$\begin{aligned}&\sup_{t\in(0,T)}\|\boldsymbol{\sigma}_{\alpha,h}^{\Delta t(,\pm)}\|_{L^2(\mathcal{D})}^2+\int_0^T\left[\alpha\|\boldsymbol{\nabla}\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+}\|_{L^2(\mathcal{D})}^2+\frac{\|\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+}-\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,-}\|_{L^2(\mathcal{D})}^2}{\Delta(t)}\right]dt\\ &\qquad+\int_0^T\left\|\mathcal{S}\frac{\partial\boldsymbol{u}_{\alpha,h}^{\Delta t}}{\partial t}\right\|_{H^1(\mathcal{D})}^{\frac{4}{\vartheta}}dt\leq C(\alpha^{-1},T),\end{aligned} \tag{3-VII.10b}$$

where $\vartheta \in (2,4)$.

We note that we have no control on the time derivative of $\boldsymbol{\sigma}_{\alpha,h}^{\Delta t}$ in (3-VII.10b). This is because if we choose $\boldsymbol{\phi} = \mathcal{P}_h \left[\mathcal{E} \left(\frac{\boldsymbol{\sigma}_{\alpha,h}^n - \boldsymbol{\sigma}_{\alpha,h}^{n-1}}{\Delta t_n} \right) \right] \in \mathrm{S}_h^1$ in the L-independent version of (3-V.46b), the terms involving $\boldsymbol{u}_{\alpha,h}^m$, $m = n-1$ and $m = n$, cannot now be controlled in the absence of the cut-off on $\boldsymbol{\sigma}_{\alpha,h}^n$. We are now in a position to prove the following convergence result for $(\mathrm{P}_{\alpha,h}^{\Delta t})$. The key difference between the following theorem and Theorem 8 for $(\mathrm{P}_{\alpha,h}^{L,\Delta t})$ is that no control on the time derivative of $\boldsymbol{\sigma}_{\alpha,h}^{\Delta t}$ in (3-VII.10b) implies no strong convergence for $\boldsymbol{\sigma}_{\alpha,h}^{\Delta t(,\pm)}$.

Theorem 10. *Let all the assumptions of Theorem 9 hold. Then there exists a subsequence of* $\{(\boldsymbol{u}_{\alpha,h}^{\Delta t}, \boldsymbol{\sigma}_{\alpha,h}^{\Delta t})\}_{h>0,\Delta t>0}$, *and functions* $\boldsymbol{u}_\alpha \in L^\infty(0,\ T;[L^2(\mathcal{D}))]^d) \cap L^2(0,T;\mathrm{V}) \cap W^{1,\frac{4}{\vartheta}}(0,T;\mathrm{V}')$ *and* $\boldsymbol{\sigma}_\alpha \in L^\infty(0,T;[L^2(\mathcal{D})]_{SPD}^{d\times d}) \cap L^2(0,T;[H^1(\mathcal{D})]_{SPD}^{d\times d})$ *such that, as* $h, \Delta t \to 0_+$,

$$\boldsymbol{u}_{\alpha,h}^{\Delta t(,\pm)} \to \boldsymbol{u}_\alpha \qquad \text{weak* in } L^\infty(0,T;[L^2(\mathcal{D})]^d), \tag{3-VII.11a}$$

$$\boldsymbol{u}_{\alpha,h}^{\Delta t(,\pm)} \to \boldsymbol{u}_\alpha \qquad \text{weakly in } L^2(0,T;[H^1(\mathcal{D})]^d), \tag{3-VII.11b}$$

$$\mathcal{S}\frac{\partial \boldsymbol{u}_{\alpha,h}^{\Delta t}}{\partial t} \to \mathcal{S}\frac{\partial \boldsymbol{u}_\alpha}{\partial t} \qquad \text{weakly in } L^{\frac{4}{\vartheta}}(0,T;\mathrm{V}), \tag{3-VII.11c}$$

$$\boldsymbol{u}_{\alpha,h}^{\Delta t(,\pm)} \to \boldsymbol{u}_\alpha \qquad \text{strongly in } L^2(0,T;[L^r(\mathcal{D})]^d), \tag{3-VII.11d}$$

and

$$\boldsymbol{\sigma}_{\alpha,h}^{\Delta t(,\pm)} \to \boldsymbol{\sigma}_\alpha \qquad \text{weak* in } L^\infty(0,T;[L^2(\mathcal{D})]^{d\times d}), \tag{3-VII.12a}$$

$$\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} \to \boldsymbol{\sigma}_\alpha \qquad \text{weakly in } L^2(0,T;[H^1(\mathcal{D})]^{d\times d}), \tag{3-VII.12b}$$

$$\Lambda_{m,p}(\boldsymbol{\sigma}_{\alpha,h}^{\Delta t(,\pm)}) \to \boldsymbol{\sigma}_\alpha \boldsymbol{I}_{mp} \qquad \text{weakly in } L^2(0,T;[L^2(\mathcal{D})]^{d\times d}), \quad m,p = 1,\dots,d, \tag{3-VII.12c}$$

where $\vartheta \in (2,4)$ *and* $r \in [1,\infty)$.

Furthermore, $(\boldsymbol{u}_\alpha, \boldsymbol{\sigma}_\alpha)$ *solve the following problem :*

($\mathbf{P}_\alpha$*)* *Find* $\boldsymbol{u}_\alpha \in L^\infty(0,T;[L^2(\mathcal{D})]^d) \cap L^2(0,T;\mathrm{V}) \cap W^{1,\frac{4}{\vartheta}}(0,T;\mathrm{V}')$ *and* $\boldsymbol{\sigma}_\alpha \in L^\infty(0,T;[L^2(\mathcal{D})]_{SPD}^{d\times d}) \cap L^2(0,T;[H^1(\mathcal{D})]_{SPD}^{d\times d})$ *such that*

$$\int_0^T \mathrm{Re}\left\langle \frac{\partial \boldsymbol{u}_\alpha}{\partial t}, \boldsymbol{v} \right\rangle_{\mathrm{V}} dt + \int_0^T \int_{\mathcal{D}} [(1-\varepsilon)\boldsymbol{\nabla}\boldsymbol{u}_\alpha : \boldsymbol{\nabla}\boldsymbol{v} + \mathrm{Re}[(\boldsymbol{u}_\alpha \cdot \boldsymbol{\nabla})\boldsymbol{u}_\alpha] \cdot \boldsymbol{v}]\, dt \tag{3-VII.13a}$$

$$= \int_0^T \langle \boldsymbol{f}, \boldsymbol{v} \rangle_{H_0^1(\mathcal{D})}\, dt - \frac{\varepsilon}{\mathrm{Wi}} \int_0^T \int_{\mathcal{D}} \boldsymbol{\sigma}_\alpha : \boldsymbol{\nabla}\boldsymbol{v}\, dt \qquad \forall \boldsymbol{v} \in L^{\frac{4}{4-\vartheta}}(0,T;\mathrm{V}),$$

$$-\int_0^T \int_{\mathcal{D}} \boldsymbol{\sigma}_\alpha : \frac{\partial \boldsymbol{\phi}}{\partial t}\, dt - \int_{\mathcal{D}} \boldsymbol{\sigma}^0 : \boldsymbol{\phi} \tag{3-VII.13b}$$

$$+ \int_0^T \int_{\mathcal{D}} [(\boldsymbol{u}_\alpha \cdot \boldsymbol{\nabla})\boldsymbol{\sigma}_\alpha : \boldsymbol{\phi} + \alpha \boldsymbol{\nabla}\boldsymbol{\sigma}_\alpha :: \boldsymbol{\nabla}\boldsymbol{\phi}]\, dt$$

$$= \int_0^T \int_{\mathcal{D}} \left[2(\boldsymbol{\nabla}\boldsymbol{u}_\alpha)\boldsymbol{\sigma}_\alpha - \frac{1}{\mathrm{Wi}}(\boldsymbol{\sigma}_\alpha - \boldsymbol{I}) \right] : \boldsymbol{\phi}\, dt$$

$$\forall \boldsymbol{\phi} \in L^2(0,T;[H^1(\mathcal{D})]_S^{d\times d}) \cap W_0^{1,1}(-T,T;[L^2(\mathcal{D})]_S^{d\times d});$$

and

$$\lim_{t\to 0_+} \int_{\mathcal{D}} (\boldsymbol{u}_\alpha(t,\boldsymbol{x}) - \boldsymbol{u}^0(\boldsymbol{x})) \cdot \boldsymbol{v} = 0 \qquad \forall \boldsymbol{v} \in \mathrm{H}. \tag{3-VII.13c}$$

Proof. The results (3-VII.11a–d) and (3-VII.12a,b) follow immediately from the bounds (3-VII.10a,b), as in the proof of Theorem 8. Similarly, the proof of positive definiteness of $\boldsymbol{\sigma}_\alpha$ follows as in Theorem 8 ; that is, (3-VI.32) and the weak convergence (3-VII.12a) is adequate for this. The result (3-VII.12c) follows from (3-VII.12a), (3-V.60) and (3-VII.10b) and the fact that $\boldsymbol{\sigma}_\alpha \in L^\infty(0,T;[L^2(\mathcal{D})]_{SPD}^{d\times d})$.

It follows from (3-V.2), (3-VII.10a,b), (3-VII.11a–d), (3-VII.12a), (3-III.8b), (3-VI.10) and (3-III.11) that we may pass to the limit, $h, \Delta t \to 0_+$, in the L-independent version of (3-V.58) to obtain that $(\boldsymbol{u}_\alpha, \boldsymbol{\sigma}_\alpha)$ satisfy (3-VII.13a). It also follows from (3-V.20a), (3-V.2), (3-VII.11c,d) and as V is dense in H that $\boldsymbol{u}_\alpha(0,\cdot) = \boldsymbol{u}^0(\cdot)$ in the required sense ; see (3-VII.13c) and [Tem84, Lemma 1.4, p179].

It follows from (3-VII.12a–c), (3-VII.11d), (3-VI.8), (3-VII.10a,b), (3-VI.1a,b), (3-I.4a) and as $\boldsymbol{u}_\alpha \in L^2(0,T;\mathrm{V})$ that we may pass to the limit $h, \Delta t \to 0_+$ in the L-independent version of (3-V.59) with $\boldsymbol{\chi} = \pi_h \boldsymbol{\phi}$ to obtain (3-VII.13b) for any $\boldsymbol{\phi} \in C_0^\infty(-T,T;[C^\infty(\overline{\mathcal{D}})]_S^{d\times d})$. For example, in order to pass to the limit on the first and third terms in the L-independent version of (3-V.59), we note that

$$\begin{aligned}
&\int_0^T \int_{\mathcal{D}} \pi_h \left[\left(\frac{\partial \boldsymbol{\sigma}_{\alpha,h}^{\Delta t}}{\partial t} + \frac{1}{\mathrm{Wi}} \boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} \right) : \pi_h \boldsymbol{\phi} \right] dt \qquad \text{(3-VII.14a)} \\
&\quad = \int_0^T \int_{\mathcal{D}} \left[\frac{1}{\mathrm{Wi}} \boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} : \pi_h \boldsymbol{\phi} - \boldsymbol{\sigma}_{\alpha,h}^{\Delta t} : \pi_h \left[\frac{\partial \boldsymbol{\phi}}{\partial t} \right] \right] dt - \int_D \pi_h \left[\boldsymbol{\sigma}_{\alpha,h}^{\Delta t} : \pi_h \boldsymbol{\phi} \right](0,\cdot) \\
&\qquad\qquad + \int_0^T \int_{\mathcal{D}} (\pi_h - \boldsymbol{I}) \left[\frac{1}{\mathrm{Wi}} \boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} : \pi_h \boldsymbol{\phi} - \boldsymbol{\sigma}_{\alpha,h}^{\Delta t} : \pi_h \left[\frac{\partial \boldsymbol{\phi}}{\partial t} \right] \right] dt, \\
&\int_{\mathcal{D}} \boldsymbol{\nabla} \boldsymbol{u}_{\alpha,h}^{\Delta t,+} : \pi_h [\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} \pi_h \boldsymbol{\phi}] \qquad \text{(3-VII.14b)} \\
&\quad = \int_{\mathcal{D}} \boldsymbol{\nabla} \boldsymbol{u}_{\alpha,h}^{\Delta t,+} : (\pi_h - \boldsymbol{I}) [\boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} \pi_h \boldsymbol{\phi}] \\
&\qquad\qquad - \int_{\mathcal{D}} \left\{ \left((\boldsymbol{\nabla} \pi_h \boldsymbol{\phi}) \boldsymbol{u}_{\alpha,h}^{\Delta t,+} \right) : \boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} + \boldsymbol{u}_{\alpha,h}^{\Delta,+} \cdot \left((\pi_h \boldsymbol{\phi}) \operatorname{div} \boldsymbol{\sigma}_{\alpha,h}^{\Delta t,+} \right) \right\};
\end{aligned}$$

where $((\boldsymbol{\nabla} \pi_h \boldsymbol{\phi}) \boldsymbol{u}_{\alpha,h}^{\Delta t,+})(t,\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ with $[(\boldsymbol{\nabla} \pi_h \boldsymbol{\phi}) \boldsymbol{u}_{\alpha,h}^{\Delta t,+}]_{ij} = \sum_{k=1}^d \frac{\partial (\pi_h \boldsymbol{\phi})_{ik}}{\partial \boldsymbol{x}_j} [\boldsymbol{u}_{\alpha,h}^{\Delta t,+}]_k$. The desired result (3-VII.13b) then follows from noting that $C_0^\infty(-T,T;[C^\infty(\overline{\mathcal{D}})]_S^{d\times d})$ is dense in $W_0^{1,1}(0,T;\, [H^1(\mathcal{D})]_S^{d\times d})$. □

We have the analogue of Remark 12.

Remark 13. *It follows from (3-VII.10a,b), (3-VII.11a,b) and (3-VII.12a,b) that*

$$\sup_{t\in(0,T)} \|\boldsymbol{u}_\alpha\|_{L^2(\mathcal{D})}^2 + \int_0^T \|\boldsymbol{\nabla} \boldsymbol{u}_\alpha\|_{L^2(\mathcal{D})}^2 dt \le C, \qquad \text{(3-VII.15a)}$$

$$\sup_{t\in(0,T)} \|\boldsymbol{\sigma}_\alpha\|_{L^2(\mathcal{D})}^2 + \alpha \int_0^T \|\boldsymbol{\nabla} \boldsymbol{\sigma}_\alpha\|_{L^2(\mathcal{D})}^2 dt \le C(\alpha^{-1}, T). \qquad \text{(3-VII.15b)}$$

Hence, although we have introduced diffusion with a positive coefficient α into the stress equation (3-VII.13b) compared to the standard Oldroyd-B model; the bound (3-VII.15a) on the velocity $\boldsymbol{u}_\alpha$ is independent of the parameter α, where $(\boldsymbol{u}_\alpha, \boldsymbol{\sigma}_\alpha)$ solves (P_α), (3-VII.13a–c).

Deuxième partie

Applications d'une méthode numérique de réduction de bases à des problèmes multi-échelles

Chapitre 4

Une méthode de réduction de bases *certifiée* avec applications à quelques problèmes multi-échelles.

Ce chapitre est une introduction à la méthode des bases réduites développée à partir du travail initial [PRV+02]. Après une exposition de la méthode dans un cas simple sont référencées les diverses généralisations de la méthode.

Dans sa version standard usuelle, la méthode des bases réduites en question permet aujourd'hui de calculer efficacement (rapidement et précisément) un champ scalaire paramétré, appelé *sortie*, pour de nombreuses valeurs du paramètre, appelé *entrée*, quand la sortie est donnée comme une fonction (linéaire) en la solution d'un problème aux bords (elliptique ou parabolique) paramétré.

On présente également des développements de la méthode, que nous avons introduits pour simuler numériquement quelques problèmes multi-échelles particulièrement gourmands en ressources informatiques.

The Certified Reduced-Basis Method for Multiscale Problems

Sébastien Boyaval[a,b],

[a]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).

[b]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

4-I An Initiation to Reduced-Basis Techniques

In multiscale models for materials, quantities of interest $u(\mu)$ at one scale are often parametrized by quantities μ from another scale [LB05]. To compute the influence of one scale on another one, one thus needs to evaluate the parametrized quantity $u(\mu)$ and some *output* $s(\mu)=F(u(\mu))$ many times, for many values of the *input* parameter μ. If the computation of $u(\mu)$ and $s(\mu)$ for one given parameter μ already invokes elaborate algorithms, then the numerical simulation of $u(\mu)$ and $s(\mu)$ for many μ in multiscale models becomes a computationally very-demanding task. Reducing the cost of such parametrized computations is thus a challenge to the numerical simulation of multiscale models. Among many possible techniques to reduce the computations in the numerical simulation of multiscale models, we focus here on those we have developed from the *Reduced-Basis* (RB) approach initiated in [PRV$^+$02]. Different kinds of multiscale problems will be treated herein, see the next Sections 4-I-B, 4-II and 4-III. But first, as appetizer, we give an introduction to the standard RB method in Section 4-I-A, starting with a simple paradigmatic problem. (See also [PR07b, Qua09] for introductions to the standard RB method, though with different perspectives).

Let us start our exposition with an overview of the main RB ideas. We need to specify the variable of interest $u(\mu)\in X$ as an element of a Hilbert space X (typically of infinite dimension) with inner product $(\cdot,\cdot)_X$ and norm $\|\cdot\|_X$, the output $s(\mu)=F(u(\mu))\in\mathbb{R}$ as a scalar quantity where $F:X\to\mathbb{R}$ is a smooth function (typically linear) and $\mu\in\Lambda\subset\mathbb{R}^P$ as a P-dimensional parameter in a fixed given range. Typically, the RB method is well-suited for the cases where $u(\mu)$ is solution to a coercive elliptic μ-parametrized Boundary Value Problem (μ-BVP), for which there exist accurate approximation schemes (like Finite-Element discretizations) and *a posteriori* estimators. So, for the sake of simplicity, let us first assume $u(\mu)$ is solution to a variational formulation :

$$\text{Find } u(\mu)\in X \text{ solution to } a(u(\mu),v;\mu)=l(v),\ \forall v\in X, \tag{4-I.1}$$

where $a(\cdot,\cdot;\mu)$ is a μ-parametrized symmetric bilinear form, continuous and coercive on X, and where $l(\cdot)$ is a linear form, continuous on X. For all $\mu\in\Lambda$, $a(\cdot,\cdot;\mu)$ thus defines an inner product in X and the existence of $u(\mu)$ then simply owes to the Riesz-Fréchet representation theorem for linear forms, continuous on X (a simple case in the Lax-Milgram lemma because of symmetry). We denote by $\|\cdot\|_\mu$ the corresponding norm equivalent to $\|\cdot\|_X$, which is usually termed the energy norm. In the sequel, we denote by $u_{\mathcal{N}}(\mu)\in X_{\mathcal{N}}$ an accurate Galerkin approximations for $u(\mu)$ in a linear subspace $X_{\mathcal{N}}\subset X$ of dimension $\mathcal{N}\gg 1$ and by $s_{\mathcal{N}}(\mu)=F(u_{\mathcal{N}}(\mu))$ the corresponding approximation for the output $s(\mu)$. For that choice of $X_{\mathcal{N}}$, we will assume that the approximation error $|s(\mu)-s_{\mathcal{N}}(\mu)|$ is uniformly sufficient small for all $\mu\in\Lambda$ to use $s_{\mathcal{N}}(\mu)$ as a good approximation of the output $s(\mu)$ in practical applications. The goal of the RB method is to construct a linear[1] subspace $X_{\mathcal{N},N}\subset X_{\mathcal{N}}$ of dimension $N\ll\mathcal{N}$ as the span of a few "snapshot" (approximate) solutions to the μ-BVP (4-I.1)

$$X_{\mathcal{N},N}=\mathbf{Span}\left(u_{\mathcal{N}}(\mu_n^N),n=1,\dots,N\right), \tag{4-I.2}$$

chosen in the manifold $\mathcal{M}_{\mathcal{N}}=\{u_{\mathcal{N}}(\mu),\mu\in\Lambda\}$ for well selected parameter values $(\mu_n^N)_{1\le n\le N}\in\Lambda^N$. For all $\mu\in\Lambda$, the RB approximation $u_{\mathcal{N},N}(\mu)$ is computed as the linear combination of snapshots $u_{\mathcal{N}}(\mu_n^N)$ solution to :

$$\text{Find } u_{\mathcal{N},N}(\mu)\in X_{\mathcal{N},N} \text{ solution to } a(u_{\mathcal{N},N}(\mu),v;\mu)=l(v),\ \forall v\in X_{\mathcal{N},N}. \tag{4-I.3}$$

Notice that in this simple framework, $u_{\mathcal{N},N}(\mu)$ can also be seen as a projection in $X_{\mathcal{N},N}\subset X_{\mathcal{N}}\subset X$:

$$u_{\mathcal{N},N}(\mu)=\underset{v\in X_{\mathcal{N},N}}{\operatorname{arginf}}\|u_{\mathcal{N}}(\mu)-v\|_\mu=\underset{v\in X_{\mathcal{N},N}}{\operatorname{arginf}}\|u(\mu)-v\|_\mu. \tag{4-I.4}$$

[1]Note yet the on-going development of nonlinear versions [EPR].

We will write $s_{\mathcal{N},N}(\mu)=F(u_{\mathcal{N},N}(\mu))$ the corresponding approximation of the output $s(\mu)$ and $\Delta_N^s(\mu)$ the *a posteriori* estimator for the output approximation error $|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)|$.

The RB method will yield good approximations $s_{\mathcal{N},N}(\mu)$ of $s_{\mathcal{N}}(\mu)$ when the dependence of the output on the input parameter μ is sufficiently smooth. As a consequence, optimal choices for the approximation space $X_{\mathcal{N},N}$ should take into account the regularity in μ. More precisely, our *goal-oriented* RB method should select parameter values $(\mu_n^N)_{1\leq n\leq N}\in\Lambda^N$ with a view to controlling the right norm of the output approximation error $|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)|$ as a function of μ. For most applications, it is enough to consider the L^∞ topology for $\mu\in\Lambda$ (see nevertheless the Remark 14), and our goal-oriented RB technique thus ideally chooses $X_{\mathcal{N},N}$ such that :

$$(\mu_n^N)_{1\leq n\leq N}\in \operatorname*{arginf}_{(\mu_n)_{1\leq n\leq N}\in\Lambda^N}\left(\operatorname*{esssup}_{\mu\in\Lambda}|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)|\right) \tag{4-I.5}$$

Unfortunately, it is very difficult to compute (4-I.5) in practice. So the RB method initiated in [PRV+02] suggests a pragmatic remedy to the computation of (4-I.5). It proposes a practical and quite general procedure to build good approximations of (4-I.5) using (i) a (computationally cheap) estimator $\Delta_N^s(\mu)$ for $|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)|$ and (ii) a greedy algorithm to approximate minimizers of

$$\operatorname*{arginf}_{(\mu_n)_{1\leq n\leq N}\in\Lambda^N}\left(\sup_{\mu\in\Lambda_{\text{trial}}}\Delta_N^s(\mu)\right) \tag{4-I.6}$$

after discretizing Λ into a very large trial sample of parameters $\Lambda_{\text{trial}}\subset\Lambda$. We recall that approximating (4-I.6) with a *greedy* algorithm means that the parameter values μ_n^N, $n=1,\ldots,N$, are selected incrementally ; in particular, this is interesting when N is not known in advance, since then the computation of approximate μ_n^N $(1\leq n\leq N)$ with a greedy algorithm does not depend on N and runs until the infimum in (4-I.6) is below some terminal condition for a sufficiently large N. In the sequel, since $\mu_n^{N+1}=\mu_n^N$ for $n=1,\ldots,N$ and all $1\leq N\leq\mathcal{N}-1$, we denote μ_n, $n=1,\ldots,N$, the parameter values retained by the greedy algorithm and used for the definition (4-I.2) of $X_{\mathcal{N},N}$.

Of course, the computation of approximations to (4-I.6) with a *greedy* algorithm can still be expensive, because a very large trial sample of parameters $\Lambda_{\text{trial}}\subset\Lambda$ has to be explored. So, in any case, the use of the RB method justifies when the solution to the μ-parametrized problem (4-I.1) has to be computed many times for a large number of input parameter values μ in the end : this is termed a *many-query* computational framework for the parametrized problem (4-I.1). (Notice that the RB method is not a competitor to the usual discretization methods ; in fact, it rather collaborates with good discretization methods through correct choices of $X_{\mathcal{N}}$ to speed up the reiterated computations calling to the same discrete scheme.) In many-query frameworks well-suited for computational reductions, the RB method will deploy in a two-stage procedure :

- first, in a so-called *offline* stage (possibly computationally expensive), one "learns" from a very large trial (or training) sample of parameters $\Lambda_{\text{trial}}\subset\Lambda$ how to choose a small number N of parameter values using a greedy algorithm – accurate approximations $u_{\mathcal{N}}(\mu_n)$ for solutions $u(\mu_n)$ to (4-I.1) are computed only at those few parameter values μ_n, $n=1,\ldots,N$, selected incrementally by the greedy algorithm – ;
- second, in a so-called *online* stage, computationally cheap approximations $u_{\mathcal{N},N}(\mu)$ for solutions $u(\mu)$ to (4-I.1) are computed for many values $\mu\in\Lambda$ of the parameter not necessarily in the offline trial sample Λ_{trial}, and yield supposedly good approximations $s_{\mathcal{N},N}(\mu)$ for the output $s(\mu)$ (here we use again the cheap estimator $\Delta_N^s(\mu)$ to check this).

In a nutshell, the core of the RB ideas, well-adapted for many-query frameworks with reiterated computations of input-output relationships, is an efficient offline-online decomposition of the computations based on :

- a (fast sharp) *a posteriori* error estimator for the output, which is both used online for *certification* of the quality of the RB approximations and offline in a (greedy) selection process, and
- a greedy algorithm to build hierarchical RB approximation spaces $X_{\mathcal{N},N}\subset X_{\mathcal{N},N+1}$.

Then, for the parametrized problems with a large number of output queries at many input parameter values, the RB method should compensate for the cost of the offline computations (construction of $X_{\mathcal{N},N}$) by accelerated online computations without loss of precision in the output.

Remark 14. *For some applications, it might be enough to consider an other topology than the L^∞ topology for $\mu\in\Lambda$. In particular, the L^2 topology, which is weaker on the bounded parameter domain Λ, might be enough (see also the Remark 18 in Section 4-II). Besides, the L^2 topology is indeed the basis of many existing reduction techniques (similar to one another but with many different names) : PCA (principal component analysis), POD (proper orthogonal decomposition) [KV02], quantization [PP05], Karhunen-Loève decomposition [New96], SVD (singular value decomposition) [LBLM08] and low-rank approximations... These reduction techniques have proved*

efficient in some applications. But as soon as one wants to compute with them, it is clear that one considers a different path than ours for the reduction of computations : the many-query frameworks we will consider next here are usually not ideal fields for efficient computational reductions with the L^2 techniques. That is why we will not compare our RB techniques with these techniques here, since they belong to different projects. Nevertheless, for some specific applications (like the one in the Section 4-II), it might still be useful to compare elsewhere, and maybe combine, ideas from the two different approaches (the L^∞ and the L^2), provided both are possible.

4-I-A Paradigmatic example with one-dimensional parameter

For the sake of simplicity, let us start with a paradigm for the development of the RB method : a *compliant* coercive elliptic scalar problem parametrized by a single *affine* parameter (the significance of the specifications compliant and affine will be made precise later in Section 4-I-B).

4-I-A-a Mathematical setting for a linear scalar second-order elliptic problem

Let us use only *one* scalar quantity $\mu \in \Lambda$ ($P=1$), with $\Lambda = [\mu_{\min}, \mu_{\max}] \subset \mathbb{R}_{\geq 0}$, to parametrize the coefficients of a linear scalar elliptic Partial Differential Equation (PDE) satisfied by $u(\mu)$ solution to the following Dirichlet problem with homogeneous Boundary Conditions (BC) :

$$\begin{cases} -\operatorname{div}\left(\underline{\underline{A}}(\mu)\boldsymbol{\nabla} u(\mu)\right) = f \text{ in } \mathcal{D}, \\ u(\mu) = 0 \text{ on } \partial\mathcal{D}, \end{cases} \tag{4-I.7}$$

where $\mathcal{D} \subset \mathbb{R}^d$ with $d=2,3$ and the matrix $\underline{\underline{A}}(\mu)$ depends on μ in an *affine* way :

$$\underline{\underline{A}}(\mu) = \underline{\underline{A_0}} + \mu\, \underline{\underline{A_1}},\ \forall \mu \in \Lambda. \tag{4-I.8}$$

We choose $\mathcal{D}$ an open bounded connected domain with Lipschitz boundary $\partial\mathcal{D}$, $f \in L^2(\mathcal{D})$ in the Lebesgue space of square integrable functions, and $\underline{\underline{A}}(\mu)$ a function in $[L^\infty(\mathcal{D})]^{d\times d}$ with *symmetric* matrix values, definite positive almost everywhere (a.e.) in $\mathcal{D}$ (each entry is in $L^\infty(\mathcal{D})$, and for all $\underline{X} \in \mathbb{R}^d$, $\underline{\underline{A}}(\mu)\underline{X}\cdot\underline{X} > 0$ a.e. in $\mathcal{D}$). Without loss of generality, we can assume $\underline{\underline{A_0}}$ is a.e. symmetric definite positive, and $\underline{\underline{A_1}}$ is a.e. symmetric positive. Then, for all $\mu \in \Lambda$, the following mapping is well-defined :

$$u(\cdot) : \mu \in \Lambda \to u(\mu) \in H^1_0(\mathcal{D}).$$

The Hilbert space X can thus be chosen equal to the Sobolev space $H^1_0(\mathcal{D})$, endowed with its usual Hilbert structure $(\cdot,\cdot)_X = (\cdot,\cdot)_{H^1_0(\mathcal{D})}$ (with norm $\|\cdot\|_X = \|\cdot\|_{H^1_0(\mathcal{D})}$). And $u(\mu)$ is equivalently solution to the variational formulation (4-I.1) with :

- $a(w,v;\mu) = \displaystyle\int_{\mathcal{D}} \underline{\underline{A}}(\mu)\nabla w \cdot \nabla v,\ \forall w,v \in X,\ \mu \in \Lambda$ and
- $l(v) = \displaystyle\int_{\mathcal{D}} f\, v, \forall v \in X,$

We are interested in efficiently computing, for many values of $\mu \in \Lambda$, the output :

$$s(\mu) = F(u(\mu)) := \int_{\mathcal{D}} f\, u(\mu) = l(u(\mu)). \tag{4-I.9}$$

Moreover, if we assume $\mathcal{D}$ is polygonal for instance, there exist many well-established discretization methods that allow to compute Galerkin approximations $u_{\mathcal{N}}(\mu)$ of $u(\mu)$ in finite dimensional linear subspaces $X_{\mathcal{N}}$ of X for any fixed parameter value $\mu \in \Lambda$, like the Finite-Element (FE) method [SF73a, Cia78]. Let us denote by $(\phi_n)_{1\leq n\leq \mathcal{N}}$ a Galerkin basis of $X_{\mathcal{N}}$. Then, for each parameter value $\mu \in \Lambda$, the numerical computation of $u_{\mathcal{N}}(\mu) = \sum_{n=1}^{\mathcal{N}} U_n(\mu)\phi_n$ is achieved by solving a large $\mathcal{N}\times\mathcal{N}$ (sparse) linear system for the vector $U(\mu) = (U_n(\mu))_{1\leq n\leq \mathcal{N}} \in \mathbb{R}^{\mathcal{N}}$:

$$\text{Find } U(\mu) \in \mathbb{R}^{\mathcal{N}} \text{ solution to } \underline{\underline{B}}(\mu)U(\mu) = b,$$

where $b = (l(\phi_n))_{1\leq n\leq \mathcal{N}}$ is also a vector in $\mathbb{R}^{\mathcal{N}}$ and $\underline{\underline{B}}(\mu) = \underline{\underline{B_0}} + \mu\, \underline{\underline{B_1}}$ is a $\mathcal{N}\times\mathcal{N}$ real invertible matrix. So, for each parameter value μ, the entries of the latter matrix $\underline{\underline{B}}(\mu)$ can be computed in $O(\mathcal{N}^2)$ operations (multiplications

and additions) using the precomputed integrals $\underline{\underline{B_q}}_{ij} = \int_{\mathcal{D}} \underline{\underline{A_q}} \nabla\phi_i \cdot \nabla\phi_j$, $i,j = 1,\ldots,\mathcal{N}$ for $q = 0,1$. (Note that the assumption of affine parametrization has been essential here, see Section 4-I-B-a.) And the evaluation of $U(\mu)$ for many $J \gg 1$ parameter values μ using iterative solvers costs $J \times O(\mathcal{N}^k)$ operations with $2 < k \leq 3$ [GvL96].

The goal of the RB approach is to build a smaller finite dimensional approximation space $X_{\mathcal{N},N} \subset X_{\mathcal{N}}$ sufficiently good for all $\mu \in \Lambda$, with $N \ll \mathcal{N}$, so that the computational cost is roughly reduced to $N \times O(\mathcal{N}^k) + J \times O(N^3)$, with $N \times O(\mathcal{N}^k)$ the cost of offline computations and $J \times O(N^3)$ the cost of online computations, the latter being *independent of* $\mathcal{N}$!

4-I-A-b Goal-oriented *a posteriori* error estimates

To proceed further with our goal-oriented RB method, we need sharp and computationally inexpensive *a posteriori* error estimators $\Delta_N^s(\mu)$ for the output RB approximation error $|s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)|$, $\forall \mu \in \Lambda$. For the coercive elliptic problem (4-I.7), such goal-oriented *a posteriori* error estimators are quite simple to devise, based on a global *a posteriori* error estimator $\Delta_N(\mu)$ for $\|u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)\|_\mu$ using a technique with residus. For other problems, this may be an issue. We refer to [VRP02, VPRP03a, HP07b, Dep08, PR07a, NRHP09, NRP09] for examples of goal-oriented *a posteriori* error estimations in various applied settings of the certified RB method.

We first need, for all $\mu \in \Lambda$, a computable lower bound $\alpha_{LB}(\mu)$ for the coercivity constant of $a(\cdot,\cdot;\mu)$, such that :

$$0 < \alpha_{LB}(\mu) \leq \alpha_c(\mu) = \inf_{w \in X \setminus \{0\}} \frac{a(w,w;\mu)}{\|w\|_X^2}, \ \forall \mu \in \Lambda, \tag{4-I.10}$$

which can be given by the *a priori* analysis, for instance after checking the coercivity assumption for $a(\cdot,\cdot;\mu)$ (see also Remark 16 in Section 4-I-B-b). We also define the residual bilinear form $g(\cdot,\cdot;\mu)$ as

$$g(w,v;\mu) = a(w,v;\mu) - l(v), \ \forall w,v \in X, \ \forall \mu \in \Lambda,$$

and a (computable) affine operator $G(\mu) : X_{\mathcal{N}} \to X_{\mathcal{N}}$ such that

$$g(w,v;\mu) = \left(G(\mu)\, w, v\right)_X, \ \forall w,v \in X_{\mathcal{N}}, \ \forall \mu \in \Lambda.$$

Then we have :

Proposition 18. *For any linear subspace $X_{\mathcal{N},N}$ of $X_{\mathcal{N}}$, there exists a computable error bound $\Delta_N^s(\mu)$ such that :*

$$|s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)| \leq \Delta_N^s(\mu) := \frac{\|G(\mu)\, u_{\mathcal{N},N}(\mu)\|_X^2}{\alpha_{LB}(\mu)}, \ \forall \mu \in \Lambda. \tag{4-I.11}$$

Proof. Since $X_{\mathcal{N},N}$ is a linear subspace of $X_{\mathcal{N}}$, the Galerkin orthogonality property holds :

$$a(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu), v) = 0, \ \forall v \in X_{\mathcal{N},N}. \tag{4-I.12}$$

On noting the linearity of $F = l$, the fact that $u_{\mathcal{N}}(\mu)$ is solution to a variational problem in $X_{\mathcal{N}} \ni u_{\mathcal{N},N}(\mu)$, the symmetry of $a(\cdot,\cdot;\mu)$ and (4-I.12), the output RB approximation error reads :

$$\begin{aligned} |s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)| &= |F(u_{\mathcal{N}}(\mu)) - F(u_{\mathcal{N},N}(\mu))| \\ &= |l(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu))| \\ &= |a(u_{\mathcal{N}}(\mu), u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu);\mu)| \\ &= |a(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu), u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu);\mu)| = \|u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)\|_\mu^2. \end{aligned} \tag{4-I.13}$$

Finally, on noting $a(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu), v;\mu) = -g(u_{\mathcal{N},N}(\mu), v;\mu)$ and $\sqrt{\alpha_{LB}(\mu)}\|v\|_X \leq \|v\|_\mu$ for all $v \in X_{\mathcal{N}}$, the choice $v = u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)$ yields a computable error bound $\Delta_N(\mu)$ as :

$$\|u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)\|_\mu \leq \Delta_N(\mu) := \frac{\|G(\mu)\, u_{\mathcal{N},N}(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}}, \tag{4-I.14}$$

which concludes the proof by combining (4-I.13) with (4-I.14).

□

Notice the quadratic effect $\Delta_N^s(\mu)=\Delta_N(\mu)^2$ in the goal-oriented error, which is essentially due to the symmetry of the bilinear form $a(\cdot,\cdot;\mu)$ here. In more general situations, it will be important for the reduction of computations to retain that quadratic effect if possible (see Section 4-I-B-c).

Note also the following inverse inequality :

$$\Delta_N^s(\mu)\leq\left(\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}\right)^2|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)| \tag{4-I.15}$$

where we have introduced the continuity constant $\gamma(\mu)<\infty$ of the bilinear form :

$$0<\gamma(\mu)=\sup_{w\in X\setminus\{0\}}\sup_{v\in X\setminus\{0\}}\frac{a(w,v;\mu)}{\|w\|_X\|v\|_X},\ \forall\mu\in\Lambda. \tag{4-I.16}$$

The inequality (4-I.15) ensures sharpness of the *a posteriori* estimator (4-I.11), depending on the quality of the lower-bound $\alpha_{LB}(\mu)$. (This can be an issue in practical computations, see Section 4-I-B-b.)

4-I-A-c Offline stage : a greedy algorithm to select the parameters in a trial sample

The RB method initiated in [PRV+02] suggests to build a linear subspace $X_{\mathcal{N},N}\subset X_{\mathcal{N}}$ incrementally with a *greedy* algorithm applied to a finite trial or training sample Λ_{trial} nested in Λ :

choose $\mu_1\in\Lambda_{\text{trial}}$ randomly, or take $\mu_1\in\text{argmax}\{|s(\mu)|,\ \mu\in\Lambda_{\text{smalltrial}}\subset\Lambda_{\text{trial}},\ |\Lambda_{\text{smalltrial}}|\ll|\Lambda_{\text{trial}}|\}$
compute $u_{\mathcal{N}}(\mu_1)$ to define $X_{\mathcal{N},1}=\mathbf{Span}(u_{\mathcal{N}}(\mu_1))$
for $n=2$ to N **do**
  choose $\mu_n\in\text{argmax}\{\Delta_{n-1}^s(\mu),\mu\in\Lambda_{\text{trial}}\}$
  compute $u_{\mathcal{N}}(\mu_n)$ to define $X_{\mathcal{N},n}=\mathbf{Span}(u_{\mathcal{N}}(\mu_m),\ m=1,\ldots,n)$
end for

where $\Lambda_{\text{smalltrial}}$ is typically a very small trial sample nested in Λ (much smaller than $\Lambda_{\text{trial}}\subset\Lambda$). Notice that the RB greedy algorithm above is usually supplied with an explicit tolerance $\varepsilon>0$ on the output RB approximation error rather than a fixed number of iterations $N\geq 2$ given *a priori*. Then $N=N(\varepsilon)$ is implicitly determined as a function of ε by the relation $\sup\{\Delta_N^s(\mu),\mu\in\Lambda_{\text{trial}}\}\leq\varepsilon$. We use N instead of ε here to control the loop of iterations only because we have not specified the context which requires numerical evaluations of the output $s(\mu)$ at a precision level ε.

We have not made precise yet how to choose the trial sample Λ_{trial} (and similarly, the smaller one $\Lambda_{\text{smalltrial}}$). In practice, Λ_{trial} is often simply chosen as a random sample in Λ (and similarly for $\Lambda_{\text{smalltrial}}$). Of course, this first guess may be insufficient to reach the required accuracy level ε in $\Delta_N^s(\mu)$ when $\mu\in\Lambda\supset\Lambda_{\text{trial}}$. But fortunately, if the computation of $\Delta_N^s(\mu)$ for any $\mu\in\Lambda$ is cheap enough, one can check this accuracy online for each query in μ. Then, in the case where $\Delta_N^s(\mu)>\varepsilon$ for some online queried μ, one can still compute $u_{\mathcal{N}}(\mu)$ for that same μ and enrich $X_{\mathcal{N},N}$ to reach the required accuracy level ε at that μ. (We explain the methodology for fast computations of $\Delta_N^s(\mu)$ in the next Section 4-I-A-d dealing with online computations.)

Besides, for the sake of consistency of the offline selection procedure with the online goal, the greedy algorithm above also needs to compute the same estimator $\Delta_N^s(\mu)$ for all $\mu\in\Lambda_{\text{trial}}$. And since the computation of $\Delta_N^s(\mu)$ is fast for all $\mu\in\Lambda\supset\Lambda_{\text{trial}}$, this allows in return the exploration of a very large training sample Λ_{trial} offline (with cardinal $|\Lambda_{\text{trial}}|\gg 1$).

To sum up, the game for efficient computational reductions with the RB method and without a precise knowledge of what Λ_{trial} should be to reach the required accuracy level is as follows. One first chooses Λ_{trial} as large as possible in the offline stage, to explore Λ extensively, insofar as the computation remains modest at each iteration $n=1,\ldots,N$ of the greedy algorithm (many cheap *a posteriori* estimations and only one accurate computation $u_{\mathcal{N}}(\mu_{n+1})$). Then, one can use the small RB space $X_{\mathcal{N},N}$ trained on the initial guess Λ_{trial} to compute cheap Galerkin approximations $u_N(\mu)$ in the online stage plus cheap *a posteriori* estimator $\Delta_N^s(\mu)$ at any $\mu\in\Lambda$. This bootstrap approach is what we term the *certified* RB approach. It provides significant computational reductions in the online stage provided the RB approximation space $X_{\mathcal{N},N}$ does not need to be enriched too often online.

We know neither about any systematic procedure to build good initial guesses Λ_{trial}, nor about *a priori* results that quantify how good is an initial guess Λ_{trial} for a precise problem. This is somewhat unsatisfactory : it is generally difficult to predict when the RB approach is successful. As a palliative to this lack of *a priori* knowledge, the RB approach proposes a general method that can *a posteriori* check the quality of, and possibly improve the quality of, a crude initial guess Λ_{trial} (coming from physical or mathematical intuition), which is

first filtered out by a greedy algorithm using some knowledge specific to the problem at stake (buried in *ad hoc* estimators $\Delta_N^s(\mu)$). Numerous numerical evidences support that viewpoint [VRP02, VPRP03a, HP07b, PR07b, PR07a, Dep08, NRP09, NRHP09].

Last, notice that the cost of offline computations scales as $W_{\text{offline}} = O(|\Lambda_{\text{trial}}|) \times \left(\sum_{n=1}^{N-1} w_{\text{online}}(n)\right) + N \times O(\mathcal{N}^k)$ where $w_{\text{online}}(n)$ is the marginal cost of one online-type computation for $u_{\mathcal{N},n}(\mu)$ and $\Delta_n^s(\mu)$ at a selected parameter value $\mu \in \Lambda_{\text{trial}}$ (where $1 \le n \le N-1$), and $O(|\Lambda_{\text{trial}}|)$ includes a max-search in Λ_{trial}. (Recall that $2 < k \le 3$ depends on the solver used for large sparse linear systems.)

4-I-A-d Online stage : fast computations including *a posteriori* estimators

We now explain how to compute fast $u_{\mathcal{N},n}(\mu)$, $s_{\mathcal{N},n}(\mu)$ and $\Delta_n^s(\mu)$ once the RB approximation space $X_{\mathcal{N},n}$ has been constructed. This procedure applies to the many offline computations when $\mu \in \Lambda_{\text{trial}}$ explores the trial sample in order to find $\mu_{n+1} \in \text{argmax}\{\Delta_n^s(\mu), \mu \in \Lambda_{\text{trial}}\}$ at each iteration $n = 1, \ldots, N-1$ of the greedy algorithm, and to the many online computations for a large number $J \gg 1$ of queried parameter values $\mu \in \Lambda$ when $n = N$. We present the procedure corresponding to the online stage in $X_{\mathcal{N},N}$ only. (It is straightforward to adapt for the offline stage in $X_{\mathcal{N},n}$, $n = 1, \ldots, N-1$.)

Recall that $X_{\mathcal{N},N} = \mathbf{Span}(u_{\mathcal{N}}(\mu_n),\ n = 1, \ldots, N)$. Assuming $\Delta_{N-1}^s(\mu_N) > 0$, it is then clear that $\dim(X_{\mathcal{N},N}) = N$. So $\left(u_{\mathcal{N}}(\mu_n^N)\right)_{1 \le n \le \mathcal{N}}$ is a basis in $X_{\mathcal{N},N}$. For any $\mu \in \Lambda$, we would then like to compute the RB approximation $u_{\mathcal{N},N}(\mu) = \sum_{n=1}^N U_{N,n}(\mu) u_{\mathcal{N}}(\mu_n)$, which can be achieved by solving a small $N \times N$ (full) linear system for the vector $U_N(\mu) = (U_{N,n}(\mu))_{1 \le n \le N} \in \mathbb{R}^N$:

$$\text{Find } U_N(\mu) \in \mathbb{R}^N \text{ solution to } \underline{\underline{C}}(\mu) U_N(\mu) = c,$$

where $c = (l(u_{\mathcal{N}}(\mu_n)))_{1 \le n \le N}$ is also a vector in $\mathbb{R}^N$ and $\underline{\underline{C}}(\mu) = \underline{\underline{C_0}} + \mu \underline{\underline{C_1}}$ is a $N \times N$ real invertible matrix. Note yet that the matrix $\underline{\underline{C}}(\mu)$ may be close to singular (that is, its singular values decrease fast) in the sense that it is the Grammian of a weighted least-squares problem $\underline{\underline{C}}(\mu) = \underline{\underline{D}}^T \underline{\underline{B}}(\mu) \underline{\underline{D}}$ where $\underline{\underline{D}}$ is a typically a nearly-rank-deficient $\mathcal{N} \times N$ real matrix. (The entries of one column of $\underline{\underline{D}}$ are the coefficients of one vector $u_{\mathcal{N}}(\mu_n)$, $1 \le n \le N$, in the Galerkin basis $(\phi_n)_{1 \le n \le \mathcal{N}}$ of $X_{\mathcal{N}}$.) So, in practice, one does not solve directly the linear system above for $U_N(\mu)$ (corresponding to the normal equations of the least-squares problem). But we rather compute the RB approximation as $u_{\mathcal{N},N}(\mu) = \sum_{n=1}^N \tilde{U}_{N,n}(\mu) \zeta_n$ where $(\zeta_n)_{1 \le n \le N}$ denotes a basis of $X_{\mathcal{N},N}$ which is orthonormal for the inner-product $(\cdot,\cdot)_X$ (or any equivalent inner-product $a(\cdot,\cdot;\mu_0)$ with fixed $\mu_0 \in \Lambda$). Simple or Modified Gram-Schmidt procedures may suffice to compute $(\zeta_n)_{1 \le n \le N}$ since N is small here[2].

To compute the RB approximation as $u_{\mathcal{N},N}(\mu) = \sum_{n=1}^N \tilde{U}_{N,n}(\mu) \zeta_n$, one has to solve for each $\mu \in \Lambda$ a small $N \times N$ (full) linear system for the vector $\tilde{U}_N(\mu) = \left(\tilde{U}_{N,n}(\mu)\right)_{1 \le n \le N} \in \mathbb{R}^N$:

$$\text{Find } \tilde{U}_N(\mu) \in \mathbb{R}^N \text{ solution to } \underline{\underline{\tilde{C}}}(\mu) \tilde{U}_N(\mu) = \tilde{c}, \tag{4-I.17}$$

where $\tilde{c} = (l(\zeta_n))_{1 \le n \le N}$ is a vector in $\mathbb{R}^N$ and $\underline{\underline{\tilde{C}}}(\mu) = \underline{\underline{\tilde{C}_0}} + \mu \underline{\underline{\tilde{C}_1}}$ is a $N \times N$ real invertible matrix. So, for each parameter value $\mu \in \Lambda$, the entries of the latter matrix $\underline{\underline{\tilde{C}}}(\mu)$ can be computed in $O(N^2)$ operations (multiplications and additions) using the precomputed integrals $\underline{\underline{\tilde{C}_q}}_{ij} = \int_{\mathcal{D}} \underline{\underline{A_q}} \nabla \zeta_i \cdot \nabla \zeta_j$, $i,j = 1, \ldots, N$ for $q = 0,1$. (Note that the assumption of affine parametrization is essential here). And the evaluation of $\tilde{U}_N(\mu)$ for many $J \gg 1$ parameter values $\mu \in \Lambda$ finally costs $J \times O(N^3)$ operations using exact solvers for symmetric problems like Cholevsky [GvL96].

For each $\mu \in \Lambda$, the output $s_{\mathcal{N},N}(\mu) = F(u_{\mathcal{N},N}(\mu))$ can also be computed very fast in $O(N)$ on noting that F is linear and $F(u_{\mathcal{N},N}(\mu_n))$, $n = 1, \ldots, N$ can be precomputed offline. Then, one would also like to compute fast the corresponding *a posteriori* estimator $\Delta_N^s(\mu)$ given by (4-I.11). Now, on noting the affine dependence of

[2] Gram-Schmidt procedures in floating point arithmetic are famously unstable when N grows [GvL96]. So one may want to use the SVD or the QR decomposition of the matrix $\underline{\underline{D}}$ in some cases. Though, one should keep in mind that the SVD or the QR decompositions will be less suitable to online basis enrichments. If the accuracy level ε is not reached for some $\mu \in \Lambda$ online and one should enrich the basis by computing a new accurate approximation $u_{\mathcal{N}}(\mu)$ for that same μ, then the full SVD or the full QR decomposition of the new $\mathcal{N} \times (N+1)$ matrix $\underline{\underline{D}}$ should be recomputed at a cost $O(\mathcal{N} \times (N+1)^2)$ which is less favourable than the $O(\mathcal{N} \times (N+1))$ operations needed to compute an additional normalized orthogonal direction ζ_{N+1} with Gram-Schmidt.

$\underline{\underline{A}}(\mu)$ on μ, a similar affine dependence $G(\mu)=G_0+\mu G_1$ holds for all $\mu\in\Lambda$, where $G_q : X_\mathcal{N}\to X_\mathcal{N}$, $q=0,1$, are defined as :

$$(G_0\,w,v)_X=\int_\mathcal{D}\underline{\underline{A_0}}\nabla w\cdot\nabla v-\int_\mathcal{D} f\,v \quad\text{ and }\quad (G_1\,w,v)_X=\int_\mathcal{D}\underline{\underline{A_1}}\nabla w\cdot\nabla v,\qquad \forall w,v\in X_\mathcal{N}\,.$$

So one can evaluate fast the following quadratic form for $\mu\in\Lambda$:

$$\|G(\mu)\,u_{\mathcal{N},N}(\mu)\|_X^2=\|G_0\,u_{\mathcal{N},N}(\mu)\|_X^2+2\mu(G_0\,u_{\mathcal{N},N}(\mu),G_1\,u_{\mathcal{N},N}(\mu))_X+\mu^2\|G_1\,u_{\mathcal{N},N}(\mu)\|_X^2 \tag{4-I.18}$$

in $O(N^2)$ operations, by computing, for each 2-uple $i,j=0,1$, the numbers $(G_i\,u_{\mathcal{N},N}(\mu),G_j\,u_{\mathcal{N},N}(\mu))_X$ as bilinear forms in the N coefficients $U_{N,n}(\mu)$, $n=1,\ldots,N$. (The $4N^2$ real numbers $(G_i\,u_{\mathcal{N},N}(\mu_n),G_j\,u_{\mathcal{N},N}(\mu_m))_X$, $i,j=0,1$, $n,m=1,\ldots,N$ can be precomputed offline and stored.) Assuming that the lower-bound $\alpha_{\rm LB}(\mu)$ used in (4-I.11) is known from *a priori* analysis (see also Remark 16), the computation of the *a posteriori* estimator $\Delta_N^s(\mu)$ itself is thus also very fast. Notice that this two-stage *construction-evaluation decomposition* of the computations for the *a posteriori* residual error estimations is possible because we chose a so-called *affine* parametrization of the form (4-I.8).

Finally, the marginal cost of one online-type computation for one parameter value μ is $w_{\rm online}(n)=O(n^3)$, with $n=1,\ldots,N$. So, assuming that no basis enrichment is necessary during the online stage using the RB approximation space $X_{\mathcal{N},N}$ (that is, $\Delta_N^s(\mu)<\varepsilon$ for all the parameter values μ queried online), the total online cost for many $J\gg 1$ parameter values μ scales as $W_{\rm online}=J\times O(N^3)$. And, the total cost of computations with the RB approach is then $W_{\rm offline}+W_{\rm online}=N\times O(\mathcal{N}^k)+(J+O(|\Lambda_{\rm trial}|))\times O(N^3)$, which has to be compared to $J\times O(\mathcal{N}^k)$ operations for a direct approach (with $2<k\leq 3$ depending on the solver used for large sparse linear systems).

4-I-A-e Some elements of analysis of the reduced-basis method

After many successful numerical experiments in various applications [VRP02, VPRP03a, HP07b, PR07b, PR07a, Dep08, NRP09, NRHP09], there is still little theoretical understanding of the RB method. As a matter of fact, there might be little theory to be expected for the RB method in the usual *a priori* way, since, as it has already been explained, the method has rather been designed to *a posteriori* adapt to practical (often complicated, computationally demanding) settings. A minimal understanding of simple paradigmatic examples like (4-I.7) is nevertheless desirable to guide our intuition for the practice of the RB method (in particular as regards the choice of $\Lambda_{\rm trial}$) : let us briefly mention the few known results.

First, by adapting the classical Lagrange interpolation theory to the context of μ-BVP, one can obtain an upper bound of (4-I.5) when $u(\mu)$ is solution to (4-I.7) and the matrix $\underline{\underline{A_1}}$ is non-negative (hence $\Delta_N^s(\mu)=\|u_\mathcal{N}(\mu)-u_{\mathcal{N},N}(\mu)\|_\mu^2\leq(1+\omega_1\mu_{\max})^2\|u_\mathcal{N}(\mu)-u_{\mathcal{N},N}(\mu)\|_0^2$ with a constant $\omega_1>0$) :

Proposition 19. *For all parameter ranges $\Lambda:=[\mu_{\min},\mu_{\max}]\subset\mathbb{R}_{>0}$, there exists $N_0\in\mathbb{N}$ with $N_0=O\left(\ln\left(\frac{\mu_{\max}}{\mu_{\min}}\right)\right)$ as $\frac{\mu_{\max}}{\mu_{\min}}\to+\infty$, and a constant $c>0$ independent of Λ such that, for all $N\geq N_0\geq 2$, there exist N parameter values $\mu_{\min}=:\lambda_1^N<\ldots<\lambda_n^N<\lambda_{n+1}^N<\ldots<\lambda_N^N:=\mu_{\max}$, $n=2,\ldots,N-2$, sastisfying (recall $\|\cdot\|_0=\|\cdot\|_\mu$ with $\mu=0$ is an Hilbertian norm on X) :*

$$\sup_{\mu\in\Lambda}\left(\inf\left\{\|u_\mathcal{N}(\mu)-w\|_0\,,w\in\mathbf{Span}\left(u_\mathcal{N}(\lambda_n^N),n=1,\ldots,N\right)\right\}\right)\leq e^{-\frac{c}{N_0-1}(N-1)}\sup_{\mu\in\Lambda}\|u_\mathcal{N}(\mu)\|_0\,. \tag{4-I.19}$$

The essential tool in the proof of Proposition 19 is a uniform bound, valid for all ranges $\Lambda:=[\mu_{\min},\mu_{\max}]\subset\mathbb{R}_{>0}$, of the $C^0(\ln\Lambda,X)$-norm of the derivatives of $\ln\mu\to u_\mathcal{N}(\mu)$. Owing to the simplicity of the specific problem (4-I.7), this uniform bound can be computed explicitly here (see [MPT02a]). Then, after choosing a log-distributed sample $(\lambda_n^N)_{1\leq n\leq N}$ the proof proceeds by explicitly building an upper bound of (4-I.5), (in fact, of $\inf\left\{\|u_\mathcal{N}(\mu)-w\|_0\,,w\in\mathbf{Span}\left(u_\mathcal{N}(\lambda_n^N),n=1,\ldots,N\right)\right\}$) expressed as the remaining term in a Taylor-Lagrange formula, which can be bounded above thanks to the uniform bound mentioned above to yield (4-I.19) (see [MPT02b, MPT02a, PR07b, BL09a]). Of course, the approximation space $\mathbf{Span}\left(u_\mathcal{N}(\lambda_n^N),n=1,\ldots,N\right)$ used for the proof of Proposition 19 is different from the RB approximation space $X_{\mathcal{N},N}$ built by the RB greedy algorithm. Numerical experiments even suggest that it is a much less good choice than $X_{\mathcal{N},N}$ (see [PR07b]), so one would like to better understand what the RB greedy algorithm does, which we expect to be closer to minimizers of (4-I.5).

Recall that the concept of *greedy* algorithm appears in many numerical approaches for problems of approximation. It typically consists in an iterative procedure approximating an optimal solution to a complex problem by sub-optimal solutions ; the sub-optimal solutions are improved incrementally : each iteration starts with the solution of the previous iteration as an initial guess to be improved. In the theory of approximation of functions in particular [Dev93, V.N08], greedy algorithms are used to incrementally compute the combinations of well-selected functions (from a given dictionnary) which nearly-best approximate some function in a general Hilbert or Banach space. The RB greedy algorithm has a different viewpoint : it incrementally computes for $N \in \mathbb{N}_{>0}$ some basis functions $u_{\mathcal{N}}(\mu_n)$, $n=1,\dots,N$, spanning a linear space $X_{\mathcal{N},N}$ that nearly-best approximates a smooth parametrized family of functions $u_{\mathcal{N}}(\mu)$, $\forall \mu \in \Lambda$. However, if we forget for the moment being the μ-BVP underpinning the approximation problem, then the RB greedy algorithm has a flavour similar to other greedy algorithms that typically build (nearly-)best-approximants in general classes of functions. To do this carefully and next show the convergence of a greedy algorithm very close to our RB greedy algorithm, we introduce the concept of Kolmogorov width : this is one tool among others (Gelf'and width, Kolmogorov entropy, etc.) which allows to characterize "blindly" the set of functions to approximate [Pin85].

Assume one is looking for a linear space of approximation F_N with dimension $\dim(F_N) \leq N$ which is uniformly good for all the functions within a bounded compact set $\mathcal{F}$ of some Hilbert space X. Moreover, we would like to build F_N using snapshots of $\mathcal{F}$ itself :

$$F_N := \mathbf{Span}(f_k,\ k=1,\dots,N)\,,\ \text{with}\ f_1, f_2, \dots, f_N \in \mathcal{F}.$$

This is a simplification of the context for the RB method : there is no μ-BVP associated with elements of $\mathcal{F}$ here. In particular, any function f from the given family $\mathcal{F}$ should now be approximated in F_N using the *same* procedure for all $f \in \mathcal{F}$. Since F_N is a (finite-dimensional, thus closed, and linear, thus convex) subspace of X, we can define the orthogonal projector $P_{F_N} : X \to F_N$. And mimicking the Galerkin orthogonality in the RB greedy algorithm using Galerkin approximations of solutions to a μ-BVP, we choose to approximate any function $f \in \mathcal{F}$ by $P_{F_N} f \in F_N$. An optimal choice for F_N is thus a minimizer of

$$\eta_N(\mathcal{F}) := \min_{f_1, f_2, \dots, f_N \in \mathcal{F}} \max_{f \in \mathcal{F}} \|f - P_{F_N} f\|_X\,,\ \forall N \in \mathbb{N}_{>0}\,, \tag{4-I.20}$$

where for any $f \in \mathcal{F}$, the same norm $\|\cdot\|_X$ is used to evaluate the approximation error $f - P_{F_N} f$ in the ambiant Hilbert space X, as opposed to (4-I.4) in the μ-BVP context. (This is equivalent to forgetting about the μ-dependence of the energy norm.) Note that minimizers and maximizers in (4-I.20) are reached by compactness of the set $\mathcal{F}$, but there is no reason why they should be unique – $\mathcal{F}$ is not necessarily convex –! Note also that, in this Hilbert framework, $\|f - P_{F_N} f\|_X$ coincides with the distance $\mathrm{dist}(f, \mathbf{Span}(f_k,\ k=1,\dots,N))_X$, where $\mathrm{dist}(f,E)_X := \min_{g \in E} \|f-g\|_X \equiv \|f - P_E f\|_X$ is defined for any $f \in X$ and any closed subset $E \subset X$. In practice, mimicking the RB greedy algorithm, we select $f_1, f_2, \dots, f_N \in \mathcal{F}$ incrementally as :

$$f_1 \in \operatorname*{argmax}_{f \in \mathcal{F}} \|f\|_X\,, \qquad f_{k+1} \in \operatorname*{argmax}_{f \in \mathcal{F}} \|f - P_{F_k} f\|_X \quad \forall k \in \mathbb{N}_{>0}\,, \tag{4-I.21}$$

so that F_N is a computable approximation to minimizers of $\eta_N(\mathcal{F})$. The compactness of $\mathcal{F}$ implies :

$$\eta_N(\mathcal{F}) \leq \delta_N(\mathcal{F}) := \|f_{N+1} - P_{F_N} f_{N+1}\|_X \xrightarrow[N \to 0]{} 0\,.$$

Let us denote by $\mathcal{E}_N$ the set of all linear subspaces of the normed linear space X with dimension less than, or equal to N. One way to characterize $\mathcal{F}$ is to give its Kolmogorov N-widths :

$$\epsilon_N(\mathcal{F}) := \min_{E \in \mathcal{E}_N} \max_{f \in \mathcal{F}} \|f - P_E f\|_X\,,\ \forall N \in \mathbb{N}_{>0}\,. \tag{4-I.22}$$

Clearly, $\epsilon_N(\mathcal{F}) \leq \eta_N(\mathcal{F})$, and one can exhibit good choices $f_1, f_2, \dots, f_N \in \mathcal{F}$ such that [CDD] :

Proposition 20. *If $\mathcal{F}$ is a bounded compact subset of the Hilbert space X, then, for all $N \in \mathbb{N}_{>0}$:*

$$\eta_N(\mathcal{F}) \leq (N+1)\epsilon_N(\mathcal{F})\,.$$

The proof of Proposition 20 is reproduced in Appendix 4-V-A. It basically consists in comparing a specific choice $\{f_k,\ k=1,\dots,N\}$ to a minimizer of the Kolmogorov N-width :

$$E_N(\mathcal{F}) \in \operatorname*{argmin}_{E \in \mathcal{E}_N} \max_{f \in \mathcal{F}} \mathrm{dist}(f,E)_X\,.$$

Furthermore, the Kolmogorov N-widths characterize the performance of the greedy algorithm (4-I.21) (as mentioned in [Mad06, PR07b]) :

Proposition 21. *If $\mathcal{F}$ is a bounded compact subset of the Hilbert space X, then N iterations of the greedy algorithm* (4-I.21) *define a linear space $F_N = \mathbf{Span}(f_k,\ k=1,\dots,N)$ such that $\forall f \in \mathcal{F}$:*

$$\|f - P_{F_N} f\|_X \leq \delta_N(\mathcal{F}) \leq \left(2^{N+1}\sqrt{N+1}\right) \epsilon_N(\mathcal{F}). \tag{4-I.23}$$

The proof of Proposition 21 is reproduced in Appendix 4-V-B (see also [MNPP09]).

Unfortunately, Proposition 21 shows the convergence of the greedy algorithm (4-I.21) for those $\mathcal{F}$ with very fast decaying Kolmogorov widths only, such that $\lim_{N\to\infty} \epsilon_N(F)\left(2^{N+1}\sqrt{N+1}\right) = 0$. This decay assumption is not only very difficult to check in practice, it is also very stringent : one could expect a better rate than the pessimistic 2^{N+1}. In particular, when $\mathcal{F} = (u_{\mathcal{N}}(\mu))_{\mu\in\Lambda}$ is given by a simple μ-BVP like (4-I.7), the use of μ-independent projections $P_{X_{\mathcal{N},N}} u_{\mathcal{N}}(\mu)$ instead of the Galerkin solutions $u_{\mathcal{N},N}(\mu)$ is not a strong modification of the RB greedy algorithm. Then, the estimate (4-I.19) can be considered as an upper-bound for the Kolmogorov N-width of $\mathcal{F}$, and the Proposition 21 proves the convergence of the RB greedy algorithm only when the factor $\frac{c}{N_0-1}$ in (4-I.19) is sufficiently large, while the estimate (4-I.23) (possibly multiplied by a constant) remains correct. This is a very stringent condition on the dependence of $u(\mu)$ on μ (with a very stringent impact on the uniform bound which we require in the proof of Proposition 19).

4-I-B The certified reduced-basis method for parametrized elliptic problems

Here we generalize the scope of the RB method introduced in the previous Section 4-I-A to a larger class of elliptic problems. We recall that these generalizations apply to the frameworks with many queries in the input parameter $\mu \in \Lambda$ which have the following features. The output $s(\mu) = F(u(\mu))$ is a linear function of a solution $u(\mu) \in X$ to a μ-BVP with variational formulation in a Hilbert space X :

$$\text{Find } u(\mu) \in X \text{ solution to } g(u(\mu), v; \mu) = 0,\ \forall v \in X, \tag{4-I.24}$$

where $g(\cdot,\cdot;\mu)$ is a μ-parametrized form in $X \times X$, and $u(\mu)$ can be numerically approximated as fine as necessary using Galerkin discretizations $u_{\mathcal{N}}(\mu) \in X_{\mathcal{N}} \subset X$ to reach a required accuracy level $|s(\mu) - s_{\mathcal{N}}(\mu)| < \varepsilon$ in the output. Starting with the paradigmatic problem (4-I.7) where $g(\cdot,\cdot;\mu) = a(\cdot,\cdot;\mu) - F(\cdot)$ with $a(\cdot,\cdot;\mu)$ a coercive symmetric continuous bilinear form in $X \times X$ and $F(\cdot)$ a continuous linear form in X, we briefly review some well-established generalizations of the RB approach by extending (4-I.7) little by little to a larger class of μ-BVPs with outputs. Because the goal-oriented *a posteriori* error estimator (4-I.11) built for the paradigmatic example (4-I.7) was an essential ingredient of the certified RB method deployed in Section 4-I-A, the scope of the generalizations will be essentially limited by the possibility of extending this fast computable *a posteriori* error estimation.

4-I-B-a Affine parameters

The RB method can be straightforwardly generalized to the μ-BVPs whose weak form (4-I.24) has an *affine* parametrization using a P-dimensional parameter $\mu = (\mu_p)_{1\leq p\leq P} \in \Lambda \subset \mathbb{R}^P$ and Q smooth functions of the parameter $\Theta_q : \Lambda \to \mathbb{R}$, $q = 1,\dots,Q$:

Definition 2. *The form $g(\cdot,\cdot;\mu)$ on $X \times X$ is said to be with affine parametrization when it is affine in functions $(\Theta_q)_{1\leq q\leq Q}$ of the parameter $\mu \in \Lambda$ and writes as the sum of parameter-independent forms $(g_q(\cdot,\cdot))_{1\leq q\leq Q}$ on $X \times X$ multiplied by coefficients $(\Theta_q(\mu))_{1\leq q\leq Q}$:*

$$g(w, v; \mu) = \sum_{q=1}^{Q} \Theta_q(\mu) g_q(w, v), \quad \forall w, v \in X,\ \forall \mu \in \Lambda. \tag{4-I.25}$$

In particular, it is easy to see that the weak form (4-I.1) of the paradigmatic μ-BVP (4-I.7) has an affine parametrization when the form $a(w, v; \mu) = \int_{\mathcal{D}} \underline{\underline{A}}(\mu) \nabla w \cdot \nabla v$ is defined $\forall w, v \in X$ with a matrix :

$$\underline{\underline{A}}(\mu) = \sum_{q=1}^{Q} \Theta_q(\mu) \underline{\underline{A_q}},\ \forall \mu \in \Lambda, \tag{4-I.26}$$

using Q matrices $\underline{\underline{A_q}}$, $q=1,\ldots,Q$. Then, $\forall \mu \in \Lambda$, a construction-evaluation decomposition of

$$\|G(\mu)\, u_{\mathcal{N},N}(\mu)\|_X^2 = \sum_{q,q'=1}^{Q} \Theta_q(\mu)\Theta_{q'}(\mu)(G_q\, u_{\mathcal{N},N}(\mu), G_{q'}\, u_{\mathcal{N},N}(\mu))_X \tag{4-I.27}$$

can be used to compute fast the *a posteriori* error estimator (4-I.11). Another straightforward extension of (4-I.7) is an affine parametrization of the boundary operators, see Section 4-II.

Note yet that the affine parameter (4-I.26), either in a domain or in a boundary differential operator, has the following important restriction : to each scalar function Θ_q of the parameter ($q=1,\ldots,Q$) should correspond a fixed subdomain of $\mathcal{D}$ or $\partial\mathcal{D}$, independent of the parameter μ, such that $\Theta_q(\mu)$ is a constant coefficient in the PDE or the BC satisfied by $u(\mu)$ on that subdomain whatever $\mu \in \Lambda$. Fortunately, the RB approach can also be extended to some cases where $\Theta_q(\mu)$ acts on a subdomain that depends on μ. Clearly, there are many-query frameworks relevant for the reduction of computations which deal with parametrized geometries like the numerical optimization of parametrized shapes. In particular, this is possible when the variations of the geometry can be described after introducing a few additional *geometric* parameters. Then, one rewrites the parametrized problem : it is mapped with piecewise affine functions from a reference configuration such that only affine parameters enter the new parametrized problem defined on the reference configuration, see [Boy08, Qua09] and Section 4-I-C.

Remark 15 (Curse of high dimensions in the limit $Q \gg 1$)**.** *Of course, difficulties are expected to occur when the dimension of the parametrized manifold $\mathcal{M} = \{u(\mu), \mu \in \Lambda\}$ approximated by the linear space $X_{\mathcal{N},N}$ grows, thus when the number of parameters $Q \gg 1$ is large in particular. The theory of approximation [Dev93, V.N08] for the approximation of general sets of functions using N-dimensional linear spaces indeed forecasts that N should grow exponentially with the dimension of the space ($\simeq Q$) if one wants to approximate all the functions within this general set at a constant level of accuracy. Yet, numerical experiments have shown that the RB approach still yields interesting computational gains in examples using up to twenty $Q \simeq 20$ parameters [Sen08, BBM+09].*

4-I-B-b Non-coercive symmetric linear elliptic problems

The RB approach can be extended to the case where the symmetric continuous bilinear form $a(\cdot,\cdot;\mu)$ on $X \times X$ is not coercive but only *inf-sup stable* :

$$0 < \beta(\mu) := \inf_{w \in X\setminus\{0\}} \sup_{v \in X\setminus\{0\}} \frac{a(w,v;\mu)}{\|w\|_X \|v\|_X}, \ \forall \mu \in \Lambda, \tag{4-I.28}$$

with an inf-sup stable discretization on $X_{\mathcal{N}} \times X_{\mathcal{N}}$:

$$0 < \beta^{\mathcal{N}}(\mu) := \inf_{w \in X_{\mathcal{N}}\setminus\{0\}} \sup_{v \in X_{\mathcal{N}}\setminus\{0\}} \frac{a(w,v;\mu)}{\|w\|_X \|v\|_X}, \ \forall \mu \in \Lambda, \tag{4-I.29}$$

see *e.g.* the Helmholtz problem treated in [SVH+06]. Then introducing a computable lower bound $\beta_{LB}^{\mathcal{N}}(\mu) \leq \beta^{\mathcal{N}}(\mu)$, the *a posteriori* analysis of Section 4-I-A-b is extended to the non-coercive case :

Proposition 22. *For any linear subspace $X_{\mathcal{N},N}$ of $X_{\mathcal{N}}$, there exists a computable error bound $\Delta_N^s(\mu)$ such that :*

$$|s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)| \leq \Delta_N^s(\mu) := \frac{\|G(\mu)\, u_{\mathcal{N},N}(\mu)\|_X^2}{\beta_{LB}^{\mathcal{N}}(\mu)}, \ \forall \mu \in \Lambda. \tag{4-I.30}$$

Proof. On noting the linearity of F and the fact that $u_{\mathcal{N}}(\mu)$ is solution to a variational problem in $X_{\mathcal{N}} \ni u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)$, the output RB approximation error still reads :

$$\begin{aligned} |s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)| &= |F(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu))| \\ &= |a(u_{\mathcal{N}}(\mu), u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu); \mu)|. \end{aligned} \tag{4-I.31}$$

We conclude the proof using the symmetry of $a(\cdot,\cdot;\mu)$ with the bound :

$$\|u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu)\|_X \beta^{\mathcal{N}}(\mu) \leq \sup_{v \in X_{\mathcal{N}}} a(u_{\mathcal{N}}(\mu) - u_{\mathcal{N},N}(\mu), v; \mu),$$

on noting that :

$$a(u_{\mathcal{N},N}(\mu) - u_{\mathcal{N}}(\mu), v; \mu) = g(u_{\mathcal{N},N}(\mu), v; \mu), \forall v \in X_{\mathcal{N}}.$$

□

Then, the only change in offline and online computations is $\beta_{LB}^{\mathcal{N}}(\mu)$ substituting for $\alpha_{LB}(\mu)$.

Remark 16 (Effectivity of estimations using computable lower bounds $\alpha_{LB}(\mu)$ and $\beta_{LB}^{\mathcal{N}}(\mu)$)**.** *To compute the error estimator $\Delta_N^s(\mu)$ in* (4-I.11) *and* (4-I.30)*, we only need lower bounds $\alpha_{LB}(\mu)$ and $\beta_{LB}^{\mathcal{N}}(\mu)$ respectively for the constants $\alpha_c(\mu)$ and $\beta^{\mathcal{N}}(\mu)$. But of course, the better the lower bounds are, the better the reduced basis $X_{\mathcal{N},N}$ is expected to be : in practice, one expects the effectivity of the estimator*

$$\eta(\mu) = \frac{\Delta_N^s(\mu)}{|s_{\mathcal{N}}(\mu) - s_{\mathcal{N},N}(\mu)|}$$

to be uniformly bounded in Λ. Now, one may know lower bounds $\alpha_{LB}(\mu)$ and $\beta_{LB}^{\mathcal{N}}(\mu)$ from a priori analysis, but the latter may either degenerate or their effectivity grow too much in some regions of Λ, which could result in an inadequately chosen reduced basis. On the other hand, computing sharp lower approximations of $\alpha_c(\mu)$ and $\beta^{\mathcal{N}}(\mu)$ for all $\mu \in \Lambda$, after discretization in $X_{\mathcal{N}}$, is computationally expensive. It corresponds to the computation of, the smallest (positive) eigenvalue of the matrix $\underline{\underline{B}}(\mu)$ for $\alpha_c(\mu)$ and, the squareroot of the smallest (positive) singular value of $\underline{\underline{B}}(\mu)\underline{\underline{B}}(\mu)^T$ for $\beta^{\mathcal{N}}(\mu)$. That is why part of the current research deals with designing lower bounds that are both sharp and fast to compute, see [HRSP07a, HRSP07b, HKY+09].

4-I-B-c Non-compliant linear elliptic problems

In (4-I.9), the particular choices of $F = l$ for the output and of symmetric matrices $\underline{\underline{A}}(\mu)$ for the definition of the (thus symmetric) bilinear form $a(\cdot,\cdot;\mu)$ correspond to a particular type of BVP with output (among different classes of goal-oriented problems), termed *compliant*. It is nevertheless possible to treat simply non-compliant linear elliptic problems, that is the cases where, for some $\mu \in \Lambda$ at least, either $u(\mu)$ is solution to a weak form (4-I.24) with $g(v,w;\mu) = a(v,w;\mu) - l(w)$, $\forall v,w \in X$ and the bilinear form $a(\cdot,\cdot;\mu)$ is not symmetric, or the output is $s(\mu) = F(u(\mu)) \neq l(u(\mu))$ with any linear continuous function $F : X \to \mathbb{R}$.

When the bilinear form $a(\cdot,\cdot;\mu)$ is not symmetric, we shall assume that the problem is not necessarily coercive, but the bilinear form $a(\cdot,\cdot;\mu)$ on $X \times X$ is continuous and inf-sup stable, see (4-I.28). For all $\mu \in \Lambda$, we also assume the dual inf-sup stability condition :

$$0 < \beta^\star(\mu) = \inf_{w \in X\setminus\{0\}} \sup_{v \in X\setminus\{0\}} \frac{a(v,w;\mu)}{\|w\|_X \|v\|_X}, \ \forall \mu \in \Lambda, \tag{4-I.32}$$

and the dual inf-sup stability condition after discretization :

$$0 < \beta^{\mathcal{N},\star}(\mu) \inf_{w \in X_{\mathcal{N}}\setminus\{0\}} \sup_{v \in X_{\mathcal{N}}\setminus\{0\}} \frac{a(v,w;\mu)}{\|w\|_X \|v\|_X}, \ \forall \mu \in \Lambda. \tag{4-I.33}$$

We then define a computable lower bound $\beta_{LB}^{\mathcal{N},\star}(\mu) \leq \beta^{\mathcal{N},\star}(\mu)$, given by the *a priori* analysis or computed using efficient procedures[3] (see Remark 16). Then we define $\psi(\mu) \in X$ as the solution to the well-posed adjoint problem[4] :

$$\text{Find } \psi(\mu) \in X \text{ solution to } a(v,\psi(\mu);\mu) = -F(v), \ \forall v \in X, \tag{4-I.34}$$

[3] In fact, notice that after discretizing X into $X_{\mathcal{N}}$ with the Galerkin method, the singular values of $\underline{\underline{B}}(\mu)\underline{\underline{B}}(\mu)^T$ are equal to the singular values of $\underline{\underline{B}}(\mu)^T\underline{\underline{B}}(\mu)$, hence one can use the same lower bound for $\beta^{\mathcal{N},\star}(\mu)$ and $\beta^{\mathcal{N}}(\mu)$. In addition, the stability condition (4-I.33) is in fact strictly equivalent to the stability condition (4-I.29), see *e.g.* [PR07b]. Besides, the generalization of the RB method to Petrov-Galerkin discretizations is not obvious.

[4] Assuming (4-I.28) suffices for the well-posedness of (4-I.24), and assuming (4-I.32) suffices for the well-posedness of the adjoint problem (4-I.34). So these assumptions naturally also cover the paradigmatic symmetric problem, but are not generalization when the bilinear form is symmetric. Then, any inf-sup constant, (4-I.28) or (4-I.32), indeed reduces to the coercivity assumption for $a(\cdot,\cdot;\mu)$. Let us recall a simple proof of this. By symmetry, the linear Fréchet differential operator D_w applied to the functional $J^\mu : w \in X \to J^\mu(w) = \|w\|_\mu^2 \in \mathbb{R}_{\geq 0}$ at any point $w \in X$ indeed satisfies :

$$D_w J^\mu : v \in X \to D_w J^\mu v = a(w,v;\mu) + a(v,w;\mu) = 2\,a(w,v;\mu), \forall \mu \in \Lambda.$$

Now, the functional J^μ is continuous, thus reaches its minimum $\alpha_c(\mu)$ on the sphere $\{w \in X, \|w\|_X = 1\}$: let us call w_0 a minimizer on the unit sphere. At w_0, the sphere is tangent to the hyperplane $\{v \in X,\ (w_0,v)_X = 0\}$ at w_0. By the Lagrange theorem on the minimization of convex functionals under equality constraints, the latter tangent hyperplane is thus necessarily parallel to the hyperplane $\{v \in X,\ D_{w_0} J^\mu v = 2\,a(w_0,v;\mu) = 0\}$. So w_0 satisfies a generalized eigenvalue problem for some real eigenvalue $\alpha \neq 0$: $a(w_0,v;\mu) = \alpha\,(w_0,v)_{\boldsymbol{X}}$, $\forall v \in X$; in particular, when $v = w_0$, $\alpha = J^\mu(w_0) = \alpha_c(\mu)$. This implies $\sup_{\|v\|=1} \|a(w_0,v;\mu)\|_X = \alpha_c(\mu)\|w_0\|_X$, thus $\alpha_c(\mu) = \inf_{\|w_0\|_X=1} \sup_{\|v\|_X=1} \|a(w_0,v;\mu)\|_X = \beta(\mu)$.

and $\psi_{\mathcal{N}}(\mu)\in X_{\mathcal{N}}$ as the corresponding Galerkin discretizations. Finally, we deploy the RB method after introducing an additional RB approximation space $X^{\star}_{\mathcal{N},N^{\star}}\subset X_{\mathcal{N}}$ of dimension $N^{\star}\ll\mathcal{N}$ for the adjoint problem (4-I.34). Then, the output RB approximation is computed as $s_{\mathcal{N},N,N^{\star}}(\mu)=F(u_{\mathcal{N},N}(\mu))+g(u_{\mathcal{N},N}(\mu),\psi_{\mathcal{N},N^{\star}}(\mu);\mu)$, where $\psi_{\mathcal{N},N^{\star}}(\mu)\in X^{\star}_{\mathcal{N},N^{\star}}$ is solution to :

$$\text{Find } \psi_{\mathcal{N},N^{\star}}(\mu)\in X^{\star}_{\mathcal{N},N^{\star}} \text{ solution to } a(v,\psi_{\mathcal{N},N^{\star}}(\mu);\mu)=-F(v),\ \forall v\in X^{\star}_{\mathcal{N},N^{\star}}. \tag{4-I.35}$$

(Clearly $g(u_{\mathcal{N}}(\mu),\psi_{\mathcal{N}}(\mu);\mu)=0$ so the accurate approximation for the output $s(\mu)$ reads $s_{\mathcal{N}}(\mu)=F(u_{\mathcal{N}}(\mu))+g(u_{\mathcal{N}}(\mu),\psi_{\mathcal{N}}(\mu);\mu)$.) Last, defining the dual residual form $g^{\star}(\cdot,\cdot;\mu)$ as

$$g^{\star}(w,v;\mu)=a(w,v;\mu)+F(w),\ \forall w,v\in X,\ \forall\mu\in\Lambda,$$

and the computable linear operator $G^{\star}:X_{\mathcal{N}}\to X_{\mathcal{N}}$ such that

$$g^{\star}(w,v;\mu)=\left(w,G^{\star}(\mu)\,v\right)_X,\ \forall w,v\in X_{\mathcal{N}},\ \forall\mu\in\Lambda,$$

the analysis of Section 4-I-A-b can be simply transfered to the non-compliant case in :

Proposition 23. *For any linear subspaces $X_{\mathcal{N},N}$ and $X^{\star}_{\mathcal{N},N^{\star}}$ of $X_{\mathcal{N}}$, there exists a computable error bound $\Delta^{s}_{N,N^{\star}}(\mu)$ such that :*

$$|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N,N^{\star}}(\mu)|\le\Delta^{s}_{N,N^{\star}}(\mu):=\frac{\|G(\mu)\,u_{\mathcal{N},N}(\mu)\|_X\,\|G^{\star}(\mu)\,\psi_{\mathcal{N},N^{\star}}(\mu)\|_X}{\beta^{\mathcal{N},\star}_{LB}(\mu)},\ \forall\mu\in\Lambda. \tag{4-I.36}$$

Proof. On noting the linearity of F and the fact that $\psi_{\mathcal{N}}(\mu)$ is solution to a variational problem in $X_{\mathcal{N}}\ni u_{\mathcal{N}}(\mu)-u_{\mathcal{N},N}(\mu)$, the output RB approximation error reads :

$$\begin{aligned}|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N,N^{\star}}(\mu)|&=|F(u_{\mathcal{N}}(\mu)-u_{\mathcal{N},N}(\mu))-g(u_{\mathcal{N},N}(\mu),\psi_{\mathcal{N},N^{\star}}(\mu);\mu)|\\&=|g(u_{\mathcal{N},N}(\mu),\psi_{\mathcal{N}}(\mu)-\psi_{\mathcal{N},N^{\star}}(\mu);\mu)|.\end{aligned} \tag{4-I.37}$$

We conclude the proof with a bound for $\|\psi_{\mathcal{N}}(\mu)-\psi_{\mathcal{N},N^{\star}}(\mu)\|_X$ combining (4-I.32) with the fact that :

$$a(v,\psi_{\mathcal{N},N^{\star}}(\mu)-\psi_{\mathcal{N}}(\mu);\mu)=a(v,\psi_{\mathcal{N},N^{\star}}(\mu);\mu)+F(v)=\left(v,G^{\star}(\mu)\,\psi_{\mathcal{N},N^{\star}}(\mu)\right)_X,\forall v\in X_{\mathcal{N}}.$$

□

So the RB method still applies for the non-compliant problems. Note that the compliant problems are particular non-compliant cases where $\psi(\mu)=-u(\mu)$. However, compared to Section 4-I-A, the two stages are slightly modified. First, in the offline stage, one has to build two RB approximation spaces $X_{\mathcal{N},N}$ and $X^{\star}_{\mathcal{N},N^{\star}}$. If we use the error estimator (4-I.36) in the RB greedy algorithm, then one has to choose the same parameters μ_n for $X_{\mathcal{N},N}=\mathbf{Span}(u_{\mathcal{N}}(\mu_n),n=1,\dots,N)$ and $X^{\star}_{\mathcal{N},N^{\star}}=\mathbf{Span}(\psi_{\mathcal{N}}(\mu_n),n=1,\dots,N)$. In particular, $X_{\mathcal{N},N}$ and $X^{\star}_{\mathcal{N},N^{\star}}$ will thus have same dimension $N=N^{\star}$ in the non-degenerate cases. But in view of (4-I.36), one can also use a slightly different version of the RB greedy algorithm of Section 4-I-A (with required accuracy ε) :

$N=1$, choose $\mu_1\in\Lambda$ randomly, or take $\mu_1\in\text{argmax}\{|s(\mu)|,\ \mu\in\Lambda_{\text{smalltrial}}\}$
compute $u_{\mathcal{N}}(\mu_1)$ to define $X_{\mathcal{N},1}=\mathbf{Span}(u_{\mathcal{N}}(\mu_1))$
compute $\psi_{\mathcal{N}}(\mu_1)$ to define $X^{\star}_{\mathcal{N},1}=\mathbf{Span}(\psi_{\mathcal{N}}(\mu_1))$
while $\Delta^{s}_{N,N^{\star}}\ge\epsilon$ **do**
  if $\max\{\|G(\mu)\,u_{\mathcal{N},N}(\mu)\|_X/\sqrt{\beta^{\mathcal{N},\star}_{LB}(\mu)},\mu\in\Lambda_{\text{trial}}\}\ge\sqrt{\epsilon}$ **then**
    $N=N+1$
    choose $\mu_N\in\text{argmax}\{\|G(\mu)\,u_{\mathcal{N},N}(\mu)\|_X/\sqrt{\beta^{\mathcal{N},\star}_{LB}(\mu)},\mu\in\Lambda_{\text{trial}}\}$
    compute $u_{\mathcal{N}}(\mu_N)$ to define $X_{\mathcal{N},N}=\mathbf{Span}(u_{\mathcal{N}}(\mu_n),\ n=1,\dots,N)$
  end if
  if $\max\{\|G^{\star}(\mu)\,\psi_{\mathcal{N},N^{\star}}(\mu)\|_X/\sqrt{\beta^{\mathcal{N},\star}_{LB}(\mu)},\mu\in\Lambda_{\text{trial}}\}\ge\sqrt{\epsilon}$ **then**
    $N^{\star}=N^{\star}+1$
    choose $\mu^{\star}_{N^{\star}}\in\text{argmax}\{\|G^{\star}(\mu)\,\psi_{\mathcal{N},N^{\star}}(\mu)\|_X/\sqrt{\beta^{\mathcal{N},\star}_{LB}(\mu)},\mu\in\Lambda_{\text{trial}}\}$
    compute $\psi_{\mathcal{N}}(\mu^{\star}_{N^{\star}})$ to define $X_{\mathcal{N},N^{\star}}=\mathbf{Span}(\psi_{\mathcal{N}}(\mu^{\star}_n),\ n=1,\dots,N^{\star})$
  end if

15: **end while**

which balances the selection procedure between the two source of errors. Second, since the adjoint problem (4-I.34) used for the error estimation (4-I.36) has the same structure in parameter μ than the problem (4-I.24), a fast construction-evaluation decomposition is possible for the online-type computations of $\|G^\star(\mu)\,\psi_{\mathcal{N},N^\star}(\mu)\|_X$ provided it is possible for $\|G(\mu)\,u_{\mathcal{N},N}(\mu)\|_X$. Finally, the amount of computations offline and online is doubled : this is the price to pay to get the "square effect" in the error estimation (4-I.36), hence faster convergence with N.

4-I-B-d Non-affine parameters

When coefficients in the PDE or the BC of the μ-BVP are parametrized through transcendental functions of the parameter μ and the variable x, like $e^{\mu x}$ or $\cos(\mu x)$, the difficulty is to compute fast the Galerkin projections $u_{\mathcal{N},N}(\mu)$, the outputs $s_{\mathcal{N},N}(\mu)$ and the *a posteriori* estimators $\Delta^s_N(\mu)$ for each new parameter value $\mu\in\Lambda$. As opposed to the μ-BVP with affine parameters, which have a construction-evaluation decomposition, one cannot use precomputations with non-affine parameters in order to fast assemble the linear system (4-I.17) at each new value μ.

The RB method naturally suggests to approximate the μ-BVP using non-affine parameters by a different μ-BVP using affine parameters like (4-I.25). For instance, if the coefficients in the PDE or the BC of the μ-BVP write with *analytic* transcendental functions of x and μ like $e^{\mu x}$ and $\cos(\mu x)$, we could use truncated versions of the series $\sum_{k=0}^\infty \frac{(\mu x)^k}{k!}$ or $\sum_{k=0}^\infty (-1)^k \frac{(\mu x)^{2k}}{(2k)!}$ respectively. This truncation procedure yields affine approximations for non-affine parameters. Then, one can deploy the standard RB method for the modified μ-BVP with affine parametrization (and thus with a fast construction-evaluation decomposition).

But as a matter of fact, there exist many ways of computing affine approximations to non-affine parameters. Recall in particular the case in Section 4-I-B-a of non-affine piecewise constant coefficients with a few discontinuities only, which can often be rewritten as a new *exactly* equivalent problem with affine parametrization, at the price of using a few additional (affine) *geometric* parameters (see also Section 4-I-C). Of course, one needs to know in advance where the discontinuities are to do this. In Section 4-II, we will use another dedicated manner of computing an affine approximation for a non-affine parameter, which is particularly well-adapted for a specific class of parameters : the case of coefficients in a BVP that are square-integrable stochastic fields. Again, this requires some knowledge about the nature of the coefficient (and about mathematical theory related to stochastic fields). In this section, we would like to present a general practical method for rewriting a problem with non-affine parametrization into another close one with affine parametrization. Of course, this general method also has its limitations. In particular, it will probably never be optimal compared to other choices dedicated to the specific nature of the non-affine parameter. So this is rather to be considered as a remedy to better choices in absence of knowledge about the specific nature of the non-affine parameter. In any case, a few questions arise when we are confronted to non-affine parameters, and we recommend to investigate these questions carefully before embarking on modifying a parametrized problem :

- how difficult is it to build an affine approximation ?
- can we evaluate the new approximation error on the μ-BVP ?
- is the new μ-BVP still well-posed ?
- can we take into account the new approximation error in the subsequent RB steps ?

The computation of good approximations $\sum_{m=1}^M \beta_m^M(\mu) q_m(x)$ for scalar functions in a set $\{x\in\mathcal{D}\to g(x;\mu)\in\mathbb{R}, \mu\in\Lambda\}$ is a general problem of approximation, very close to the one investigated in the Proposition 21 of Section 4-I-A-e for functions in a bounded compact subset of a Hilbert functional space. Except that here, the functions $g(\cdot;\mu)$ are coefficients in a PDE or a BC, and thus typically belong to the Banach space $L^\infty(\mathcal{D})$. Hence, it was noted in [BNMP04] that a $L^\infty(\mathcal{D})$-modification of the greedy algorithm (4-I.21) is a good candidate for the fast computation of so-called *collateral RB* approximations of order M to $g(\cdot;\mu)$ as $\mathcal{I}_M[g(\cdot;\mu)] := \sum_{m=1}^M \tilde\beta_m^M(\mu) g(\cdot;\mu_m^g)$. For well selected sets $(\mu_m^g)_{m=1,\dots,M}$, the coefficients $\left(\tilde\beta_m^M(\mu)\right)_{m=1,\dots,M}$ can indeed be computed fast by interpolation at the *magic points* $(x_m)_{m=1,\dots,M}$ of the set $(g(\cdot;\mu_m^g))_{m=1,\dots,M}$ (given in this order)

$$x_1\in\operatorname*{argmax}_{x\in\mathcal{D}}|g(\cdot;\mu_1^g)|, \qquad x_m\in\operatorname*{argmax}_{x\in\mathcal{D}}|g(\cdot;\mu_m^g)-\mathcal{I}_{m-1}[g(\cdot;\mu_m^g)]| \qquad \forall m=2,\dots,M.$$

In [BNMP04, GMNP07, MNPP09], one uses a greedy algorithm combined with the *empirical interpolation* described above to incrementally select parameters $(\mu_m^g)_{m=1,\dots,M}$ close to minimizers of

$$\inf_{(\mu_m^g)_{m=1,\dots,M}\in\Lambda^M} \sup_{\mu\in\Lambda} \|g(\cdot;\mu)-\mathcal{I}_M[g(\cdot;\mu)]\|_{L^\infty(\mathcal{D})}. \tag{4-I.38}$$

The Lebesgue constant of these approximations $\mathcal{I}_M[g(\cdot;\mu)]$ is numerically observed to behave well with M, for different functions $g(\cdot;\mu)\in L^\infty(\mathcal{D})$ with smooth dependence [5] on $\mu\in\Lambda$ and various domains $\mathcal{D}$. Unfortunately, the theory is still largely not understood [6]. It is also observed that the linear system for $\left(\tilde{\beta}_m^M(\mu)\right)_{m=1,\ldots,M}$ is ill-conditioned, just like (4-I.17) was. So in practice, one computes the collateral RB approximations as :

$$\mathcal{I}_M[g(x;\mu)]=\sum_{m=1}^{M}\beta_m^M(\mu)q_m(x),\ \forall x\in\mathcal{D},\ \forall\mu\in\Lambda, \tag{4-I.39}$$

with $q_m(x)=\frac{g(x;\mu_m^g)-\mathcal{I}_{m-1}[g(x;\mu_m^g)}{\|g(\cdot;\mu_m^g)-\mathcal{I}_{m-1}[g(\cdot;\mu_m^g)]\|_{L^\infty(\mathcal{D})}}$, $\forall x\in\mathcal{D}$, $m=1,\ldots,M$, at a computational cost of $O(M^2)$ operations for each parameter value $\mu\in\Lambda$.

Most often, the new μ-BVP using $\mathcal{I}_M[g(\cdot;\mu)]$ instead of $g(\cdot;\mu)$ as coefficient will still be well-posed if $\|g(\cdot;\mu)-\mathcal{I}_M[g(\cdot;\mu)]\|_{L^\infty(\mathcal{D})}$ is sufficiently small (thus M sufficiently large). But there is neither a theory that predicts how large M should be for this, nor rigorous *a posteriori* estimators for this interpolation error. In practice, one can check this for instance by estimating $\|g(\cdot;\mu)-\mathcal{I}_M[g(\cdot;\mu)]\|_{L^\infty(\mathcal{D})}$ through $|g(x_{M+1};\mu)-\mathcal{I}_M[g(\cdot;\mu)](x_{M+1})|$, where x_{M+1} is the $(M+1)$-th magic point computed for a trial sample of parameter values with the greedy algorithm approximating minimizers of (4-I.38). Moreover, $|g(x_{M+1};\mu)-\mathcal{I}_M[g(\cdot;\mu)](x_{M+1})|$ can often be integrated in *a posteriori* estimators taking into account both the collateral RB approximation error and the standard RB approximation error in the modified μ-BVP with fixed collateral RB approximations of order M. It is then expected (and numerically observed) that the error due to the collateral RB approximations dominates the error in the standard RB approximation error of the modified μ-BVP at a fixed order M when $N\geq N_{\text{crit}}(M)$ is beyond a critical value $N_{\text{crit}}(M)$ increasing with M.

4-I-B-e Non-linear elliptic problems

Recall that it is essential to the RB method that there exist (sharp and fast) goal-oriented *a posteriori* error estimators $\Delta_N^s(\mu)$ for the output RB approximation error $|s_{\mathcal{N}}(\mu)-s_{\mathcal{N},N}(\mu)|$, typically computed from global *a posteriori* error estimator $\Delta_N(\mu)$ for $\|u_{\mathcal{N}}(\mu)-u_{\mathcal{N},N}(\mu)\|_X$ using a technique with residus. To our knowledge, there is no general theory for such computable goal-oriented *a posteriori* error estimators using residus. The development of (possibly goal-oriented) *a posteriori* error estimators has been a very active field of research for a decade, in particular to design adaptive strategies of computations like mesh refinement techniques in the FE method and many *a posteriori* error estimators are now available for linear problems. The situation is more complicated for nonlinear problems. We refer to [AO00, BR01] and various successful applications of the RB method implemented in the past [VRP02, VPRP03a, HP07b, Dep08, PR07a, NRHP09, NRP09] for tracks about how to further extend the residual technique of Section 4-I-A-b in order to design efficient *a posteriori* error estimators for specific application of the RB method to a non-linear problem.

Moreover, once sharp *a posteriori* estimators have been designed for a nonlinear problem, they should still be computed fast, just like the RB approximations of the solutions to μ-BVP and their outputs. To achieve this, one can for instance invoke the empirical interpolation procedure introduced in Section 4-I-B-d, which has indeed proved numerically performant in low-dimensional nonlinear problems [GMNP07]. (Note that in practice, the solutions $u_{\mathcal{N}}(\mu)$ to nonlinear problems are typically solved by fixed-point procedures applied to a sequence of linearized versions of the nonlinear problem. The empirical interpolation procedure invoked for nonlinear problems is different in nature to that linearization process. It is applied to the functions nonlinear in the solution, and finally yields a parametrized problem with affine parametrization which is still nonlinear in the solution.)

4-I-B-f Semi-discretized parabolic problems

After discretization of the time variations, parabolic problems can be viewed as a collection of elliptic problems parametrized by the time variable. The standard RB method can thus be applied to the collection of parametrized elliptic problems where the time is an additional parameter for the semi-discretized parametrized parabolic problems. The numerical results are satisfactory for parabolic problems with time-independent

[5] We recall that when discontinuities of $g(\cdot;\mu)$ could be clearly identified as functions of μ, one should rather rewrite the μ-BVP with additional *geometric* parameters for the position of the discontinuities, see Sections 4-I-B-a and 4-I-C for instance. We are thus not interested in non-smooth functions here.

[6] One should except here a version of the Proposition 21 of Section 4-I-A-e where the Hilbert space X and the projection operator are replaced with $X=L^\infty(\mathcal{D})$ and the empirical interpolation at magic points, see [MNPP09].

coefficients [GP05], even in some nonlinear cases [GMNP07]. Yet, the nature of the regularity in time is different from that in μ, and a different selection procedure has also been proposed recently to take this into account [HO08, NRHP09].

4-I-C Numerical performance in a benchmark testcase

As explained in the previous Section 4-I-A-e, the theory for the RB method is not well-established yet. But practice has already shown good performance of the RB approach in many different problems [VRP02, VPRP03a, HP07b, Dep08, PR07a, NRHP09, NRP09]. Here we illustrate the standard performance expected for elliptic problems through the numerical results obtained in [Boy08] for a prototypical multiscale problem (the numerical homogenization of second-order linear elliptic scalar PDEs with fast-oscillating coefficients).

In the subsequent Sections 4-II and 4-III, we will show the succesful application of the RB ideas extended and adapted to new many-query frameworks :

- in Section 4-II, we show a simple extension of the RB method presented so far to treat problems with stochastic parameters and various statistical quantities as outputs (essentially, only the *a posteriori* error estimation is new there), and
- in Section 4-III, we show a new RB framework for parametrized Stochastic (ordinary) Differential Equations (SDEs) in the Itô sense, where the strategy for computational reductions focuses on the variance reduction of Monte-Carlo estimations.

4-I-C-a Definition of the problem

Let $Y=[0,1]^2$ be the unit square in $\mathbb{R}^2$. Denoting $(e_i)_{i=1,2}$ the canonical basis of $\mathbb{R}^2$, we want to compute the periodic solutions $u_i(x,\cdot)\in X$ ($X=H^1_{\text{periodic}}(Y)$) to the symmetric scalar linear elliptic problems :

$$-\text{div}_y(\underline{\underline{A}}(x,y)\cdot[e_i+\nabla_y u_i(x,y)])=0,\quad \forall y\in Y,\ i=1,2, \tag{4-I.40}$$

parametrized by $x\in\mathcal{D}$ through :

$$\underline{\underline{A}}(x,y)=\theta(x)\mathbf{1}_{\mathbf{Q(x)}}(y)\underline{\underline{I_2}},\quad \forall y\in Y, \tag{4-I.41}$$

where $\underline{\underline{I_2}}$ is the 2×2 identity matrix, $\mathbf{1}_{\mathbf{Q(x)}}$ is a test function which is one if $y\in Q(x)=\{(y_1,y_2)\in[0,1]^2|0<b_i(x)\le y_i\le c_i(x)<1, i=1,2\}$ and zero otherwise, and $\theta(x)$ is a scalar intensity factor varying in a range ensuring coercivity. The two constants δ and θ^0 being fixed respectively in $]0;.25[$ and $]0;1[$, it is clear that the five-dimensional parametrization :

$$x\in\mathcal{D}\to\mu(x):=(b_1(x),c_1(x),b_2(x),c_2(x),\theta(x))\in\Lambda:=[.25-\delta;.25+\delta]^2\times[.75-\delta;.75+\delta]^2\times[\theta^0;1],$$

affects the geometry. So we have to choose the parameter carefully in order for the PDE (4-I.40) to yield a variational formulation with affine parameter (recall Section 4-I-B-d). More precisely, we would like to map the equation (4-I.40) on $Y\subset\bigcup_{k=1}^K\overline{Y_k(x)}$ using K nonoverlapping open subsets $Y_k(x)$ from a reference partition $\bigcup_{k=1}^K\overline{Y_k(x_0)}$, $x_0\in\mathcal{D}$, with piecewise affine functions such that $\forall x\in\mathcal{D}$:

- the cell Y can be partitioned into K nonoverlapping open subsets $Y_k(x)$ satisfying $Y\subset\bigcup_{k=1}^K\overline{Y_k}(x)$,
- there exist a piecewise affine homeomorphism $\Phi(x,\cdot):Y\to Y$ such that the restrictions $\Phi(x,\cdot):Y_k(x_0)\to Y_k(x)$, $1\le k\le K$, are affine homeomorphisms, and
- for every $1\le k\le K$, the family of coefficients $(\underline{\underline{A}}(x,\Phi(x,\cdot)))_{x\in\mathcal{D}}$ restricted to $Y_k(x_0)$ has the affine form (see Definition 2)

$$\underline{\underline{A}}(x,\Phi(x,y))=\sum_{q=1}^{Q_k}\Theta_q(x)\underline{\underline{\underline{\underline{A_q}}}}(y),\quad \forall y\in Y_k(x_0). \tag{4-I.42}$$

Here we divide Y with $K=9$ affine mappings (one for each subdomain delineated with dashlines in Fig. 4.1). Then $\underline{\underline{A}}(x,\Phi(x,y))\equiv\underline{\underline{I_2}}$ for all $y\in Y_k(x_0)$ and k except the subdomain $Y_k(x_0)$ corresponding to the inclusion $Q(x)$ where $\underline{\underline{A}}(x,\Phi(x,y))=\theta(x)\underline{\underline{I_2}}$. The output is the second-order tensor ($x\in\mathcal{D}$) :

$$\underline{\underline{S}}_{i,j}(x)=\int_Y\underline{\underline{A}}(x,y)\nabla_y u_i(x,y)\cdot e_j\ dy,\quad i,j=1,2. \tag{4-I.43}$$

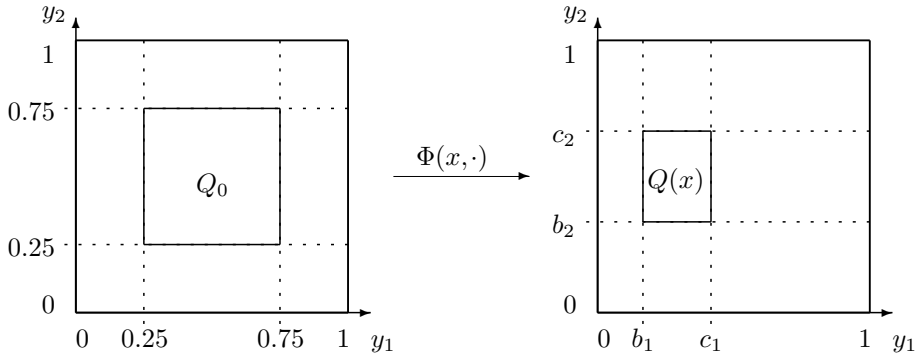


FIG. 4.1 – For each parameter value x, the cell with inclusion $Q(x)$ (on the right) is mapped through the piecewise affine homeomorphism $\Phi(x,\cdot)$ from a reference cell with inclusion Q_0 (on the left).

So each $\underline{\underline{S}}_{i,i}(x)$ for $i=1,2$ coincides with the compliance associated with one of the symmetric PDE (4-I.40), but the full problem is not compliant (see Section 4-I-B-c). We nevertheless avoid the computation of dual problems in this example of goal-oriented problems with many outputs by building only one reduced basis for the whole manifold $\{u_i(x,\cdot),\ x\in\mathcal{D},\ i=1,2\}$ (see [Boy08]).

4-I-C-b Discretization and numerical results

For the numerical results, we fix $\delta=.1$ and $\theta^0=.01$. As accurate Galerkin approximations for (4-I.40) in $X_{\mathcal{N}}$, we use continuous piecewise affine simplicial FE on a regular mesh for Y divided into 200 isosceles triangles with edge of size .1 ($\mathcal{N}=110$). (Note that when δ is close to .25, a regular FE mesh for the reference $Y_k(x_0)$ will be strongly distorted by the mapping $\Phi(x,\cdot)$, so one may also want to choose more than $K=9$ subdomains to limit this phenomenon, see [Qua09] for a discussion). Reduced bases $X_{\mathcal{N},N}$ are built incrementally for N increasing with a greedy algorithm trained on a trial sample of size $|\Lambda_{\text{trial}}|=50$ randomly chosen in Λ (offline stage). Then, they are tested for N increasing on another randomly chosen sample $\Lambda_{\text{test}}\subset\Lambda$ of size 1000 (online stage).

In Fig. 4.2, we observe that the RB greedy algorithm performs very well here, even with a small training sample. The reduced bases $X_{\mathcal{N},N}$ obtained yield very fast (in fact, exponentially) decaying RB approximation errors when N increases. In particular, the RB method yields interesting computational gains with reduced bases of size as small as $N=10$ and a very limited loss in accuracy. (Notice that the RB approximation error for the outputs $\underline{\underline{S}}_{ij}$, $i,j=1,2$, indeed scales as the square of the error for $u_i(x,\cdot)$; the output is precise up to the first six digits when $N=10$.)

Remark 17 (Multiscale interpretation of the problem)**.** *Assume we want to approximate (when $\epsilon\ll 1$) the scalar function u^ϵ solution in $\mathcal{D}=\{x=(x_1,x_2)\in\mathbb{R}^2,\ 0<x_i<1,\ i=1,2\}$ to :*

$$-\text{div}(\underline{\underline{A}}^\epsilon(x)\nabla u^\epsilon(x))=0,\quad \forall x\in\mathcal{D}, \tag{4-I.44}$$

supplied with mixed BC ($\underline{n}$ being the outward unit vector normal to the boundary) :

$$u^\epsilon(1,x_2)=0=u^\epsilon(x_1,1),\quad \underline{\underline{A}}^\epsilon\nabla u^\epsilon\cdot\underline{n}|_{(0,x_2)}=+1=\underline{\underline{A}}^\epsilon\nabla u^\epsilon\cdot\underline{n}|_{(x_1,0)},\quad \forall x_i\in(0,1),\ i=1,2, \tag{4-I.45}$$

where $\underline{\underline{A}}^\epsilon=\underline{\underline{A}}(x,\epsilon^{-1}x)$ are fast-oscillating coefficients, $\underline{\underline{A}}$ being like in (4-I.41) *(recall $\epsilon\ll 1$).*

To fix ideas, the matrix $\underline{\underline{A}}^\epsilon$ typically stands for conductivity (or diffusion) coefficients of a heterogeneous materials with "background value" $\underline{\underline{I_2}}$ perturbed by small inclusions within cells of periodicity $\epsilon\ll 1$ (see Fig. 4.3 for an example). One would like to now the average behaviour of this material.

Since the assumption $\underline{\underline{A}}^\epsilon=\underline{\underline{A}}(x,\epsilon^{-1}x)$ implies a separation of scales in the limit $\epsilon\to 0$, the homogenization theory yields an explicit approximation $u^\star$ for u^ϵ when $\epsilon\ll 1$, solution to the PDE :

$$-\text{div}(\underline{\underline{A}}^\star(x)\nabla u^\star(x))=0,\quad \forall x\in\mathcal{D}, \tag{4-I.46}$$

supplied with the same BC (4-I.45)*, where $\underline{\underline{A}}^\star$ are homogenized coefficients defined as :*

$$\underline{\underline{A}}^\star_{i,j}(x)=\int_Y \underline{\underline{A}}_{i,j}(x,y)\,dy+\underline{\underline{S}}_{i,j}(x),\quad \forall x\in\mathcal{D},\ i,j=1,2. \tag{4-I.47}$$

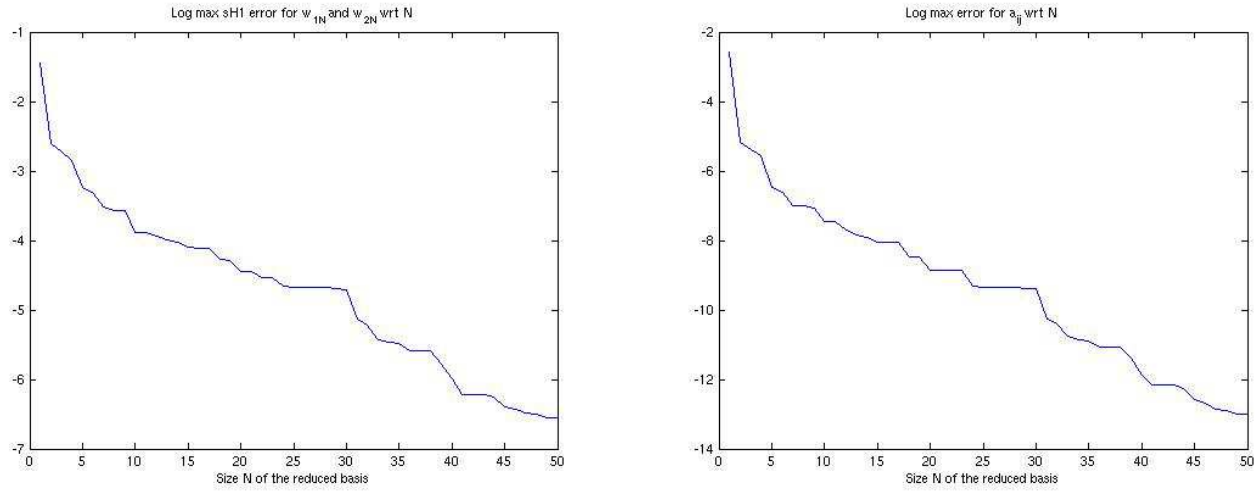


FIG. 4.2 – Maxima of the RB approximation errors $\max\{\|u_{i,\mathcal{N}}(x,\cdot)-u_{i,\mathcal{N},N}(x,\cdot)\|_X,\ 1\le i\le 2\}$ (left) and $\max\{|\underline{\underline{S}}_{ij,\mathcal{N}}(x)-\underline{\underline{S}}_{ij,\mathcal{N},N}(x)|,\ 1\le i,j\le 2\}$ (right), using a log scale with respect to the size N of the RB approximation space $X_{\mathcal{N},N}$, in the random test sample of the online stage ($x\in\Lambda_{\rm test}$).

The computation of $u^\star$ thus requires the evaluation of the output $\underline{\underline{S}}_{i,j}(x)$ at many material points $x\in\mathcal{D}$: this defines a many-query framework well-suited for a RB approach (see [Boy08] for details).

4-II RB Approach for Boundary Value Problems with Stochastic Coefficients

In this section, we apply the certified RB method to the computation of statistical outputs of a stochastic field solution to a BVP parametrized by stochastic coefficients in the BC.

4-II-A Position of the problem

Let us denote by $(\Omega,\mathcal{F},\mathbb{P})$ a probability space, and by $\omega\in\Omega$ the stochastic variable. We define a stochastic field $U(\cdot,\omega)$ as the $\mathbb{P}$-almost sure (a.s.) solution to the PDE :

$$-\mathrm{div}\left(\underline{\underline{A}}(x)\nabla U(x,\omega)\right)=0,\ \forall x\in\mathcal{D}, \tag{4-II.1}$$

supplied with a random Robin (or third-type) BC parametrized by a stochastic input field $\mathrm{Bi}(\cdot,\omega)$

$$\underline{n}(x)\cdot\underline{\underline{A}}(x)\nabla U(x,\omega)+\mathrm{Bi}(x,\omega)U(x,\omega)=g(x),\ \forall x\in\partial\mathcal{D}\ . \tag{4-II.2}$$

The matrix $\underline{\underline{A}}(x)=\kappa(x)\underline{\underline{I_2}}$ is assumed symmetric positive definite ($0<\kappa(x)<\infty$) for a.e. $x\in\mathcal{D}$. The scalar random field $\mathrm{Bi}(\cdot,\omega)$ is taken non-degenerate positive on some subset $\Gamma_{\rm B}\subset\partial\mathcal{D}$ with non-zero measure : $\mathbb{P}$-a.s. $0<\bar{b}_{\min}\le\mathrm{Bi}(\cdot,\omega)\le\bar{b}_{\max}<\infty$ a.e. in $\Gamma_{\rm B}$. So the problem is well-posed, $\underline{n}$ being the outward unit normal at the boundary of the smooth domain $\mathcal{D}$ chosen like in Fig. 4.4.

The boundary is divided into three non-overlapping open subsets : $\partial\mathcal{D}\subset\left(\overline{\Gamma_{\rm N}}\cup\overline{\Gamma_{\rm R}}\cup\overline{\Gamma_{\rm B}}\right)$ (see Fig. 4.4). The boundary source term is constant non-zero on $\Gamma_{\rm R}$ only : $g(x)=1_{\Gamma_{\rm R}},\forall x\in\partial\mathcal{D}$. Note that on $\Gamma_{\rm N}$, Eq. (4-II.2) thus reduces to homogeneous Neumann conditions. The physical interpretation is simple : $U(\cdot,\omega)$ is the steady-state temperature field in a heat sink with geometry $\mathcal{D}$ (comprised of an isotropic material of thermal conductivity κ) subject to zero heat flux on boundary $\Gamma_{\rm N}$, constant flux at boundary $\Gamma_{\rm R}$ (heat source), and a convective heat transfer at boundary $\Gamma_{\rm B}$. The Biot number Bi is a fashion for decoupling the solid conduction problem from the exterior fluid convection problem [LL02].

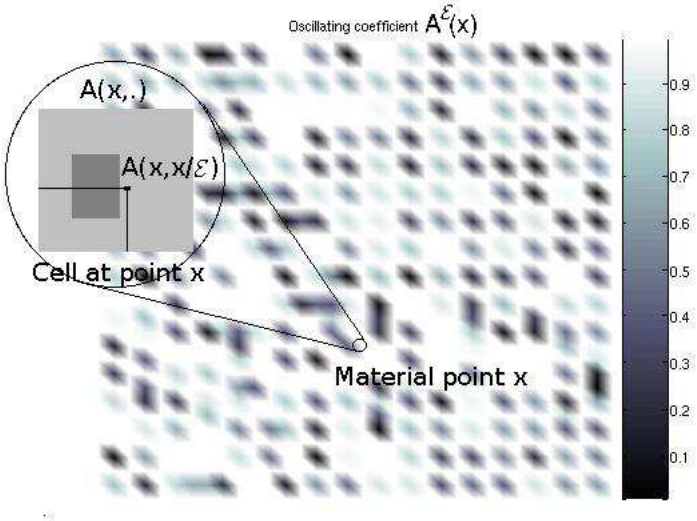


FIG. 4.3 – Example of fast oscillating coefficient $\underline{\underline{A}}^{\epsilon}_{i,i} : x \in \mathcal{D} \to \underline{\underline{A}}^{\epsilon}_{i,i}(x) \in (0,1)$, $i = 1,2$.

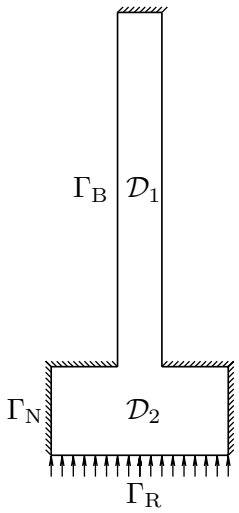


FIG. 4.4 – $\mathcal{D}$ has the geometry of a (piece of) heat sink : a spreader $\mathcal{D}_2$ with a fin $\mathcal{D}_1$ on top.

We consider statistical outputs such as the expected value $\mathbf{E}_{\mathbb{P}}(S)$ and the variance $\mathbf{Var}_{\mathbb{P}}(S)$ of a random variable $S(\omega)=\mathcal{E}(U(\cdot,\omega))$ given by a linear functional $\mathcal{E}$ of the trace of $U(\cdot,\omega)$ on Γ_{R} :

$$\mathcal{E}(U(\cdot,\omega))=\int_{\Gamma_{\mathrm{R}}} U(\cdot,\omega).$$

Coming back to physical interpretation, Bi is typically an engineering approximation, that is some average not reflecting all the environmental details. It thus makes sense to model the unknown Bi variations as a random field $\mathrm{Bi}(\cdot,\omega)$ and compute the sensitivity of $S(\omega)$ in order to assess the quality of the numerical simulation : this is typically a problem of Uncertainty Quantification (UQ). Now, UQ has been identified as a demanding computational task for several years [BTZ05, DBO01, MK05, DNP$^+$04], hence the interest of a method like our RB approach for reducing its cost.

It is often the case in UQ that robust and long-qualified codes exist for the numerical simulation of the BVP when stochastic coefficients assume deterministic values (here, for the problem (4-II.1-4-II.2) with a given realization of the random function $\mathrm{Bi}(\cdot,\omega)$). It thus seems natural for someone who does not want to invest too much (time and effort) in new algorithmic developments to use those old codes in UQ. One simple way of doing this is to evaluate $\mathbf{E}_{\mathbb{P}}(S)$ and $\mathbf{Var}_{\mathbb{P}}(S)$ with the Monte-Carlo (MC) method using M independent random variables $(S^m)_{1\le m\le M}$ with the same distribution law as S :

$$E_M[S]=\frac{1}{M}\sum_{m=1}^{M} S^m, \qquad V_M[S]=\frac{1}{M-1}\sum_{m=1}^{M}\left(E_M[S]-S^m\right)^2 , \tag{4-II.3}$$

(with Bessel's correction factor $M/M-1$ to get an unbiased estimator $V_M[S]$ of the variance). It is what we will do here. (Note yet that there exist other techniques for UQ and that our RB approach also combines with most of them as an accelerator, see [BBM$^+$09] for a discussion).

Finally, the computational problem reads as such : one often needs very large M to reach good accuracies in the approximation of $\mathbf{E}_{\mathbb{P}}(S)$ and $\mathbf{Var}_{\mathbb{P}}(S)$ through $E_M[S]$ and $V_M[S]$. Now, for each $m=1,\dots,M$, one must solve a BVP, that is typically compute FE approximations $U^m_{\mathcal{N}}\approx U^m$ for a given realization of the parameter function Bi^m. This is a (computationally demanding) many-query framework, well suited for the deployment of a RB approach if we can deal efficiently with the non-affine parameter Bi, and outputs $\mathbf{E}_{\mathbb{P}}(S)$ and $\mathbf{Var}_{\mathbb{P}}(S)$ (in fact, $E_M[S]$ and $V_M[S]$) that are sums over many realizations of the parameter.

4-II-B Discretization of the problem

First, we take a stochastic coefficient with a Karhunen–Loève (KL) expansion [Kar46, Loè78, ST06] :

$$\mathrm{Bi}(x,\omega)=\overline{\mathrm{Bi}}\left(G(x)+\sum_{k=1}^{\mathcal{K}}\Phi_k(x)\,Y_k(\omega)\right)\ ,\ \forall x\in\partial\mathcal{D}, \tag{4-II.4}$$

where $\mathcal{K}$ is the rank of the covariance operator for $\mathrm{Bi}(\cdot,\omega)$ with eigenvectors $(\Phi_k)_{1\le k\le\mathcal{K}}$ and eigenvalues $(\lambda_k)_{1\le k\le\mathcal{K}}$ in decreasing order, where the random variables $(Y_k)_{1\le k\le\mathcal{K}}$ are mutually uncorrelated in $L^2_{\mathbb{P}}(\Omega)$ with zero mean, and where $\overline{\mathrm{Bi}}=\int_\Omega d\mathbb{P}(\omega)\int_{\partial\mathcal{D}}\mathrm{Bi}(\cdot,\omega)$ is an intensity factor.

Second, we define a deterministic function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ parametrized by $\overline{\mathrm{Bi}}\in\mathbb{R}_{>0}$ and a sequence $y=(y_1,y_2,\dots)\in\mathbb{R}^{\mathcal{K}}$:

$$\mathrm{bi}(x;\overline{\mathrm{Bi}},y)=\overline{\mathrm{Bi}}\left(G(x)+\sum_{k=1}^{\mathcal{K}}\Phi_k(x)y_k\right),\ \forall x\in\partial\mathcal{D}. \tag{4-II.5}$$

For all $\overline{\mathrm{Bi}}\in\mathbb{R}_{>0}$, the function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ is well defined when $y\in\Lambda^y\subset\mathbb{R}^{\mathcal{K}}$, Λ^y being the range of the sequence $Y=(Y_k)_{1\le k\le\mathcal{K}}$ of random variables in (4-II.4).

Third, for any positive integer $K\le\mathcal{K}$, we define truncated versions $y^K=(y_1,y_2,\dots,y_K,0,0,\dots)$ (resp. $Y^K=(Y_1,Y_2,\dots,Y_K,0,0,\dots)$) for the sequence y (resp. Y). Then, the solution $U_K(\cdot,\omega)$ to the BVP (4-II.1-4-II.2) in which $\mathrm{Bi}(\cdot,\omega)\equiv\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},Y(\omega))$ is replaced by a truncated KL expansion $\mathrm{Bi}_K(\cdot,\omega):=\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},Y^K(\omega))$ can be mapped from $u_K(\cdot;y^K)$ solution to the BVP $(\forall y^K\in\Lambda^y)$:

$$\begin{cases} -\mathrm{div}\left(\underline{\underline{A}}(x)\nabla u_K(x;y^K)\right)=0,\ \forall x\in\mathcal{D}\ ,\\ \underline{n}(x)\cdot\underline{\underline{A}}(x)\nabla u_K(x;y^K)+\mathrm{bi}(x;\overline{\mathrm{Bi}},y^K)u_K(x;y^K)=g(x),\ \forall x\in\partial\mathcal{D}\ , \end{cases} \tag{4-II.6}$$

through the relation $U_K(x,\omega) \equiv u_K(x;Y^K(\omega))$, which holds for almost all $(x,\omega) \in \mathcal{D} \times \Omega$.

The strategy is now clear. For a given integer $K \leq \mathcal{K}$, we approximate the random variable $S(\omega)$ by $S_K(\omega) := \mathcal{E}(U_K(\cdot,\omega))$ and the statistical outputs $\mathbf{E}_{\mathbb{P}}(S_K(\omega))$ and $\mathbf{Var}_{\mathbb{P}}(S_K(\omega))$ by

$$E_M[S_K] = \frac{1}{M}\sum_{m=1}^{M} S_K^m, \qquad V_M[S_K] = \frac{1}{M-1}\sum_{m=1}^{M} \left(E_M[S_K] - S_K^m\right)^2 , \tag{4-II.7}$$

using M realizations $S_K^m := \mathcal{E}\left(u_K(\cdot;Y_m^K)\right)$, $m = 1,\ldots,M$, obtained from M independent realizations $\left(Y_m^K\right)_{1\leq m\leq M}$ of the random vector Y^K. In practice, $u_K(\cdot;Y_m^K)$ is approximated by $u_{\mathcal{N},K}(\cdot;Y_m^K)$ typically computed with a FE method (with $\mathcal{N} \gg 1$ degrees of freedom). So we would like to apply our RB approach to the problem (4-II.6), parametrized by $y^K \in \Lambda^y$, which has to be computed for (many) M parameter values.

Note that the problem (4-II.6) clearly casts into the scope of problems usually treated by the RB approach. In particular, it is affine in the input parameter y^K thanks to the KL expansion of Bi, which decouples the dependence on x and ω. However, the outputs (4-II.7) are not classical (they are computed only once for M parameter values) ; new developments of the RB method are thus necessary.

Note that our UQ strategy lies upon an approximation of the input parameter given as a stochastic field. So the error due to truncation of the KL expansion should be assessed. Moreover, the problem (4-II.6) after truncation need not be well-posed anymore. That is why we restrict to stochastic coefficients with a KL expansion (4-II.4) that is positive for any truncation order K ($1 \leq k \leq K \leq \mathcal{K}$) and that converges absolutely a.e. in $\partial\mathcal{D}$ when $K \to \mathcal{K}$. More specifically, we require for $k = 1,\ldots,\mathcal{K}$ a uniform bound $\|\Phi_k\|_{L^\infty(\Gamma_B)} \leq \phi$ and $Y_k := \Upsilon\sqrt{\lambda_k}Z_k$ with independent random variables Z_k uniformly distributed in the range $(-\sqrt{3},\sqrt{3})$, Υ being a positive coefficient. And we also ask $\sum_{k=1}^{\mathcal{K}} \sqrt{\lambda_k} < \infty$. (In particular, this imposes a very fast decay of the eigenvalues λ_k with k increasing.) We will see in the numerical results of Section 4-II-D that, in addition to the well-posedness of the truncated problems, the fast-decay assumption will also be important for the succes of our RB approach. (The ranges of the parameter random variables Y_k indeed scale in proportion to the eigenvalues λ_k of the covariance operator for $\mathrm{Bi}(\cdot,\omega)$, since the Z_k have a uniform range, and thus decrease very fast.)

4-II-C Reduced-Basis ingredients

As explained before in Section 4-I-A, essential ingredients in the RB method are : (fast, sharp) *a posteriori* estimators and (efficient) greedy selection procedures. Here, like in most specific applications of the RB method, both ingredients have to be adapted to the specificities of the UQ context.

The statistical outputs (4-II.7) require new *a posteriori* estimators. Moreover, the statistical outputs can only be computed after M queries Y_m^K, $m = 1,\ldots,M$, in the parameter y^K, so one cannot use these new *a posteriori* estimators in the offline parameter selection procedure.

In our RB greedy algorithm, we use an *a posteriori* estimation $|S_{K,\mathcal{N}}^m - S_{K,\mathcal{N},N}^m| \leq \Delta_{N,K}^s(Y_m^K)$ for the error between the FE approximation $S_{K,\mathcal{N}}^m := \mathcal{E}(u_{K,\mathcal{N}}(\cdot;Y_m^K))$ and the RB approximation $S_{K,\mathcal{N},N}^m := \mathcal{E}(u_{K,\mathcal{N},N}(\cdot;Y_m^K))$ of S_K^m at a fixed truncation order K, for any realization $Y_m^K \in \Lambda^y$. This is standard [Boy08, NVP05b, RHP08] and similar to the example in Section 4-I-A-b, see [BBM+09] for details. Note yet that the following variational formulation of the problem (4-II.6) :

$$\text{Find } u(\cdot;y^K) \in H^1(\mathcal{D}) \text{ s.t. } \int_{\mathcal{D}} \kappa \nabla u(\cdot;y^K)\cdot\nabla v + \int_{\Gamma_B} \mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y^K)u(\cdot;y^K)v = \int_{\Gamma_R} v, \forall v \in H^1(\mathcal{D}). \tag{4-II.8}$$

invokes a bilinear form with a coercivity constant that depends on K· To avoid the additional computation of the coercivity constant at each K, we use a uniform lower bound for the coercivity constant, after imposing $\mathrm{bi}(x;\overline{\mathrm{Bi}},y^K) \geq \overline{\mathrm{Bi}}G(x)/2$, $\forall x \in \Gamma_B$. (In practice, this imposes a limit $0 < \Upsilon \leq \Upsilon_{\max}$ on the intensity factor in the ranges of the random variables Y_k, thus on the random fluctuations of the stochastic coefficient.)

An *a posteriori* estimation for the truncation error $|S_{\mathcal{N}}^m - S_{K,\mathcal{N}}^m| = |\mathcal{E}(u_{\mathcal{N}}(\cdot;Y_m^K)) - \mathcal{E}(u_{K,\mathcal{N}}(\cdot;Y_m^K))| \leq \Delta_{N,K}^t(Y_m^K)$ is also designed in [BBM+09]. Note that in contrast to *a posteriori* estimators of the collateral RB approximation error due to the empirical interpolation of non-affine parameters (see Section 4-I-B-d), it is rigorous owing to the assumptions on the stochastic coefficient Bi.

Last, the errors $\Delta_{N,K}^t(Y_m^K)$ and $\Delta_{N,K}^s(Y_m^K)$ are combined for $m = 1,\ldots,M$ with the outputs $S_{K,\mathcal{N},N}^m$ to yield global error bounds for the RB approximation and the KL truncation errors in the MC estimations of the statistical outputs : $|E_M[S_{K,\mathcal{N},N}] - E_M[S_{\mathcal{N}}]| \leq \Delta_E^0(S_{K,\mathcal{N},N})$ and $|V_M[S_{K,\mathcal{N},N}] - V_M[S_{\mathcal{N}}]| \leq \Delta_V^0(S_{K,\mathcal{N},N})$.

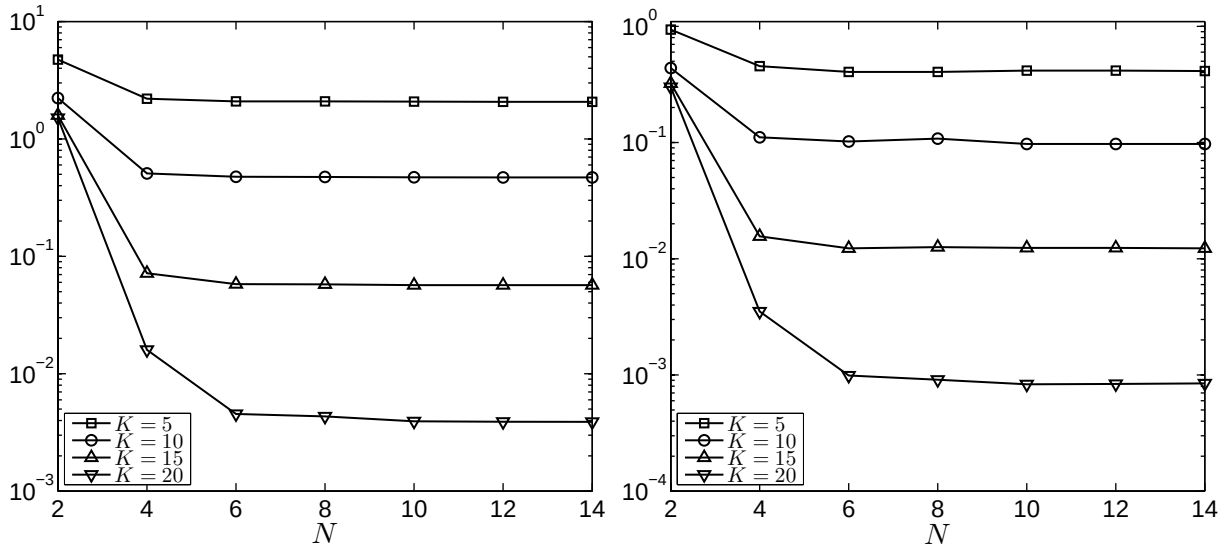


FIG. 4.5 – Global error bounds for the RB approximation error and the KL truncation error in the output expectation (left : $\Delta_E^0(S_{K,\mathcal{N},N})$) and the output variance (right : $\Delta_V^0(S_{K,\mathcal{N},N})$), as functions of the size $N = 2,\ldots,14$ of the reduced basis, at different truncation orders $K = 5,10,15,20$.

4-II-D Numerical results

We consider the steady heat conduction problem (4-II.1-4-II.2) inside the T-shaped heat sink $\mathcal{D} \subset \overline{\mathcal{D}_1} \cup \overline{\mathcal{D}_2}$ pictured in Fig. 4.4. The heat sink comprises a 2×1 rectangular substrate (spreader) $\mathcal{D}_2 \equiv (-1,1) \times (0,1)$ and a 0.5×4 thermal fin $\mathcal{D}_1 \equiv (-0.25,0.25) \times (1,5)$ on top. The diffusion coefficient is piecewise constant, $\kappa = 1_{\mathcal{D}_1} + \kappa_0 1_{\mathcal{D}_2}$, where $1_{\mathcal{D}_i}$ is the characteristic function of domain $\mathcal{D}_i$ $(i = 1,2)$. We show in Fig. 4.5 numerical results obtained with a regular mesh and quadratic finite elements using $\mathcal{N} = 6{,}882$ degrees of freedom for $\kappa_0 = 2.0$. (These have been previously published in [BBM+09].) Our random input field $\mathrm{Bi}(\cdot,\omega)$ was built like in (4-II.4) with $\overline{\mathrm{Bi}} = 0.5$, $G(x) \equiv 1$, $\Upsilon = 0.058$ and with the first $\mathcal{K} = 25$ terms in the KL expansion of the autocovariance function with Gaussian kernel $(\overline{\mathrm{Bi}}\Upsilon)^2 \exp(-(x-y)^2/\delta^2)$, for a correlation length $\delta = 0.5$.

After training the reduced basis offline with our RB greedy algorithm on a trial sample of size $|\Lambda_{\mathrm{trial}}| = 10000$, the RB approximation error in the output MC sums $E_M[S_{K,\mathcal{N},N}]$ and $V_M[S_{K,\mathcal{N},N}]$ decays very fast (in fact, exponentially) with the size $N = 1,\ldots,14$ of the reduced basis, see Fig. 4.5 with $K = 20$. (Note that $M = 10000$ for the MC sums.) But, it is classically observed that the RB approximation error for truncated problems at a fixed order K is quickly dominated by the truncation error beyond a critical value $N \geq N_{\mathrm{crit}}(K)$ increasing with K (recall Section 4-I-B-d).

So, for the problems where the rank $\mathcal{K}$ is finite, one can take $K = \mathcal{K}$ and our RB approach of UQ yields very interesting computational gains. The RB approximation error is controlled as usual by *a posteriori* estimation techniques. The RB computational gains for UQ can even be combined with an RB approach for other parametric variations in the context of design optimization under UQ : in (4-II.1-4-II.2), one may want to vary $\overline{\mathrm{Bi}}$ or κ for instance. See [BBM+09] for such complements.

When $\mathcal{K}$ is infinite, the control of the KL truncation error may be difficult : the problem is common to the UQ strategies that invoke a decomposition of the stochastic coefficient and similar to the approximation of non-affine parametrization through affine parametrization (recall Section 4-I-B-d). One may then want to choose K as large as possible to correctly describe all the random fluctuations, and thus require a large-dimensional parameter. Our RB approach is still efficient in some regimes with large K, which moderates the infamous *curse of dimensionality* (see Remark 15). In particular, a fast decay of the ranges of the parameters $(y_k)_{1 \leq k \leq K}$ facilitates the exploration of Λ^y by the greedy algorithm, which allows in return to treat large K when the eigenvalues λ_k decay very fast with k. In [BBM+09], we decreased the correlation length to $\delta = 0.2$ and could treat up to $K = 45$ parameters at a speed till 50 times faster than direct FE computations. But of course, further developments may yield far better results, in particular other (affine) parametrizations of the input noise with a faster decay rate of the volume to be explored, and/or a less stringent greedy selection procedure, which would not be based on the $L^\infty(\Lambda^y)$ norm but on the weaker $L^2(\Lambda^y)$ norm (sufficient for the control of the first two moments, but see Remark 18). One may also want to combine our RB approach with other UQ strategies than

the MC method, like the (pseudo)-Spectral Finite Elements method [GS91, BNT07].

Remark 18 (L^∞- *vs.* L^2-norm in the parameter range)**.** *It is usual in the many applications of the standard RB method that one should take care of all the specificities of the problem. Clearly, in the greedy procedure, the distance used to select parameters within the set of the least-successful RB approximations should reflect the specific nature of the BVP. This is usually done by adequately designing the a posteriori estimator. Now, one may also want to take into account the specific nature of the output, through the definition of what least-successful RB approximations are. In UQ in particular, a natural topology induced by the output on the parameter range Λ is not $L^\infty(\Lambda)$, but $L^2(\Lambda)$. So a question naturally arises : for such applications, is it not more efficient to construct approximation spaces using the $L^2(\Lambda)$ norm rather than the $L^\infty(\Lambda)$ norm ? Many existing techniques [New96, KV02, BGL06, Nou07, Nou08] indeed use the L^2-norm for model reduction, not only with applications in UQ. But the problem with the L^2-norm is to design efficient procedures retaining the essential features of our RB method. Most existing L^2-techniques are different in essence and scope from our RB method : the definition of the approximation space typically invokes the solution to a computationally expensive eigenvalue problem and the approximations leading to computational reductions are often difficult to certify.*

4-III RB approach of Variance Reduction for Monte-Carlo estimations using Parametrized Itô Stochastic Processes

In this section, we present an extension of the RB ideas to the context of parametrized Monte-Carlo estimations using RB approximations for control variates, which has been proposed recently in [BL09a].

4-III-A Position of the problem

We would like to compute many MC estimations of the expectation $\mathbf{E}_{\mathbb{P}}\left(Z^\lambda\right)$ of a functional

$$Z^\lambda = g^\lambda(X_T^\lambda) - \int_0^T f^\lambda(s, X_s^\lambda)\, ds \tag{4-III.1}$$

of the solutions $\left(X_t^\lambda, t\in[0,T]\right)$ to the Stochastic Differential Equation (SDE)

$$X_t^\lambda = x + \int_0^t b^\lambda(s, X_s^\lambda) ds + \int_0^t \sigma^\lambda(s, X_s^\lambda) dB_s \tag{4-III.2}$$

parametrized by $\lambda\in\Lambda$, for many values of the parameter λ. Here, we assume b^λ, σ^λ and the d-dimensional standard Brownian motion $\left(B_t\in\mathbb{R}^d, t\in[0,T]\right)$ are such that, for every $\lambda\in\Lambda$, the Itô processes $\left(X_t^\lambda\in\mathbb{R}^d, t\in[0,T]\right)$ are well defined. (Notice the deterministic initial condition $X_0^\lambda = x\in\mathbb{R}^d$.) In addition, f^λ and g^λ are also assumed smooth, so $Z^\lambda\in L^2_{\mathbb{P}}(\Omega)$ in particular.

Such parametrized problems are encountered in numerous applications, like the calibration of volatilities in finance, or the molecular simulation of Brownian particles in material physics. In rheology for instance, the viscoelastic mechanical properties of a flowing polymeric fluid can be determined from a non-Newtonian stress tensor $\mathbf{E}_{\mathbb{P}}\left(Z^\lambda\right)$ given by the Kramers formula :

$$Z^\lambda = X_T^\lambda \otimes \boldsymbol{F}(X_T^\lambda), \tag{4-III.3}$$

when the presence of many Brownian polymers diluted in the fluid is modelled by a Langevin equation :

$$dX_t^\lambda = \left(\lambda X_t^\lambda - \boldsymbol{F}(X_t^\lambda)\right) dt + dB_t \tag{4-III.4}$$

at each position of the fluid domain, the parameter $\lambda\in\mathbb{R}^{d\times d}$ being the local instantaneous value of the velocity gradient field ($d=2$ or 3 here). In (4-III.4), the orientation and the strectch of a molecule modelled as a "dumbbell" X_t^λ instantaneously equilibrates under a hydrodynamic force λX_t^λ (using the matrix λ of null trace since most polymeric fluids are incompressible), Brownian collisions B_t against the solvent molecules, and an entropic force $\boldsymbol{F}(X_t^\lambda)$ specific to the polymer molecule. Typically, this entropic force reads either $\boldsymbol{F}(X_t^\lambda) = X_t^\lambda$ (for Hookean dumbbells), or $\boldsymbol{F}(X_t^\lambda) = \frac{X_t^\lambda}{1-|X_t^\lambda|^2/b}$ (for Finitely-Extensible Nonlinear Elastic or FENE dumbells such that $|X_t^\lambda| < \sqrt{b}$). [7] Note yet that numerical simulations of the flow evolution of a polymeric fluid typically

[7] Notice that the material time-derivative of quantities attached to polymers let also appear an advection term in the left-hand-side of (4-III.4). We have neglected this term here, assuming it can be solved in a separate step of the numerical scheme using the characteristic method.

segregate, on many successive time slots $[0,T]$, (i) the computation of (4-III.4) from (ii) the computation of a new velocity gradient field in the fluid domain, so the problem (4-III.3-4-III.4) theoretically differs from (4-III.1-4-III.2). In particular, dumbbell models X_t^λ for polymers should rather invoke probabilistic than determinsitic initial conditions at the beginning of each time slot $[0,T]$. [8] The determinsitic assumption $X_0^\lambda = x$ is nevertheless already an interesting first step for the simulation of polymeric fluids with micro-macro models, see [BL09a] for a discussion. So finally, the simulation of micro-macro models for polymeric fluids requires to compute, on each time slot of a time-evolving simulation, the expectation $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$ for many parameter values λ corresponding to many spatial positions in the fluid domain; this is typically a many-query framework for the input-output relationship $\lambda \to \mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$.

We next consider the general form (4-III.1-4-III.2) of the problem with the MC estimation $\mathrm{E}_\mathrm{M}[Z^\lambda]$ parametrized by $\lambda \in \Lambda$ as output. Recalling the Central Limit Theorem (CLT) and the Chebyshev inequality, our strategy for reducing the amount of computations is, as usual for MC evaluations [Aro04, MO95, OvdBH97, BP99], *variance reduction* [HDH64, MT06]. We focus on one particular variance reduction technique :
Find a control variate $Y^\lambda \in L^2_\mathbb{P}(\Omega)$ such that in $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right) = \mathbf{E}_\mathbb{P}\left(Z^\lambda - Y^\lambda\right) + \mathbf{E}_\mathbb{P}\left(Y^\lambda\right)$, $\mathbf{E}_\mathbb{P}\left(Y^\lambda\right)$ can be easily evaluated, while the expectation $\mathbf{E}_\mathbb{P}\left(Z^\lambda - Y^\lambda\right)$ is approximated by MC estimations $\mathrm{E}_\mathrm{M}[Z^\lambda - Y^\lambda]$ that have smaller statistical error than direct MC estimations $\mathrm{E}_\mathrm{M}[Z^\lambda]$ of $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$ (hence $\mathbf{Var}_\mathbb{P}\left(Z^\lambda\right) \geq \mathbf{Var}_\mathbb{P}\left(Z^\lambda - Y^\lambda\right)$). In the following, we consider only centered control variates Y^λ (hence $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right) = \mathbf{E}_\mathbb{P}\left(Z^\lambda - Y^\lambda\right)$). Clearly, an optimal choice of control variate is, $\forall \lambda \in \Lambda$:

$$Y^\lambda = Z^\lambda - \mathbf{E}_\mathbb{P}\left(Z^\lambda\right), \tag{4-III.5}$$

(then, $\mathbf{Var}\left(Z^\lambda - Y^\lambda\right) = 0$). Unfortunately, the result $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$ itself is necessary to compute Y^λ as $Z^\lambda - \mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$. Equivalently, the optimal control variate (4-III.5) also writes :

$$Y^\lambda = \int_0^T \nabla u^\lambda(s, X_s^\lambda) \cdot \sigma^\lambda(s, X_s^\lambda) dB_s, \tag{4-III.6}$$

using Itô formula with the solution $u^\lambda(t,y) \in C^1\left([0,T], C^2(\mathbb{R}^d)\right)$ to the backward Kolmogorov equation (4-III.7) that satisfies the same (polynomial) growth assumptions at infinity than f^λ and g^λ :

$$\begin{cases} \partial_t u^\lambda + b^\lambda \cdot \nabla u^\lambda + \dfrac{1}{2}\sigma^\lambda (\sigma^\lambda)^T : \nabla^2 u^\lambda = f^\lambda \ , \\ u^\lambda(T,\cdot) = g^\lambda(\cdot), \end{cases} \tag{4-III.7}$$

where $\nabla u^\lambda \equiv \nabla_y u^\lambda(t,y)$ and $\sigma^\lambda(\sigma^\lambda)^T : \nabla^2 u^\lambda \equiv \sum_{i,j,k=1}^d \sigma_{ik}^\lambda(t,y)\sigma_{jk}^\lambda(t,y)\partial^2_{y_i,y_j} u^\lambda(t,y)$. But numerically solving the PDE (4-III.7) is still at least as difficult as computing $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$.

Now, there are many cases where one can use discretizations of either (4-III.5) or (4-III.6) to compute good approximations $\tilde{Y}^\lambda$ of Y^λ which yield good variance reduction even when d is quite large [New94]. In the many-query contexts where $\mathbf{E}_\mathbb{P}\left(Z^\lambda\right)$ has to be computed for a large number of parameter values $\lambda \in \Lambda$, we would like to exploit the (supposedly smooth) dependence $\lambda \to Y^\lambda$ to develop a RB approach for the manifold $\{Y^\lambda, \lambda \in \Lambda\} \subset L^2_\mathbb{P}(\Omega)$ (in fact, $\{\tilde{Y}^\lambda, \lambda \in \Lambda\}$) and thus further improve the computational gains for many variance-reduced MC evaluations at many λ. (Note that in contexts where the computation of $\tilde{Y}^\lambda$ for one λ is very expensive but the dependence on λ is smooth, a RB approach might not only improve, but in fact permit, variance-reduced MC estimations.)

Remark 19. *The computation of the expectation of* (4-III.1) *can also be achieved with quadrature formulas in a fully deterministic setting using the probability density functional of* (X_t^λ)*, solution to the Fokker-Planck (or Kolmogorov forward) equation of the SDE* (4-III.2) *on* $[0,T] \times \mathbb{R}^d$*. See in particular the recent work [KP09] where the standard RB method for the parametrized (semi-discretized, parabolic) PDEs has been used to solve the FENE Fokker-Planck discretized by dedicated deterministic methods. The approach here is different since we consider a stochastic discretization, perhaps more suitable for an extension to the high-dimensional settings* $d \geq 4$*.*

[8] Moreover, because of an additional advection step in the numerical scheme in between (i) and (ii) on each time slot $[0,T]$, the SDEs (4-III.4) at different position in the fluid domain would also naturally couple together after a few time slots. So the manifold described by the collection of SDEs (4-III.4) in a real fluid simulation has typically a complex structure. (The parameter λ is not random.) But we will not consider such particular structures for the parameter λ here. To retain some generality in our approach, we further assume that λ has no particular structure and belongs to some hypercube Λ. Yet, note that the case of a polymer flow simulation could in fact be much more favourable to a RB approach, since the right solution manifold to consider is then often thinner and more regular.

4-III-B Discretization of the problem

We assume from now on that realizations of the stochastic process (4-III.2) and the functional (4-III.1) can be computed at any desired accuracy level, for any $\lambda \in \Lambda$, using classical discretizations in the field of Itô calculus [KP00] for a given realization of the Brownian motion.

Considering either (4-III.5) or (4-III.6), we will make use of two approximations for Y^λ, either

$$\text{(i) } \tilde{Y}^\lambda = Z^\lambda - \mathrm{E}_{M_{\text{large}}}[Z^\lambda] \text{ or (ii) } \tilde{Y}^\lambda = \int_0^T \nabla \tilde{u}^\lambda(s, X_s^\lambda) \cdot \sigma^\lambda(s, X_s^\lambda) dB_s \tag{4-III.8}$$

respectively, using either a precomputed MC estimation $\mathrm{E}_{M_{\text{large}}}[Z^\lambda]$ with a very large number M_{large} of realizations or an approximation $\tilde{u}^\lambda$ of u^λ solution to (4-III.7). (Note that the Itô integral in (4-III.8-ii) is also assumed classically discretized, inline with the discretizations of (4-III.2) and (4-III.1) since one realization of (4-III.2–4-III.1–4-III.8-ii) is indeed computed with only one realization of the Brownian motion (B_t).)

The expectations $\mathbf{E}_{\mathbb{P}}\left(Z^\lambda - \tilde{Y}^\lambda\right)$ are approximated with M independent copies of $Z^\lambda - \tilde{Y}^\lambda$ as

$$\mathrm{E}_M[Z^\lambda - \tilde{Y}^\lambda] := \frac{1}{M} \sum_{m=1}^{M} (Z_m^\lambda - \tilde{Y}_m^\lambda) \xrightarrow[M\to\infty]{\mathbb{P}-a.s.} \mathbf{E}_{\mathbb{P}}\left(Z^\lambda - \tilde{Y}^\lambda\right) \tag{4-III.9}$$

by virtue of the Kolmogorov's strong law of large numbers. It is controlled by confidence intervals as equivalent of the *a posteriori* estimates in the standard RB method. Computable MC estimators

$$\mathrm{Var}_M\left(Z^\lambda - \tilde{Y}^\lambda\right) := \mathrm{E}_M\left(\left(Z^\lambda - \tilde{Y}^\lambda - \mathrm{E}_M(Z^\lambda - \tilde{Y}^\lambda)\right)^2\right) \xrightarrow[M\to\infty]{\mathbb{P}-a.s.} \mathbf{Var}\left(Z^\lambda\right) \tag{4-III.10}$$

for the variance yield a computable CLT by virtue of the Slutsky theorem : for all $a > 0$,

$$\mathbb{P}\left(\left|\mathbf{E}_{\mathbb{P}}\left(Z^\lambda - \tilde{Y}^\lambda\right) - \mathrm{E}_M\left(Z^\lambda - \tilde{Y}^\lambda\right)\right| \le a\sqrt{\frac{\mathrm{Var}_M\left(Z^\lambda - \tilde{Y}^\lambda\right)}{M}}\right) \xrightarrow[M\to\infty]{} \int_{-a}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \tag{4-III.11}$$

Thus, our true output is the parametrized couple $\left(\mathrm{E}_M\left(Z^\lambda - \tilde{Y}^\lambda\right), \mathrm{Var}_M\left(Z^\lambda - \tilde{Y}^\lambda\right)\right)$, and possible strategies for a RB approach arise. In particular, two algorithms have been proposed in [BL09a] :

in **Alg. 1** , a RB approach to compute fast linear approximations for $\tilde{Y}^\lambda$ as

$$\tilde{Y}_N^\lambda = \sum_{n=1}^{N} \alpha_n(\lambda) \tilde{Y}^{\lambda_n}, \ \forall \lambda \in \Lambda, \tag{4-III.12}$$

in **Alg. 2** , a RB approach to compute fast linear approximations for $\tilde{u}^\lambda$ as

$$\tilde{u}_N^\lambda = \sum_{n=1}^{N} \alpha_n(\lambda) \tilde{u}^{\lambda_n}, \ \forall \lambda \in \Lambda \tag{4-III.13}$$

and then use it in (4-III.8-ii) to compute $\tilde{Y}_N^\lambda \approx \tilde{Y}^\lambda$.

In each case, our paradigm, the two-stage standard RB method, suggests for computational efficiency to (i) choose N well-selected parameter values λ_n, $n = 1, \ldots, N$, in an offline stage, and (ii) for $\lambda \in \Lambda$ and $\{\tilde{Y}^{\lambda_n},\ n = 1, \ldots, N\}$ (resp. $\{\tilde{u}^{\lambda_n},\ n = 1, \ldots, N\}$) given, compute fast the coefficients $(\alpha_n(\lambda))_{n=1,\ldots,N}$ in the linear combination (4-III.12) (resp. (4-III.13)) as minimizers of the least-squares problem :

$$\min_{(\alpha_n) \in \mathbb{R}^N} \mathrm{Var}_M[Z^\lambda - \tilde{Y}_N^\lambda] \ . \tag{4-III.14}$$

In the context of variance reduction, the reliability of the RB approximation (4-III.12) (resp. (4-III.13)) can be easily certified online simply by checking $\mathrm{Var}_M[Z^\lambda - \tilde{Y}_N^\lambda] \le \mathrm{Var}_M[Z^\lambda]$.

4-III-C Reduced-Basis ingredients

For the application of the two RB approaches above, one has to specify (i) a fast technique for computing an approximation of the output, along with an error estimator suitable both for the offline selection of parameters and for the online certification of the output, and (ii) a greedy algorithm for the effective selection of good parameters λ_n, $n=1,\dots,N$, in a large (finite) sample Λ_{trial} nested in Λ.

In the context of variance reduction (recall (4-III.11)), the empirical variance (4-III.10) will be our error estimator for the selection of parameters with a greedy algorithm thus approximating a minimizer of

$$\inf_{(\lambda_n)\in\Lambda^N} \sup_{\lambda\in\Lambda} \inf_{(\alpha_n)\in\mathbb{R}^N} \mathbf{Var}_{\mathbb{P}}\left(Z^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right), \tag{4-III.15}$$

using our previously introduced discretizations of Λ, $\mathbf{Var}_{\mathbb{P}}(\cdot)$ and Y^λ. Note that because of

$$\mathbf{Var}_{\mathbb{P}}\left(Z^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right) = \mathbf{Var}_{\mathbb{P}}\left(Y^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right) = \mathbf{E}_{\mathbb{P}}\left(\left|Y^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right|^2\right),$$

the computation of $(\alpha_n)\in\mathbb{R}^N$ as a minimizer of (4-III.14) is actually a least-squares problem, which can be very-ill conditionned. So, in practice, one should choose dedicated algorithms for this [GvL96]. Note also that, in view of (4-III.11), one may want to use another greedy algorithm, for minimizers of

$$\inf_{(\lambda_n)\in\Lambda^N} \sup_{\lambda\in\Lambda} \inf_{(\alpha_n)\in\mathbb{R}^N} \mathbf{Var}_{\mathbb{P}}\left(Z^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right) \Big/ \mathbf{E}_{\mathbb{P}}\left(Z^\lambda - \sum_{n=1}^N \alpha_n Y^{\lambda_n}\right)^2. \tag{4-III.16}$$

As regards the speed of computations, both algorithms allow to precompute offline many quantities in order to fast solve (4-III.14) and then fast evaluate $\left(\mathrm{E}_M\left(Z^\lambda - \tilde{Y}_N^\lambda\right), \mathrm{Var}_M\left(Z^\lambda - \tilde{Y}_N^\lambda\right)\right)$ at each new $\lambda\in\Lambda$. Though the Algorithm 2 appears to be strongly more demanding than Algorithm 1, essentially because it requires the evaluation of the Itô integral (4-III.8-ii) (see [BL09a]).

4-III-D Numerical Results

The numerical results are taken from [BL09a]. The SDE (4-III.4) for FENE dumbbells when $d=2$ is discretized with the Euler-Maruyama scheme using 100 iterations with a constant time step $\Delta t = 10^{-2}$ starting from a (deterministic) initial condition $x=(1,1)$, with reflecting boundary conditions at the boundary of the ball with radius $\sqrt{b}$. For $b=16$ and $|\Lambda_{\text{trial}}|=100$ trial parameter values randomly chosen in the cubic range $\Lambda=[-1,1]^3$ (the traceless matrix $\underline{\underline{\lambda}}$ has entries $(\lambda_{11}=-\lambda_{22},\lambda_{12},\lambda_{21})$), a greedy algorithm is used to incrementally select $N=20$ parameter values after solving $|\Lambda_{\text{trial}}|=100$ least-squares problems (4-III.14) of dimension $M=1000$ at each step of the greedy algorithm (one for each of the trial parameter values $\lambda\in\Lambda_{\text{trial}}$). Then, the $N=20$ selected parameter values are used online for variance reduction within a test sample of $|\Lambda_{\text{test}}|=1000$ random parameter values in Λ.

The variance reduction obtained online by Algorithm 1 with $M_{\text{large}}=100\,M$ is very interesting, of about 4 orders of magnitude (whatever (4-III.15) or (4-III.16) used for the greedy selection). For the Algorithm 2, we use the (exact) solution $\tilde{u}^\lambda$ to the Kolmogorov backward equation for Hookean dumbells as an approximation to u^λ solution to (4-III.7). This also yields satisfying variance reduction though apparently not as good as in Algorithm 1. The Algorithm 2 is also more computationally demanding, since $\tilde{Y}_N^{\lambda_n}$, $n=1,\dots,N$, has to be recomputed for each new $\lambda\in\Lambda$ (recall (4-III.6)) before solving (4-III.14) and computing (4-III.13). In return, Algorithm 2 seems to be slightly more robust than Algorithm 1 when some online sample test $\Lambda_{\text{testwide}}$ uniformly distributed in $[-2,2]^3$ extrapolates the trial sample used offline (see Fig.4.6).

To sum up, the reiterated computations of parametrized MC estimations seem to be a promising opportunity of applications for new RB approaches. Much remains to be done theoretically : the situation is similar to the standard RB method. And many improvements of the initial work [BL09a] are also certainly possible on the numerical side, possibly dedicated to specific applications. Among the possible directions for further exploration of the method, high-dimensional problems seem particularly interesting ($d\geq 4$), which may require probabilistic approximations of u^λ (see [New94]).

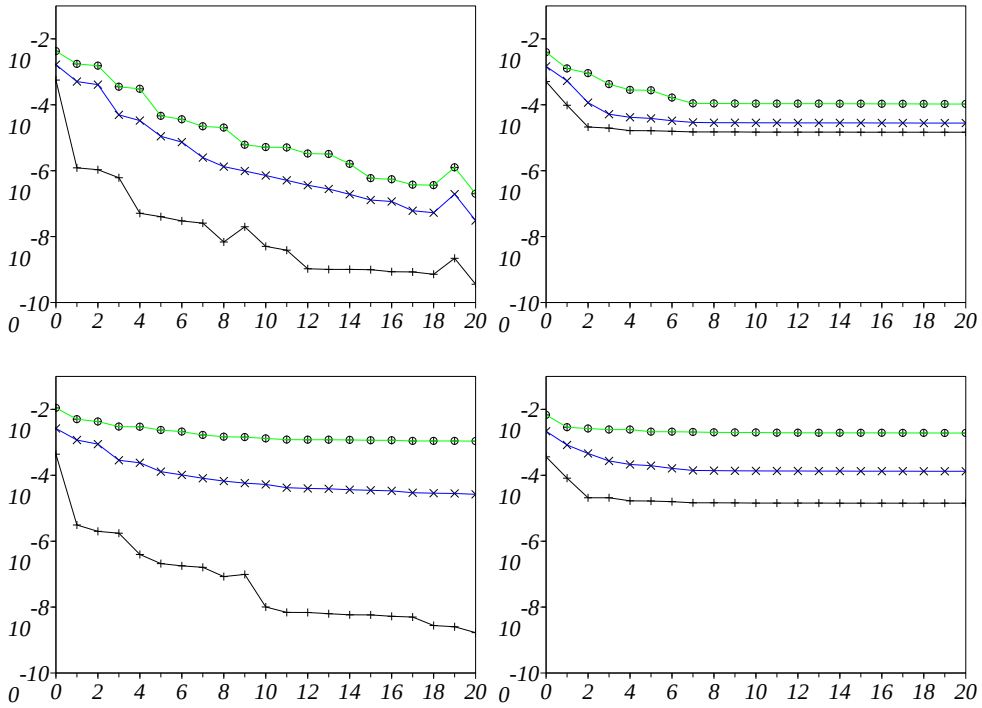


FIG. 4.6 – Algorithm 1 (left) and 2 (right) for FENE model with $b = 16$: Minimum $+$, mean $\times$ and maximum $\circ$ of $\mathrm{Var}_M[Z^\lambda - \tilde{Y}_N^\lambda]/\mathrm{E}_M[Z^\lambda - \tilde{Y}_N^\lambda]^2$ in online test samples $\Lambda_{\mathrm{test}} \subset \Lambda$ (top) and $\Lambda_{\mathrm{testwide}} \supset \Lambda$ (bottom) of parameters, with respect to the size N of the reduced basis.

4-IV Conclusion and Perspectives

We have recalled the maturity of RB techniques for the efficient computation of a parametrized functional at many values of the parameter, defined as the output of a solution to a parametrized BVP. Multiscale problems are one interesting many-query framework of application for the standard RB method among many (including real-time, optimization,etc.), because they usually invoke the parametrization of one scale by another one.

In some multiscale problems, the parameter has the right affine form required by the standard RB method. (Recall the many parametrized cell problems occuring in a two-scale homogenization procedure, Remark 17.) A large class of such multiscale problems should thus benefit from computational reductions when combined to the standard RB method. The scope of the RB method is indeed being generalized little by little, by further extending the domain of application of the (fast) *a posteriori* estimation with residus of Section 4-I-A-b to more and more complicated BVPs (nonlinear elliptic or parabolic, and possibly hyperbolic).

In some other multiscale problems, the parameter does not have the right affine form required for the standard RB method. Equivalently, the difficulty there is the high-dimensionality of the parameter. (Recall Section 4-II.) Additional knowledge can thus be necessary in order to rewrite the parameter so that a greedy algorithm can efficiently explore its (possibly truncated) range.

Last, there are multiscale problems which do not cast in the standard RB framework, but for which the essence of the standard RB method can be transferred :

- the decomposition into offline precomputations and fast online computations,
- a greedy algorithm to construct offline a linear approximation space spanned by snapshots of the same smooth manifold to be approximated,
- and a certification procedure to assess the quality of the RB approximation.

In the case of parametrized MC estimations (Section 4-III), RB approximations were computed as least-squares approximations rather than as the usual Galerkin projections for μ-BVP, and the quality of RB approximations for the controlled variates $Z^\lambda - \tilde{Y}^\lambda$ were estimated thanks to the CLT (4-III.11).

In the future, we hope that the scope of RB ideas can be further extended to different contexts of application requiring a large number of input-output relationships which are based on stochastic computations : indeed, it seems to us that this is a promising path toward efficient computations in high-dimensional spaces ($d \geq 4$).

4-V Appendix to the Chapter 4

4-V-A Appendix A. Proof of Proposition 20

Proof. For a given N such that $\epsilon_N(\mathcal{F})>0$, assume $\{f_k,\ k=1,\dots,N\}$ have been selected. Then one can define $e_k=P_{E_N(\mathcal{F})}f_k$, $k=1,\dots,N$. Now, we choose $\{f_k,\ k=1,\dots,N\}$ such that $|\det(e_1,\dots,e_N)|$ is maximal among all such determinants when $\{f_k,\ k=1,\dots,N\}$ explores $\mathcal{F}^N$. In particular, $\epsilon_N(\mathcal{F})>0$ implies $|\det(e_1,\dots,e_N)|\neq 0$. Hence $\{e_k,\ k=1,\dots,N\}$ is a basis for $E_N(\mathcal{F})$. For all $f\in\mathcal{F}$, we write $e:=P_{E_N(\mathcal{F})}f=\sum_{k=1}^N\alpha_k(e)e_k$ with

$$\alpha_k(e)=\frac{\det(e_1,\dots,e_{k-1},e,e_{k+1},\dots,e_N)}{\det(e_1,\dots,e_{k-1},e_k,e_{k+1},\dots,e_N)}.$$

Now, because of our choice of $\{f_k,\ k=1,\dots,N\}$, we have $|\alpha_k|<1$, $k=1,\dots,N$, and $\forall f\in\mathcal{F}$

$$\begin{aligned}\|f-P_{\mathbf{Span}(f_k,\ k=1,\dots,N)}f\|_X &\le \|f-\sum_{k=1}^N\alpha_k(e)f_k\|_X\\ &\le \|f-e\|_X+\|e-\sum_{k=1}^N\alpha_k(e)e_k\|_X+\sum_{k=1}^N|\alpha_k(e)|\|e_k-f_k\|_X\\ &\le 0+\epsilon_N(\mathcal{F})+N\,\epsilon_N(\mathcal{F}).\end{aligned}$$

If $\epsilon_N(\mathcal{F})=0$, then the result is obvious provided $\mathbf{Span}(f_k,\ k=1,\dots,N)=\mathbf{Span}(\mathcal{F})$.

□

4-V-B Appendix B. Proof of Proposition 21

Proof. Let us define an orthogonal sequence $(\xi_k)_{k\in\mathbb{N}_{>0}}$ in the Hilbert space X as :

$$\xi_1=f_1,\qquad \xi_{k+1}=f_{k+1}-P_{F_k}f_{k+1}\quad k\in\mathbb{N}_{>0}.$$

(Note that ξ is not necessarily in $\mathcal{F}$, which is indeed not necessarily a linear subspace of X.) Then the elements selected by the RB greedy algorithm read as linear combinations in the orthogonal basis $(\xi_k)_{k\in\mathbb{N}_{>0}}$, using collections of scalars $(\alpha_i^k)_{i=1,\dots,k}\in\mathbb{R}^k$,

$$f_k=\sum_{i=1}^k\alpha_i^k\xi_i\quad\text{ with }\alpha_k^k=1,\quad\forall k\in\mathbb{N}_{>0}. \tag{4-V.1}$$

In addition, by orthogonality, the Cauchy-Schwarz inequality and (4-I.21), we have for all $i=1,\dots,k$

$$\begin{aligned}\alpha_i^k&=\frac{(f_k,\xi_i)_X}{(\xi_i,\xi_i)_X}=\frac{(f_k-P_{F_{i-1}}f_k,\xi_i)_X}{(\xi_i,\xi_i)_X}\\ &\le\frac{\|f_k-P_{F_{i-1}}f_k\|_X}{\|\xi_i\|_X}=\frac{\|f_k-P_{F_{i-1}}f_k\|_X}{\|f_i-P_{F_{i-1}}f_i\|_X}\le 1.\end{aligned} \tag{4-V.2}$$

Now, let $N\in\mathbb{N}_{>0}$ be fixed. On the one hand, on noting $F_{j-1}\subset F_N$ for all $j=1,\dots,N+1$, the following bound holds for all $f\in\mathcal{F}$

$$\|f-P_{F_N}f\|_X\le\|f-P_{F_{j-1}}f\|_X\le\|f_j-P_{F_{j-1}}f_j\|_X\le\|\xi_j\|_X. \tag{4-V.3}$$

On the other hand, by defintion of $E_N(\mathcal{F})$, there exists a sequence $(v_k)_{k\in\mathbb{N}_{>0}}$ in $E_N(\mathcal{F})$ (which thus depends on N) such that

$$\|f_k-v_k\|\le\epsilon_N(\mathcal{F}),\ \forall k\in\mathbb{N}_{>0}. \tag{4-V.4}$$

To evaluate $\|\xi_j\|_X$ in (4-V.3) with the help of (4-V.4), we incrementally define a sequence $(\zeta_k)_{k\in\mathbb{N}_{>0}}$ in the linear space $E_N(\mathcal{F})$ such that

$$v_k=\sum_{i=1}^k\alpha_i^k\zeta_i,\quad\forall k\in\mathbb{N}_{>0},$$

using the same collections of scalars $(\alpha_i^k)_{i=1,\dots,k}\in\mathbb{R}^k$ as above in (4-V.1). Since $\dim E_N(\mathcal{F})=N$, the sequence $(\zeta_k)_{k=1,\dots,N+1}$ in particular is linearly dependent, thus :

$$\exists(\beta_k)_{k=1,\dots,N+1}\in\mathbb{R}^{N+1}/\ \text{(a)}\ \sum_{k=1}^{N+1}\beta_k\zeta_k=0\ \text{and}\ \ \text{(b)}\ \sum_{k=1}^{N+1}\beta_k^2=1. \tag{4-V.5}$$

First, recalling the orthogonality of the sequence $(\xi_k)_{k\in\mathbb{N}_{>0}}$, (4-V.5-a) with the Cauchy-Schwarz inequality and (4-V.5-b), we obtain a bound for all j such that $\beta_j\neq 0$, $1\leq j\leq N+1$:

$$\begin{aligned}\|\xi_j\|_X &\leq \beta_j^{-1}\|\sum_{k=1}^{N+1}\beta_k\xi_k\|_X\\ &\leq \beta_j^{-1}\|\sum_{k=1}^{N+1}\beta_k(\xi_k-\zeta_k)\|_X\\ &\leq \beta_j^{-1}\left(\sum_{k=1}^{N+1}\|\xi_k-\zeta_k\|_X^2\right)^{\frac{1}{2}}.\end{aligned} \tag{4-V.6}$$

Second, on noting the definition (4-V.1), the Cauchy-Schwarz inequality and (4-V.4), $\forall k=2,\dots,N+1$:

$$\begin{aligned}\|\xi_k-\zeta_k\|_X &= \|f_k-v_k+\sum_{i=1}^{k-1}\left(\sum_{\substack{l=0,\dots,k-i-1\\ i=j_0<j_1<j_2<\dots<j_l<j_{l+1}=k}}(-\alpha_{j_l}^{j_{l+1}})\times(-\alpha_{j_{l-1}}^{j_l})\times\dots\times(-\alpha_{j_0}^{j_1})\right)(f_i-v_i)\|_X,\\ &\leq \|f_k-v_k\|_X+\sum_{i=1}^{k-1}\left|\sum_{\substack{l=0,\dots,k-i-1\\ i=j_0<j_1<j_2<\dots<j_l<j_{l+1}=k}}(-\alpha_{j_l}^{j_{l+1}})\times(-\alpha_{j_{l-1}}^{j_l})\times\dots\times(-\alpha_{j_0}^{j_1})\right|\|f_i-v_i\|_X,\\ &\leq \epsilon_N(\mathcal{F})+\sum_{i=1}^{k-1}\left|\sum_{\substack{l=0,\dots,k-i-1\\ i=j_0<j_1<j_2<\dots<j_l<j_{l+1}=k}}(-\alpha_{j_l}^{j_{l+1}})\times(-\alpha_{j_{l-1}}^{j_l})\times\dots\times(-\alpha_{j_0}^{j_1})\right|\epsilon_N(\mathcal{F}) \qquad (4\text{-V.7})\\ &\leq \left(1+\sum_{i=1}^{k-1}\sum_{l=0}^{k-i-1}\binom{k-i-1}{l}\right)\epsilon_N(\mathcal{F})=\left(1+\sum_{i=1}^{k-1}2^{k-i-1}\right)\epsilon_N(\mathcal{F})=2^{k-1}\epsilon_N(\mathcal{F}). \qquad (4\text{-V.8})\end{aligned}$$

Last, on noting (4-V.5), there exists an integer j such that $|\beta_j|>\frac{1}{\sqrt{N+1}}>0$, $1\leq j\leq N+1$, and using (4-V.3) combined with (4-V.6) and (4-V.8) for that j, we obtain the desired result for all $f\in\mathcal{F}$:

$$\|f-P_{F_N}f\|_X\leq\sqrt{N+1}\left(\sum_{k=1}^{N+1}4^{k-1}\right)^{\frac{1}{2}}\epsilon_N(\mathcal{F})\leq 2^{N+1}\sqrt{N+1}\epsilon_N(\mathcal{F}). \tag{4-V.9}$$

□

The bound (4-V.9) is very pessimistic, mainly because of the upper-bound (4-V.8) for (4-V.7). (This is the source of the bad factor 2^{N+1} in the final rate of convergence.) We hope that better results can be achieved. In particular, to improve the bound (4-V.8) for (4-V.7), one could try to use less than 2^{k-1} terms $\|f_i-v_i\|_X$ $(i=1,\dots,k)$ to estimate the error $\|\xi_k-\zeta_k\|_X$ in (4-V.7), for instance either by using another orthogonal basis (see the proof of Section 4-V-A) or by specifying the coefficients (4-V.2).

Chapitre 5

Approche par bases réduites de l'homogénéisation au-delà du cas périodique

Ce chapitre est la reproduction de [Boy08]. Nous nous y intéressons au calcul de coefficients moyennés pour l'homogénéisation d'équations aux dérivées partielles elliptiques. Comme de nombreux problèmes multiéchelles, ce problème nécessite, à l'échelle microscopique, une grande quantité de calculs similaires entre eux, qui sont paramétrés par l'échelle macroscopique (voir aussi par exemple le Chapitre 7). Un tel cadre se prête très bien aux tentatives de réduction d'ordre.

Le but de ce travail est de montrer comment une approche par bases réduites permet d'accélérer le calcul d'un grand nombre de problèmes de cellules sans perte de précision. Les composants essentiels de cette approche par bases réduites sont :

- une estimation d'erreur *a posteriori*, qui fournit des bornes d'erreur précises pour les quantités finales intéressantes,
- une procédure d'approximation efficacement divisée entre des étapes "offline" et "online", ce qui découple la construction de l'espace d'approximation de son utilisation pour les projections de Galerkin, et
- un pré-calcul "offline" des quantités coûteuses à calculer et qui sont ensuite utilisées de nombreuses fois "online" (pourvu que la paramétrisation s'écrive sous forme affine et le permette).

Reduced-Basis Approach for Homogenization beyond the Periodic Setting

Sébastien Boyaval[a,b],

[a]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).

[b]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

The work of this chapter considers the computation of averaged coefficients for the homogenization of elliptic partial differential equations. In this problem, like in many multiscale problems, a large number of similar computations parameterized by the macroscopic scale is required at the microscopic scale. This is a framework very well suited for model reduction attempts. The purpose of this work is to show how the reduced-basis approach allows one to speed up the computation of a large number of cell problems without any loss of precision. The essential components of this reduced-basis approach are the a posteriori error estimation, which provides sharp error bounds for the outputs of interest, and an approximation process divided into offline and online stages, which decouples the generation of the approximation space and its use for Galerkin projections.

Keywords : Homogenization, Reduced-Basis Method, *A Posteriori* Estimates.

5-I Introduction

In this work, we study the numerical homogenization of *linear scalar elliptic* partial differential equations (PDEs) such as those encountered in the problems of thermal diffusion and electrical conduction. *Oscillating test functions*, also called *correctors*, are computed through a *reduced-basis* (RB) approach for parameterized *cell problems* supplied with periodic boundary conditions. Numerical results have been obtained with some prototypical parameterizations of the oscillating coefficients and are shown in a two-dimensional case with one single varying rectangular inclusion inside rectangular cells. The method applies to all numerical homogenization strategies that require one to solve a large number of parameterized cell problems, provided the approximation error for the homogenized coefficients can be a posteriori estimated from approximation errors for the cell problems.

In periodic homogenization, only *one* cell problem has to be solved in order to completely determine the averaged coefficients to be used in the homogenized (macroscopic) equation. In sharp contrast, nonperiodic homogenization theoretically requires one to solve *infinitely* many (microscopic) cell problems, that is, a *very large* number in practice, in order to compute averaged coefficients in the whole macroscopic domain (see [BLP78, JKO94], and section 5-II-B-b). Consequently, as opposed to the periodic case where the computation is light and exact, the nonperiodic case asks for a computationally demanding and approximate-in-nature task. This is why the design of a fast and accurate numerical homogenization method is considered as an important issue for the treatment of nonperiodic heterogeneous structures (see [HW97, MBS00, AB05, Glo06b, EEL$^+$07] and discussion in section 5-II-B-c). Model reduction techniques such as the RB method seem particularly well suited to this framework (see [YH07] for another recent reduction attempt in homogenization, using proper orthogonal decomposition (POD)).

The article is organized as follows. In section 2, we give a detailed presentation of the setting of the problem. For the sake of consistency and the convenience of the reader, we also briefly outline the main relevant issues in homogenization and RB theories. In section 3, we detail the RB approach for parameterized cell problems by providing a posteriori error bounds and an analysis of the convergence of the RB method in the homogenization context. We also notably deal with implementation issues that are crucial to the efficiency of our approach. Numerical results for the prototypical example of rectangular cells with one single rectangular inclusion are presented in section 4. Possible extensions of our work are discussed in the final section 5.

5-II Setting of the problem, elements of homogenization theory, and the RB approach

5-II-A Formulation of the problem

The mathematical problem under consideration throughout this article reads as follows. We are interested in the behavior of a sequence of scalar functions u^ϵ that satisfy

$$-\mathrm{div}(\bar{\bar{A}}^\epsilon(x)\nabla u^\epsilon(x)) = f(x) \ \forall x \in \Omega \tag{5-II.1}$$

in a connected bounded open set $\Omega \subset \mathbb{R}^n$ for a sequence of scalars $\epsilon > 0$. More precisely, we would like to derive the asymptotic limit of the sequence u^ϵ when $\epsilon \to 0$ along with approximations for u^ϵ when ϵ is small.

For the sake of simplicity, the scalar source term f is chosen in $L^2(\Omega)$, and we supply (5-II.1) with the following boundary conditions (BC) on the smooth (say, Lipschitz) boundary $\partial\Omega = \Gamma_D \bigcup \Gamma_N$ of Ω

$$(\mathrm{BC}) \begin{cases} u^\epsilon|_{\Gamma_D} = 0, \\ \bar{\bar{A}}^\epsilon \nabla u^\epsilon \cdot \bar{n}|_{\Gamma_N} = 1. \end{cases} \tag{5-II.2}$$

As a matter of fact, it is well known that the homogenization results are local in nature and do not depend on the boundary conditions, except for what regards error estimations close to the boundary [BLP78, JKO94]. Nor do the homogenization results depend on the source term f. Hence the generality of the assumptions (BC) and $f \in L^2(\Omega)$ chosen here only such as to give a precise mathematical frame to the numerical experiments.

To fix ideas, the unknown u^ϵ could be thought of either as a temperature or as an electric field in a macroscopic domain Ω. The tensorial coefficients for $\bar{\bar{A}}^\epsilon(x)$ would respectively be thought of either as temperature diffusivities or as electric conductivities.

Next, let us define, for any $\epsilon > 0$, the family $\bar{\bar{A}}^\epsilon \in L^\infty(\Omega, \mathcal{M}_{\alpha_A,\gamma_A})$ of functions from Ω to the set $\mathcal{M}_{\alpha_A,\gamma_A}$ of uniformly positive definite $n \times n$ matrices (second order tensors) with uniformly positive definite inverses, that is, matrices $\bar{\bar{A}}^\epsilon$ satisfying, for all $x \in \Omega$,

$$\begin{cases} 0 < \alpha_A |u|^2 \leq & \bar{\bar{A}}^\epsilon(x) u \cdot u \\ 0 < \gamma_A |u|^2 \leq & \bar{\bar{A}}^\epsilon(x)^{-1} u \cdot u \end{cases} \ \forall u \in \mathbb{R}^n. \tag{5-II.3}$$

Under such conditions, problems (5-II.1)–(5-II.2) are well posed in the sense of Hadamard. That is, for every $\epsilon > 0$, there exists a unique solution u^ϵ in

$$H^1_{\Gamma_D}(\Omega) = \left\{ u \in H^1(\Omega), u|_{\Gamma_D} = 0 \right\}$$

that continuously depends on the "data" f :

$$\|u^\epsilon\|_{H^1(\Omega)} \leq C(\Omega) \|f\|_{L^2(\Omega)}, \tag{5-II.4}$$

the constant $C(\Omega)$ depending only on Ω.

Moreover, the sequence of solutions u^ϵ is bounded in $H^1_{\Gamma_D}(\Omega)$, so that some subsequence ϵ' weakly converges to a limit $u^\star \in H^1_{\Gamma_D}(\Omega)$ when $\epsilon' \to 0$. We are specifically interested in estimating the behavior of this weakly convergent subsequence.

In a typical frame for the homogenization theory, the coefficients $\bar{\bar{A}}^\epsilon$ are assumed to oscillate very rapidly on account of numerous small heterogeneities in the domain Ω. For example, ϵ typically denotes the ratio of the mean period for microscopic fast oscillations of $\bar{\bar{A}}^\epsilon$ divided by the mean period for macroscopic slow oscillations of $\bar{\bar{A}}^\epsilon$ in Ω. Moreover, it is usually assumed that macroscopic and microscopic scales "separate" when ϵ is sufficiently small, which allows for the oscillating coefficients to be explicitly *homogenized* in the limit $\epsilon \to 0$, as exposed in sections 5-II-B-b and 5-II-B-c.

5-II-B General context for homogenization

As announced above, this section includes some basics of homogenization theory for linear scalar elliptic PDEs. Also, since the purpose of this summary is only to collect some elementary results, readers familiar with the homogenization theory may then like to skip this section and proceed to section 5-II-C, which introduces the RB theory.

5-II-B-a Abstract homogenization results

The following abstract homogenization result is the basis for many studies that aim at computing a numerical approximation for u^ϵ when ϵ is small (see [HW97, MBS00, AB05, Glo06b, EEL+07] for some of the many possible numerical strategies). It shows that, in the limit $\epsilon \to 0$, the small oscillating scale "disappears" from the macroscopic point of view. That is, the microscopic and macroscopic behaviors asymptotically separate. This implies that the limit problem is easier to solve than (5-II.1) for some small ϵ, since the former does not require one to resolve microscopic details. Moreover, a tractable approximation of u^ϵ when ϵ is small enough can be computed from the asymptotic limit when $\epsilon \to 0$.

More precisely, $u^\star$ can be obtained as the solution to the *H-limit* equation for (5-II.1) (see (5-II.7) below). The function $u^\star$ is then an L^2 approximation for u^ϵ when ϵ is small, as the asymptotic L^2 limit of u^ϵ when $\epsilon \to 0$. Moreover, an improved H^1 approximation for u^ϵ when ϵ is small can also be computed with $u^\star$ after "correction" of the gradient $\nabla u^\star$.

The homogenization of the sequence of equations (5-II.1) is the mathematical process which allows one to define the H-limit equation and the H^1 approximation for u^ϵ. It is performed by using the following abstract tools [JKO94] :

- a sequence of n *oscillating test functions* $z_i^\epsilon \in H^1(\Omega)$ such that, for every direction e_i of the ambient physical space $\mathbb{R}^n$ $(1 \leq i \leq n)$, we have $z_i^\epsilon \rightharpoonup x_i$ in $H^1(\Omega)$ and

$$-\mathrm{div}(\bar{\bar{A}}^\epsilon \nabla z_i^\epsilon) = -\mathrm{div}(\bar{\bar{A}}^\star e_i) \text{ in } H^{-1}(\Omega),$$

- a *homogenized tensor* $\bar{\bar{A}}^\star$ defined by

$$\bar{\bar{A}}^\epsilon \nabla z_i^\epsilon \rightharpoonup \bar{\bar{A}}^\star e_i \text{ in } [L^2(\Omega)]^n, \tag{5-II.5}$$

- a subsequence $u^{\epsilon'}$ of solutions for (5-II.1) that satisfies

$$\begin{cases} u^{\epsilon'} \rightharpoonup u^\star \text{ in } H^1_{\Gamma_D}(\Omega), \\ \bar{\bar{A}}^{\epsilon'} \nabla u^{\epsilon'} \rightharpoonup \bar{\bar{A}}^\star \nabla u^\star \text{ in } [L^2(\Omega)]^n, \end{cases} \tag{5-II.6}$$

  where $u^\star$ is solution for the H-limit or *homogenized* equation

$$-\mathrm{div}(\bar{\bar{A}}^\star(x) \nabla u^\star(x)) = f(x) \ \forall x \in \Omega, \tag{5-II.7}$$

  supplied with the BC,
- and an asymptotic approximation for a subsequence ϵ' of ϵ that satisfies

$$\left\| u^{\epsilon'} - u^\star \right\|_{L^2(\Omega)} \overset{\epsilon' \to 0}{\longrightarrow} 0, \tag{5-II.8}$$

$$\left\| \nabla u^{\epsilon'} - \sum_{i=1}^n \nabla z_i^{\epsilon'} \partial_i u^\star \right\|_{[L^1_{loc}(\Omega)]^n} \overset{\epsilon' \to 0}{\longrightarrow} 0, \tag{5-II.9}$$

  where $\partial_i u^\star$ are the components of $\nabla u^\star$ in each direction e_i.

Note that the latter convergence result (5-II.9) for $\nabla u^{\epsilon'}$ also holds in $[L^2_{loc}(\Omega)]^n$ if $u^\star \in W^{1,\infty}(\Omega)$. So, if $u^\star \in H^2(\Omega)$, the *corrector result* states that u^ϵ can be approximated with the following : formula,

$$u^\epsilon = u^\star + \sum_{i=1}^n (z_i^\epsilon - x_i) \partial_i u^\star + r_\epsilon, \tag{5-II.10}$$

where the remainder term r_ϵ converges strongly to zero in $W^{1,1}_{loc}(\Omega)$.

In a nutshell, the homogenization of the sequence of equations (5-II.1) has allowed us to derive an abstract homogenized problem (5-II.5)–(5-II.7), the solution $u^\star$ of which can be corrected with (5-II.9) into an H^1 approximation of u^ϵ in the limit $\epsilon \to 0$.

But to this point, we still lack an explicit expression for the homogenized tensor $\bar{\bar{A}}^\star$ in order to get an explicit asymptotic limit $u^\star$. That is why, though it is not required by the previous abstract theory, the scale separation in the behavior of the oscillating coefficients $\bar{\bar{A}}^\epsilon$ is often assumed to be explicitly encoded, using some specific postulated form for $\bar{\bar{A}}^\epsilon$. This allows one to derive an explicit expression of the homogenized problem, and even an error estimate in terms of ϵ for the *error correction* r_ϵ in (5-II.10), which allows one to quantify the homogenization approximation error.

5-II-B-b The explicit two-scale homogenization

To get explicit expressions for the homogenized problem, some particular dependence of the family $\bar{\bar{A}}^\epsilon$ on the space variable x is often assumed, such as in two-scale homogenization, for instance [All92]. Namely, on account of the scale separation assumption and the local dependence of the homogenization process, one of the most common assumptions is the *local periodicity* for $\bar{\bar{A}}^\epsilon$, which can be made precise as follows.

It is assumed that tensors $\bar{\bar{A}}^\epsilon$ are traces of functions of two coupled variables on the set locally defined by a fast (microscopic) variable $\epsilon^{-1}x$ linearly coupled with the slow (macroscopic) variable x in Ω :

$$\bar{\bar{A}}^\epsilon(x) = \bar{\bar{A}}\left(x, \frac{x}{\epsilon}\right), \tag{5-II.11}$$

where, for any $x \in \Omega$, the function $\bar{\bar{A}}(x,\cdot)$

$$\bar{\bar{A}}(x,\cdot) : y \in \mathbb{R}^n \to \bar{\bar{A}}(x,y) \in \mathbb{R}^{n\times n}$$

is 1-periodic in each of the n directions $(e_i)_{1\le i\le n}$. The domain $Y = [0,1]^n$ of the periodic pattern is called the *cell* and is identified with the n-dimensional torus. $\bar{\bar{A}}(x,\cdot)$ is said to be *Y-periodic*. Note that the property (5-II.3) of the tensors $\bar{\bar{A}}^\epsilon$ implies that $\bar{\bar{A}} \in L^\infty(\Omega, L^\infty(Y, \mathcal{M}_{\alpha_A,\gamma_A}))$.

Now, under the assumption (5-II.11) of local periodicity, one possible manner to get explicit expressions for the homogenized problem is to perform a formal two-scale analysis with the following ansatz :

$$u^\epsilon(x) = u_0\left(x, \frac{x}{\epsilon}\right) + \epsilon u_1\left(x, \frac{x}{\epsilon}\right) + \epsilon^2 u_2\left(x, \frac{x}{\epsilon}\right)\ldots, \tag{5-II.12}$$

where, for any $x \in \Omega$, the functions $u_i(x,\cdot)$ are Y-periodic. The first two terms of the ansatz (5-II.12) are shown to coincide with the H^1 approximation (5-II.10) for u^ϵ [BLP78, All92].

Inserting the ansatz (5-II.12) into (5-II.1) gives the following explicit expressions for the objects previously defined by the abstract homogenization result :

- The function $u_0 = u^\star(x)$ does not depend on the fast variable $\epsilon^{-1}x$ and is the L^2 approximation for u^ϵ given by the convergence result (5-II.8) ;
- the gradient $\nabla_y u_1(x,\cdot)$ linearly depends on $\nabla_x u^\star(x)$:

$$u_1\left(x, \frac{x}{\epsilon}\right) = \sum_{i=1}^n \partial_i u^\star(x) w_i\left(x, \frac{x}{\epsilon}\right) + \tilde{u}_1(x),$$

  where $(w_i(x,\cdot))_{1\le i\le n}$ are n Y-periodic *cell functions* ;
- the n cell functions $w_i(x,\cdot)$, parameterized by their macroscopic position $x \in \Omega$, are solutions to the following n *cell problems* :

$$-\mathrm{div}_y(\bar{\bar{A}}(x,y)\cdot[e_i + \nabla_y w_i(x,y)]) = 0 \ \forall y \in \mathbb{R}^n, \tag{5-II.13}$$

  and the correctors z_i^ϵ now read $z_i^\epsilon = x_i + \epsilon w_i(x, x/\epsilon)$;
- the entries $(\bar{\bar{A}}^\star(x)_{i,j})_{1\le i,j\le n}$ of the homogenized matrix $\bar{\bar{A}}^\star$ can be explicitly computed with the cell functions $w_i(x,\cdot)$:

$$\bar{\bar{A}}^\star(x)_{i,j} = \int_Y \bar{\bar{A}}(x,y)[e_i + \nabla_y w_i(x,y)]\cdot e_j \ dy; \tag{5-II.14}$$

- the H^1 approximation for u^ϵ is now tractable and is written

$$u^\epsilon = u^\star + \epsilon \sum_{i=1}^n w_i \partial_i u^\star + r_\epsilon, \tag{5-II.15}$$

  where, provided $u^\star \in W^{2,\infty}(\Omega)$, the correction error r_ϵ can be estimated to locally scale as ϵ (far enough from the boundary layer) and to globally scale as $\sqrt{\epsilon}$:

$$\|r_\epsilon\|_{H^1_{\Gamma_D}(\omega)} \le C_1 \epsilon \|u^\star\|_{W^{2,\infty}(\omega)}, \ \forall \omega \Subset \Omega, \tag{5-II.16}$$

$$\|r_\epsilon\|_{H^1_{\Gamma_D}(\Omega)} \le C_2 \sqrt{\epsilon} \|u^\star\|_{W^{2,\infty}(\Omega)}, \tag{5-II.17}$$

  with constants C_1 and C_2 depending only on Ω.

To sum up, the local periodicity assumption (5-II.11) allows one to completely determine the homogenized problem through explicit two-scale expressions. The derivation of the homogenized equation in the case of locally periodic coefficients serves as a basis for many numerical homogenization strategies.

5-II-B-c Numerical homogenization strategies

Under the local periodicity assumption (5-II.11), a possible two-scale explicit homogenization strategy for a sequence of linear scalar elliptic PDEs such as (5-II.1) simply reads as follows, in the frame of finite-element (FE) approximations for the scalar elliptic problems (5-II.13) and (5-II.7).

Algorithm 1 (two-scale homogenization strategy)**.** *To homogenize sequence* (5-II.1) *of PDEs,*

1. *solve the parameterized cell problems* (5-II.13) *at each point* $x \in \Omega$ *where the value of* $\bar{\bar{A}}^\star(x)$ *is necessary to compute the FE matrix of the homogenized problem* (5-II.7)*,*
2. *store the functions* w_i *for future computation of the* H^1 *approximation* (5-II.15)*,*
3. *assemble the FE matrix associated with the homogenized operator* $-\mathrm{div}(\bar{\bar{A}}^\star \nabla \cdot)$*,*
4. *solve the discretized (macroscopic) homogenized problem* (5-II.7)*, and*
5. *build the* H^1 *approximation* (5-II.15) *for* u^ϵ *with* $u^\star$ *and* w_i*.*

Now, on the one hand, in many practical situations, it is very common to assume that the tensors $\bar{\bar{A}}^\epsilon$ satisfy the local-periodicity assumption (5-II.11).

In practice, $\bar{\bar{A}}^\epsilon$ is often known for some given $\epsilon = \epsilon_0$ only. So, some asymptotic structure $\bar{\bar{A}}^\epsilon$ of the problem with oscillating coefficients in Ω should be constructed starting from the single member $\bar{\bar{A}}^{\epsilon_0}$ only. Now, to take advantage of the exact explicit expressions given by the two-scale analysis, it is preferable to build a family $\bar{\bar{A}}^\epsilon$ that satisfies assumption (5-II.11). Assumption (5-II.11) then seems fully justified for many applications from the practitioner's point of view. The main numerical difficulty of the two-scale homogenization is the first step of Algorithm 1, that is, the accurate computation of a large number of cell functions. This is the main issue addressed in this article.

On the other hand, for some applications where heterogeneities are highly nonperiodic, one may want to build the sequence $\bar{\bar{A}}^\epsilon$ differently or even skip the explicit construction of the sequence $\bar{\bar{A}}^\epsilon$. For example, the actual construction process of heterogeneities may suggest another sequence $\bar{\bar{A}}^\epsilon$ for which the error estimation of the H^1 approximation would then be more precise and meaningful or it may seem too difficult to explicitly build such a sequence $\bar{\bar{A}}^\epsilon$ that satisfies (5-II.11) from the knowledge of a single matrix $\bar{\bar{A}}^{\epsilon_0}$ only. In such cases, many numerical homogenization strategies have been developed to treat the numerical homogenization of oscillating coefficients that are not locally periodic.

To our knowledge, most of the existing numerical homogenization strategies with nonperiodic coefficients can be classified in one of the two following categories :

(i) Some strategies rely on different space assumptions than the local periodicity for the oscillating coefficients (e.g., reiterated homogenization [LLPW00], stochastic homogenization [BP04], deformed periodic coefficients [Bri94], stochastically deformed periodic coefficients [BLL06]) and still allow one to derive exact (but not always fully explicit) expressions for the homogenized equation and the error estimate of the approximation.

(ii) Other numerical homogenization strategies are much coarser and rely only on the assumption that some scale separation allows for the behavior of the oscillating coefficients to be numerically homogenized when $\epsilon \to 0$. Those strategies are then approximate in nature. They manage to approximate quite a large class of heterogeneous problems but may be computationally very demanding. In the absence of an assumption equivalent to local periodicity (5-II.11), they usually lack sharp error estimates beyond the periodic setting. Examples are the multiscale finite-element method (MsFEM) [HW97, AB05], the heterogeneous multiscale method (HMM) [EEL$^+$07], or the recent variational approach for nonlinear monotone elliptic operators proposed in [Glo06b].

Now, in any of the previously described situations where the numerical homogenization strategies require the computation of a large number of parameterized cell problems, the RB approach proposed hereafter is likely to bring some additional computational efficiency. As a matter of fact, most numerical approximate homogenization strategies are only slight modifications of the exact two-scale homogenization strategy proposed above in the frame of local periodicity assumption (5-II.11), and they do require the computation of a large number of parameterized cell functions.

For many mechanical applications, the similarity of most numerical approximate homogenization strategies with Algorithm 1 owes to the assumed existence of local representative volume elements (RVEs), which lead to general cell problems at each point x of the macroscopic domain [Glo06a, YH07]. Indeed, many approximate numerical homogenization strategies that require the computation of a large number of parameterized cell functions are based on some version of the following well known theorem (proved, e.g., in [JKO94]), which gives a mathematical meaning to the notion of the RVE.

Theorem 1. *Let $\bar{\bar{A}}^\epsilon$ be a sequence of matrices in $L^\infty(\Omega,\mathcal{M}_{\alpha,\gamma})$ that defines a sequence of linear scalar elliptic problems like* (5-II.1)*. The H-limit of $\bar{\bar{A}}^\epsilon$ is the homogenized tensor $\bar{\bar{A}}^\star$.*

For any $x\in\Omega$ and $\epsilon>0$, and any sufficiently small $h>0$, let us define a sequence of locally periodic matrices $\bar{\bar{A}}^\epsilon_h$ in the sense of (5-II.11) *:*

$$\bar{\bar{A}}^\epsilon_h(x)=\bar{\bar{A}}^\epsilon(x+h[\epsilon^{-1}x]), \tag{5-II.18}$$

where $[\epsilon^{-1}x]$ denotes the integer part of $\epsilon^{-1}x$.

Then, for every $1\le i\le n$, there exists a unique sequence of periodic solutions $w_i^{\epsilon,h}(x,\cdot)$ in the quotiented Sobolev space $H^1_\#(Y)/\mathbb{R}$ of Y-periodic functions in $H^1(Y)$ that satisfy the n cell problems

$$-\mathrm{div}(\bar{\bar{A}}^\epsilon(x+hy)\cdot[e_i+\nabla_y w_i^{\epsilon,h}(y)])=0\ \forall y\in Y \tag{5-II.19}$$

in the n-torus $Y=[0,1]^n$.

For each point $x\in\Omega$ and $\epsilon>0$, we define a matrix $\bar{\bar{A}}^\star_{\epsilon,h}$ made of the entries

$$\bar{\bar{A}}^\star_{\epsilon,h}(x)e_i\cdot e_j=\int_Y \bar{\bar{A}}^\epsilon(x+hy)\cdot[e_i+\nabla_y w_i^{\epsilon,h}(x,y)]\cdot[e_j+\nabla_y w_j^{\epsilon,h}(x,y)]\ dy$$

for any (i,j) in $\{1,2,\dots,n\}^2$. Then there exists a subsequence $h'\to 0$ such that

$$\lim_{h'\to 0}\lim_{\epsilon\to 0}\bar{\bar{A}}^\star_{\epsilon,h'}(x)=\bar{\bar{A}}^\star(x).$$

This theorem shows that, for any family $\bar{\bar{A}}^\epsilon$, it is always possible to approximate the exact homogenized problem through explicit expressions similar to those obtained under the local periodicity assumption (5-II.11), after some local periodization process of $\bar{\bar{A}}^\epsilon$ such as the one suggested in (5-II.18). For mechanical applications, choosing h for numerical purposes exactly translates as choosing the dimensions of a local RVE.

So, considering the landscape for the homogenization theory as described above, among alternative ways of improving the numerical homogenization strategies, we choose here to concentrate on speeding up the numerical treatment of a large number of parameterized cell functions rather than, for example, refining the approximations leading to explicit cell problems for a larger class of oscillating coefficients.

In what follows, for the sake of simplicity, we assume that the sequence of tensors $\bar{\bar{A}}^\epsilon$ satisfies the assumption (5-II.11), and we apply the RB approach to the two-scale numerical homogenization strategy of Algorithm 1. Our RB approach would apply as well with any numerical homogenization strategy that consists of first approximating $\bar{\bar{A}}^\epsilon$ by some sequence of tensors $\bar{\bar{A}}^\epsilon_h$ such as in (5-II.18), the latter leading to an explicit approximation for the homogenized problem after solving parameterized generalized cell problems such as (5-II.19). Let us now concentrate on decreasing as much as possible the computational cost of solving (5-II.13) for many parameter values $x\in\Omega$.

5-II-C The reduced-basis method

Two critical observations allow one to think that an output-oriented model reduction technique such as the RB method exposed in [NVP05a] is likely to improve the repeated numerical treatment of parameterized cell problems (5-II.13). First, only *outputs* of the cell functions are required to solve the homogenized problem, and there exist tight a posteriori estimators rendering sharp error bounds for those outputs.

Second, as extensively discussed above, numerical homogenization strategies in a nonperiodic setting often require one to *independently* solve numerous parameterized cell problems for many values of the parameter, though some parameter-independent quantities can be precomputed. Thus, a computational procedure based on an *offline-online* approach should naturally allow for a reduction of the computation time. Also, this reduction should turn out all the more relevant as the number of parameterized cell problems (5-II.13) increases, for instance, in the frame of parameter estimation and optimization problems that require one to solve many homogenized problems (5-II.7).

So the two previous observations motivate a RB approach for the parameterized cell problems (5-II.13), which should significantly decrease the expense of computations in terms of CPU time for the homogenization problems where the offline stage is short compared to the online stage or where the offline stage is not even an issue (such as in real-time engineering or with multiquery procedures, for instance). We are now going to introduce the basics of the reduced-basis method, well known to experts, who may want to directly proceed to section 5-III.

5-II-C-a The parameterized cell problem

Let X be the quotiented space $H^1_\#(Y)/\mathbb{R}$ of Y-periodic functions that belong to the Sobolev space $H^1(Y)$. The Hilbert space X is imbued with the $\tilde{H}^1(Y)$-norm

$$\|u\|_X = \left(\int_Y \nabla u \cdot \nabla u\right)^{1/2}$$

induced by the inner product $\big(u,v\big)_X = \int_Y \nabla u \cdot \nabla v$ for any $(u,v) \in X \times X$. In the dual space X' of X, the dual norm is defined for any $g \in X'$ by

$$\|g\|_{X'} = \sup_{v \in X} \frac{g(v)}{\|v\|_X}.$$

We also define, with a tensor $\bar{\bar{A}}$ defined like in (5-II.11) :

- a continuous and coercive bilinear form in $X \times X$ parameterized by $x \in \Omega$

$$a(u,v;x) = \int_Y \bar{\bar{A}}(x,y)\nabla u(y) \cdot \nabla v(y) dy \ \forall (u,v) \in X \times X,$$

  for which α_A and γ_A^{-1} are, respectively, coercivity and continuity constants,
- and n continuous linear forms in X also parameterized by $x \in \Omega$:

$$f_i(v;x) = -\int_Y \bar{\bar{A}}(x,y) e_i \cdot \nabla v(y) dy \ \forall v \in X, 1 \leq i \leq n.$$

Now, for all $1 \leq i \leq n$, the ith cell problem (5-II.13) for the cell function $w_i(x,\cdot)$ is rewritten in the following weak form : Find a $w_i(x,\cdot) \in X$ solution for

$$a(w_i(x,\cdot),v;x) = f_i(v;x) \ \forall v \in X, \tag{5-II.20}$$

where $x \in \Omega$ plays the role of a parameter.

We set $\mathcal{M}_i = \{w_i(x,y),\ x \in \Omega\}$ as the solution subspace of the ith cell problem (5-II.20) induced by the variations of x in Ω and

$$\mathcal{M} = \{w_i(x,y), x \in \Omega, 1 \leq i \leq N\} = \bigcup_{i=1}^{N} \mathcal{M}_i$$

as the global solution subspace, that is, the reunion of all solution subspaces for all cell problems.

Remark 20. Note at this point that $\mathcal{M}_i$ and $\mathcal{M}$ should be seen as spaces induced by coefficients $(\bar{\bar{A}}^\star(x,\cdot))_{x\in\Omega}$, and not induced by the values of $x \in \Omega$. It is indeed always possible (and often useful) to use other explicit quantities than x as a parameter to map $\mathcal{M}_i$ and $\mathcal{M}$, provided that the variations of this parameter inside a given range of values induce the same family of coefficients and the same corresponding cell functions as $x \in \Omega$.

For the sake of simplicity in the presentation of the RB method, the tensor $\bar{\bar{A}}(x,\cdot)$ will be assumed symmetric in the following. Thus, in the computation of the homogenized tensor $\bar{\bar{A}}^\star(x)$, the only interesting output for the n solutions $w_i(x,\cdot)$ is given by a symmetric matrix $s \in \mathbb{R}^{n\times n}$. Note that the RB approach still applies with nonsymmetric tensors $\bar{\bar{A}}(x,\cdot)$ modulo slight modifications.[1] The entries $(s_{ij})_{1\leq i,j\leq n}$ of the output matrix s, similar to compliances in the terminology of mechanics, are given by :

$$s_{ij}(x) = -f_j(w_i(x,\cdot);x) = \int_Y \bar{\bar{A}}(x,y)\nabla w_i(x,y) \cdot e_j dy. \tag{5-II.21}$$

In this frame, the purpose of the RB method is to speed up the computation of a large number of solutions $w_i(x,\cdot) \in X$ to (5-II.20) for many parameter values $x \in \Omega$ while controlling the approximation error for the output s.

[1] When the tensor $\bar{\bar{A}}(x,\cdot)$ is not symmetric, a dual problem, adjoint to the problem (5-II.20), is introduced. The dual problem can be solved similarly to the "primal" problem (5-II.20) with a RB method, in a dual RB projection space. Last, the output should be rewritten like s plus an additional term that accounts for the residual error due to the RB projection of (5-II.20). This primal-dual approach is more extensively described in [NVP05a] for instance.

5-II-C-b Principle of the reduced-basis method

Model reduction techniques for solving PDEs are usually developed in a *practical* context, with a view to optimizing numerics for specific computational purposes. This is the case of the POD method used in the reduced model multiscale method of [YH07], as well as of the RB method exposed in [VPRP03b, NVP05a, PR07b] that is used in the present work, and of other methods referenced in the three previously cited works.

Most often, model reduction techniques are formulated in a variational frame and aim at efficiently solving weak forms of PDEs through a Galerkin projection method, with a Hilbertian basis "adapted" to the practical context.

In the context of a two-scale numerical strategy of homogenization, we are interested in computing approximations for the linear outputs[2] s with accuracy scaling as some "tolerable" precision, typically of the same order of magnitude $O(\epsilon)$ as the correction error r_ϵ (5-II.15), after homogenization of the oscillating problem (5-II.1).

In our RB approach, a Hilbertian basis $(\xi_j)_{j\in\mathbb{N}}$ for X is consequently adapted to the parameterized equations (5-II.20) for all $x\in\Omega$ and $1\leq i\leq n$, if it is an orthonormal family (with respect to the ambient inner product $(\cdot,\cdot)_X$) such that :

- the ambient solution space $X\subset\overline{\text{span}\{\xi_j,\, j\in\mathbb{N}\}}$ is separable,
- and, for any "tolerable precision" $\varepsilon=O(\epsilon)$, there exists a finite-dimensional vector subspace $X_N=\text{span}\{\xi_j,\, 1\leq j\leq N(\varepsilon)\}$ of X, where, for any x in Ω and $1\leq i\leq n$, the Galerkin approximations $w_{iN}(x,\cdot)\in X_N$ for $w_i(x,\cdot)$ satisfying

$$a(w_{iN}(x,\cdot),v;x)=f_i(v;x)\ \forall v\in X_N \tag{5-II.22}$$

are sufficiently close to $w_i(x,\cdot)$ so that :

$$\operatorname*{esssup}_{x\in\Omega}|s(x)|\leq\varepsilon$$

and $N(\varepsilon)$ is a monotonically nonincreasing function of the precision ε.

Another characteristic feature of our RB methodology inspired from [VPRP03b, NVP05a, PR07b] is the possibility to explicitly assess the quality of approximations for $s(x)$ at any $x\in\Omega$ thanks to rigorous a posteriori estimates, unlike the methodology used in [YH07], for instance.

A posteriori error estimators are indeed a key ingredient in our RB methodology. They allow us to a posteriori certify the efficiency of the model reduction. But they essentially allow for an efficient offline construction of the RB vector subspace X_N that is consistent with the expected goal of achieving some tolerable precision ε for the output s and whose computational cost scales well with the dimension of the parameter space [PR07b].

5-II-C-c Practice of the reduced-basis method

Our RB methodology divides in two steps and exploits the a posteriori estimations as optimally as possible to first build a reduced basis from a *training* sample of parameters in an *offline* stage and to then use this reduced basis for multiquery purposes in the *online* stage.

So the RB method first needs a training sample $\mathcal{D}=\{x_k\in\Omega,\, 1\leq k\leq p\}$ of parameter values, in order to build a reduced basis for the subset $\mathcal{M}_p$ of $\mathcal{M}$:

$$\mathcal{M}_p=\{w_i(x_k,\cdot),1\leq k\leq p,1\leq i\leq n\},$$

which satisfies the constraint of accuracy for the output.

This is termed as the *offline* stage, where accurate approximations for some elements of $\mathcal{M}_p$ are computed by using an accurate and generic numerical method with a large number $\mathcal{N}$ of degrees of freedom. The $(n\times p)$ FE approximations $(w_{i\mathcal{N}}(x_k,\cdot))_{i,k}$ for $l(w_i(x_k,\cdot))_{i,k}$ span an $(n\times p)$-dimensional vector subspace of X that contains the approximation

$$\mathcal{M}_p^{\mathcal{N}}=\{w_{i\mathcal{N}}(x_k,y),x_k\in\mathcal{D}\}$$

of the solution subspace $\mathcal{M}$.

Typically, the elements of $\mathcal{M}_p^{\mathcal{N}}$ selected to build the reduced basis are computed through a FE method in order to span growing Galerkin projection spaces $X_k\subset X_{k+1}\subset\mathcal{M}_p^{\mathcal{N}}$, $k\in\mathbb{N}$, until the tolerable precision for the output approximation is reached at some $k=N$ for every member of the training sample.

[2] With nonsymmetric tensors $\bar{\bar{A}}$, the approximation error for linear outputs such as s can be expressed as a product of two approximation errors, one for the parameterized solutions $w_i(x,\cdot)$ and another one for some dual quantity that is the solution to the problem dual for (5-II.20). But here, because of the symmetry and of the specific nature of the output, that is, the so-called compliance in reference to mechanics, introducing a dual problem is unnecessary because it formulates exactly the same way like the cell problem [NVP05a]. The approximation error for the output can then be directly expressed as the square of the approximation error for the cell function, as will be made clearer in section 5-III-A.

At each step k ($k \leq N$) of the reduced basis construction, RB approximation errors with the RB projection space X_k for the whole training sample $\mathcal{M}_p$ are estimated, just like in the online stage for any parameter $x \in \Omega$, in order to successively select elements from $\mathcal{M}_p$ that have to be computed accurately and that span the reduced basis. So certified a posteriori error estimators are a key ingredient in our RB approach, which allow for a rigorous construction of the reduced basis, consistent with the initial practical context (see details of construction in section 5-III-B).

Note that an efficient treatment of the offline stage implies that, in practice, the RB approximation errors decrease *fast* when the size k of the Galerkin projection space X_k increases, optimally as fast as $N(\varepsilon)$. This is limited by the choice of the training sample in practice. Yet, if the initial training sample is "sufficiently" well distributed over Ω, the effective rate of decrease is often close to optimal (see section 5-IV for a numerical verification of this statement with a five-dimensional parameter space).

Then, in a so-called *online* stage, it is possible to compute Galerkin approximations in X_N, solutions to (5-II.22) for any $x \in \Omega$, for which outputs $s(x)$ satisfy the constraint of tolerable precision ε.

A reduced (orthonormal) basis $(\xi_j)_{1 \leq j \leq N}$ spans the low-dimensional vector space X_N, assumed to capture the directions of the solution manifold $\mathcal{M}$ (in fact, only a FE approximation $\mathcal{M}_p^{\mathcal{N}}$ of a "training" sample $\mathcal{M}_p$) that are the most sensitive to the evaluation of the output s (up to precision ε).

Computations for the RB approximations in the low-dimensional Galerkin space X_N ought to be fast (provided they are well prepared enough, possibly after some offline precomputations), since only small matrices are to be inverted.

In the end, the RB methodology should provide an efficient model reduction if the tolerable precision for the output approximation error is effectively reached, for most $x \in \Omega$, with a N-dimensional reduced basis when N is much smaller than $\mathcal{N}$, the number of degrees of freedom necessary for a generic numerical method like the FE method to reach the same precision.

With computationally inexpensive a posteriori estimators, it is moreover possible to control online the RB approximation error for outputs and consequently adapt the reduced basis (ξ_j) when the precision is insufficient.

5-III Reduced-basis approach for the cell problem

5-III-A Error bounds for the cell problem

A RB approach of (5-II.20) needs to a posteriori estimate the approximation error of Galerkin solutions of the cell problem and the approximation error of their output s. The purpose of this section is then to derive the error bound (5-III.8) for the cell functions solutions of (5-II.20). This allows us to a posteriori estimate the approximation error for Galerkin solutions of the cell problem and their outputs through the error bound (5-III.3).

To this end, let us introduce Galerkin approximations for the homogenized and the output matrices with a Galerkin approximation w_{iN} solution for (5-II.22) :

$$\bar{\bar{A}}^{\star}_N(x)_{i,j} = \int_Y \bar{\bar{A}}(x,y)[e_i + \nabla_y w_{iN}(x,y)] \cdot e_j \, dy, \tag{5-III.1}$$

$$s^N_{ij}(x) = \int_Y \bar{\bar{A}}(x,y) \nabla w_{iN}(x,y) \cdot e_j dy \tag{5-III.2}$$

and a linear operator $T^x : X \to X$ so that, for any $u \in X$ and $x \in \Omega$,

$$(T^x u, v)_X = a(u,v;x) \ \forall v \in X.$$

The existence of such an operator T^x directly leans on the Riesz–Fréchet representation theorem in the Hilbert space X.

Lemma 1. *Let $w_i(x,\cdot)$ and $w_j(x,\cdot)$ be two cell functions, solutions of* (5-II.20) *for $i \neq j$. Let $w_{iN}(x,\cdot)$ and $w_{jN}(x,\cdot)$ be two Galerkin approximations of the former cell functions in a N-dimensional vector space X_N. $w_{iN}(x,\cdot)$ and $w_{jN}(x,\cdot)$ are solutions for* (5-II.22). *Then the approximation error $|s_{ij}(x) - s^N_{ij}(x)|$ for the output entries can be superiorly bounded by :*

$$\Delta^s_{ij,N}(x) = \frac{\|T^x(w_i(x,\cdot) - w_{iN}(x,\cdot))\|_X \|T^x(w_j(x,\cdot) - w_{jN}(x,\cdot))\|_X}{\alpha_A}, \tag{5-III.3}$$

where $\alpha_A > 0$ defined in (5-II.3) *is a coercivity constant for the quadratic operator associated with the matrix $\bar{\bar{A}}(x,\cdot)$ for all $x \in \Omega$.*

Remark 21. The error estimator (5-III.3) can be easily sharpened through tighter local coercivity constants $\alpha_A(x)$ for each $x \in \Omega$, which would account for the parameterization of the problem. In practice, such improvements imply a trade-off between the cost of the additional computations and the effectivity of the error estimation, since they are often not essential to get an efficient reduction of the model in the end (see [PR07b], and numerical results in Figure 5.2 of section 5-IV, where only a global lower coercivity constant α_A for all $x \in \Omega$ has been used).

Proof of Lemma 1. For any $1 \leq i \leq N$ and $x \in \Omega$, the Galerkin approximation error

$$\|w_i(x,\cdot) - w_{iN}(x,\cdot)\|_X \tag{5-III.4}$$

can be bounded starting from the following equality :

$$a(w_i(x,\cdot) - w_{iN}(x,\cdot), v; x) = f_i(v;x) - a(w_{iN}(x,y), v; x) \ \forall v \in X, \tag{5-III.5}$$

which is easily obtained by substraction of (5-II.22) from (5-II.20).

Let us define the parameterized bilinear residual forms g_i in $X \times X$ such that, for all parameter values $x \in \Omega$ and $1 \leq i \leq n$,

$$g_i(u,v;x) = a(u,v;x) - f_i(v;x) \ \forall (u,v) \in X \times X.$$

Then (5-III.5) with $v = w_i(x,\cdot) - w_{iN}(x,\cdot)$ allows one to immediately derive the following estimates through the dual norm of the residual linear form for $w_i(x,\cdot)$ defined in X :

$$v \to g_i(w_{iN}(x,\cdot), v; x).$$

First, owing to the coercivity of the bilinear form a, we obtain the lower bound :

$$\alpha_A \|w_i(x,\cdot) - w_{iN}(x,\cdot)\|_X \leq \|g_i(w_{iN}(x,\cdot), v; x)\|_{X'} \tag{5-III.6}$$

for the Galerkin approximation error.

Second, in view of the continuity of the bilinear form a, we obtain the superior bound :

$$\|g_i(w_{iN}(x,\cdot), v; x)\|_{X'} \leq \gamma_A^{-1} \|w_i(x,\cdot) - w_{iN}(x,\cdot)\|_X \tag{5-III.7}$$

for the Galerkin approximation error.

Finally, note that it is possible to compute the dual norm of the linear form

$$v \to g_i(w_{iN}(x,\cdot), v; x) = -a(w_i(x,\cdot) - w_{iN}(x,\cdot), v; x)$$

by using the Riesz–Fréchet representant $T^x(w_i(x,\cdot) - w_{iN}(x,\cdot))$ in the Hilbert space X

$$\|g_i(w_{iN}(x,\cdot), v; x)\|_{X'} = \|T^x(w_i(x,\cdot) - w_{iN}(x,\cdot))\|_X$$

and that one can obtain numerical approximations for α_A and γ_A^{-1} (in this simple case, by using the spectral properties of the matrices resulting from the Galerkin projection of the bilinear form a in the large generic solution spaces of the offline stage, for instance). So the Galerkin approximation error (5-III.4) can be a posteriori bounded by using estimations (5-III.6) and (5-III.7).

For $x \in \Omega$ and $1 \leq i \leq N$, we define a posteriori estimators $\Delta_N(w_i(x,\cdot))$ for the Galerkin approximation errors (5-III.4), by using the previous superior bounds, by

$$\Delta_N(w_i(x,\cdot)) = \frac{\|a(w_i(x,\cdot) - w_{iN}(x,\cdot), \cdot; x)\|_{X'}}{\alpha_A}. \tag{5-III.8}$$

The effectivities $\eta_N(w_i(x,\cdot))$ corresponding to the estimators $\Delta_N(w_i(x,\cdot))$:

$$\eta_N(w_i(x,\cdot)) = \frac{\Delta_N(w_i(x,\cdot))}{\|w_i(x,\cdot) - w_{iN}(x,\cdot)\|_X} \tag{5-III.9}$$

satisfy the following N-independent inequalities :

$$1 \leq \eta_N(w_i(x,\cdot)) \leq \frac{\gamma_A^{-1}}{\alpha_A}, \tag{5-III.10}$$

which shows the stability of the error estimator $\Delta_N(w_i(x,\cdot))$.

The a posteriori superior bound $\Delta_N(w_i(x,\cdot))$ for the Galerkin approximation error (5-III.4) will now allow us to derive a simple superior bound for output approximation errors. Indeed, we have, for any $1\le i,j\le N$ and $x\in\Omega$,

$$\begin{aligned}|s_{ij}(x)-s_{ij}^N(x)| &= |f_j(w_i(x,\cdot)-w_{iN}(x,\cdot);x)|\\ &= |a(w_j(x,\cdot),w_i(x,\cdot)-w_{iN}(x,\cdot);x)|\\ &= |a(w_j(x,\cdot)-w_{jN}(x,\cdot),w_i(x,\cdot)-w_{iN}(x,\cdot);x)|\\ &\le \alpha_A\Delta_N(w_i(x,\cdot))\Delta_N(w_j(x,\cdot))\end{aligned}$$

since $w_j(x,\cdot)$ and $w_i(x,\cdot)$ are solutions for (5-II.20) and $w_{iN}(x,\cdot)$ is a solution for (5-II.22). Hence the a posteriori superior bound is $\Delta^s_{ij,N}(x)$ for Galerkin approximations of the output $s_{ij}(x)$.

Numerical approximations for $\Delta^s_{ij,N}(x)$ will allow us to build quickly a reduced basis for cell problems (5-II.20) and to control the online RB approximations. Note that $\Delta^s_{ij,N}(x)$ scales as the product $\Delta_N(w_i(x,\cdot))\Delta_N(w_j(x,\cdot))$, hence the interest of model order reduction techniques for solutions $w_i(x,\cdot)$ without much loss of precision for output $s(x)$.

Remark 22. Note that, for the output error bounds to scale like the square of the error bound for the cell functions, it has been essential to have the following orthogonality property for any $x\in\Omega$:

$$a(w_i(x,\cdot)-w_{iN}(x,\cdot),w_{jN}(x,\cdot);x)=0,\forall 1\le i,j\le N.$$

That is why we have chosen to build only one RB projection space X_N, spanned by all the parameterized cell functions $w_i(x_k,y)$ when $1\le i\le n$ and $x_k\in\mathcal{D}$. Yet, note that, without this choice, the same scaling can still be obtained with n distinct RB projection spaces $(X_{iN})_{1\le i\le n}$ for each of the n solution subspaces $\mathcal{M}_i$, provided one slightly modifies the definition of the output. Namely, another output matrix σ and its RB approximation σ^N should then be defined, starting from s and s^N, by adding a residual error. Their entries read, for $1\le i,j\le n$,

$$\sigma_{ij}(x)=-f_j(w_i(x,\cdot);x)+g_i(w_i(x,\cdot),w_j(x,\cdot);x), \tag{5-III.11}$$

$$\sigma^N_{ij}(x)=-f_j(w_{iN}(x,\cdot);x)+g_i(w_{iN}(x,\cdot),w_{jN}(x,\cdot);x), \tag{5-III.12}$$

where $\sigma=s$, because the tensor $\bar{\bar{A}}(x,\cdot)$ is symmetric, and $\sigma^N=s^N$ only when the RB projection space is the same for all solution subspaces $\mathcal{M}_i$ (as above). Interestingly, the same additional residual term in the output σ also arises when the tensor $\bar{\bar{A}}(x,\cdot)$ is not symmetric. It is then evaluated with *dual* cell functions, solutions for a problem dual to the cell problems (5-II.20) [NVP05a].

5-III-B The reduced-basis construction

Let $\mathcal{D}=\{x_k,\ 1\le k\le p\}$ be a "training" sample of p values for the parameter x in Ω. Unless some physical properties of the system guide the choice for $\mathcal{D}$, the parameter values x_k should be p realizations of a random variable uniformly distributed over Ω.

To accurately solve (5-II.20) at the selected parameter value $x_k\in\mathcal{D}$, we use a FE method with a $\mathcal{N}$-dimensional FE vector space. Typically, $\mathcal{N}$ is very large, in order for the FE approximations to be accurate.

In the offline stage of the RB approach, we would like to build a N-dimensional RB projection subspace $X_N\subset X$ so that elements of $\mathcal{M}$ can be approximated in X_N with sufficient precision for the resulting approximate outputs.

A *reduced basis* $(\xi_j)_{1\le j\le N}$ made of N vectors of $\mathcal{M}^{\mathcal{N}}_p$ ($N<n\times p$, and $N\ll\mathcal{N}$ for the model reduction to allow a significant gain of computation time) is computed to span the vector subspace X_N.

For any $x\in\Omega$, we write :

$$w_{ik}(x,y)=\sum_{j=1}^{k}w_{ikj}(x)\xi_j(y),$$

the RB approximation for $w_i(x,y)$ in the Galerkin projection space X_k of size $k\le N$.

The reduced basis $(\xi_j(y))_{1\le j\le N}$ of X_N is built in order to optimally control the approximation error for the outputs of the training sample. This is performed in the offline stage with the following algorithm.

Algorithm 2 (offline algorithm)**.** *We build a reduced basis* $(\xi_j(y))_{1\le j\le N}$ *as follows :*

1. *For some couple* $(k^0(1),i^0(1))$, $1\le k^0(1)\le p$, *and* $1\le i^0(1)\le n$, *we compute the accurate FE approximation* $w_{i^0(1)\mathcal{N}}(x_{k^0(1)},y)\in\mathcal{M}_p^{\mathcal{N}}$ *for* $w_{i^0(1)}(x_{k^0(1)},y)\in\mathcal{M}_p$ *;*
2. *we set* $l=1$ *and compute* $\xi_1(y)=\frac{w_{i^0(1)\mathcal{N}}(x_{k^0(1)},y)}{\|w_{i^0(1)\mathcal{N}}(x_{k^0(1)},\cdot)\|_X}$ *as the first RB basis function;*
3. *then, while* $l<N$ *:*
   (a) *we compute for every* $x_k\in\mathcal{D}$ *and* $1\le i\le n$ *the* $(n\times p)$ *RB approximations* $w_{il}(x_k,y)\in X_l=\mathrm{span}\{\xi_j,\ 1\le j\le l\}$ *for the* n *cell problems* (5-II.22),
   (b) *for* $(k^0(l+1),\ i^0(l+1))=\mathrm{argmax}_{1\le k\le p,1\le i\le n}\frac{\Delta^s_{ii,l}(x)}{|s^l(x_k)|}$, *we compute the accurate FE approximation* $w_{i^0(l+1)\mathcal{N}}(x_{k^0(l+1)},y)\in\mathcal{M}_p^{\mathcal{N}}$ *of* $w_{i^0(l+1)}(x_{k^0(l+1)},y)$,
   (c) *we set* $\xi_{l+1}(y)=\frac{R_{l+1}(y)}{\|R_{l+1}(y)\|_X}$ *the* $(l+1)$*th RB basis function, where* R_{j+1} *is the remainder of the projection on the* l*-dimensional reduced basis :*

$$R_{l+1}(y)=w_{i^0(l+1)}(x_{k^0(l+1)},y)-\sum_{k=1}^{l}(w_{i^0(l+1)}(x_{k^0(l+1)},\cdot),\xi_k)_X\xi_k(y),$$

   (d) *we do* $l=l+1$ *and go back to* (a).

5-III-C Convergence of the reduced-basis method for the cell problem

The a priori convergence of Galerkin approximations for solutions of continuous and coercive elliptic equations such as (5-II.20) is classical. It usually relies on the following lemma (see, e.g., [SF73b] for a proof).

Lemma 2 (Céa lemma)**.** *For any* $1\le i\le n$, *let* w_i *be the solution of* (5-II.20) *and* w_{iN} *its approximation in some* N*-dimensional Galerkin projection space* $X_N\subset X$. *Then we have, for any* $x\in\Omega$,

$$\|w_i(x,y)-w_{iN}(x,y)\|_X\le\sqrt{\frac{\gamma_A^{-1}}{\alpha_A}}\inf_{w(y)\in X_N}\|w_i(x,y)-w(y)\|_X.$$

To conclude that RB approximations such as $w_{iN}\in X_N$ a priori converge to $w_i\in X$ when $N\to\infty$, it would then be natural to use Lemma 2 as follows.

Lemma 3. *If there exist a dense separable subspace* $\mathcal{V}$ *of* $\mathcal{M}$ *and an application* $r_N:\mathcal{V}\to X_N$ *such that*

$$\lim_{N\to\infty}\|v-r_N(v)\|_X=0\ \forall v\in\mathcal{V},\tag{5-III.13}$$

then, by the Céa lemma, for any $x\in\Omega$ *and* $1\le i\le n$, *RB approximations* $w_{iN}(x,\cdot)$ *converge to* $w_i(x,\cdot)$ *in the following sense :*

$$\lim_{N\to\infty}\|w_i(x,y)-w_{iN}(x,y)\|_X.\tag{5-III.14}$$

That is, for all $\delta>0$, *there exists a positive integer* $N(\delta)$ *such that, for all* $x\in\Omega$ *and* $1\le i\le N$,

$$\|w_i(x,\cdot)-w_{iN}(x,\cdot)\|_X\le\delta\ \forall N\ge N(\delta)\tag{5-III.15}$$

in combination with the choice $\mathcal{V}=\mathcal{M}$ *and* r_N *as the projection operator from* $\mathcal{M}$ *to* X_N *for the inner product in* X.

Unfortunately, the convergence assumed in (5-III.13) can be shown only insofar as we have information about the sampling technique for generating $\mathcal{D}$, which amounts to knowing how the parameter values are selected to build X_N as N increases. Such an assumption is often unrealistic since, to choose the right parameter values x_k for $\mathcal{D}$, one should already know $\mathcal{M}$ or some spectral representation of it [MPT02b]. So the scope of Lemma 3 seems strongly limited, as any a priori analysis of the RB method in general.

As a matter of fact, the RB method is a *practical* method of order reduction and should only be a posteriori certified to converge, by using reliable and computationally inexpensive error bounds that can be evaluated along the RB approximations, for instance.

Remark 23. Note that, in practice, the Galerkin projection space X_N is built to converge to the solution subspace $\mathcal{M}_{\mathcal{N}} = \{w_{i\mathcal{N}}(x,y),\ x \in \Omega\}$ induced by the FE approximations $w_{i\mathcal{N}}(x,y)$. Now the FE approximations converge only pointwise in parameter space. That is, for some given parameter $x \in \Omega$, and $1 \leq i \leq n$, there exists for all $\delta > 0$ a positive integer $\mathcal{N}_i(\delta, x)$ such that :

$$\|w_i(x,y) - w_{i\mathcal{N}}(x,y)\|_X \leq \delta \ \forall \mathcal{N} \geq \mathcal{N}_i(\delta, x).$$

So, in any case, one can only hope for a pointwise convergence of the RB approximations, of the form :

$$\lim_{N \to \infty} \lim_{\mathcal{N} \to \infty} \|w_i(x,\cdot) - w_{iN}(x,\cdot)\|_X = 0, \tag{5-III.16}$$

where $w_{iN}(x,\cdot)$ implicitly depends on $\mathcal{N}$. Besides, the only RB approximation errors that are effectively estimated by using the error bounds of section 5-III-A are $\|w_{i\mathcal{N}}(x,\cdot) - w_{iN}(x,\cdot)\|_X$.

5-III-D Error estimate for the asymptotic H^1 homogenized solution

In the frame of the two-scale homogenization strategy, the asymptotic H^1 homogenized approximation for $u^\epsilon(x)$ in the limit $\epsilon \to 0$ is

$$u_0(x) + \epsilon\ u_1\left(x, \frac{x}{\epsilon}\right) = u^\star(x) + \epsilon \sum_{1 \leq i \leq n} w_i\left(x, \frac{x}{\epsilon}\right) \partial_i u^\star(x),$$

which strongly converges to $u^\epsilon(x)$ in $H^1_{\Gamma_D}(\Omega)$ when $\epsilon \to 0$ if $u^\star \in W^{2,\infty}(\Omega)$.

In this approximation, the homogenized solution $u^\star$ is the solution to the variational formulation (5-III.17) of the homogenized equation (5-II.7) :

$$\int_\Omega \bar{\bar{A}}^\star \nabla u^\star \cdot \nabla v = \int_\Omega f v + \int_{\Gamma_N} v,\ \forall v \in H^1_{\Gamma_D}(\Omega). \tag{5-III.17}$$

But in practice, one can compute only a RB approximation $\bar{\bar{A}}^\star_N$ for $\bar{\bar{A}}^\star$, with entries :

$$\left(\bar{\bar{A}}^\star_N(x)\right)_{i,j} = \left(\int_Y \bar{\bar{A}}(x,y)dy\right)_{i,j} - s^N_{ij}(x),$$

which should be taken into account to estimate the approximation error for the asymptotic H^1 homogenized approximation.

The following Lemma 4 will show how the a posteriori control of the RB output allows one to control the approximation error for the asymptotic H^1 homogenized approximation with a RB approach.

Let us first discretize (5-II.20) with an FE space of dimension $\mathcal{N}$ and (5-III.17) with $W_{h_{hom}} \subset H^1_{\Gamma_D}(\Omega)$ a discrete FE Galerkin projection space associated with a mesh of size h_{hom} for Ω. We will still denote by w_i and $u^\star$ the FE approximate solutions resulting from the previous discretizations of (5-II.20) and (5-III.17). Then, we define a RB approximation for the asymptotic H^1 homogenized approximation :

$$u^\star_N(x) + \epsilon \sum_{1 \leq i \leq n} w_{iN}\left(x, \frac{x}{\epsilon}\right) \partial_i u^\star_N(x),$$

where w_{iN} is the RB approximation for w_i as defined by (5-II.22) and $u^\star_N$ is an approximation for $u^\star$ that is a solution in $W_{h_{hom}}$ for the discrete variational problem

$$\int_\Omega \bar{\bar{A}}^\star_N \nabla u^\star_N \cdot \nabla v = \int_\Omega f v + \int_{\Gamma_N} v,\ \forall v \in W_{h_{hom}}. \tag{5-III.18}$$

We have the following result.

Lemma 4. *Assume that Γ_D is a measurable subset of $\partial\Omega$ with a positive $(n-1)$-dimensional measure (when $n > 1$) so that a Poincaré inequality holds for elements of the Sobolev space $H^1_{\Gamma_D}(\Omega) = \{v \in H^1(\Omega),\ v|_{\Gamma_D} = 0\}$.*

If the approximations $w_{iN}(x,\cdot)$ converge to $w_i(x,\cdot)$ for all parameter values x in Ω in the sense that

$$\lim_{N \to \infty} \max_{1 \leq i \leq n} \left\{ \operatorname*{esssup}_{x \in \Omega} \|w_i(x,y) - w_{iN}(x,y)\|_X \right\} = 0, \tag{5-III.19}$$

then the asymptotic L^2 homogenized approximation $u_N^\star$ converges to $u^\star$, and so does the approximation for the asymptotic H^1 homogenized approximation of u^ϵ. That is, we have the two results :

$$\lim_{N\longrightarrow\infty} \|u^\star(x)-u_N^\star(x)\|_{L^2(\Omega)}=0 \tag{5-III.20}$$

and

$$\lim_{N\longrightarrow\infty}\lim_{\epsilon\to 0}\left\|u^\star(x)-u_N^\star(x)+\epsilon\sum_{1\le i\le n}\left(w_i\left(x,\frac{x}{\epsilon}\right)\partial_i u^\star(x)-w_{iN}\left(x,\frac{x}{\epsilon}\right)\partial_i u_N^\star(x)\right)\right\|_{H^1(\Omega)}=0, \tag{5-III.21}$$

where the two successive limits cannot be inverted.

Remark 24. As explained in section 5-III-C, the assumption (5-III.19) can barely be satisfied a priori. But in practice, the error bounds derived in the a posteriori analysis of section 5-III-A allow one to check this assumption. The numerical results of section 5-IV even show that the convergence with respect to N in (5-III.19) is exponential.

Remark 25. In practice, the RB approximations w_{iN} and $\bar{\bar{A}}_N^\star$ are computed by using the FE approximations of the offline stage (see section 5-III-B). Then the limits $\lim_{N\longrightarrow\infty}$ in (5-III.19), (5-III.20), and (5-III.21) should in fact be read as the double non-invertible limit process $\lim_{N\longrightarrow\infty}\lim_{\mathcal{N}\longrightarrow\infty}$ (see section 5-III-C).

Proof of Lemma 4. To show this result, let us define two quantities :

$$E_N^{u^\star(x)}=u^\star(x)-u_N^\star(x)$$

and

$$E_N^{\nabla u^\star(x)}=\nabla_x(u^\star-u_N^\star)(x)+\sum_{i=1}^n\left(\nabla_y w_i\left(x,\frac{x}{\epsilon}\right)\partial_i u^\star(x)-\nabla_y w_{iN}\left(x,\frac{x}{\epsilon}\right)\partial_i u_N^\star(x)\right).$$

The approximation errors for the asymptotic L^2 and H^1 homogenized approximation of $u^\epsilon(x)$ now, respectively, are written

$$\|u^\star(x)-u_N^\star(x)\|_{L^2(\Omega)}=\left\|E_N^{u^\star(x)}\right\|_{L^2(\Omega)}$$

and

$$\left\|u^\star(x)-u_N^\star(x)+\epsilon\sum_{1\le i\le n}\left(w_i\left(x,\frac{x}{\epsilon}\right)\partial_i u^\star(x)-w_{iN}\left(x,\frac{x}{\epsilon}\right)\partial_i u_N^\star(x)\right)\right\|_{H^1(\Omega)}$$
$$=\sqrt{\left\|E_N^{u^\star(x)}\right\|^2_{L^2(\Omega)}+\left\|E_N^{\nabla u^\star(x)}\right\|^2_{L^2(\Omega)}}+\underset{\epsilon\to 0}{O}(\epsilon).$$

Thus, the proof consists of the two successive results

$$\lim_{N\longrightarrow\infty}\lim_{\epsilon\to 0}\left\|E_N^{u^\star}\right\|_{L^2(\Omega)}=0 \tag{5-III.22}$$

and

$$\lim_{N\longrightarrow\infty}\lim_{\epsilon\to 0}\left\|E_N^{\nabla u^\star}\right\|_{L^2(\Omega)}=0. \tag{5-III.23}$$

First, let us begin with properties of the homogenized tensor. On account of definition (5-II.14), $\bar{\bar{A}}^\star(x)$ is a positive definite and continuous matrix.

Indeed, for any $x\in\Omega$, $\bar{\bar{A}}^\star(x)$ is positive definite :

$$0<\alpha_A u\cdot u\le\alpha_A\left(u\cdot u+\int_Y\sum_{i=1}^n|u_i[\nabla_y w_i(x,y)]|^2\,dy\right)\le\bar{\bar{A}}^\star(x)u\cdot u\ \forall u\in\mathbb{R}^n$$

since $w_i(x,\cdot)$ is periodic.

Also, there exists a positive constant $\gamma^\star(x)$ such that $\gamma^\star(x)$ is a continuity bound for $\bar{\bar{A}}^\star(x)$:

$$\bar{\bar{A}}^\star(x)u\cdot u \le \gamma_A^{-1}\left(u\cdot u + \int_Y \sum_{i=1}^n |u_i[\nabla_y w_i(x,y)]|^2\right) \le \gamma^\star(x)u\cdot u \ \forall u\in\mathbb{R}^n$$

since the bilinear form in $\mathbb{R}^n\times\mathbb{R}^n$

$$(u,v)\to\int_Y\left(\sum_{i=1}^n u_i[\nabla_y w_i(x,y)]\right).\left(\sum_{j=1}^n v_j[\nabla_y w_j(x,y)]\right)dy$$

is clearly continuous.

Moreover, we have uniform continuity : there exists a real number $\gamma_{A^\star}>0$ such that, for any x in Ω, $\gamma^\star(x)\le\gamma_{A^\star}$.

Second, $u^\star\in W_{h_{hom}}$ and $u_N^\star\in W_{h_{hom}}$ satisfy the variational formulations (5-III.17) and (5-III.18). We then have the following equality :

$$\int_\Omega \bar{\bar{A}}^\star\nabla u^\star\cdot\nabla v = \int_\Omega fv + \int_{\Gamma_N} v = \int_\Omega \bar{\bar{A}}_N^\star\nabla u_N^\star\cdot\nabla v \ \forall v\in W_{h_{hom}}$$

that we rewrite with $v=(u^\star-u_N^\star)$:

$$\int_\Omega \bar{\bar{A}}^\star\nabla(u^\star-u_N^\star)\cdot\nabla(u^\star-u_N^\star) = \int_\Omega(\bar{\bar{A}}_N^\star-\bar{\bar{A}}^\star)\nabla u_N^\star\cdot\nabla(u^\star-u_N^\star).$$

Because of the coercivity of $\bar{\bar{A}}^\star(x)$, we finally have the inequality

$$\alpha_A\|\nabla(u^\star-u_N^\star)\|_{L^2(\Omega)} \le \|\bar{\bar{A}}_N^\star-\bar{\bar{A}}^\star\|_\infty\|\nabla u_N^\star\|_{L^2(\Omega)}.$$

Moreover, the Poincaré inequality for $u^\star-u_N^\star$ in $H^1_{\Gamma_D}(\Omega)$ is written as follows :

$$\|u^\star-u_N^\star\|_{L^2(\Omega)} \le \mathcal{P}\|\nabla(u^\star-u_N^\star)\|_{L^2(\Omega)},$$

with a certain constant $\mathcal{P}$ which depends only on Ω. We have established an error estimate for $\|E_N^{u^\star}\|_{L^2(\Omega)}$.

Next, since $\bar{\bar{A}}^\epsilon(x)$ is a positive definite matrix for any x in Ω, we deduce the following inequality :

$$\alpha_A\|E_N^{\nabla u^\star}\|^2_{L^2(\Omega)} \le \int_\Omega \bar{\bar{A}}\left(x,\frac{x}{\epsilon}\right)E_N^{\nabla u^\star(x)}\cdot E_N^{\nabla u^\star(x)}dx.$$

In the limit $\epsilon\to 0$, on account of the periodicity of $\bar{\bar{A}}(x,\cdot)$ and $\nabla_y w_{iN}(x,\cdot)$, the previous inequality is rewritten

$$\lim_{\epsilon\to 0}\left\|E_N^{\nabla u^\star}\right\|^2_{L^2(\Omega)} \le \iint_{\Omega\times Y}\frac{\bar{\bar{A}}(x,y)}{\alpha_A}\left[\sum_{i=1}^n(e_i+\nabla_y w_i(x,y))\partial_i u^\star(x) \ -(e_i+\nabla_y w_{iN}(x,y))\partial_i u_N^\star(x)\right]^2 dydx.$$

Last, the definition (5-II.14) of the homogenized tensor $\bar{\bar{A}}^\star$ allows us to rewrite the expression

$$\int_Y \bar{\bar{A}}(x,y)\left[\sum_{i=1}^n(e_i+\nabla_y w_i(x,y))\partial_i u^\star(x)-(e_i+\nabla_y w_{iN}(x,y))\partial_i u_N^\star(x)\right]^2 dy,$$

and we finally get the following error estimate :

$$\lim_{\epsilon\to 0}\left\|E_N^{\nabla u^\star}\right\|^2_{L^2(\Omega)} \le \int_\Omega\frac{A^\star}{\alpha_A}\nabla(u^\star-u_N^\star)\cdot\nabla(u^\star-u_N^\star)+\int_\Omega\frac{A_N^\star-A^\star}{\alpha_A}\nabla u_N^\star\cdot\nabla u_N^\star,$$

the superior bound of which is itself superiorly bounded by

$$\frac{\gamma_{A^\star}}{\alpha_A}\|\nabla(u^\star-u_N^\star)\|^2_{L^2(\Omega)}+\frac{1}{\alpha_A}\|\bar{\bar{A}}^\star-\bar{\bar{A}}_N^\star\|_\infty\|\nabla u_N^\star\|^2_{L^2(\Omega)}.$$

In the end, we have the following error estimates :

$$\lim_{\epsilon\to 0}\left\|E_N^{\nabla u^\star}\right\|^2_{L^2(\Omega)} \le \frac{1}{\alpha_A}\left(\|\bar{\bar{A}}^\star-\bar{\bar{A}}^\star_N\|_\infty\frac{\gamma_{A^\star}}{\alpha_A}+1\right)\|\bar{\bar{A}}^\star-\bar{\bar{A}}^\star_N\|_\infty\|\nabla u^\star_N\|^2_{L^2(\Omega)}, \tag{5-III.24}$$

$$\lim_{\epsilon\to 0}\left\|E_N^{u^\star}\right\|^2_{L^2(\Omega)} = \left\|E_N^{u^\star}\right\|^2_{L^2(\Omega)} \le \frac{\mathcal{P}}{\alpha_A}\|\bar{\bar{A}}^\star-\bar{\bar{A}}^\star_N\|_\infty\|\nabla u^\star_N\|^2_{L^2(\Omega)}. \tag{5-III.25}$$

They show that the asymptotic homogenized approximations converge if the approximate homogenized tensor $\bar{\bar{A}}^\star_N$ converges to $\bar{\bar{A}}^\star$.

Now recall that the homogenized tensor $\bar{\bar{A}}^\star_N$ converges to $\bar{\bar{A}}^\star$ if the approximations $w_{iN}(x,\cdot)$ converge to the cell functions $w_i(x,\cdot)$ since we have already obtained the following error estimate :

$$\|\bar{\bar{A}}^\star-\bar{\bar{A}}^\star_N\|_{[L^\infty(\Omega)]^{n\times n}} = \max_{1\le i,j\le n}\left\{\operatorname*{esssup}_{x\in\Omega}|s_{ij}(x)-s^N_{ij}(x)|\right\} \le \gamma_A^{-1}\max_{1\le i\le n}\left\{\operatorname*{esssup}_{x\in\Omega}\|w_i(x,y)-w_{iN}(x,y)\|_X\right\}^2$$

to derive error bounds for the output s. This concludes the proof of Lemma 4.

5-III-E Practical influence of the parameterization

This section is devoted to the preprocessing computed offline and used online in order to quickly assemble the matrices corresponding to projections of the variational formulation (5-II.20) onto the discrete Galerkin approximation spaces.

Indeed, for a given family $\bar{\bar{A}}(x,y)$ of tensors and a given range Ω for parameter x values, the solution subspace $\mathcal{M}$ for cell problems (5-II.20) is completely determined and fixed. Then, from the theoretical point of view, the way functions w in $\mathcal{M}$ explicitly depend on some parameter $x\in\Omega$, which we call the parameterization, should not influence the efficiency of the RB method as a model order reduction technique ; however, it induces the solution subspace $\mathcal{M}$. But, in practice, the explicit parameterization of $\mathcal{M}$ can significantly account for the efficiency of the RB method, because it greatly influences the practical assembling of the matrix and vectors in the Galerkin projection method.

Only piecewise-affine parameterizations (according to the terminology explained hereafter) are treated in this work, which allows for a fast, very accurate, and simple preprocessing of the FE and RB matrices. But the RB method also adapts to other types of parameterizations, by using the extrapolation method introduced in [BNMP04, GMNP07], for instance.

In the case of an affine parameterization, the assembling of the matrix and vectors corresponding to the Galerkin projection of cell problems is always fast and easy. By affine parameterization of the cell problems, we mean that $\bar{\bar{A}}(x,\cdot)$ depends on the parameterization in an affine manner as follows : For any $x\in\Omega$,

$$\bar{\bar{A}}(x,y) = \bar{\bar{A}}_0(y) + \sum_{q=1}^{Z}\Theta^q(x)\bar{\bar{A}}_q(y)\ \forall y\in Y, \tag{5-III.26}$$

where :

- the functions $\Theta^q:\Omega\to\mathbb{R}$ are parameter-dependent coefficient functions,
- the matrices $\bar{\bar{A}}_q(y)$ $(q=0\dots Z)$ define parameter-independent continuous bilinear forms in $X\times X$:

$$a_q(u,v) = \int_Y \bar{\bar{A}}_q(y)\nabla u(y)\cdot\nabla v(y)dy\ \forall (u,v)\in X\times X,$$

and the bilinear form a_0 is coercive.

With such affine parameterizations, the numerical RB treatment of cell problems is straightforward. Let us detail its implementation.

We follow the offline algorithm presented in section 5-III-B. At each step of the offline stage, a cell problem (5-II.13) for some parameter value in $\mathcal{D}=\{x_k,\ 1\le k\le p\}$ is to be explicitly solved in order to build the reduced basis $(\xi_j(y))_{1\le j\le N}$. For this, a FE method is used that consists of accurately solving the variational formulation (5-II.20) with conforming $\mathbb{P}_1$ Lagrange finite elements and a fine mesh for Y. That is, the solution space X is discretized into the vector space of continuous, piecewise linear functions. Let $\mathcal{T}_Y$ be a conformal mesh for the n-torus $Y=[0,1]^n$ made of $\mathcal{N}_e$ elements $(\Sigma_k)_{1\le k\le\mathcal{N}_e}$ of size h_Y. We write ϕ_k the FE (nodal) basis functions associated with the $\mathcal{N}$ nodes y_k in Y. Now, for $0\le q\le Z$, we define the FE matrices $M_q\in\mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ with entries :

$$(M_q)_{ij} = a_k(\phi_i,\phi_j)$$

for any $1 \leq i,j \leq \mathcal{N}$ and n FE vectors $F_{q,l} \in \mathbb{R}^{\mathcal{N}}$ $(1 \leq l \leq n)$ with entries :

$$(F_{k,l})_i = \int_Y \bar{\bar{A}}_k(y) e_l \cdot \nabla_y \phi_i(y) dy$$

for any $1 \leq i \leq \mathcal{N}$.

In the offline stage, after selecting parameters $x \in \mathcal{D}$ and $1 \leq i \leq n$, we easily compute the FE approximate solutions

$$w_{i\mathcal{N}}(x,y) \in X_{\mathcal{N}} = \sum_{k=1}^{\mathcal{N}} w_{i\mathcal{N}k}(x)\phi_k(y)$$

to the cell problem (5-II.20) discretized by :

$$\left(M_0 + \sum_{q=1}^{Z} \Theta^q(x) M_q \right) w_{i\mathcal{N}}(x,y_l) = \left(F_{0,i} + \sum_{k=1}^{Z} \Theta^q(x) F_{q,i} \right)_l, \ 1 \leq l \leq \mathcal{N}, \tag{5-III.27}$$

by using the FE basis functions associated to nodes y_l. So the FE problem (5-III.27) can be quickly assembled through linear combinations of the precomputed matrices $(M_q)_{0 \leq q \leq Z}$ and vectors $(F_{q,i})_{0 \leq k \leq Z}$, $1 \leq i \leq n$, while the need in memory slightly increases.

In the online stage, to treat fast cell problems (5-II.20) for any parameter value $x \in \Omega$, we moreover project the FE matrices and vectors on the RB space $X_{n \times N}$, which renders :

$$M_q^{RB} = \xi^t M_q \xi, 0 \leq q \leq Z$$

and

$$F_q^{RB} = \xi^t F_q, 0 \leq q \leq Z,$$

with ξ the $\mathcal{N} \times (nN)$ matrix with columns ξ_k, $1 \leq k \leq n \times N$. The RB approximation $w_{iN}(x,y) = \sum_{k=1}^{N} w_{iNk}(x)\xi_k(y)$ for the cell function $w(x,y)$ is a solution of the linear system :

$$\left(M_0^{RB} + \sum_{q=1}^{Z} \Theta^q(x) M_q^{RB} \right) w_{i\mathcal{N}}(x,y_l) = \left(F_{0,i}^{RB} + \sum_{k=1}^{Z} \Theta^q(x) F_{q,i}^{RB} \right)_l, \tag{5-III.28}$$

which is quickly assembled through linear combinations in the present affine case. Besides, for $1 \leq i,j \leq n$, the outputs are easily given by

$$s_{ij}(x) = \sum_{l=1}^{nN} \left(F_{0,i}^{RB} + \sum_{q=1}^{Z} \Theta^q(x) F_{q,i}^{RB} \right)_l w_{iNl}(x).$$

So affine parameterization obviously allows for a fast assembling of the RB matrix and vectors in (5-III.28). It can be considered as an ideal frame for an efficient RB method, because the CPU time for the online solution of one cell problem (5-III.28) actually scales like the CPU time for solving a linear system of size N and because the offline stage is actually very short in comparison with a direct FE computation of a large number of cell functions. The possibility of a similar gain of computation time is not so obvious in the case of nonaffine parameterizations, when (5-III.26) is not valid anymore. Then the assembling of matrices and vectors needs more elaborate techniques [BNMP04, GMNP07].

We finally treat the case of piecewise affine parameterizations that can be re-cast into the class of affine parameterizations. The preprocessing that we propose in this case relies on the fact that, in practice, the RB approximations are numerical approximations for FE approximations and not for the "true" cell functions. Then, to apply the RB method to the parameterized FE approximations, it is possible to consider a "global" enlarged parameterization for the cell problem, which is made of parameters for the oscillating coefficients plus other parameters for the FE method (for instance, the geometrical features of the mesh).

Assume that $Y \subset \bigcup_{k=1}^{d} \overline{Y_k^0}$ is covered by d nonoverlapping reference open subsets Y_k^0. For each $x \in \Omega$, we deal with oscillating coefficients $\bar{\bar{A}}(x,\cdot)$ parameterized in a piecewise affine manner as follows :

- The cell Y can be partitioned into d nonoverlapping open subsets $Y_k(x)$ such that $Y \subset \bigcup_{k=1}^{d} \overline{Y_k}(x)$,
- there exist d affine homeomorphisms $\Phi_k(x,\cdot) : Y_k^0 \to Y_k(x)$, $1 \leq k \leq d$,

- and, for every $1 \leq k \leq d$, the family of functions $(\bar{\bar{A}}(x,\Phi(x,\cdot)))_{x\in\Omega}$ restricted to Y_k^0 can be parameterized in an affine manner as defined in (5-III.26) by

$$\bar{\bar{A}}(x,\Phi(x,y)) = \bar{\bar{A}}_0(y) + \sum_{q=1}^{Z} \Theta^q(x)\bar{\bar{A}}_q(y) \ \forall y \in Y_k^0. \tag{5-III.29}$$

The function $\Phi(x,\cdot)$, defined almost everywhere in Y by

$$\Phi(x,y) = \Phi_k(x,y) \ \forall y \in Y_k^0, 1 \leq k \leq d,$$

maps a "reference" cell onto the cell with parameter value x. By using such a mapping, the family of cell problems defined with these piecewise affine oscillating coefficients can then be treated as if the parameterization was affine like in (5-III.26), provided one can easily take into account the volume deformation in subdomains $Y_k(x)$, which is expressed through the Jacobian determinants of $\Phi_k(x,\cdot)$.

For this, we define $2(Z+1)$ tensors of rank 3, $(\overline{\overline{\overline{M}}}_k)_{1\leq k\leq Z+1}$ and $(\overline{\overline{\overline{F}}}_k)_{1\leq k\leq Z+1}$, thanks to a morphism between $\mathcal{N}_e$ sums over one single element Σ_1 of the mesh $\mathcal{T}_Y$ in $(\mathbb{R}^{\mathcal{N}*\mathcal{N}})^{\mathcal{N}_e}$ and $(\mathbb{R}^{\mathcal{N}_e*\mathcal{N}*\mathcal{N}})$, f_1, $1 \leq 1 \leq \mathcal{N}_e$ being the canonical basis functions of $\mathbb{R}^{\mathcal{N}_e}$:

$$\overline{\overline{\overline{M}}}_k = \sum_{l=1}^{\mathcal{N}_e}\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{N}} \left(\int_{\Sigma_l} \bar{\bar{A}}_k(y)\nabla_y\phi_i(y)\cdot\nabla_y\phi_j(y)dy \right) f_l \otimes e_i \otimes e_j, \tag{5-III.30}$$

$$\overline{\overline{\overline{F}}}_k = \sum_{l=1}^{\mathcal{N}_e}\sum_{i=1}^{\mathcal{N}}\sum_{j=1}^{\mathcal{N}} \left(\int_{\Sigma_l} \bar{\bar{A}}_k(y)e_i\cdot\nabla_y\phi_j(y)dy \right) f_l \otimes e_i \otimes e_j. \tag{5-III.31}$$

An accurate preprocessing in the piecewise affine cases is then possible that assembles fast $\mathbb{P}_1$-FE matrices and RB projected matrices, by adding only a mapping step to the linear combinations of the affine cases. Namely, since $\Phi(x,\cdot)$ is piecewise affine for all $x \in \Omega$, with the family of vectors $(\overline{V}(x))_{x\in\Omega}$:

$$\overline{V}(x) = \sum_{l=1}^{\mathcal{N}_e} \det\left((\nabla_y\Phi(x,y))|_{\Sigma_l}\right) f_l \ \forall x \in \Omega,$$

we easily get the FE matrix for any parameter value x in Ω after the following contraction :

$$\overline{V}(x)\cdot\left(\overline{\overline{\overline{M}}}_0 + \sum_{q=1}^{Z}\Theta^q(x)\overline{\overline{\overline{M}}}_q\right),$$

and so on for the RB matrix. Note that, in the piecewise affine cases, it may also be desirable to reorthonormalize the reduced basis at each parameter value x in Ω, because the FE basis functions are not normalized anymore with respect to the L^2 inner product in Y after mapping.

5-IV Numerical results

We now show numerical results for the RB approximation of a seemingly nonaffine two-dimensional problem that is brought back to the affine setting after mapping of the cell Y. We do not show the RB treatment with *magic points* of more general piecewise continuous parameterizations in this work, but elementary results for one-dimensional problems can be found in [Boy09].

The following two-dimensional model problem is chosen here to show the efficiency of the RB method in a classical situation for the homogenization theory. To fix ideas, it consists of homogenizing the conductivity of a heterogeneous composite material in a domain Ω, where a two-dimensional matrix is full of inclusions with varying positions and conductivity properties.

5-IV-A Definition of the problem

For $n=2$ and $f=0$, we supply the problem (5-II.1) with the mixed boundary conditions

$$(\text{BC}) \ \begin{cases} u^\epsilon(1,x_2) = 0 \ = u^\epsilon(x_1,1), \\ \bar{\bar{A}}^\epsilon\nabla u^\epsilon\cdot\bar{n}|_{(0,x_2)} = +1 = \bar{\bar{A}}^\epsilon\nabla u^\epsilon\cdot\bar{n}|_{(x_1,0)}. \end{cases} \tag{5-IV.1}$$

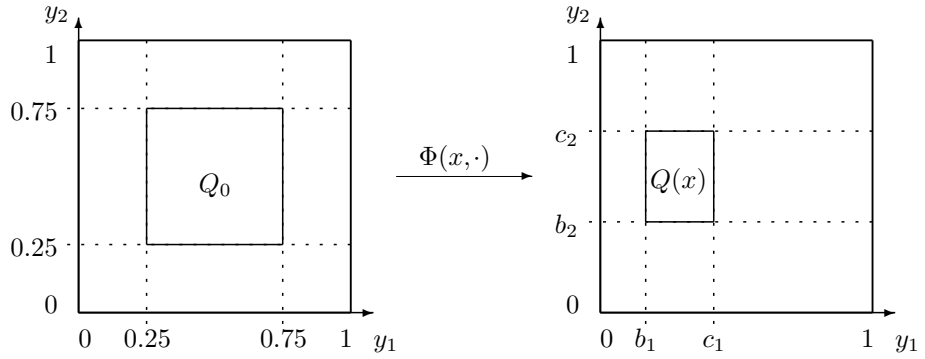


FIG. 5.1 – For each parameter value x, the cell with inclusion $Q(x)$ (on the right) is mapped through the piecewise affine homeomorphism $\Phi(x,\cdot)$ from a reference cell with inclusion Q_0 (on the left).

We define at each point $x\in\Omega$ a single rectangular inclusion $Q(x)\subset Y$ in the cell $Y=[0,1]^2$ by (see Figure 5.1) :

$$Q(x)=\{(y_1,y_2)\in[0,1]^2|0<b_i(x)\leq y_i\leq c_i(x)<1,i=1,2\}$$

We will write $\bar{\bar{I}}_2$ as the second order identity tensor and $\mathbf{1_{Q(x)}}$ the $Q(x)$-test function, which, for every $y\in Y$, $\mathbf{1_{Q(x)}}(y)$ is one if $y\in Q(x)$ and zero otherwise.

We will work with *locally periodic* oscillating coefficients for all $x\in\Omega$:

$$\bar{\bar{A}}^\epsilon(x)=\bar{\bar{I}}_2+\bar{\bar{A}}_1(x,\epsilon^{-1}x), \tag{5-IV.2}$$

where $\bar{\bar{A}}_1(x,y)=\theta(x)\mathbf{1_{Q(x)}}(y)\bar{\bar{I}}_2$ for all y in the periodic cell Y.

We are going to homogenize the problem (5-II.1) with oscillating coefficients (5-IV.2), parameterized by the five-dimensional (5-d) parameter :

$$(b_1,c_1,b_2,c_2,\theta):\Omega\rightarrow[.25-\delta;.25+\delta]^2\times[.75-\delta;.75+\delta]^2\times[-\theta^0;0],$$

where it remains to choose δ and θ^0, respectively, in $]0;.25[$ and $]0;1[$.

For the FE matrices to be easily assembled, we define a "reference" cell problem with a centered inclusion $Q_0=[0.25;0.75]^2$ (see Figure 5.1). Then, at each point $x\in\Omega$, the inclusion $Q(x)$ can be mapped to Q_0 as explained in section 5-III-E.

5-IV-B Offline computations

A FE approach is developed for mapped cell problems in Y with the reference inclusion Q_0. More precisely, we use classical $\mathbb{P}_1$ simplicial Lagrange finite elements on a quadrangular, uniform, and affine FE mesh, divided in isosceles triangles with the base along direction $y_2=-y_1$ and size h_Y in each direction e_1 and e_2. The mesh is fixed and adapted to the reference domain in the sense that the boundaries of the inclusion Q_0 are multiples of h_Y.

We choose a random initial sample $\mathcal{D}$ of $p=50$ parameter values that is uniformly distributed over the 5-d parameter range. A reduced basis is then built with selected solutions $w_i(x_k,\Phi(x_k,y))$ by using the mapping $\Phi(x,\cdot)$ and the algorithm of section 5-III-B. Numerical results are shown for $\delta=.1$ and $h_Y=.1$ in Figures 5.2 and 5.3, where the contrast between coefficients inside and outside the inclusions grows up to 99% ($\theta^0=.99$).

The relative a posteriori error bounds for the RB approximations at the parameter values of the initial sample are computed at each step of the offline algorithm. The maximal error bound in this initial sample decreases exponentially with the size N of the reduced basis (Figure 5.2). The effectivity of the a posteriori estimation is checked all along the RB construction (we found $\eta_N(w_i(x_k,\cdot))\in]1.4;3.5[$ for all $1\leq k\leq p$, $1\leq i\leq n$, and $x_k\in\Lambda$ in the numerical experiment corresponding to Figure 5.2). Note that the offline algorithm (almost always) alternatively selects cell functions for both cell problems, in directions e_1 and e_2 (in the experiment of Figure 5.2 when $N=2$, only one cell function was chosen to optimally span the reduced basis using the algorithm of section 5-III-B, which strongly amplifies the RB approximation errors for the cell problem corresponding to the other direction).

The reduced basis is then tested for another sample $\Lambda=\{x_k,\,1\leq k\leq p\}$ of parameter values in Ω. The maximal a posteriori error in this test sample still decreases very quickly, but the rate of decrease is slightly smaller than

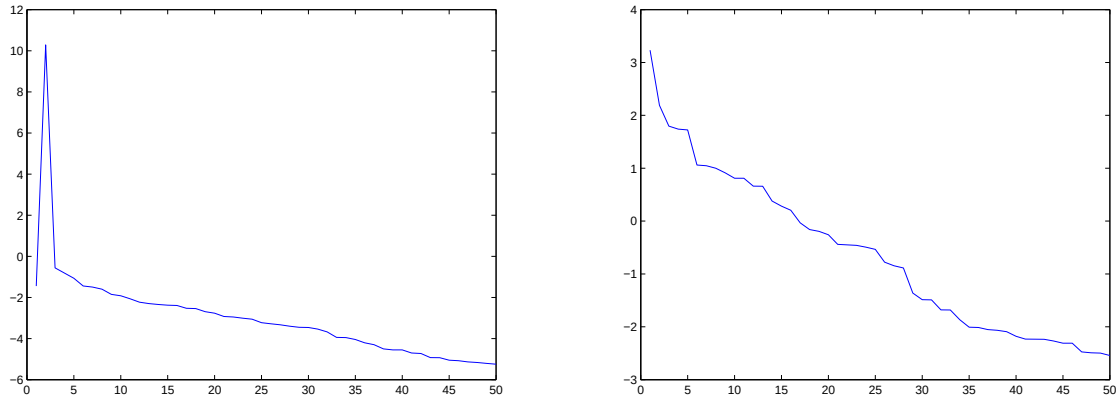


FIG. 5.2 – Maximal relative error bound $\max_{1\leq i\leq 2,\{x_k\}} \frac{\Delta_N(w_{iN}(x_k,\cdot))}{\|w_{iN}(x_k,\cdot)\|_X}$ for the RB approximations $w_{iN}(x_k,\cdot)$ when x_k belongs to the training sample $\mathcal{D}$ used along the RB construction (left picture) and when x_k belongs to a test sample $\Lambda \subset \Omega$ different from $\mathcal{D}$ (right picture). The error is shown in log scale with respect to the size N of the RB approximation space X_N.

TAB. 5.1 – CPU time (in seconds) needed by a Matlab code with an Intel Pentium IV processor (3.0 GHz/1 Go) to approximate the FE matrix for the homogenized problem either with a direct FE approach or with a RB method. In the RB approach, one has to take into account the RB construction (offline algorithm with a sample of p parameter values), the online RB computation of one homogenized solution plus the online a posteriori estimation error (hence the two terms, RB solution + estimation, in the RB online column).

Ratio $N/\mathcal{N}$	$(p=50)$ h_Y	Offline algorithm	$(\frac{h_{hom}}{\epsilon}=\frac{3}{2})$ ϵ	RB for $\bar{\bar{A}}^\star_N$ (online)	FE for $\bar{\bar{A}}^\star_{\mathcal{N}}$ (direct)
1/5	$1E^{-1}$	17 s	$2E^{-2}$	4+3 = 7 s	27 s
1/5	$1E^{-1}$	15 s	$2E^{-3}$	410+330 = 740 s	3100 s
1/20	$5E^{-2}$	42 s	$2E^{-2}$	16+10 = 26 s	520 s
1/20	$5E^{-2}$	53 s	$2E^{-3}$	1600+1000 = 2600 s	37000 s

that of the initial sample used for the RB construction (see Figure 5.2). This shows that the initial sample $\mathcal{D}$ was not an optimal choice to compute a reduced basis for all $x \in \Omega$, yet it still allows for an efficient Galerkin approximations by considering the fast rate of decrease of the a posteriori error.

Besides, the output approximation error for this test sample scales as the square of the approximation error for cell functions (see Figure 5.3, obtained for the same numerical experiment as Figure 5.2). That is, the RB approximations are all the more accurate for the outputs, and the approximation errors scale like the error bounds derived in section 5-III-A. The effectivities of the error bounds of section 5-III-A are indeed hardly bigger than one (we found $\eta_N(w_i(x_k,\cdot)) \in]1.3;3.9[$ for all $1 \leq k \leq p$, $1 \leq i \leq n$, and $x_k \in \Lambda$).[3]

5-IV-C Online computations

After building a reduced basis with the greedy algorithm from the previous FE approximations, we use the RB method to compute online RB approximations for cell functions as linear combinations of the RB basis functions. For this online stage, we develop a FE method for the homogenized problem (5-II.7) and use classical $\mathbb{P}_1$ simplicial Lagrange finite elements on a uniform quadrangular FE mesh divided into isosceles triangles with a base along direction $x_2 = -x_1$ and size h_{hom} in each direction e_1 and e_2.

The RB computations are performed in the step of the numerical homogenization strategy where the values of the homogenized coefficients are collected, as outputs of the cell functions, at some quadrature points in Ω that are necessary for the computations of the entries of the FE matrix in the homogenized problem (5-II.7). The CPU time needed for computing these outputs is compared between the RB and FE methods (Table 5.1),

[3]Note that the maximal relative a posteriori error bound and the maximal error in Figures 5.2 and 5.3 are not obtained for the same parameter values, hence the discrepancy between their ratio and the different effectivities $\eta_N(w_i(x_k,\cdot))$ we computed.

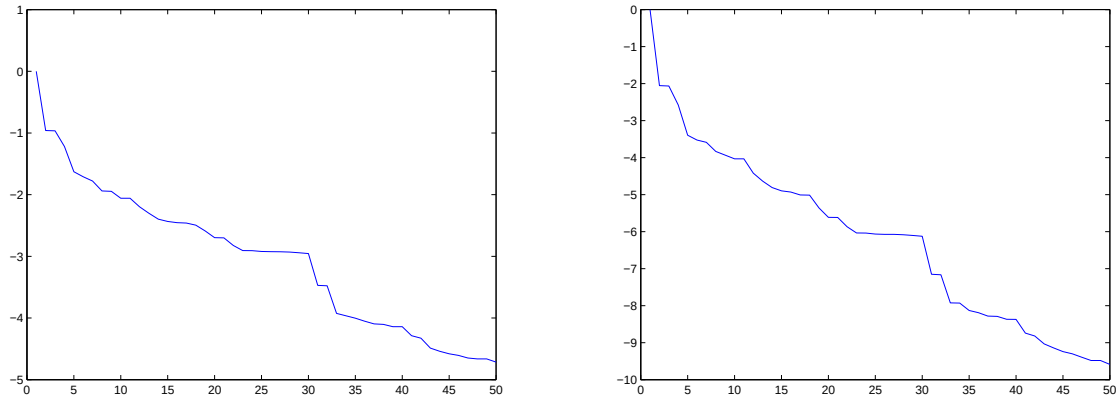


FIG. 5.3 – Maximal relative errors $\max_{1\leq i\leq 2, x_k\in\Lambda} \frac{\|w_i(x_k,\cdot)-w_{iN}(x_k,\cdot)\|_X}{\|w_i(x_k,\cdot)\|_X}$ (left picture) and $\max_{1\leq i,j\leq 2, x_k\in\Lambda} \frac{|s_{ij}(x_k)-s_{ij}^N(x_k)|}{|s_{ij}(x_k)|}$ (right picture) for the random test sample Λ shown in log scale with respect to the size N of the RB approximation space X_N.

TAB. 5.2 – Theoretical correction error for the homogenized solution and the RB numerical approximation error for the homogenized solution when $\delta = .2$, $\theta^0 = .99$, $p = 50$, and $N = 20$.

$(\frac{h_{hom}}{\epsilon} = \frac{3}{2})$	$\|u^\epsilon - u^\star\|_{L^2}$	$\|u^\star_\mathcal{N} - u^\star_N\|_{H^1}$	$\|\nabla r_\epsilon\|_{L^2}$	$\|\nabla_y(w_{i\mathcal{N}} - w_{iN})\|_{L^2}$
(theory)	$\leq C_1\epsilon$		$\leq C_2\sqrt{\epsilon}$	
$h_Y = 1E^{-1}$	$(\epsilon = 2.0E^{-2})$	$1.2E^{-4}$	$(\sqrt{\epsilon} = 1.4E^{-1})$	$2.9E^{-2}$
$h_Y = 1E^{-1}$	$(\epsilon = 2.0E^{-3})$	$4.7E^{-3}$	$(\sqrt{\epsilon} = 4.5E^{-2})$	$1.0E^{-2}$
$h_Y = 5E^{-2}$	$(\epsilon = 2.0E^{-2})$	$3.1E^{-3}$	$(\sqrt{\epsilon} = 1.4E^{-1})$	$8.6E^{-5}$
$h_Y = 5E^{-2}$	$(\epsilon = 2.0E^{-3})$	$1.1E^{-3}$	$(\sqrt{\epsilon} = 4.5E^{-2})$	$3.0E^{-2}$

where the RB method includes an online a posteriori estimation of the approximation error.

In the calculations of Table 5.1, the RB method has been applied with a reduced basis of size $N = 20$ initially trained with a parameter sample $\mathcal{D}$ of size $p = 50$. The main result of Table 5.1 is that the ratio of the RB computation time on the FE computation time scales like $N/\mathcal{N}$, the ratio of the numbers of degrees of liberty in the RB and (direct) FE methods for the cell problem.

So we can distinguish between two main regimes. The less favorable regime is the case of large ratios $N/\mathcal{N}$, which corresponds to cases where one needs only small precision for the homogenized solution and its correction (5-II.10) (large h_Y). Then the RB method is likely to be faster, but only slightly faster, than a direct FE method. So it is likely to be an interesting approach in the frame of many queries of the homogenized solutions only, when the computation time spent by the offline algorithm is not even an issue. It is then possible to enlarge the initial parameter sample $\mathcal{D}$ (take a larger p), which increases the computation time spent by the offline algorithm in the RB construction but improves the quality of the RB approximations.

On the contrary, the favorable regime corresponds to small ratios $N/\mathcal{N}$, where the correction is sought very accurately (small h_Y, or large $\mathcal{N}$). In our numerical example, for instance, taking $\mathcal{N} = 400$ ($h_Y = 5E^{-2}$) instead of $\mathcal{N} = 100$ ($h_Y = 1E^{-1}$) allows one to improve the accuracy of the homogenized coefficients from the third digit and the accuracy of the homogenized solution from the second digit (in accordance with the error estimate (5-III.24)). Then, in this regime where accurate homogenized results are wanted, the numerical results for the RB approximations of the cell functions show that there is an important gain of computation time, while there is no significant loss of numerical precision (Table 5.2).

5-V Conclusion and perspectives

We have shown in the present work that, for a prototypical class of parameterized cell problems (with piecewise affine oscillating coefficients), the reduced-basis approach applies and significantly reduces the time needed to compute a large number of parameterized cell problems in homogenization, in comparison with a

direct FE method.

But many interesting questions remain concerning the extension of the RB approach in homogenization to a larger class of parameterized cell problems. In particular, other geometries for more realistic cell problems should now be addressed, as well as other boundary conditions for the cell problems (including the treatment of oversampling techniques) and less regular oscillating coefficients (with many inclusions in varying the amount). Also, the same questions as those examined in the present work for scalar elliptic equations could be asked for the equations of linear elasticity in two- and three-dimensional contexts. Further developments, but also improvements of the RB methodology, such as the fast estimation of tight parameterized coercivity constants α_A, are then needed that may lead to interesting (fast and accurate) approaches in homogenization. A major issue in homogenization is the limitation of the time computation, and speeding up the homogenization procedures should inevitably bring new possibilities of refinements, maybe after a clever combination with oversampling techniques.

In any case, we believe that our result is interesting in the frame of many of the commonly used homogenization strategies, namely, all of those that ask for solving a computationally demanding number of parameterized cell problems. This is true provided the type of the parameterization can be handled with our RB approach. Among those homogenization strategies, the two-scale homogenization strategy is well known and much used in practice. That is why we have chosen this frame for our numerical experiments. But other homogenization strategies, which are used for nonlocally periodic oscillating coefficients, can also be treated with a RB approach.

For example, stochastic homogenization also asks for solving a large number of parameterized cell problems in the frame of local approximations of the homogenized tensor [BP04]. The homogenization of locally deformed oscillating coefficients $\bar{\bar{A}}^\epsilon(x) = \bar{\bar{A}}(\Phi_x^{-1}(\epsilon^{-1}x))$, with Φ a diffeomorphism, in the frame of deterministic homogenization [Bri94] or of stochastic homogenization [BLL06], $\bar{\bar{A}}^\epsilon(x) = \bar{\bar{A}}(\Phi_x^{-1}(\epsilon^{-1}x,\omega))$, with ω an element of a probability space, by nature, also demands to solve parameterized cell problems.

More general cases, often computationally demanding, also rely on the computation of a large number of cell problems and offer a frame for an application of the RB approach. Among those homogenization strategies, the HMM that averages over a large number of cell problems could directly make use of our RB approach when cell problems are correctly parameterized. Another one, the MsFEM, also averages over numerous cell problems. Yet the range of geometries for those cell problems is often larger, and it is still not obvious that model order reduction techniques may speed up the MsFEM computations.

Although we have not tested all of the above-mentioned possible improvements, we believe that our work is likely to improve a large number of existing homogenization strategies. Definite conclusions on the validity of our approach in such settings will hopefully be obtained soon.

Acknowledgement 2. *Thanks to C. Le Bris and A. T. Patera for their guidance in this work.*

5-VI Addendum to the Chapter 5

In this section, we present additional results that were not published in the article [Boy08].

5-VI-A One-dimensional tests

We consider the problem (4-I.44) on the one-dimensional range $\Omega =]0,1[$ with $f = 0$.

5-VI-A-a Affine parameter

We choose $a(x,y) = 1 + \phi(x)\cos(2\pi y)$. In the numerical tests, we take $\phi : \mathbb{R} \to]-1,1[$ linear $\phi(x) = \delta x$ with $\delta = .5$ for the sake of simplicity. Then we observe numerically that the RB approximation error converges very fast with respect to the size N of the reduced basis (see Fig. 5.4) : equivalently, the manifold generated by the cell functions is very thin. Note also the sharpness of the *a posteriori* estimator in this simple example, and the

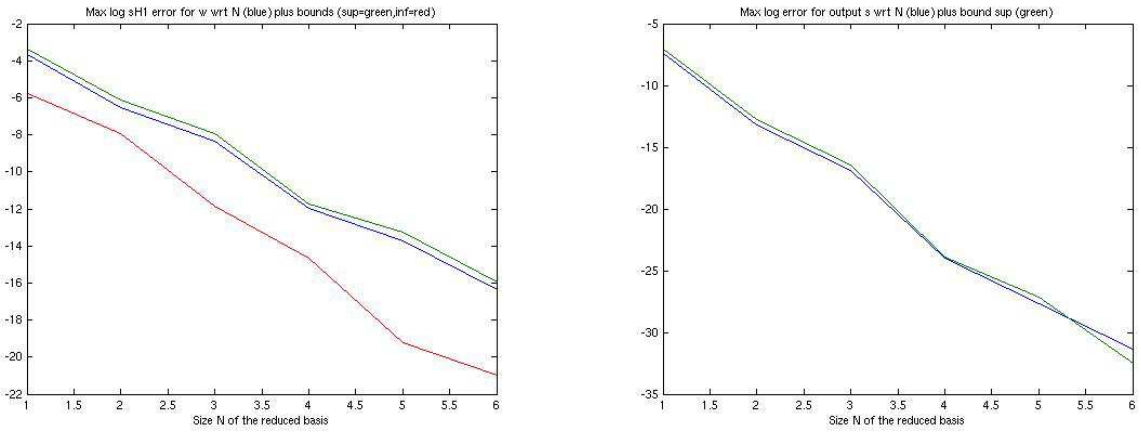


FIG. 5.4 – Left : Maximum of $\|w(x) - w^N(x)\|_X$ with respect to the size N of the reduced basis within a test sample $\Lambda_{\text{test}} \in \Omega$ of parameter values x with size 100 (blue line), with *a posteriori* lower (red) and upper (green) bounds. Right : Maximum of the output $|s(x) - s^N(x)|$ (blue) with *a posteriori* upper bound (green).

quadratic effect in the output RB approximation error compared to the RB approximation error for the cell function, which is similar to what the upper-bound predicts.

5-VI-A-b Smooth non-affine parameter

We choose $a(x,y) = 1 + \phi(x)\cos\left(\frac{2\pi}{T(x)}y\right)$ with $T(x) = 1 + .5\cos(x)$, for all $y \in \boldsymbol{Y}$. So the parameter dependence is no longer affine. [4] In the numerical tests, we take $\phi(x) = \delta x$ and $\delta = 0,05$. The RB approach is used to compute fast the $T(x)$-periodic solutions to the x-parametrized cell problem :

$$-\text{div}_x(a(x,y) \cdot [1 + \nabla_y w(x,y)]) = 0, \forall y \in Y \ ,$$

with output : $a^\star(x) = \int_Y a(x,y) \cdot [1 + \nabla_y w(x,y)] \, dy$.

We tackle the non-affine parameter dependence with the *empirical interpolation* method introduced in [BNMP04] (see Chapter 4). That is, for each parameter value where the solution to the cell function has to computed, additional operations are necessary to compute the parametrized variational formulation, which does not have the right affine decomposition for a direct fast construction. To maintain the rapidity of the computations in the online RB approximation procedure, we thus build a *collateral* reduced basis of size M for the coefficient functions $\{a(x,\cdot) \in L^\infty(Y), \ x \in \Omega\}$, and a fast interpolation procedure using M *magic points* (MP), the computation of which scales like $O(M^2)$. Typically, M is small if the dependence of the non-affine coefficients on the parameter x is smooth, thus the additional operations necessary to compute an approximation to the (discretized) variational formulation using that empirical interpolation procedure are fast. Note yet that the MP method is heuristic ; in particular, it does not ensure the well-posedness of the reduced problem ! In

[4] Note another possible choice of non-affine parameter, $a(x,y) = 1 + \exp(\phi(x)\cos(2\pi y))$ for instance.

practice, the approximated diffusion coefficient in the MP approximate problem sometimes happened to have pointwise negative values, but the discrete system obtained after projection in the subspace $X_{n\times N}$ spanned by the reduced basis was always well-posed. In Fig. 5.5, 5.7 and 5.7 we compare different approximations of the non-affine coefficient, either the L^2 projection or the MP interpolation, using the same collateral reduced basis of size M : there seems to be very little difference for our smooth coefficient. We also observe that the RB approximation error for the cell function solution to the problem approximated by a collateral reduced basis (and its output) decays until saturation at a critical size $N_{\text{crit}}(M)$ of the reduced basis, which depends on the quality of the approximation for the non-affine approximation. The larger M, the larger the critical size $N_{\text{crit}}(M)$, and the lower the RB approximation error at saturation.

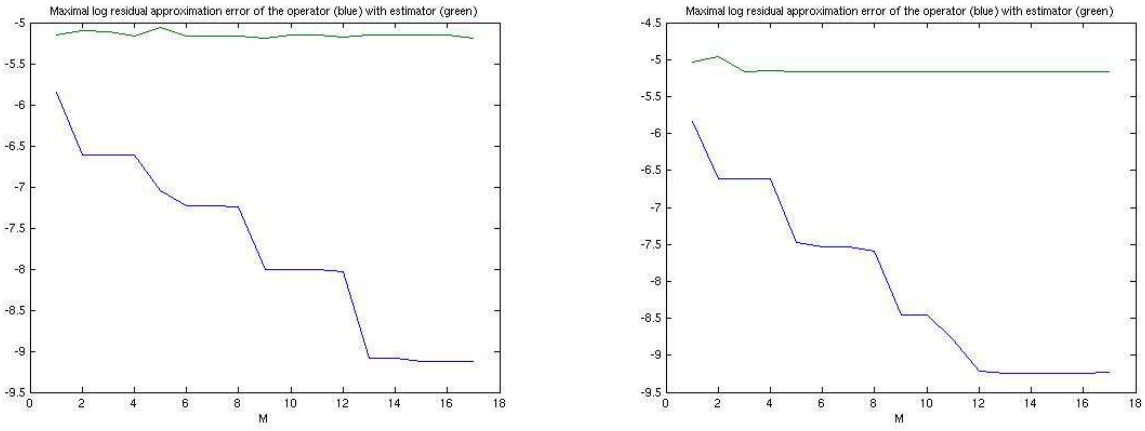


FIG. 5.5 – Left : Maximal $L^\infty(Y)$ error for the magic-point interpolants in the collateral RB space of size M inside a sample test of non-affine coefficient functions, with respect to M. Right : Maximal $L^\infty(Y)$ error for the $L^2(Y)$ projections in the collateral RB space of size M inside a sample test of non-affine coefficient functions, with respect to M. (In green, the *a posteriori* estimator at the $M+1$ magic point proposed in [GMNP07].)

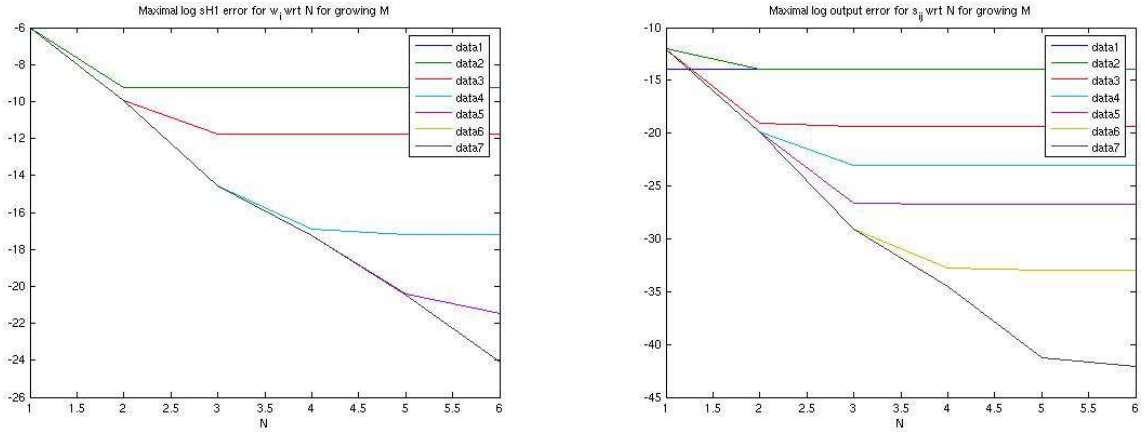


FIG. 5.6 – Left : maximal H^1 error for the RB approximations $w_{N,M}(x,\cdot)$ in a random test sample using empirical interpolation and magic points with different M, with respect to the size N of the reduced-basis. Right : maximal L^∞ error for the RB approximations $s_{N,M}(x)$ in the same random test sample of outputs $s(x)$ using empirical interpolation and magic points with different M, with respect to the size N of the reduced-basis. The line termed `data1` corresponds to $M=1$, etc. until $M=6$, and `data7` corresponds to the error of the RB approximation without empirical interpolation.

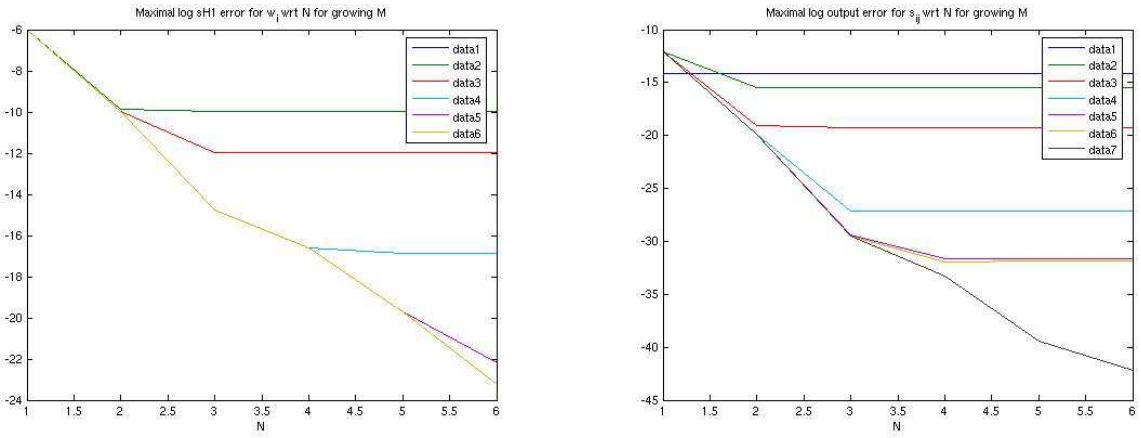


FIG. 5.7 – Left : maximal H^1 error for the RB approximations $w_{N,M}(x,\cdot)$ in a random test sample using empirical interpolation and L^2 projection with different M, with respect to the size N of the reduced-basis. Right : maximal L^∞ error for the RB approximations $s_{N,M}(x)$ in the same random test sample of outputs $s(x)$ using empirical interpolation and L^2 projection with different M, with respect to the size N of the reduced-basis. The line termed `data1` corresponds to $M=1$, etc. until $M=6$, and `data7` corresponds to the error of the RB approximation without empirical interpolation.

5-VI-A-c Non-smooth non-affine parameter

We choose $a(x,y)=1+g(x,y)$ such that for any $x\in]0,1[$

$$g(x,y)=\begin{vmatrix} 0.05x \text{ when } y-[y]\in[0.3+.15x;0.7-.15x] \\ 0 \text{ otherwise.}\end{vmatrix} \tag{5-VI.1}$$

The parameter dependence is clearly not affine : it shows two discontinuities moving with x. We could re-parametrize the discontinuities using two additional *geometric* parameters : then, a piecewise affine mapping of the problem would render an affine parametrization as it was shown previously for the 2-d case. But here, we want to test the empirical interpolation approach. Of course, the difficulty is the control of the $L^\infty(Y)$ norm of $g(x,\cdot)$, which is not a smooth function. Regularization of the coefficient functions $g(x,\cdot)$ slightly improve the numerical results for the RB approximations of the cell functions, but not for the output (see Fig. 5.8).

5-VI-B Two-dimensional tests

Most simple two-dimensional tests do not invoke re-parametrization of the geometry ; typical fast oscillating coefficients are

$$\bar{\bar{A}}^\epsilon(x)=\bar{\bar{I}}_2+\phi(x)\bar{\bar{A}}_1\left(\frac{x}{\epsilon}\right) \tag{5-VI.2}$$

where the function $\bar{\bar{A}}_1(y)=\cos(2\pi y_1)\cos(2\pi y_2)\bar{\bar{I}}_2$ is Y-periodic and where the parameter dependence is clearly affine with respect to $\phi(x)=\delta x_1 x_2$, without requiring a piecewise affine mapping. (In the numerical test, we take $\delta=0.05$, see Fig. 5.9.) Observe then in Fig. 5.10 a typical cell function : it is very regular, in contrast with cell functions where the position and the size of the inclusion change.

Since the problem here is much smoother in the parameter than the one treated in the article [Boy08], we numerically observe a very rapid convergence of the reduced basis (Fig. 5.11).

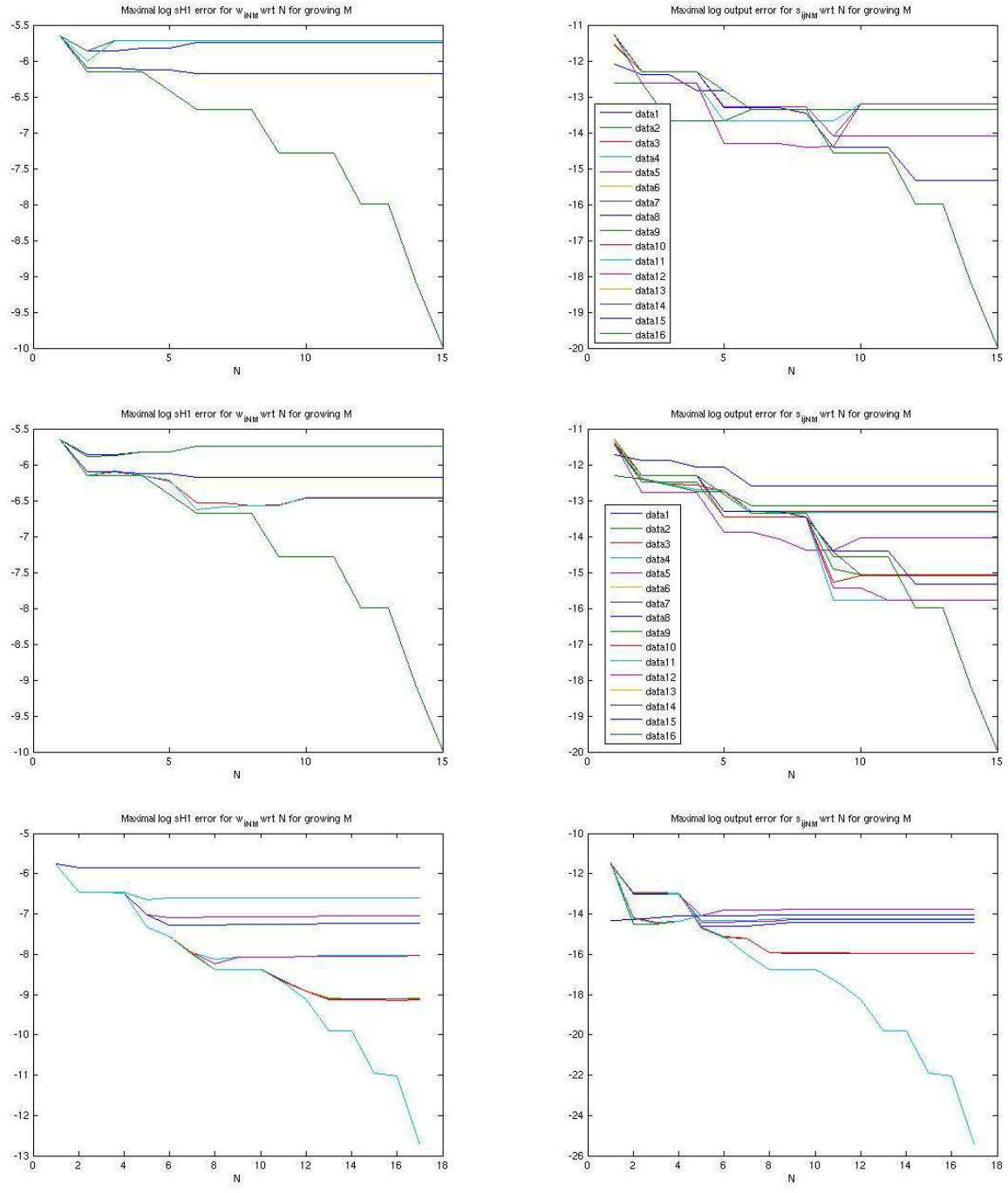


FIG. 5.8 – From top to bottom, Maximal H^1 error for the RB approximations $w_{N,M}(x,\cdot)$ (left) and L^∞ error for the RB approximations $s_{N,M}(x)$ (right) obtained with collateral reduced bases of different sizes M using with the usual MP interpolation (top), L^2 projection (middle) and the usual MP interpolation after regularizing the discontinuities with convolutions (bottom).

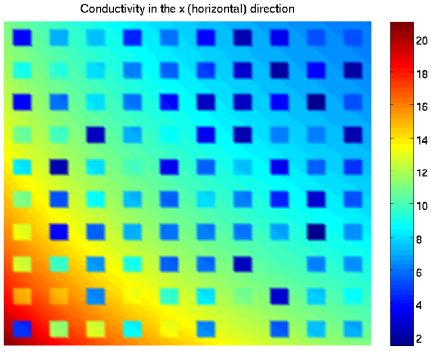


FIG. 5.9 – Fast-oscillating coefficient $\bar{\bar{A}}^{\epsilon}_{11}(x_1,x_2)$ as defined in $(x_1,x_2)\in\Omega$ by (5-VI.2).

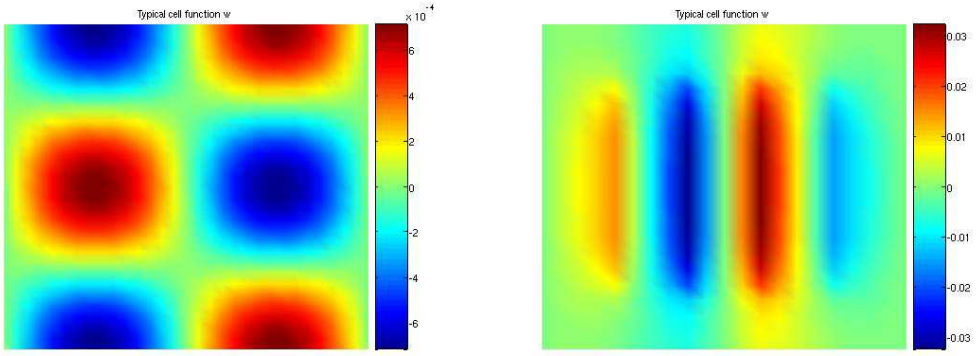


FIG. 5.10 – Left : typical cell function obtained when $\bar{\bar{A}}_1(y)=\cos(2\pi y_1)\cos(2\pi y_2)\bar{\bar{I}}_2$. Right : example of cell function much less regular, obtained when using a moving inclusion (in size and length) to define $\bar{\bar{A}}_1(y)$.

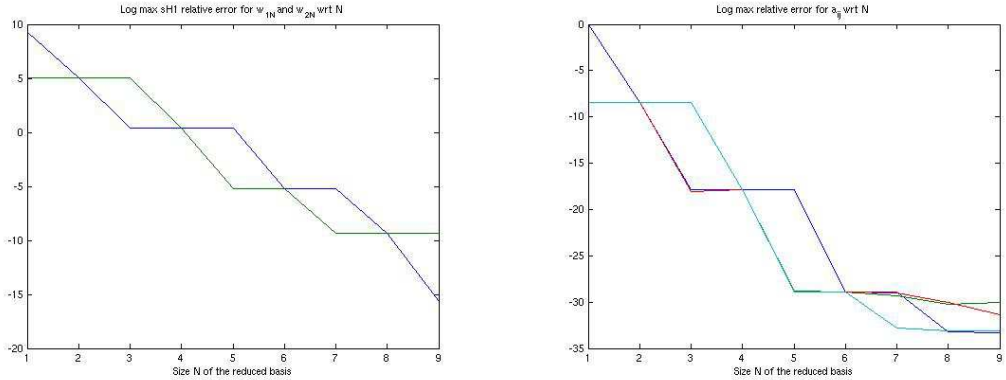


FIG. 5.11 – Maximal relative RB approximation errors $\|w_i(x)-w_i^N(x)\|_X$ (left, blue : $i=1$, green : $i=2$) and $|s_{ij}(x)-s_{ij}^N(x)|$ (right, for the four components), with respect to the size N of the reduced basis, within a test sample Λ_{test}.

Chapitre 6

Approche par bases réduites de problèmes variationnels avec paramètres stochastiques : application à un problème de conduction de chaleur avec un coefficient de Robin variable.

Ce chapitre est la reproduction de [BBM+09]. Les résultats ont été obtenus en collaboration avec C. Le Bris, Y. Maday, N. C. Nguyen et A. T. Patera.

On y utilise une approche Bases réduites (RB) pour résoudre un grand nombre de Problèmes aux bords (BVPs) paramétrés par une donnée d'entrée stochastique — exprimée comme un développement de Karhunen-Loève — en vue de calculer des données de sortie qui sont régulières en le champ aléatoire solution. La méthode RB proposée ici pour des problèmes variationnels paramétrés par des coefficients stochastiques est très similaire à l'approche RB standard développée antérieurement pour des problèmes déterministes. Cependant, le cadre stochastique requiert le développement de nouveaux estimateurs *a posteriori* pour des données de sortie "statistiques" — par exemple ici, les deux premiers moments de fonctionnelles intégrales du champ aléatoire solution, estimés en fonction d'un nombre variable (à ajuster selon la précision voulue) de données d'entrées "stochastiques". Ces bornes d'erreurs permettent donc un échantillonnage efficace des paramètres d'entrée stochastiques et un calcul rapide et fiable des données de sortie, en particulier quand le calcul des données de sortie est réitéré pour de nombreuses valeurs des autres paramètres du problème (distincts des coefficients stochastiques).

A Reduced Basis Approach for Variational Problems with Stochastic Parameters : Application to Heat Conduction with Variable Robin Coefficient

Sébastien Boyaval[a,b], Claude Le Bris[a,b], Yvon Maday[c,d], Ngoc Cuong Nguyen[e], Anthony T. Patera[e]

[a]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).

[b]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

[c]Université Pierre et Marie Curie, Laboratoire Jacques-Louis Lions UMR 7598, (Université Paris 6, F-75005, Paris, France).

[d]Brown University, Division of Applied Mathematics (Providence, RI 02912 USA).

[e]Massachusetts Institute of Technology, Dept. of Mechanical Engineering, (Cambridge, MA 02139 USA).

In this work, a Reduced Basis (RB) approach is used to solve a large number of boundary value problems parametrized by a stochastic input — expressed as a Karhunen–Loève expansion — in order to compute outputs that are smooth functionals of the random solution fields. The RB method proposed here for variational problems parametrized by stochastic coefficients bears many similarities to the RB approach developed previously for deterministic systems. However, the stochastic framework requires the development of new *a posteriori* estimates for "statistical" outputs — such as the first two moments of integrals of the random solution fields ; these error bounds, in turn, permit efficient sampling of the input stochastic parameters and fast reliable computation of the outputs in particular in the many-query context.

Keywords : Stochastic Parameterized Partial Differential Equations ; Karhunen-Loève ; Monte Carlo ; Reduced Basis Method ; *A posteriori* error estimation.

6-I Introduction

6-I-A Overview

Let $U(\cdot,\omega)$ be a scalar random field solution to a (presumed well-posed) Boundary Value Problem (BVP) involving a Stochastic Partial Differential Equation (SPDE). For instance, if variations in the probability space $(\Omega,\mathcal{F},\mathbb{P})$ are denoted by the variable ω, we take $U(\cdot,\omega)$ as the $\mathbb{P}$-almost sure (a.s.) solution to the Partial Differential Equation (PDE) in a (smooth) physical domain $\mathcal{D}$

$$-\mathrm{div}\,(\boldsymbol{a}(x)\nabla U(x,\omega))=0\,,\ \forall x\in\mathcal{D} \tag{6-I.1}$$

supplied with a stochastic Robin Boundary Condition (BC) on the boundary $\partial\mathcal{D}$ parametrized by a random input field $\mathrm{Bi}(\cdot,\omega)$

$$\boldsymbol{n}(x)^{\mathrm{T}}\,\boldsymbol{a}(x)\nabla U(x,\omega)+\mathrm{Bi}(x,\omega)\,U(x,\omega)=g(x)\,,\ \forall x\in\partial\mathcal{D}\ . \tag{6-I.2}$$

Here, $\boldsymbol{a}$ takes symmetric positive definite matrix values, the random field $\mathrm{Bi}(\cdot,\omega)$ (Biot number [LL02]) is non-zero (non-degenerate positive) on some subset $\Gamma_{\mathrm{B}}\subset\partial\mathcal{D}$ (with non-zero measure), $\boldsymbol{n}(x)$ is the outward unit normal at $x\in\partial\mathcal{D}$ and T denotes the transpose.

We consider the rapid and reliable computation of statistical outputs associated with $U(\cdot,\omega)$ such as the expected value $\mathbf{E}_{\mathbb{P}}(S)$ and the variance $\mathbf{Var}_{\mathbb{P}}(S)$ of a random variable $S(\omega)=\mathcal{E}(U(\cdot,\omega))$ given by a linear (scalar) functional $\mathcal{E}$ of the trace of $U(\cdot,\omega)$ on $\Gamma_{\mathrm{R}}\subset\partial\mathcal{D}$ (where $\Gamma_{\mathrm{R}}\cap\Gamma_{\mathrm{B}}=\emptyset$)

$$\mathcal{E}(U(\cdot,\omega))=\int_{\Gamma_{\mathrm{R}}}U(\cdot,\omega)\ . \tag{6-I.3}$$

One possible strategy is to evaluate the statistical outputs as Monte-Carlo (MC) sums of the random variable S

$$E_M[S]=\frac{1}{M}\sum_{m=1}^{M}S^m,\qquad V_M[S]=\frac{1}{M-1}\sum_{m=1}^{M}\left(E_M[S]-S^m\right)^2\ , \tag{6-I.4}$$

using M independent random variables $(S^m)_{1\leq m\leq M}$ with the same distribution law as S. But M can be very large, and hence these MC evaluations can be very demanding (for each m, one must solve a BVP PDE in

$\mathcal{D}$). Furthermore, in actual practice, and as developed subsequently in this paper, we are often interested in evaluating our statistical outputs for different values of deterministic parameters, say ϱ — which even further increases the computational challenge. For this reason we develop a Reduced Basis (RB) approach : to decrease the computational cost of the M realizations of the Finite Element (FE) approxiations $U_{\mathcal{N}}(\cdot,\omega)\approx U(\cdot,\omega)$ required in the Monte-Carlo sums.

Toward this goal, we first rewrite the parametrization of the BVP using a Karhunen–Loève (KL) expansion of the random input field (see Section 2 for details)

$$\mathrm{Bi}(x,\omega)=\overline{\mathrm{Bi}}\left(G(x)+\sum_{k=1}^{\mathcal{K}}\Phi_k(x)\,Y_k(\omega)\right)\ ,\ \forall x\in\partial\mathcal{D}, \tag{6-I.5}$$

where $\mathcal{K}$ is the rank (possibly infinite) of the covariance operator for $\mathrm{Bi}(\cdot,\omega)$ with eigenvectors $(\Phi_k)_{1\le k\le\mathcal{K}}$, the positive number $\overline{\mathrm{Bi}}=\int_\Omega d\mathbb{P}(\omega)\int_{\partial\mathcal{D}}\mathrm{Bi}(\cdot,\omega)$ is an intensity factor and the random variables $(Y_k)_{1\le k\le\mathcal{K}}$ are mutually uncorrelated in $L^2_{\mathbb{P}}(\Omega)$ with zero mean. Next, we define a function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ parametrized by $\overline{\mathrm{Bi}}\in\mathbb{R}_{>0}$ and the (possibly infinite) sequence $y=(y_1,y_2,\ldots)\in\Lambda^y\subset\mathbb{R}^{\mathcal{K}}$

$$\mathrm{bi}(x;\overline{\mathrm{Bi}},y)=\overline{\mathrm{Bi}}\left(G(x)+\sum_{k=1}^{\mathcal{K}}\Phi_k(x)y_k\right)\ ,\ \forall x\in\partial\mathcal{D}, \tag{6-I.6}$$

such that for all $\overline{\mathrm{Bi}}\in\mathbb{R}_{>0}$ and $y\in\Lambda^y$ the parametrized function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ is well defined ; we also define truncated versions $y^K=(y_1,y_2,\ldots,y_K,0,0,\ldots)\in\Lambda^y$ up to order $K\le\mathcal{K}$ of the deterministic parameter sequence y.

For any positive integer $K\le\mathcal{K}$, we then define a solution $U_K(\cdot,\omega)$ to the BVP in which the KL expansion of $\mathrm{Bi}(\cdot,\omega)$ in the Robin BCs is replaced by a truncated version at order K

$$\mathrm{Bi}_K(\cdot,\omega)=\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},Y^K(\omega)),$$

using truncated versions Y^K (with $K\le\mathcal{K}$) of the (possibly infinite) sequence $Y=(Y_k)_{1\le k\le\mathcal{K}}$ of random variables. For almost all fixed $x\in\mathcal{D}$, the random variable $U_K(x,\cdot)$ is clearly $\sigma\left(Y^K\right)$-measurable and, by the Doob-Dynkin lemma [Oks03], we have $U_K(x,\omega)=u_K(x;Y^K(\omega))$ for almost all $(x,\omega)\in\mathcal{D}\times\Omega$, where $u_K(\cdot;y^K)$ solves a y^K-parametrized BVP PDE problem ($y^K\in\Lambda^y$) :

$$\begin{cases}-\mathrm{div}\left(\boldsymbol{a}(x)\nabla u_K(x;y^K)\right)=0\,,\ \forall x\in\mathcal{D}\ ,\\ \boldsymbol{n}(x)^{\mathrm{T}}\boldsymbol{a}(x)\nabla u_K(x;y^K)+\mathrm{bi}(x;\overline{\mathrm{Bi}},y^K)u_K(x;y^K)=g(x)\,,\ \forall x\in\partial\mathcal{D}\ .\end{cases} \tag{6-I.7}$$

The problem (6-I.7) is well-posed under standard hypotheses for all $y^K\in\Lambda^y$ in the range of Y.

The statistical outputs for $S_K(\omega)=\mathcal{E}\left(U_K(\cdot,\omega)\right)$ obtained after truncation of the KL expansion

$$E_M[S_K]=\frac{1}{M}\sum_{m=1}^{M}S_K^m,\qquad V_M[S_K]=\frac{1}{M-1}\sum_{m=1}^{M}\left(E_M[S_K]-S_K^m\right)^2\ , \tag{6-I.8}$$

can then be obtained as, respectively, $E_M[s_K(Y^K)]$ and $V_M[s_K(Y^K)]$, using $s_K(y^K)=\mathcal{E}\left(u_K(\,\cdot\,;y^K)\right)$ and M independent random vectors $\left(Y_m^K\right)_{1\le m\le M}$ with the same distribution law as Y^K. Clearly, the error in these outputs due to truncation of the KL expansion must be assessed ; we discuss this issue further below. (We must also ensure that M is large enough ; we address this question in the context of our numerical results.)

In Section 6-III, we develop a reduced basis (RB) approach [ASB78, FR83, NP80, Por85] for the parametrized (deterministic) BVP (6-I.7) and outputs (6-I.8) for the case in which the random variables Y_k, $1\le k\le K$ ($\le\mathcal{K}$), are bounded (uniformly if $\mathcal{K}=+\infty$) such that the KL expansion is positive for any truncation order K (and converges absolutely a.e. in $\partial\mathcal{D}$ when $\mathcal{K}=+\infty$) ; the latter ensures well-posedness of the BVPs obtained after truncation at any order $1\le K\le\mathcal{K}$. We shall present numerical results for a random input field $\mathrm{Bi}(\cdot,\omega)$ whose spatial autocovariance function is a Gaussian kernel such that the KL spectrum decays rapidly.

In particular, we shall show that our RB approach significantly reduces the computational cost of the MC evaluations with no sensible loss of accuracy compared to a direct Finite Element (FE) approach. For instance, with truncated KL expansions of order $K\le 20$, the RB computational time for solutions to (6-I.7) is reduced by a factor of $\frac{1}{45}$ relative to direct FE, and the (relative) approximation error in the expectation due to both RB and KL truncation is controlled *and* certified to 0.1% (for $K=20$). Our RB approach thus also straightforwardly permits rapid exploration of the dependence of the outputs $E_M[s_K(Y^K)]$ and $V_M[s_K(Y^K)]$ on variations in additional *deterministic* parameters ϱ entering the problem. (In the limit of many evaluations at different ϱ, computational savings relative to FE can be as much as $O(200)$.)

6-I-B Relation to Prior Work

The computation of BVPs involving SPDEs has been identified as a demanding task [BTZ05, DBO01, MK05, DNP+04] for several years, whatever the numerical approach used to discretize the SPDE. For instance, among those numerous numerical approaches, the popular *spectral (stochastic) Galerkin* discretizations [GS91], based on a (generalized) Polynomial Chaos (PC) expansion of the solution [Wie38, XK02], consists in solving a variational problem in a *high-dimensional* tensor-product functional space on $\mathcal{D}\times\Lambda^y$, which is computationally (very) expensive. Hence several reduction techniques have been proposed recently for the spectral Galerkin approach, in particular :

- sparse/adaptive methods [FST05, TS07b],
- efficient iterative algorithms for the fully discretized problems, using parallel solvers, preconditioners and/or Krylov projections [PG00, KM05], sometimes termed "stochastic RB" Krylov methods [NK02, SNK06, MNK08],
- POD approaches for PC discretizations of the functions in the stochastic variable (combined with a two-scale approach in the physical space) [DGRH07, GSD07],
- POD approaches for PC-FE discretizations of the functions defined on the whole tensor-product space, termed "generalized spectral decomposition" [Nou07, Nou08],
- and stochastic collocation approaches [XH05, BNT07, NTW08].

These reduction techniques have shown good performance on test cases. However, the sparse/adaptive methods require substantial implementation efforts, the Krylov methods and the POD approaches do not yet provide rigorous *a posteriori* analysis to control the output approximation error, and the stochastic collocation method still invokes numerous (expensive) FE solutions — at each collocation point. The RB method described here — albeit for a limited class of problems — focuses on simple implementation, rigorous *a posteriori* error bounds, and parsimonious appeal to the FE "truth".

The formulation of the RB method presented herein can be straightforwardly applied to discretizations of the SPDE that lead to the solution of many *decoupled* variational formulations of the BVP on $\mathcal{D}$ for many fixed given values of the random input in Λ^y (like (6-I.7)). In the present work, the RB method is only applied to Monte-Carlo/Galerkin (in fact, Finite-Element) discretizations of the SPDE, as described earlier in this introduction. That is, the statistical outputs like mean and variance of some functional of the random variable solution to the SPDE are computed through Monte-Carlo (MC) evaluations of the random variable $S_K = s_K(Y^K)$, and not through quadrature or collocation formulæ for the (weighted) integration of the function $y^K \to s_K(y^K)$ over $y^K \in \Lambda^y$.

However, the RB method could be applied as well to many numerical approaches where integration in the stochastic space is discretized by collocation at many points in the range of the random input, where at each of these points one has to solve a PDE parametrized *only* by the value of the random input at the same point. In particular, the RB method proposed in this paper can be viewed as an *accelerator* of the stochastic collocation approach described in [BNT07], where a basis of orthogonal polynomials in the stochastic variables is substituted for the standard PC basis. As a matter of fact, the stochastic collocation approach is just a *pseudo-spectral* Galerkin discretization : it applies quadrature formulæ for the computation of the ouputs $\mathbf{E}_{\mathbb{P}}\big(s_K(Y^K)\big)$ and $\mathbf{Var}_{\mathbb{P}}\big(s_K(Y^K)\big)$ so as to split the variational formulation for $u_K(\cdot;\cdot)$ on the high-dimensional tensor-product space $(x,y^K)\in\mathcal{D}\times\Lambda^y$ into many variational formulations on the lower-dimensional space $\mathcal{D}$ parametrized by $y^K\in\Lambda^y$. Clearly, we may replace s_K by a (certified) RB approximation to further reduce the computational effort[1] ; equivalently, we may replace the MC sums of our current approach with the quadrature rules developed in [BNT07, NTW08]. Future work will investigate this promising opportunity.

Compared with numerical approaches developed previously for SPDEs, the main features of our RB approach are the following :

(a) the solution $U_K(\cdot,\omega)$ to the original *stochastic* BVP is mapped to the distribution of Y^K,

$$U_K(x,\omega) = u_K(x;Y^K(\omega)) \text{ for almost every (a.e.) } x\in\mathcal{D} \text{ and } \mathbb{P}\text{-a.e. outcome } \omega\in\Omega,$$

through the solution $u_K(\cdot;y^K)$ to a *deterministic* BVP, the variational formulation of which must have an *affine* parametrization[2] (*affine* in the sense that the weak form can be expressed as a sum of products of parameter-dependent functions and parameter-independent forms) — as typically provided by a KL

[1] In [NTW08], it is even shown that one can minimize the number of collocation points, which correspond to zeros of the family of orthogonal polynomials substituted for the PC basis, with a view to "optimally" describing the range of the random input.

[2] Non-affine (but piecewise smooth) parametrizations can also be treated by the so-called *magic points* to "empirically" interpolate the coefficients entering the variational formulation [BNMP04, GMNP07].

expansion of the random input field which decouples the dependencies on the probability and physical spaces;

(b) a large number of variational approximations for the solutions $u_K(\cdot;y^K)$ to the *deterministic* BVP, defined over the (relatively) low-dimensional physical space $\mathcal{D}$ and parametrized by y^K, must be computed for each MC evaluation of the statistical outputs (and for each value of the additional parameter ϱ) — as opposed to spectral Galerkin variational methods in which $u_K(\cdot;\cdot)$ is discretized on the high-dimensional tensor-product space $(x,y^K)\in\mathcal{D}\times\Lambda^y$ such that only one, very expensive, solution is required (for each value of the additional parameter ϱ);

(c) the "deterministic" RB approach [MMO+00, PRV+02, RHP08] is then applied to the deterministic BVP to yield — based on a many-query Offline-Online computational strategy — greatly reduced computational cost at little loss in accuracy or, thanks to rigorous *a posteriori* bounds, certainty.

Of course our approach also bears many similarities to earlier proposals, most notably reliance on the Kolmogorov strong law of large numbers (for the MC evaluations to converge), on the KL expansion of the random input field, and on smoothness with respect to the parameter y^K.

Note that the usual RB method can be extended to the SPDE framework thanks to new error bounds (to take into account the effect of the truncation of the KL expansion, and to assess the efficiency of the reduction, that is to control the RB error in outputs that are sums over many parameter realizations). But the idea behind the RB method remains the same as in the usual case of parametrized (deterministic) PDEs, even though SPDEs typically result in *many* ($>K$) deterministic parameters (y^K,ϱ). The rapid convergence of the RB method we observe here — that does not break but at least moderates the curse of dimensionality — relies heavily not only on the smoothness of $u_K(\cdot;y^K)$ with respect to y^K, but also on the limited range of the y_k component of y^K when $k\gg 1$; the latter, in turn, derives from the (assumed) smoothness of the autocovariance function (rapid decay of the eigenvalues of the Hilbert-Schmidt integral operator with the autocovariance function as kernel). It is imperative to choose K as small as possible.

6-II Variational Formulation of a Boundary Value Problem with Stochastic Parameters

6-II-A Stochastic Partial Differential Equations

The modeling of multiscale problems in science and engineering is often cast into the following framework. At the macroscopic scale at which important quantities must be computed, a (possibly multi-dimensional) field variable $U(\cdot,\omega)$ is assumed to satisfy a PDE on a physical domain $\mathcal{D}\subset\mathbb{R}^d$ ($d=2$, 3, or 4 for common applications)

$$A(\cdot,\omega)\,U(\cdot,\omega)=f(\cdot,\omega)\text{ in }\mathcal{D}\ , \tag{6-II.1}$$

supplied with Boundary Conditions (BC) on the (sufficiently smooth) boundary $\partial\mathcal{D}$,

$$B(\cdot,\omega)\,U(\cdot,\omega)=g(\cdot,\omega)\text{ in }\partial\mathcal{D}\ ; \tag{6-II.2}$$

here the differential operators $A(\cdot,\omega),B(\cdot,\omega)$ and the source terms $f(\cdot,\omega),g(\cdot,\omega)$ are parametrized at each point of the physical domain by a variable ω describing the state of some generalized local microstructure. We shall not discuss other possible formulations for multiscale problems, such as integral equations; furthermore, the formulation above will be assumed well-posed in the sense of Hadamard for the case in which A,B,f and g vary with the microstructure ω (extensions of this work to distributions, that is, generalized functions of ω, are not straightforward).

To model the "fluctuations" of the underlying microstructure, whose impact on the macroscopic quantities of interest is to be evaluated, we can assume — without invoking detailed information about the microstructure — that the input is random. To this aim, one can introduce an abstract probability space to model the fluctuations, the latter being then described through variations within the set of elementary events $\omega\in\Omega$ (similar arguments are often developed to model material properties [3], see *e.g.* [OSW99, Xu07]). The outputs of such models are then also random by nature. The equations (6-II.1),(6-II.2) are then generally called Stochastic PDEs (SPDEs).

[3] We note that by choosing the microscopic fluctuations as stationary ergodic random fields, the numerical treatment of averaged outputs for SPDEs also applies to many situations considered in stochastic homogenization theory [BLL06, JG04], in which a powerful and elegant analysis of (weak) convergence allows one to reduce the modeling of complex multiscale problems to a more tractable set of sub-problems. Note that the RB approach has been applied to efficient numerical treatment of multiscale problems with locally periodic fluctuations within the context of deterministic homogenization theory [Boy08].

SPDEs are useful when one cannot, or does not want to, describe precisely the microstructure. Examples include uncertainty quantification for structures in civil engineering [CGRdB04, SBL06], for complex flows in fluid dynamics [MHZ05], or for multiphase flows in porous media [GD98].

6-II-B Problem Statement : Stochastic Robin Boundary Condition

The RB method has been introduced earlier for the many-query evaluation of outputs for various parametrized variational problems [MMO$^+$00, PRV$^+$02, RHP08] in a deterministic framework (deterministic PDE and BC). In this work, we shall choose only one (simple) example to illustrate the stochastic case ; however, it should be clear that the approach admits a general abstraction applicable to a wide class of problems.[4] We now pose our particular problem.

We shall let $(\Omega,\mathcal{F},\mathbb{P})$ be a complete probability space where Ω is the set of outcomes ω, $\mathcal{F}$ is the σ-algebra of events among all subsets of Ω, and $\mathbb{P}$ is a probability measure (notice that this definition itself is often a practical issue for the modeller). And we shall let the physical domain $\mathcal{D}$ be an open, bounded, connected subset of $\mathbb{R}^2$ $(d=2)$ with Lipschitz polyhedral boundary, which we classically equip with the usual Borel σ-algebra and the Lebesgue measure. We recall that random fields are collections of scalar random variables that can be mapped to a physical domain ; for instance, functions are defined on $\partial\mathcal{D}$ and take values in $L^2_{\mathbb{P}}(\Omega)$ — the space of square-integrable functions on the probability space $(\Omega,\mathcal{F},\mathbb{P})$.

Let us introduce some further notations :

- $L^2(\mathcal{D})$ the Hilbert space of Lebesgue square integrable functions in $\mathcal{D}$;
- $H^1(\mathcal{D})$ the usual Sobolev space (with Hilbert structure) of functions in $L^2(\mathcal{D})$ that have gradient in $[L^2(\mathcal{D})]^2$, imbued with the usual Hilbert norm $\|\cdot\|_{1,\mathcal{D}}$;
- $L^2(\partial\mathcal{D})$ the Hilbert space of the Lebesgue square integrable functions in the manifold $\partial\mathcal{D}$ equipped with its Borel σ-algebra, imbued with the Hilbert norm $\|\cdot\|_{0,\partial\mathcal{D}}$;
- $L^\infty(\partial\mathcal{D})$ the Banach space of essentially bounded functions on the manifold $\partial\mathcal{D}$, imbued with its usual norm $\|\cdot\|_{\infty,\partial\mathcal{D}}$.

We also recall that functions $v\in H^1(\mathcal{D})$ have a trace $v\in L^2(\partial\mathcal{D})$ on $\partial\mathcal{D}$ that satisfies

$$\|v\|_{0,\partial\mathcal{D}}\le\tilde{\gamma}\|v\|_{1,\mathcal{D}}\ , \tag{6-II.3}$$

where $\tilde{\gamma}\equiv\tilde{\gamma}(\mathcal{D})$ is a constant positive real number that depends only on $\mathcal{D}$.

In the following, we shall deal with SPDEs in which only the boundary differential operator $B(\omega)$ is parametrized by a random scalar input field, in particular $\mathrm{Bi}(\cdot,\cdot):\partial\mathcal{D}\times\Omega\to\mathbb{R}$. We identify in (6-II.1), (6-II.2)

$$A(x,\omega)=-\mathrm{div}(\boldsymbol{a}(x)\nabla\cdot),\qquad f(x,\omega)=0,\qquad\forall x\in\mathcal{D}\ ,$$
$$B(x,\omega)=\boldsymbol{n}^{\mathrm{T}}(x)\,\boldsymbol{a}(x)\,\nabla\cdot+\mathrm{Bi}(x,\omega)\cdot,\ g(x,\omega)=g(x),\ \forall x\in\partial\mathcal{D}\ .$$

The case in which the other terms also depend on a single scalar random field $\mathrm{Bi}(\cdot,\omega)$ is a straightforward extension, provided the problem (6-II.1),(6-II.2) remains well-posed in the sense of Hadamard with respect to the variations $\omega\in\Omega$. Note that the divergence div and gradient ∇ operators imply differentiations with respect to the physical variable x only, and not with respect to the probability variable ω. The scalar random field $U(\cdot,\omega)$ with $x\in\mathcal{D}$ is defined as the $\mathbb{P}$-a.s. solution to the Robin BVP (6-I.1), (6-I.2). The deterministic (strictly positive) diffusion matrix $\boldsymbol{a}$ is assumed isotropic though non-constant for all $x\in\mathcal{D}$ (the function κ is specified below to get a simple "additional" deterministic parameter ϱ),

$$\boldsymbol{a}(x)=\begin{bmatrix}\kappa(x) & 0\\ 0 & \kappa(x)\end{bmatrix},\qquad\forall x\in\mathcal{D}\ .$$

We shall assume $0<\kappa_{\min}\le\kappa(x)\le\kappa_{\max}<+\infty$ for well-posedness. The boundary $\partial\mathcal{D}$ is divided into three non-overlapping open subsets

$$\partial\mathcal{D}\subset\left(\overline{\Gamma_{\mathrm{N}}}\cup\overline{\Gamma_{\mathrm{R}}}\cup\overline{\Gamma_{\mathrm{B}}}\right)\ .$$

[4] We shall limit attention to those simple SPDEs which are not generalizations of Stochastic Differential Equations (SDEs) to multi-dimensional derivatives — where outcomes of the random input are distributions (generalized functions). Such interesting cases will be the subject of future work.

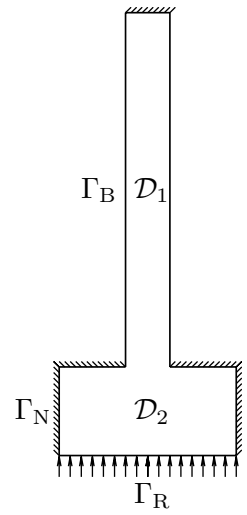


FIG. 6.1 – Geometry of the heat sink : a spreader $\mathcal{D}_2$ and a fin $\mathcal{D}_1$.

The boundary (Root) source term g is taken as deterministic (constant), non-zero on Γ_R only,

$$g(x) = 1_{\Gamma_R}, \qquad \forall x \in \partial\mathcal{D} \ ,$$

while the *Biot number* Bi is taken as a positive random field, non-degenerate on Γ_B only,

$$\mathrm{Bi}(x,\omega) = \mathrm{Bi}(x,\omega)\ 1_{\Gamma_B}, \qquad \forall x \in \partial\mathcal{D} \ .$$

Note that on Γ_N, Eq. (6-I.2) thus reduces to homogeneous Neumann conditions.

The physical interpretation is simple : if T_0 is the constant temperature of the ambient medium, $T_0 + U$ is the steady-state temperature field in a domain $\mathcal{D}$ (comprised of an isotropic material of thermal conductivity κ) subject to zero heat flux on boundary Γ_N (either by contact with a thermal insulator or for reasons of symmetry), constant flux at boundary Γ_R (contact with a heat source), and a random heat transfer coefficient Bi at boundary Γ_B (contact with a convective fluid medium). Note that the Biot number Bi is a fashion for decoupling the solid conduction problem from the exterior fluid convection problem : it is at best an engineering approximation, and at worst a rough average — often not reflecting the environmental details ; it thus makes sense to model the unknown Bi variations as a random (but typically rather smooth) field $\mathrm{Bi}(\cdot,\omega)$ in order to understand the sensitivity of output quantities to heat transfer coefficient uncertainties.

With a view to specify parameters which will then be used in the numerical application of Section 6-III, we shall more precisely consider the steady heat conduction problem (6-I.1), (6-I.2) inside the T-shaped heat sink $\mathcal{D}$ as shown in Figure 6.1. The heat sink comprises a 2×1 rectangular substrate (spreader) $\mathcal{D}_2 \equiv (-1,1)\times(0,1)$ on top of which is situated a 0.5×4 thermal fin $\mathcal{D}_1 \equiv (-0.25,0.25)\times(1,5)$. (In effect, all lengths will be nondimensionalized relative to the side-length of the substrate.) We also specify the diffusion coefficient, which we shall take as a (normalized) piecewise constant

$$\kappa(x) = 1_{\mathcal{D}_1} + \kappa 1_{\mathcal{D}_2}, \forall x \in \mathcal{D} \ ,$$

where $1_{\mathcal{D}_i}$ is the characteristic function of domain $\mathcal{D}_i$ ($i=1,2$). On Γ_B, the two sides of the fin, we shall impose a *stochastic* convection/Robin BC with a non-zero random Biot number Bi (built as a random field $\mathrm{Bi}(\cdot,\omega)$ with *a priori* known mean and autocovariance function, see Section 6-II-D-a) ; on Γ_R, the root, we impose unit flux $g(x) = 1$; and on Γ_N, we impose zero flux.

We recall that the outputs of interest will be the first two moments of a (scalar) linear functional $\mathcal{E}$ of the random solution field $U(\cdot,\omega)$ defined in (6-I.3) as the (random) integrated trace $S(\omega) = \mathcal{E}(U(\cdot,\omega))$ on the edge Γ_R of the domain $\mathcal{D}$ (corresponding to the location of the heat source — the point at which we wish to control the temperature) :

$$\mathbf{E}_{\mathbb{P}}(S) := \int_\Omega S(\omega)\, d\mathbb{P}(\omega) \ , \qquad \mathbf{Var}_{\mathbb{P}}(S) := \int_\Omega S(\omega)^2\, d\mathbb{P}(\omega) - \mathbf{E}_{\mathbb{P}}(S)^2 \ , \tag{6-II.4}$$

provided the random variable S is sufficiently regular (for instance in $L^2_{\mathbb{P}}(\Omega)$).

Remark 26 (Outputs)**.** *It is possible to consider other (and multiple) outputs within the RB approach. Essentially these outputs should be empirical estimations for functionals of $U(\cdot,\omega)$ that are continuous with respect to some $L^p_{\mathbb{P}}\big(\Omega, H^1(\mathcal{D})\big)$ topology $(1\le p\le+\infty)$. Note that interesting cases such as $p=+\infty$ above, and pointwise values of a cumulative distribution function*

$$\mathbb{P}\{\omega\in\Omega|\mathcal{E}(U(\cdot,\omega))\le\mathcal{E}_0\}$$

for some finite numbers $\mathcal{E}_0\in\mathbb{R}$, are covered by this first RB approach. Indeed, assuming smoothness in ω, one can bin the range of the random variable $\mathcal{E}(U(\cdot,\omega))$, and use a tree algorithm to account for the variations inside the confidence interval obtained for each realization S^m $(1\le m\le M)$ of $S(\omega)=\mathcal{E}(U(\cdot,\omega))$ using the RB approach. If a confidence interval Δ^0_m is associated to each realization S^m and overlaps $n_m\in\mathbb{N}$ bins, then computing the confidence interval for the output cumulative distribution function amounts to a search for the extreme variations in the output among the $\big(\Pi_{m=1}^M n_m\big)$ leaves of the tree.

In the numerical application of Section 6-III, the statistical outputs (6-II.4) (expected value and variance of the integrated temperature at the bottom surface $\Gamma_{\rm R}$ of the heat sink) will be explored in a many-query context (of design optimization for instance) as functions of the "additional" deterministic parameter $\varrho=(\kappa,\overline{\rm Bi})$ in the range Λ^ϱ, where

$$\overline{\rm Bi}:=\frac{1}{|\Gamma_{\rm B}|}\int_{\Gamma_{\rm B}}\mathbf{E}_{\mathbb{P}}({\rm Bi})\ .$$

A detailed stochastic description of the random field ${\rm Bi}(\cdot,\omega)$ used for numerics is given in Section 2.4.

6-II-C Different Discretization Formulations

Much attention has been devoted recently to the development and the numerical analysis of various numerical approaches for BVPs involving SPDEs *e.g.* [BNT07, BTZ05, DBO01, FST05, GS91, KM05, MK05, NTW08, SNK06, VJ05, WK05, WK09, XK02]. Our RB approach specifically aims at reducing the number of computations in many of the previously developed frameworks without any loss in precision by (i) splitting the computations into *Offline* and *Online* steps, and (ii) maintaining accuracy control through *a posteriori* error estimation of the outputs. The RB approach applies to those formulations that are variational with respect to variables in the physical space $\mathcal{D}$, which we denote $\mathcal{D}$-weak formulations, and can be combined with different treatments of the probabilistic dependence. The latter fall into two main categories : the Ω-strong/$\mathcal{D}$-weak formulations ; and the Ω-weak/$\mathcal{D}$-weak formulations. Although we shall only deal with Ω-strong/$\mathcal{D}$-weak formulations in the rest of this paper, our RB approach applies equally well to many Ω-weak/$\mathcal{D}$-weak formulations, as already discussed in the introduction. It is for this reason that we briefly summarize the principles of each of the different formulations so as to make it clear how our RB approach would adapt to Ω-weak/$\mathcal{D}$-weak formulations. (Both formulations have been studied extensively before, though typically by different authors ; a few studies already compare both formulations [MK05, BTZ05], but it may be interesting to reevaluate such comparisons between formulations from the viewpoint of our RB approach.)

6-II-C-a Strong-Weak Formulations

If the Biot number ${\rm Bi}(\cdot,\omega)$ is a non-degenerate positive random field on the (non-negligible) subset $\Gamma_{\rm B}$ of $\partial\mathcal{D}$, that is if there exist two constants $0<\bar b_{\min}<\bar b_{\max}<+\infty$ such that

$${\rm Bi}(\cdot,\omega)\in(\bar b_{\min},\bar b_{\max})\ \text{a.e. in}\ \Gamma_{\rm B}\,, \tag{6-II.5}$$

or equivalently ${\rm Bi}(\cdot,\omega),{\rm Bi}^{-1}(\cdot,\omega)\in L^\infty_{\mathbb{P}}(\Omega,L^\infty_{\mathbb{P}}(\Gamma_{\rm B}))$, then, by virtue of the Lax-Milgram theorem, there exists a unique (weak) solution $U(\cdot,\omega)\in H^1(\mathcal{D})$ to (6-I.1),(6-I.2), satisfying (6-II.6) $\mathbb{P}$-a.s. :

$$\int_{\mathcal{D}_1}\nabla U(\cdot,\omega)\cdot\nabla v+\kappa\int_{\mathcal{D}_2}\nabla U(\cdot,\omega)\cdot\nabla v+\int_{\Gamma_{\rm B}}{\rm Bi}(\cdot,\omega)\,U(\cdot,\omega)v=\int_{\Gamma_{\rm R}}v,\qquad\forall v\in H^1(\mathcal{D})\ . \tag{6-II.6}$$

Furthermore, from (6-II.5), we have the stability result :

$$\|U(\cdot,\omega)\|_{1,\mathcal{D}}\le\frac{C_1(\mathcal{D})}{\min\big\{1,\kappa_{\min},\bar b_{\min}\big\}}\ , \tag{6-II.7}$$

and $\|U(\cdot,\omega)\|_{1,\mathcal{D}}\in L^\infty_{\mathbb{P}}(\Omega)$ (with $C_1(\mathcal{D})$ a constant positive real number that depends only on $\mathcal{D}$).

Strong-weak formulations then use the fact that we also have $S \in L^\infty_{\mathbb{P}}(\Omega) \subset L^2_{\mathbb{P}}(\Omega)$, where the functional $S(\omega) = \mathcal{E}(U(\cdot,\omega))$ makes sense since, using (6-II.3) and (6-II.7), the trace of $U(\cdot,\omega)$ on the boundary segment $\Gamma_{\rm R}$ is well-defined. The outputs $\mathbf{E}_{\mathbb{P}}(S), \mathbf{Var}_{\mathbb{P}}(S)$ are thus approximated as the empirical Monte-Carlo estimations (6-I.4) where $\{S^m, m=1,\ldots,M\}$ are M independent copies (with same law) of the random variable S, and with the following convergence properties (by virtue of the Strong Law of Large Numbers)

$$E_M[S] \xrightarrow[M\to+\infty]{\mathbb{P}\text{-a.s.}} \mathbf{E}_{\mathbb{P}}(S) \ , \qquad V_M[S] \xrightarrow[M\to+\infty]{\mathbb{P}\text{-a.s.}} \mathbf{Var}_{\mathbb{P}}(S) \ . \tag{6-II.8}$$

Hence a major advantage of the Ω-strong/$\mathcal{D}$-weak formulations is to permit the direct application of classical computational procedures (in particular, FE) for the numerical approximation of deterministic BVPs such as (6-II.6) in their usual form, without any modification. Many (many ...) computations of such parametrized approximate solutions can then be combined — according to (the numerical simulation of) the law of the random field parameter $\mathrm{Bi}(\cdot,\omega)$ — to form the MC evaluations. Such formulations are thus very simple from the implementation viewpoint, presuming that we can readily simulate the law of $\mathrm{Bi}(x_k,\omega)$ at those discrete (e.g., quadrature or nodal) points x_k in the physical domain $\mathcal{D}$ required by the numerical approximation of (6-II.6). Note that the latter point is of course true for all formulations, but seems less stringent for the Ω-strong/$\mathcal{D}$-weak formulation (see Section 6-II-D-a).

However, the convergence (in probability) of SLLN will be slow — the rate of convergence for $E_M[S]$ is governed by the ratio of the variance of S (or its MC counterpart $V_M[S]$) to $\sqrt{M}$ by virtue of the Central Limit Theorem (CLT). This slow convergence is a strong limitation in the application of Ω-strong/$\mathcal{D}$-weak formulations. Variance reduction techniques, such as Quasi-Monte-Carlo (QMC) methods based on low-discrepancy sequences of random numbers [VJ05], have been developed to reduce the statistical error of the empirical estimations (6-I.4). And the RB approach itself brings new possibilities to addressing this slow convergence problem, not by directly reducing the number of necessary outcomes in the MC sums, but rather by improving the numerical treatment of many slow-varying outcomes.

In Section 6-III, we shall show how to apply our RB approach to the numerical approximation of Ω-strong/$\mathcal{D}$-weak formulations by taking advantage of the parametrized character of the BVP. We first map outcomes of stochastic coefficients to deterministic values of the parameters; we then reduce the computational cost of numerical approximations of the BVP for many values of the parameter by splitting the computations into Offline-Online steps; finally, we introduce *a posteriori* error control on the accuracy of the RB-KL approximations (relative to very accurate approximations in high-dimensional discretization-probability space). (We do not consider here variance reduction strategies.)

6-II-C-b Weak-Weak Formulations

Assuming (6-II.5) again for well-posedness, the Ω-weak/$\mathcal{D}$-weak formulations discretize a variational formulation of the original BVP on the full tensor-product space $\Omega \times \mathcal{D}$

$$\begin{aligned}\int_\Omega d\mathbb{P}(\omega) \int_{\mathcal{D}_1} \nabla U(\cdot,\omega)\cdot\nabla v(\cdot,\omega) + \kappa \int_\Omega d\mathbb{P}(\omega) \int_{\mathcal{D}_2} \nabla U(\cdot,\omega)\cdot\nabla v(\cdot,\omega) \\ + \int_\Omega d\mathbb{P}(\omega) \int_{\Gamma_{\rm B}} \mathrm{Bi}(\cdot,\omega) U(\cdot,\omega) v(\cdot,\omega) = \int_\Omega d\mathbb{P}(\omega) \int_{\Gamma_{\rm R}} v(\cdot,\omega), \quad \forall v(\cdot,\omega) \in L^2_{\mathbb{P}}\big(\Omega, H^1(\mathcal{D})\big)\end{aligned} \tag{6-II.9}$$

to compute an approximation of a weak solution $U(\cdot,\omega) \in L^2_{\mathbb{P}}\big(\Omega, H^1(\mathcal{D})\big)$ satisfying (6-II.9), typically through Galerkin projections over tensor-product approximation subspaces of the Hilbert space $L^2_{\mathbb{P}}\big(\Omega, H^1(\mathcal{D})\big)$ defined over the (high-dimensional) domain $\Omega \times \mathcal{D}$. The computations of $\mathbf{E}_{\mathbb{P}}(S)$ and $\mathbf{Var}_{\mathbb{P}}(S)$ are then effected by quadrature (or collocation) formulæ in $\Omega \times \mathcal{D}$ once discrete approximations for $U(\cdot,\omega)$ have been computed.

The weak-weak formulations may thus require less regularity (in fact, this seems very useful for input random fields that do not fulfill (6-II.5) but only a weaker assumption for well-posedness), although it also seems essential to the Ω-weak/$\mathcal{D}$-weak formulations that $\mathrm{Bi}(\cdot,\omega)$ be compatible with tensor-product approximations (see Section 6-II-D-a : this adds condition on $\mathrm{Bi}(\cdot,\omega)$ in comparison with the Ω-strong/$\mathcal{D}$-weak formulations). The weak-weak formulations essentially provide greatly improved convergence relative to SLLN (in fact, convergence is often improved only for small dimensions, where numerical approaches for this formulation are sufficiently simple).

For instance, after substituting in (6-II.9) a truncated version (6-I.6) of the KL expansion (6-I.5) of $\mathrm{Bi}(\cdot,\omega)$ using K (with $1 \le K \le \mathcal{K}$) independent identically distributed (i.i.d.) random variables in a complete set $\{Z_k,\ k \in \mathbb{N}\}$ of $L^2_{\mathbb{P}}(\Omega)$, the seminal work [GS91] used so-called spectral (stochastic) Galerkin methods, in which

$L^2_{\mathbb{P}}\big(\Omega, H^1(\mathcal{D})\big)$ is discretized by tensor products of classical discrete approximations for the variational formulation of a BVP in $H^1(\mathcal{D})$ (such as FE) multiplied by orthogonal polynomials $\{H_n, n\in\mathbb{N}\}$ of the random variables $\{Z_k, k\in\mathbb{N}\}$

$$H_0, \quad H_1(Z_k(\omega)), \quad H_2(Z_{k_1}(\omega), Z_{k_2}(\omega)), \ldots, \qquad k, k_1, k_2 \in \mathbb{N},\ k_1 \geq k_2 \geq 0, \ldots \ .$$

(In the original Polynomial Chaos (PC) expansion of Wiener [Wie38] for $L^2_{\mathbb{P}}(\Omega)$, the H_n are Hermite polynomials and the variates Z_k are Gaussian ; this expansion has then been generalized to other couples of polynomials and probability distributions [XK02, SG04].) The Galerkin projections in the stochastic variable that truncate the PC expansions at polynomial order $L\in\mathbb{N}_{>0}$ ($L\geq K$), hence using $D=K+L-1$ i.i.d variates $Z_k(\omega)$, then result in a p-dimensional vector space

$$\mathbf{Span}\left(H_l(Z_{k_1},\ldots,Z_{k_l}) \,|\, 0\leq l\leq L, K+L-1\geq k_l>\ldots>k_1\geq 1, \{k_1,\ldots,k_l\}\cap\{1,\ldots,K\}\neq\emptyset\right)$$

with $p=1+\sum_{l=1}^{L}\sum_{k=1}^{l}\binom{K}{k}\binom{L-1}{l-k}$. Equivalently, the variational formulation (6-II.9) is projected onto the (very high) $(d+D)$-dimensional domain in which $(x, Z_1, \ldots, Z_D)$ take its values. (Alternatively, the discretization level in each direction of the tensor-product Galerkin approximations can be tailored to achieve rapid convergence with respect to the number of degrees of freedom (d.o.f.). In fact, *a posteriori* error indicators and reduced spaces — though quite different from the error bounds and reduced basis spaces presented in the present paper — can serve to identify efficient truncations [WK09].)

A major limitation of such spectral Galerkin methods is the high-dimensionality of the approximation spaces for (truncated) PC expansions (p increases rapidly with K and L), which necessitates complicated (though certainly often efficient) numerical strategies in order to maintain sparsity on the discretization grid [BTZ05, FST05, MK05, TS07b, WK09]. There are many approaches to this *curse of dimensionality*, most of which have already been mentioned in the introduction. The essential features of our RB approach compared to the other reduction techniques previously applied to SPDEs have also been discussed in the introduction. Clearly, the efficiency of the reduction methods — which are not necessarily incompatible between one another and may thus be combined in future studies — only makes sense in a precise context, where it is clear what has to be reduced, why, and for what purpose.

6-II-D Random Input Field

6-II-D-a Karhunen–Loève Expansions of Random Fields

To develop efficient numerical procedures for SPDEs, it has been noted in the above Section 6-II-C that it was essential to discretize the (scalar) random input field $\mathrm{Bi}(\cdot,\omega)$ consistently with the discretization of the BVP problem (whatever the formulation). Besides, the (de)coupling of variations of $\mathrm{Bi}(x,\omega)$ on the space variable $x\in\mathcal{D}$ and on the probability variable $\omega\in\Omega$ is also an important feature of the variational problems resulting from our numerical approach. It indeed leads to a parametrized weak form where the parametrization is *affine* (see Section 6-I-B for a definition). We thus do not only need to assume the non-degeneracy of the random field $\mathrm{Bi}(\cdot,\omega)$ on Γ_{B} for well-posedness of the BVP, but also the possibility to rewrite it in a decoupled manner like in the KL expansion (6-I.5).

In the present work, we introduce general random input fields $\mathrm{Bi}(x,\omega)$ at a continuous level, defined by an *infinite* collection of correlated random numbers mapped to an *infinite* number of points in the physical domain $\mathcal{D}$. This is typically a situation where the fluctuations are modeled following physical assumptions (statistical mechanics for instance). More precisely, we deal with a random process $(\mathrm{Bi}(x,\cdot))_{x\in\partial\mathcal{D}}$ where $\mathrm{Bi}(x_1,\cdot)$ and $\mathrm{Bi}(x_2,\cdot)$ are not necessarily decorrelated when $x_1\neq x_2$.

For well-posedness of the BVP, we only consider random input fields that satisfy (6-II.5). Now, such random fields are in $L^2_{\mathbb{P}}\big(\Omega, L^2(\partial\mathcal{D})\big)$. Thus, assuming (6-II.5), the random input fields $\mathrm{Bi}(\cdot,\omega)$ in this work always have a KL expansion and can always be generated by decoupled variations in x and ω (possibly asymptotically if $\mathcal{K}$ is infinite) after the well-known Proposition 24 (recalled below). Yet, in cases where there is no other motivation like well-posedness for assuming (6-II.5), one should still keep in mind that specific assumptions may be necessary to fulfill the requirement of decoupling – by the way, other expansions than KL might also fulfill that requirement.

Note that in practical engineering situations, $\mathrm{Bi}(\cdot,\omega)$ is often not given but rather constructed from a few measurements only, after solving an inverse problem to assimilate (or calibrate) statistical data (see *e.g.* [JZ08]). Since the inverse problem is solved at the discrete level [5], this yields a *finite* collection of random numbers mapped to a *finite* number of points in the physical domain $\mathcal{D}$, and the assumptions may be simplified.

[5] It is interesting to note that inverse problems are usually solved through optimization algorithms that define a typical many-query context where a RB approach for parametrized PDEs is well motivated.

Proposition 24. *Random fields* $\mathrm{Bi}(\cdot,\omega)\in L^2_{\mathbb{P}}\big(\Omega,L^2(\partial\mathcal{D})\big)$ *are in one-to-one correspondence with couples* $(\mathbf{E}_{\mathbb{P}}(\mathrm{Bi}),\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi}))\in L^2(\partial\mathcal{D})\times L^2(\partial\mathcal{D}\times\partial\mathcal{D})$ *supplied with a collection of mutually uncorrelated random variables* $\{Z_k(\omega);1\le k\le\mathcal{K}\}$ *in* $L^2_{\mathbb{P}}(\Omega)$ *with zero mean and unit variance*

$$\mathbf{E}_{\mathbb{P}}(Z_k)=0\quad \mathbf{E}_{\mathbb{P}}(Z_kZ_{k'})=\delta_{k,k'}\quad \forall 1\le k,k'\le\mathcal{K}\ (\textit{with Kronecker notations, hence } \mathbf{Var}_{\mathbb{P}}(Z_k)=1),$$

when the kernel $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})$ *defines a positive, self-adjoint, trace class linear operator*

$$\widetilde{T}\in\mathcal{L}\big(L^2(\partial\mathcal{D}),L^2(\partial\mathcal{D})\big),\qquad (\widetilde{T}f)(x)=\int_{\partial\mathcal{D}}\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)\,f(y)\,dy,\qquad \forall\, f\in L^2(\partial\mathcal{D})\ , \tag{6-II.10}$$

of (possibly infinite) rank $\mathcal{K}$*. Furthermore, random fields* $\mathrm{Bi}(\cdot,\omega)\in L^2_{\mathbb{P}}\big(\Omega,L^2(\partial\mathcal{D})\big)$ *have the following Karhunen–Loève expansion [Loè78]*

$$\mathrm{Bi}(x,\omega)=\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(x)+\sum_{k=1}^{\mathcal{K}}\sqrt{\tilde\lambda_k}\,\Phi_k(x)\,Z_k(\omega),\ x\in\partial\mathcal{D}, \tag{6-II.11}$$

where $\{\tilde\lambda_k;1\le k\le\mathcal{K}\}$ *are the positive eigenvalues (in descending order) of the positive, self-adjoint, trace class operator* $\widetilde{T}$ *associated with eigenvectors* $\{\Phi_k(x)\in L^2(\partial\mathcal{D});1\le k\le\mathcal{K}\}$

$$(\widetilde{T}f)(x)=\sum_{1\le k\le\mathcal{K}}\tilde\lambda_k\left(\int_{\partial\mathcal{D}}\Phi_k(y)\,f(y)\,dy\right)\Phi_k(x),\qquad \forall\, f\in L^2(\partial\mathcal{D})\ ,$$

(orthonormal in the $L^2(\partial\mathcal{D})$*-inner-product), and the random variables* $\{Z_k\}$ *are defined by*

$$Z_k(\omega)=\frac{1}{\sqrt{\tilde\lambda_k}}\int_{\partial\mathcal{D}}(\mathrm{Bi}(\cdot,\omega)-\mathbf{E}_{\mathbb{P}}(\mathrm{Bi}))\,\Phi_k,\qquad \forall 1\le k\le\mathcal{K}\ .$$

Since $L^2(\partial\mathcal{D})$ and $L^2_{\mathbb{P}}(\Omega)$ are Hilbert spaces, the Proposition 24 can be easily proved using Riesz representation theorem, and the Hilbert-Schmidt theorem for bounded (linear) operators of the trace class (then compact) like $\widetilde{T}$ (see *e.g.* [ST06]).

In the following, we rewrite the usual representation (6-II.11) with a scaling parameter $\widetilde{\Upsilon}>0$,

$$\widetilde{\Upsilon}^2:=\int_{\Gamma_{\mathrm{B}}}\int_{\Gamma_{\mathrm{B}}}\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)dxdy=\int_{\Gamma_B}\mathbf{Var}_{\mathbb{P}}(\mathrm{Bi})=\mathrm{tr}(\widetilde{T})=\sum_{1\le k\le\mathcal{K}}\tilde\lambda_k,$$

and then re-scale the collection of positive eigenvalues as

$$\lambda_k:=\frac{\tilde\lambda_k}{\widetilde{\Upsilon}^2},\qquad \forall 1\le k\le\mathcal{K}\ ,$$

to obtain the following KL expansion from Proposition 24

$$\mathrm{Bi}(x,\omega)=\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(x)+\widetilde{\Upsilon}\sum_{k=1}^{\mathcal{K}}\sqrt{\lambda_k}\,\Phi_k(x)\,Z_k(\omega),\ x\in\partial\mathcal{D}.$$

Lastly, when $\mathcal{K}$ is infinite or too large, numerical approaches exploit, instead of the full KL expansion, KL truncations of order K ($K\in\mathbb{N}$, $0<K<\mathcal{K}$) which we write as

$$\mathrm{Bi}_K(x,\omega)=\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(x)+\widetilde{\Upsilon}\sum_{k=1}^{K}\sqrt{\lambda_k}\,\Phi_k(x)\,Z_k(\omega),\ x\in\partial\mathcal{D}\ .$$

The truncation error satisfies

$$\mathbf{E}_{\mathbb{P}}\Big((\mathrm{Bi}-\mathrm{Bi}_K)^2\Big)=\widetilde{\Upsilon}^2\sum_{k=K+1}^{\mathcal{K}}\lambda_k\Phi_k^2(x)\underset{K\to\mathcal{K}}{\longrightarrow}0\quad \text{in } L^1(\partial\mathcal{D}). \tag{6-II.12}$$

6-II-D-b Additional Assumptions on the Random Input Field

In the numerical applications of the next section, we shall require (6-II.5) for well-posedness of the BVP. This implies $\mathrm{Bi}(\cdot,\omega)\in L^\infty_{\mathbb{P}}(\Omega,L^\infty(\Gamma_{\mathrm{B}}))$, thus $\mathrm{Bi}(\cdot,\omega)$ is fully determined by (Proposition 24)

(i) an expected value function $\mathbf{E}_{\mathbb{P}}(\mathrm{Bi}) : x\in\Gamma_{\mathrm{B}}\to\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(x)\in\mathbb{R}$ in $L^\infty(\Gamma_{\mathrm{B}})\subset L^2(\Gamma_{\mathrm{B}})$,

(ii) a covariance function $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi}) : (x,y)\in\Gamma_{\mathrm{B}}\times\Gamma_{\mathrm{B}}\to\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)\in\mathbb{R}$ in $L^2(\Gamma_{\mathrm{B}}\times\Gamma_{\mathrm{B}})$, thus the kernel of a positive self-adjoint trace class operator of rank $\mathcal{K}$ with eigenpairs $(\widetilde{\Upsilon}^2\lambda_k,\Phi_k)$ $(\lambda_k\geq\lambda_{k+1}>0, 1\leq k\leq\mathcal{K})$ satisfying $\sum_{k=1}^{\mathcal{K}}\lambda_k=1$ and

$$\int_{\Gamma_{\mathrm{B}}}\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)\,\Phi_k(y)\,dy=\widetilde{\Upsilon}^2\lambda_k\Phi_k(x),\qquad\forall x\in\Gamma_{\mathrm{B}}\ , \tag{6-II.13}$$

(iii) and mutually uncorrelated random variables $\{Z_k\in L^\infty_{\mathbb{P}}(\Omega)\subset L^2_{\mathbb{P}}(\Omega);1\leq k\leq\mathcal{K}\}$ with zero mean and unit variance,

through the Karhunen-Loève (KL) expansion

$$\mathrm{Bi}(x,\omega)=\overline{\mathrm{Bi}}\left(G(x)+\Upsilon\sum_{k=1}^{\mathcal{K}}\sqrt{\lambda_k}\,\Phi_k(x)\,Z_k(\omega)\right)\ , \tag{6-II.14}$$

where $G\in L^\infty(\Gamma_B)$ is a prescribed (deterministic) positive function such that $\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(\cdot)=\overline{\mathrm{Bi}}G(\cdot)$,

$$\text{and }\frac{1}{|\Gamma_B|}\int_{\Gamma_B}G(x)dx=1,\text{using the scaling parameters }\quad\overline{\mathrm{Bi}}=\frac{1}{|\Gamma_B|}\int_{\Gamma_B}\mathbf{E}_{\mathbb{P}}(\mathrm{Bi})(x)\,dx\ \text{ and }\ \Upsilon=\frac{\widetilde{\Upsilon}}{\overline{\mathrm{Bi}}}.$$

For all nonnegative integer $1\leq K\leq\mathcal{K}$, we introduce the truncation of KL expansion (6-II.14)

$$\mathrm{Bi}_K(x,\omega)=\overline{\mathrm{Bi}}\left(G(x)+\Upsilon\sum_{k=1}^{K}\sqrt{\lambda_k}\,\Phi_k(x)\,Z_k(\omega)\right). \tag{6-II.15}$$

For the sake of consistency of the numerical discretization, we require

$$\|\mathrm{Bi}(\cdot,\omega)-\mathrm{Bi}_K(\cdot,\omega)\|_{L^\infty_{\mathbb{P}}(\Omega,L^\infty(\Gamma_{\mathrm{B}}))}\xrightarrow{K\to\mathcal{K}}0\ , \tag{6-II.16}$$

which is stronger than (6-II.12) and can be achieved for instance by choosing

(H1) a smooth covariance function $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})$ such that

(H1a) the eigenvectors are uniformly bounded by some positive real number $\phi>0$

$$\|\Phi_k\|_{L^\infty(\Gamma_{\mathrm{B}})}\leq\phi,\qquad 1\leq k\leq\mathcal{K}\ , \tag{6-II.17}$$

(H1b) the eigenvalues decay sufficiently rapidly,

$$\sum_{k=1}^{\mathcal{K}}\sqrt{\lambda_k}<\infty\ , \tag{6-II.18}$$

(H2) uniformly bounded random variables (say) $\{Z_k;|Z_k(\omega)|<\sqrt{3},\mathbb{P}\text{-a.s.}\}$.

In the numerical results we shall consider Gaussian covariances $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)=(\overline{\mathrm{Bi}}\Upsilon)^2e^{-\frac{(x-y)^2}{\delta^2}}$, with δ a positive real constant, which complies with the requirements above [FST05]. The fast decay of the eigenvalues in the Gaussian case play an important role in the fast convergence of any numerical discretization based on KL expansions of the input random field ; as we shall see, this is true also for our RB approach — the eigenvalues determine the ranges of the parameters, which in turn affect the dimension of the RB space. Next, we shall also insist upon

(H3) *independent* (thus mutually uncorrelated) random variables $\{Z_k;1\leq k\leq\mathcal{K}\}$,

(H4) Z_k, $1\leq k\leq K$, i.i.d. according to the uniform density with respect to the Lebesgue measure on $\mathbb{R}$ in the range $(-\sqrt{3},\sqrt{3})$,

(H5) Υ chosen such that

$$\tau_0:=\sqrt{3}\Upsilon\sum_{k=1}^{\mathcal{K}}\sqrt{\lambda_k}\,\|\Phi_k\|_{L^\infty(\Gamma_{\mathrm{B}})}\leq\frac{\min_{x\in\Gamma_{\mathrm{B}}}G(x)}{2}\ . \tag{6-II.19}$$

Then, under our assumptions, the truncation error is bounded above $\forall 1 \le K \le \mathcal{K}$:

$$\|\mathrm{Bi}(\cdot,\omega) - \mathrm{Bi}_K(\cdot,\omega)\|_{L^\infty_{\mathbb{P}}(\Omega, L^\infty(\Gamma_\mathrm{B}))} \le \overline{\mathrm{Bi}}\,\tau_K\,,$$

$$\tau_K := \sqrt{3}\Upsilon \sum_{k=K+1}^{\mathcal{K}} \sqrt{\lambda_k}\,\|\Phi_k\|_{L^\infty(\Gamma_\mathrm{B})}\,, \tag{6-II.20}$$

and furthermore for $0 < \bar{b}_{\min} \le \frac{\overline{\mathrm{Bi}}}{2}\left(\min_{x\in\Gamma_\mathrm{B}} G(x)\right)$ we have $\mathbb{P}$-a.s.

$$\mathrm{Bi}_K(\cdot,\omega) \ge \bar{b}_{\min} > 0 \quad \text{a.e. in } \mathcal{D}, \qquad 1 \le K \le \mathcal{K}. \tag{6-II.21}$$

Remark 27 (Choice of the random variables $\{Z_k\}$)**.** *Note that there are many other interesting cases where, for a given smooth covariance function, the random variables $\{Z_k\}$ are not uniformly distributed. These cases will be considered in future studies as they necessitate refinements that would complicate this first exposition of our viewpoint.*

6-III Reduced Basis Approach for Monte-Carlo Evaluations

6-III-A Discretization of a Test Problem in Strong-Weak Formulation

We now equip the Sobolev space $X := H^1(\mathcal{D})$ with the following inner product for all $w, v \in X$

$$(w,v)_X = \int_{\mathcal{D}_1} \nabla w \cdot \nabla v + \int_{\mathcal{D}_2} \nabla w \cdot \nabla v + \int_{\Gamma_\mathrm{B}} wv\ , \tag{6-III.1}$$

and induced norm $\|v\|_X = \sqrt{(v,v)_X}$. It is a standard result that the norm $\|\cdot\|_X$ is equivalent to the usual norm $\|\cdot\|_{1,\mathcal{D}}$ defined previously. We also introduce a finite element (FE) subspace $X_\mathcal{N} \subset X$ of dimension $\mathcal{N}$ which inherits the inner product and norm of X. For functions $v \in X_\mathcal{N}$, it is possible to define a trace $v \in L^2(\Gamma_\mathrm{B})$ which satisfies

$$\|v\|_{0,\Gamma_\mathrm{B}} \le \gamma_\mathcal{N} \|v\|_X\ , \tag{6-III.2}$$

where the constant $\gamma_\mathcal{N}$ depends only on $\mathcal{D}$ and is bounded above for all $\mathcal{N}$ since

$$\gamma_\mathcal{N} \equiv \gamma_\mathcal{N}(\mathcal{D}) = \sup_{v\in X_\mathcal{N}} \frac{\sqrt{\int_{\Gamma_\mathrm{B}} v^2}}{\|v\|_X} \le \gamma \equiv \sup_{v\in X} \frac{\sqrt{\int_{\Gamma_\mathrm{B}} v^2}}{\|v\|_X}\ . \tag{6-III.3}$$

(Note $\tilde{\gamma}$ of (6-II.3) differs from γ of (6-III.3) only because of the choice of norm.)

For a given positive scalar κ and a given random input field $\mathrm{Bi}(\cdot,\omega)$, we define

(a) the temperature distribution $U(\cdot,\omega) \in X$ in $\mathcal{D}$,

(b) a FE approximation $U_\mathcal{N}(\cdot,\omega) \in X_\mathcal{N}$ to the temperature distribution in $\mathcal{D}$,

as the respective solutions to the following variational formulations (6-III.4),

$$\int_{\mathcal{D}_1} \nabla U_{(\mathcal{N})}(\cdot,\omega)\cdot\nabla v + \kappa \int_{\mathcal{D}_2} \nabla U_{(\mathcal{N})}(\cdot,\omega)\cdot\nabla v + \int_{\Gamma_\mathrm{B}} \mathrm{Bi}(\cdot,\omega) U_{(\mathcal{N})}(\cdot,\omega) v = \int_{\Gamma_\mathrm{R}} v, \quad \forall v \in X_{(\mathcal{N})} \quad , \tag{6-III.4}$$

and, when $\mathrm{Bi}(\cdot,\omega)$ is approximated by $\mathrm{Bi}_K(\cdot,\omega)$,

(c) an approximation $U_{,K}(\cdot,\omega) \in X$ to $U(\cdot,\omega)$,

(d) and a FE approximation $U_{\mathcal{N},K}(\cdot,\omega) \in X_\mathcal{N}$ to $U_\mathcal{N}(\cdot,\omega)$

as the respective solutions to the following variational formulations (6-III.5)

$$\int_{\mathcal{D}_1} \nabla U_{(\mathcal{N}),K}(\cdot,\omega)\cdot\nabla v + \kappa \int_{\mathcal{D}_2} \nabla U_{(\mathcal{N}),K}(\cdot,\omega)\cdot\nabla v + \int_{\Gamma_\mathrm{B}} \mathrm{Bi}_K(\cdot,\omega) U_{(\mathcal{N}),K}(\cdot,\omega) v = \int_{\Gamma_\mathrm{R}} v, \forall v \in X_{(\mathcal{N})}, \tag{6-III.5}$$

where the same subscripts into brackets $(\cdot)$ are simultaneously active or not, which means

in $(a), (b)$ that (6-III.4) holds for $U(\cdot,\omega)$ and X, or $U_\mathcal{N}(\cdot,\omega)$ and $X_\mathcal{N}$ respectively,

in $(c), (d)$ that (6-III.5) holds for $U_{,K}(\cdot,\omega)$ and X, or $U_{\mathcal{N},K}(\cdot,\omega)$ and $X_\mathcal{N}$ respectively.

With a similar use of the subscripts in $(\cdot)$, we also define (intermediate) outputs as

$$S_{(\mathcal{N})(,K)}(\omega) := \mathcal{E}\left(U_{(\mathcal{N})(,K)}(\cdot,\omega)\right) = \int_{\Gamma_{\mathrm{R}}} U_{(\mathcal{N})(,K)}(\cdot,\omega)\,. \tag{6-III.6}$$

We are interested in evaluating the expected value and variance of the integrated temperature $S_{(\mathcal{N})(,K)}(\cdot,\omega)$, which are our (final) statistical outputs :

$$\mathbf{E}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right) = \int_{\Omega} S_{(\mathcal{N})(,K)}(\omega) d\mathbb{P}(\omega)\ , \tag{6-III.7}$$

$$\mathbf{Var}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right) = \int_{\Omega} \left(\mathbf{E}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right) - S_{(\mathcal{N})(,K)}(\cdot,\omega)\right)^2 d\mathbb{P}(\omega)\ . \tag{6-III.8}$$

Since $\mathrm{Bi}_K(\cdot,\omega)$ is $\mathbb{P}$-a.s. strictly positive on Γ_{B} and every $1 \le K \le \mathcal{K}$ (by assumption), the variational problems (6-III.4) and (6-III.5) are well-posed, and the solutions satisfy the following bound $\mathbb{P}$-a.s.

$$\|U_{(\mathcal{N})(,K)}(\cdot,\omega)\|_X \le \frac{C_1'(\mathcal{D})}{\min\left\{1,\kappa,\bar{b}_{\min}\right\}} \tag{6-III.9}$$

for some positive constant $C_1'(\mathcal{D})$. In addition, we have

Proposition 25. *Under standard regularity hypotheses (as $\mathcal{N}\to\infty$) on the family of FE spaces $X_{\mathcal{N}}$, the FE approximation converges as $\mathcal{N}\to\infty$. Furthermore, under the hypotheses of Section 6-II-D-b, the KL approximation converges as $K\to\mathcal{K}$. Finally, the following convergence holds $\mathbb{P}$-a.s.*

$$\begin{array}{ccccc} & & S_{\mathcal{N},K}(\omega) & \xrightarrow{\mathcal{N}\to\infty} & S_{,K}(\omega) & & \\ K\to\mathcal{K} & & \downarrow & & \downarrow & & K\to\mathcal{K}\ . \\ & & S_{\mathcal{N}}(\omega) & \xrightarrow{\mathcal{N}\to\infty} & S(\omega) & & \end{array} \tag{6-III.10}$$

Proof.
First, for any fixed $1 \le K \le \mathcal{K}$, the $\mathbb{P}$-a.s. convergence, as $\mathcal{N}\to 0$, of $U_{\mathcal{N}(,K)}(\cdot,\omega) \to U_{(,K)}(\cdot,\omega)$ in X as $\mathcal{N}\to\infty$ follows under standard hypotheses on the FE spaces $X_{\mathcal{N}}$. Then, by subtracting the variational formulation (6-III.5) for $U_{(\mathcal{N}),K}(\cdot,\omega)$ from (6-III.4) for $U_{(\mathcal{N})}(\cdot,\omega)$ (in $X_{(\mathcal{N})}$) with $v = U_{(\mathcal{N})}(\cdot,\omega) - U_{(\mathcal{N}),K}(\cdot,\omega)$, we get $\mathbb{P}$-a.s.

$$\|U_{(\mathcal{N})}(\cdot,\omega) - U_{(\mathcal{N}),K}(\cdot,\omega)\|_{H^1(\mathcal{D})} \le C_2(\mathcal{D},\bar{b}_{min}) \|\mathrm{Bi}(\cdot,\omega) - \mathrm{Bi}_K(\cdot,\omega)\|_{L^\infty(\Gamma_{\mathrm{B}})} \|U_{(\mathcal{N}),K}(\cdot,\omega)\|_{L^2(\Gamma_{\mathrm{B}})} \tag{6-III.11}$$

for some positive real number $C_2(\mathcal{D},\bar{b}_{min})$ depending only on $\mathcal{D}$ and $\bar{b}_{min}$. By compactness of the trace mapping from $H^1(\mathcal{D})$ into $L^2(\partial\mathcal{D})$, the uniform bound (6-III.9) for all K and the continuity (6-II.16) of $\mathrm{Bi}(\cdot,\omega)$ with respect to the $L^\infty(\Gamma_{\mathrm{B}})$ norm, we get the $\mathbb{P}$-a.s. convergence of $U_{(\mathcal{N})}(\cdot,\omega) \to U_{(\mathcal{N}),K}(\cdot,\omega)$ in X as $K\to\mathcal{K}$. So the following diagram of convergence holds :

$$\begin{array}{ccccc} & & U_{\mathcal{N},K}(\cdot,\omega) & \xrightarrow{\mathcal{N}\to\infty} & U_{,K}(\cdot,\omega) & & \\ K\to\mathcal{K} & & \downarrow & & \downarrow & & K\to\mathcal{K} \text{ in } L^\infty_{\mathbb{P}}(\Omega,X)\ . \\ & & U_{\mathcal{N}}(\cdot,\omega) & \xrightarrow{\mathcal{N}\to\infty} & U(\cdot,\omega) & & \end{array} \tag{6-III.12}$$

Finally, because $S_{(\mathcal{N})(,K)}(\omega)$ are linear functionals of $U_{(\mathcal{N})(,K)}(\cdot,\omega)$ and by continuity of the trace of $U_{(\mathcal{N})(,K)}(\cdot,\omega) \in H^1(\mathcal{D})$ on Γ_{B}, the diagram of convergences (6-III.10) holds. □

Proposition 26. *Under the same standard regularity hypotheses (as $\mathcal{N}\to\infty$) on the family of FE spaces $X_{\mathcal{N}}$ as in Proposition 26, the following convergence holds*

$$\begin{array}{ccccc} & & (\mathbf{E}_{\mathbb{P}}(S_{\mathcal{N},K}), \mathbf{Var}_{\mathbb{P}}(S_{\mathcal{N},K})) & \xrightarrow{\mathcal{N}\to\infty} & (\mathbf{E}_{\mathbb{P}}(S_{,K}), \mathbf{Var}_{\mathbb{P}}(S_{,K})) & & \\ K\to\mathcal{K} & & \downarrow & & \downarrow & & K\to\mathcal{K}\ ; \\ & & (\mathbf{E}_{\mathbb{P}}(S_{\mathcal{N}}), \mathbf{Var}_{\mathbb{P}}(S_{\mathcal{N}})) & \xrightarrow{\mathcal{N}\to\infty} & (\mathbf{E}_{\mathbb{P}}(S), \mathbf{Var}_{\mathbb{P}}(S)) & & \end{array} \tag{6-III.13}$$

Proof.
Because $S_{(\mathcal{N})(,K)}(\omega) \in L^\infty_{\mathbb{P}}(\Omega) \subset L^2_{\mathbb{P}}(\Omega)$, we simply use the following estimates which hold for any two linear functionals S_1, S_2 of random fields $U_1(\cdot,\omega), U_2(\cdot,\omega)$ in $L^\infty_{\mathbb{P}}(\Omega,X)$ and some positive constant C_0,

$$|\mathbf{E}_{\mathbb{P}}(S_1) - \mathbf{E}_{\mathbb{P}}(S_2)| \le \int_{\Omega} d\mathbb{P}(\omega) \int_{\Gamma_{\mathrm{R}}} |U_1(\cdot,\omega) - U_2(\cdot,\omega)| \le |\Gamma_{\mathrm{R}}| \|U_1(\cdot,\omega) - U_2(\cdot,\omega)\|_{0,\partial\mathcal{D}\times\Omega}\ , \tag{6-III.14}$$

$$|\mathbf{Var}_{\mathbb{P}}(S_1) - \mathbf{Var}_{\mathbb{P}}(S_2)| \le C_0 \max_{i=1,2} \|U_i(\cdot,\omega)\|_{0,\partial\mathcal{D}\times\Omega} \|U_1(\cdot,\omega) - U_2(\cdot,\omega)\|_{0,\partial\mathcal{D}\times\Omega}\ , \tag{6-III.15}$$

as well as the uniform bound (6-III.9) for all $U_{(\mathcal{N})(,K)}(\cdot,\omega)$, $1\le K\le\mathcal{K}$, and the compactness of the trace mapping from $H^1(\mathcal{D})$ into $L^2(\partial\mathcal{D})$. □

Lastly, for all positive integer M, we define, akin to (6-I.4), M i.i.d. copies $\left(S^m_{(\mathcal{N})(,K)}\right)_{1\le m\le M}$ of $S_{(\mathcal{N})(,K)}$ and empirical estimators for the expected values $\left(\mathbf{E}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right),\mathbf{Var}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right)\right)$ as

$$E_M\left[S_{(\mathcal{N})(,K)}\right]=\frac{1}{M}\sum_{m=1}^{M}S^m_{(\mathcal{N})(,K)}\ , \tag{6-III.16}$$

$$V_M\left[S_{(\mathcal{N})(,K)}\right]=\frac{1}{M-1}\sum_{m=1}^{M}\left(S^m_{(\mathcal{N})(,K)}-E_M\left[S_{(\mathcal{N})(,K)}\right]\right)^2\ . \tag{6-III.17}$$

The results in (6-III.13) for real numbers $\left(\mathbf{E}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right),\mathbf{Var}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right)\right)$ also clearly hold $\mathbb{P}$-a.s. for the discrete sums $(E_M[S_{(\mathcal{N})(,K)}],V_M[S_{(\mathcal{N})(,K)}])$ for any $M>0$; and by SLLN, it also $\mathbb{P}$-a.s. holds :

$$\left(E_M[S_{(\mathcal{N})(,K)}],V_M[S_{(\mathcal{N})(,K)}]\right)\xrightarrow[M\to\infty]{\mathbb{P}-a.s.}\left(\mathbf{E}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right),\mathbf{Var}_{\mathbb{P}}\left(S_{(\mathcal{N})(,K)}\right)\right).$$

Now, assume sufficient regularity on the PDE data such that the FE approximations $U_{\mathcal{N}}(\cdot,\omega)$ are $\mathbb{P}$-a.s. sufficiently close to $U(\cdot,\omega)$ (for some large $\mathcal{N}$), and that furthermore the accuracy required in the evaluation of the outputs $\mathbf{E}_{\mathbb{P}}\left(S_{(,K)}\right),\mathbf{Var}_{\mathbb{P}}\left(S_{(,K)}\right)$ (respectively $E_M[S_{(,K)}],V_M[S_{(,K)}]$) is provided by $\mathbf{E}_{\mathbb{P}}\left(S_{\mathcal{N}(,K)}\right),\mathbf{Var}_{\mathbb{P}}\left(S_{\mathcal{N}(,K)}\right)$ (respectively $E_M[S_{\mathcal{N}(,K)}],V_M[S_{\mathcal{N}(,K)}]$). Even then, the empirical estimations (6-III.16),(6-III.17) will still typically converge slowly : many evaluations of the FE approximation are required (M should be large) for the empirical estimations to be good approximations of the required statistical outputs.

In addition, even if, for a given (supposedly large) M, empirical estimations (6-III.16),(6-III.17) are assumed both sufficiently close to the required outputs and accessible to numerical computation for given parameters κ and $\mathrm{Bi}(\cdot,\omega)$, the evaluation of $E_M[S_{\mathcal{N}}]$ and $V_M[S_{\mathcal{N}}]$ for many values of these parameters in a many-query context is arguably prohibitive for a direct FE method.

In summary, the FE method with large $\mathcal{N}$ is too expensive to permit the rapid evaluation of empirical estimations (6-III.16),(6-III.17), first for a given large M, and second for many values of the (deterministic and stochastic) parameters κ and $\mathrm{Bi}(\cdot,\omega)$ in a many-query context in which M is fixed (presumably large).

Our Reduced Basis approach aims at reducing the computational cost of multiple (many) FE computations — without sacrificing certified accuracy — by exploiting the parametric structure of the problem through Offline-Online decompositions.

6-III-B Reduced-Basis Approximation

6-III-B-a A Deterministic Parametrized Problem

As mentioned in the introduction, we would like to map the sequence of random variables $(Z_k)_{1\le k\le\mathcal{K}}$ in (6-II.14) to random solution fields $U_{(\mathcal{N})(,K)}(\cdot,\omega)$, through the solutions $u_{(\mathcal{N})(,K)}(\cdot;y^{(K)})$ of deterministic BVP PDE problems parametrized by deterministic coefficients $y^{(K)}$, invoking the Doob-Dynkin lemma [Oks03].

Moreover, we would like to study variations of the statistical outputs on an "additional" deterministic parameter ϱ, corresponding to many given values of the (deterministic and stochastic) parameters κ and $\mathrm{Bi}(\cdot,\omega)$; this has also been mentioned previously. In the following, we take as "additional" deterministic parameter

$$\varrho=(\kappa,\overline{\mathrm{Bi}})\in\Lambda^{\varrho}.$$

Recall that truncations at order K of $Y=(Y_k)_{1\le k\le\mathcal{K}}$ $(1\le K\le\mathcal{K})$ have been defined in the introduction as

$$Y^K(\omega):=(Y_1(\omega),\dots,Y_K(\omega),0,0,\dots),\qquad\text{where } Y_k(\omega)=\Upsilon\sqrt{\lambda_k}\,Z_k(\omega),\ 1\le k\le K.$$

We also recall that has been set

$$\begin{aligned}
y&:=(y_1,y_2\dots)\in\Lambda^y\subset\mathbb{R}^{\mathbb{N}}\text{ such that for all finite positive integer }1\le K\le\mathcal{K},\\
y^K&:=(y_1,\dots,y_K,0,0,\dots)\in\Lambda^y\text{ and the range }\Lambda^y\text{ is the cylinder}\\
\Lambda^y&:=\left[-\sqrt{3}\Upsilon\sqrt{\lambda_1},+\sqrt{3}\Upsilon\sqrt{\lambda_1}\right]\times\left[-\sqrt{3}\Upsilon\sqrt{\lambda_2},+\sqrt{3}\Upsilon\sqrt{\lambda_2}\right]\times\cdots\subset\mathbb{R}^{\mathcal{K}}\ .
\end{aligned}$$

It is important to note that when the eigenvalues λ_k decay rapidly with k, the extent $2\sqrt{3}\Upsilon\sqrt{\lambda_k}$ of the intervals $\left[-\sqrt{3}\Upsilon\sqrt{\lambda_k},+\sqrt{3}\Upsilon\sqrt{\lambda_k}\right]$ will also shrink rapidly. This small range in the y_k for larger k is one of the reasons the RB approximation developed subsequently will converge quickly. [6] A function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ has been defined on the boundary, parametrized by $\overline{\mathrm{Bi}}$ and by the deterministic parameters $y_k \in \left[-\sqrt{3}\Upsilon\sqrt{\lambda_k},+\sqrt{3}\Upsilon\sqrt{\lambda_k}\right]$ $(1 \le k \le K \le \mathcal{K})$

$$\mathrm{bi}(x;\overline{\mathrm{Bi}},y) := \overline{\mathrm{Bi}}\left(G(x)+\sum_{k=1}^{\mathcal{K}} y_k \Phi_k(x)\right),\ \forall x \in \partial\mathcal{D}; \tag{6-III.18}$$

note that the function $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)$ is well defined since, by assumption, the series (6-III.18) absolutely converges in $L^\infty(\Gamma_{\mathrm{B}})$ for a.e. $y \in \Lambda^y$ (see Section 6-II-D-b). Lastly, we denote the full parameter as $\mu := (\kappa,\overline{\mathrm{Bi}},y) \in \Lambda^\mu$ with countably (possibly infinite) entries, and truncated versions with $K+2$ entries (for any finite integer $1 \le K \le \mathcal{K}$)

$$\mu_K := (\kappa,\overline{\mathrm{Bi}},y^K) \in \Lambda^\mu \equiv \Lambda^\varrho \times \Lambda^y$$

where $\Lambda^\varrho \subset \mathbb{R}^2_{>0}$ denotes the range of $\varrho = (\kappa,\overline{\mathrm{Bi}})$ (at this point, there is no *a priori* assumption on Λ^ϱ : it is some subset of $\mathbb{R}^2_{>0}$ that will be made precise later in the numerical part).

Let us now introduce a deterministic BVP PDE problem parametrized by the deterministic (full) parameter $\mu \in \Lambda^\mu$. For every $\mu \in \Lambda^\mu$, with notations obviously in accordance with the previous Section 6-III-A, we define $u(\mu), u_{,K}(\mu_K) \in X$ and $u_{\mathcal{N}}(\mu), u_{\mathcal{N},K}(\mu_K) \in X_{\mathcal{N}}$ as solutions to the respective variational formulations

$$a(u_{(\mathcal{N})(,K)}(\mu_{(K)}),v;\mu_{(K)}) = f(v), \qquad \forall v \in X_{(\mathcal{N})}\ , \tag{6-III.19}$$

where the subscripts $(\mathcal{N})$ and (K) are simultaneoulsy active everywhere or not, and where the functional $f(\cdot)$ and the parametrized bilinear form $a(\cdot,\cdot;\mu)$ are given by :

$$f(v) = \int_{\Gamma_{\mathrm{R}}} v, \quad \forall v \in X\,, \tag{6-III.20}$$

$$a(w,v;\mu) = \int_{\mathcal{D}_1} \nabla w \cdot \nabla v + \kappa \int_{\mathcal{D}_2} \nabla w \cdot \nabla v + \int_{\Gamma_{\mathrm{B}}} \mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y) wv, \quad \forall w,v \in X\,. \tag{6-III.21}$$

We may then define our realization output as

$$s_{(\mathcal{N})(,K)}(\mu_{(K)}) = f\left(u_{(\mathcal{N})(,K)}(\mu_{(K)})\right)\ . \tag{6-III.22}$$

Clearly, there exists a sequence $\mathcal{M}$ of random variable in $L^\infty_{\mathbb{P}}(\Omega)$, with range Λ^μ such that for a.e. ω in Ω it holds $\mathcal{M}(\omega) = (\kappa,\overline{\mathrm{Bi}},Y(\omega))$. We then define truncations such that $\mathbb{P}$-a.s., $\forall 1 \le K \le \mathcal{K}$

$$\mathcal{M}_K(\omega) = (\kappa,\overline{\mathrm{Bi}},Y^K(\omega))\,,$$

which implies in return, provided $U_{(\mathcal{N})(,K)}(\cdot,\omega)$ is well defined, that $\mathbb{P}$-a.s. holds

$$u_{(\mathcal{N})(,K)}(\mathcal{M}_{(K)}(\omega)) = U_{(\mathcal{N})(,K)}(\cdot,\omega)\ ,$$
$$s_{(\mathcal{N})(,K)}(\mathcal{M}_{(K)}(\omega)) = S_{(\mathcal{N})(,K)}(\omega)\ .$$

Moreover, for each $M > 0$, we define M i.i.d. copies $(\mathcal{M}^m)_{1 \le m \le M}$ of the random variable $\mathcal{M}$ such that the empirical estimations

$$E_M[s_{(\mathcal{N})(,K)}(\mathcal{M}_{(K)})] = \frac{1}{M}\sum_{m=1}^{M} s_{(\mathcal{N})(,K)}(\mathcal{M}^m_{(K)})\ , \tag{6-III.23}$$

$$V_M[s_{(\mathcal{N})(,K)}(\mathcal{M}_{(K)})] = \frac{1}{M-1}\sum_{m=1}^{M}\left(E_M[s_{(\mathcal{N})(,K)}] - s_{(\mathcal{N})(,K)}(\mathcal{M}^m_{(K)})\right)^2\ , \tag{6-III.24}$$

[6] Note we can treat with a single RB many different covariance functions of varying smoothness if we introduce the parameters y_k in the interval (say) $\left[-\sqrt{3}\Upsilon,\sqrt{3}\Upsilon\right]$ independent of k such that $y \equiv (y_1,\ldots,y_K) \in \mathcal{L}^y_K \equiv \left[-\sqrt{3}\Upsilon,\sqrt{3}\Upsilon\right]^K \subset \mathbb{R}^K$. However, in this case the reduced basis approximation will converge much more slowly since the parameter space $\mathcal{L}^y_K$ is much larger.

coincide $\mathbb{P}$-a.s. with $E_M[S_{(\mathcal{N})(,K)}]$ and $V_M[S_{(\mathcal{N})(,K)}]$ as statistical approximations of the expected value and variance $\mathbf{E}_{\mathbb{P}}\big(S_{(\mathcal{N})(,K)}\big)$ and $\mathbf{Var}_{\mathbb{P}}\big(S_{(\mathcal{N})(,K)}\big)$, respectively. Note that all the convergence results established in the previous Section 6-III-A for $\mathcal{N}, K \to \infty$ still hold for $s_{(\mathcal{N})(,K)}(\mu_{(K)})$ and a fixed parameter value μ.

In the following, we shall develop a reduced basis (RB) approximation and associated *a posteriori* error estimator which will permit rapid and reliable evaluation of the empirical approximations (6-III.23) and (6-III.24) for the outputs of interest (the expected value and variance $(\mathbf{E}_{\mathbb{P}}(S), \mathbf{Var}_{\mathbb{P}}(S))$). Our RB approximation will be based upon, and the RB error will be measured relative to, the FE approximation $u_{\mathcal{N},K}(\mu_K)$ of (6-III.19), for a fixed parameter value $\mu \in \Lambda^\mu$. Note we assume that $\mathcal{N}$ is chosen sufficiently large *a priori* to provide the desired accuracy relative to the exact solution ; we shall thus concentrate our *a posteriori* estimation and control on the RB approximation and on the KL truncation (note it is very simple to change the order of KL truncation in a strong-weak formulation). As we shall see, the total RB cost (Offline and Online, see Section 6-III-D) will actually depend rather weakly on $\mathcal{N}$, and hence $\mathcal{N}$ may be chosen conservatively.

6-III-B-b RB Approximation

Let $N_{\max}$ X-orthonormalized basis functions $\zeta_n \in X_{\mathcal{N}}$, $1 \le n \le N_{\max}$ ($N_{\max} \le \mathcal{N}$) be given, and define the associated hierarchical Lagrange [Por85] RB spaces $X_N \subset X_{\mathcal{N}}$, $1 \le N \le N_{\max}$, as

$$X_N = \text{span}\{\zeta_n, 1 \le n \le N\}, \quad N = 1, \ldots, N_{\max} \ . \tag{6-III.25}$$

In practice (see Section 6-III-D), the spaces X_N will be generated by a Greedy sampling procedure [NVP05b, RHP08] ; for our present purpose, however, X_N can in fact represent any sequence of (low-dimensional) hierarchical approximation spaces. Let the KL expansion of the random input field be truncated at some finite order K, the (N,K)-RB approximation of the problem (6-III.19) then reads :
Given $\mu \in \Lambda^\mu$, we look for an RB approximation $u_{N,K}(\mu_K) \in X_N$ such that

$$a_K(u_{N,K}(\mu_K), v; \mu_K) = f(v), \qquad \forall v \in X_N \, . \tag{6-III.26}$$

We then calculate the RB realization output as

$$s_{N,K}(\mu_K) = \int_{\Gamma_{\mathrm{R}}} u_{N,K}(\mu_K) \ . \tag{6-III.27}$$

The RB output will be evaluated in the Online stage, by the procedure described in Section 6-III-D, with a computational cost depending on N and K but *not* on $\mathcal{N}$: hence, for small N and K, the RB approximation can be significantly less expensive than the FE approximation.

We shall use this RB approximation to approximate the expected value and variance of the output of interest, for sufficiently large integer $M > 0$, through the empirical estimations

$$E_M[s_{N,K}(\mathcal{M}_K)] = \frac{1}{M} \sum_{m=1}^{M} s_{N,K}(\mathcal{M}_K^m) \ , \tag{6-III.28}$$

$$V_M[s_{N,K}(\mathcal{M}_K)] = \frac{1}{M-1} \sum_{m=1}^{M} \big(E_M[s_{N,K}(\mathcal{M}_{(K)})] - s_{N,K}(\mathcal{M}_K^m) \big)^2 \ . \tag{6-III.29}$$

In the next section we develop rigorous *a posteriori* bounds for these quantities relative to $E_M[s_{(\mathcal{N}),(\mathcal{K})}(\mathcal{M}_{(K)})]$ and $V_M[s_{(\mathcal{N}),(\mathcal{K})}(\mathcal{M}_{(K)})]$, respectively.

6-III-C *A Posteriori* Error Estimation

6-III-C-a Error Bounds for the RB Output

We note from (6-III.26) that, for any $\mu \in \Lambda^\mu$, the residual $r(v;\mu_K)$ associated with $u_{N,K}(\mu_K)$ reads

$$r(v;\mu_K) = f(v) - a(u_{N,K}(\mu_K), v; \mu_K), \qquad \forall v \in X_{\mathcal{N}} \ ; \tag{6-III.30}$$

the dual norm of the residual (defined over the FE "truth" space) is given by

$$\|r(\cdot;\mu_K)\|_{X'_{\mathcal{N}}} = \sup_{v \in X_{\mathcal{N}}} \frac{r(v;\mu_K)}{\|v\|_X} \ . \tag{6-III.31}$$

We next introduce a bilinear form parametrized by the deterministic parameter $\varrho=(\kappa,\overline{\mathrm{Bi}})$ but independent of the parameter y,

$$a_C(w,v;(\kappa,\overline{\mathrm{Bi}}))=\int_{\mathcal{D}_1}\nabla w\cdot\nabla v+\kappa\int_{\mathcal{D}_2}\nabla w\cdot\nabla v+\frac{\overline{\mathrm{Bi}}}{2}\int_{\Gamma_\mathrm{B}}G(x)wv,\qquad\forall w,v\in X_\mathcal{N},\tag{6-III.32}$$

such that, since $\mathrm{Bi}_K(x,y^K)\geq\overline{\mathrm{Bi}}G(x)/2$, $\forall x\in\Gamma_B$, by (6-II.19) (assumption H5)

$$a_C(v,v;(\kappa,\overline{\mathrm{Bi}}))\leq a(v,v;\mu_K)\,,\ \forall\mu\in\Lambda^\mu,\forall v\in X_\mathcal{N},\forall 1\leq K\leq\mathcal{K}.$$

Denoting $\alpha(\mu_K)$ the coercivity constant associated with $a(\cdot,\cdot;\mu_K)$, it follows

$$\alpha_C(\kappa,\overline{\mathrm{Bi}})=\inf_{v\in X_\mathcal{N}}\frac{a_C(v,v;(\kappa,\overline{\mathrm{Bi}}))}{\|v\|_X^2}\leq\alpha(\mu_K):=\inf_{v\in X_\mathcal{N}}\frac{a(v,v;\mu_K)}{\|v\|_X^2},\qquad\forall\mu\in\Lambda^\mu\ .\tag{6-III.33}$$

It should be noted that $\alpha_C(\kappa,\overline{\mathrm{Bi}})$ depends only on the deterministic parameters κ and $\overline{\mathrm{Bi}}$, *not* on the (ultimately mapped to a random) parameter y^K ! The following result is standard [Boy08, NVP05b, RHP08].

Proposition 27. *Given a computable lower bound α_LB for $\alpha_C(\kappa,\overline{\mathrm{Bi}})$, thus also for $\alpha(\mu_K)$, $\forall\mu\in\Lambda^\mu$, the following a posteriori estimates hold for all positive integers $N,\mathcal{N},K$*

$$\|u_{\mathcal{N},K}(\mu_K)-u_{N,K}(\mu_K)\|_X\leq\Delta_{N,K}(\mu_K)\equiv\frac{\|r(\cdot;\mu_K)\|_{X'_\mathcal{N}}}{\alpha_\mathrm{LB}}\ ,\tag{6-III.34}$$

$$|s_{\mathcal{N},K}(\mu_K)-s_{N,K}(\mu_K)|\leq\Delta^s_{N,K}(\mu_K)\equiv\frac{\|r(\cdot;\mu_K)\|^2_{X'_\mathcal{N}}}{\alpha_\mathrm{LB}}\ .\tag{6-III.35}$$

6-III-C-b Error Bounds for the KL Truncation Effect

We now bound the error $|s_\mathcal{N}(\mu)-s_{\mathcal{N},K}(\mu_K)|$ due to the truncation of the KL expansion for any $\mu\in\Lambda^\mu$, where μ_K is the truncated version that retains the $K+2$ first entries of μ.

Proposition 28. *With the same lower bound α_LB as in Proposition 27, $\forall\mu\in\Lambda^\mu$, holds for all positive integer $N,\mathcal{N},K$*

$$|s_\mathcal{N}(\mu)-s_{\mathcal{N},K}(\mu_K)|\leq\Delta^t_{N,K}(\mu_K)\equiv\frac{\overline{\mathrm{Bi}}\tau_K\gamma_\mathcal{N}}{\alpha_\mathrm{LB}}\|f\|_{X'_\mathcal{N}}(\|u_{N,K}(\mu_K)\|_X+\Delta_{N,K}(\mu_K))\ ,\tag{6-III.36}$$

where $\Delta_{N,K}(\mu)$ is the error bound defined above in (6-III.34) *for $\|u_{\mathcal{N},K}(\mu_K)-u_{N,K}(\mu_K)\|_{X_\mathcal{N}}$ and τ_K is the bound introduced in* (6-II.20).

Proof.
First note that

$$\begin{aligned}|s_\mathcal{N}(\mu)-s_{\mathcal{N},K}(\mu_K)|&=|f(u_\mathcal{N}(\mu)-u_{\mathcal{N},K}(\mu_K))|\\&\leq\|f\|_{X'_\mathcal{N}}\|u_\mathcal{N}(\mu)-u_{\mathcal{N},K}(\mu_K)\|_X\ .\end{aligned}\tag{6-III.37}$$

Then, to get (6-III.36), we now show that the last term is bounded by

$$\|u_\mathcal{N}(\mu)-u_{\mathcal{N},K}(\mu_K)\|_X\leq\frac{\overline{\mathrm{Bi}}\tau_K\gamma_\mathcal{N}}{\alpha_\mathrm{LB}}(\|u_{N,K}(\mu_K)\|_X+\Delta_{N,K}(\mu_K))\ ,\tag{6-III.38}$$

where $\overline{\mathrm{Bi}}\tau_K$ is the error bound for $\|\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)-\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y^K)\|_{L^\infty(\Gamma_\mathrm{B})}$ introduced in (6-II.20), and $\gamma_\mathcal{N}$ is the continuity constant for the trace application $X_\mathcal{N}\to\Gamma_\mathrm{B}$ already defined in (6-III.2).

To prove (6-III.38), we substract the truncated and full problems (6-III.19) after FE discretization, and choose $v=e_{\mathcal{N},K}(\mu)=u_\mathcal{N}(\mu)-u_{\mathcal{N},K}(\mu_K)$ as test function. We obtain

$$a(e_{\mathcal{N},K}(\mu),e_{\mathcal{N},K}(\mu);\mu)=-\int_{\Gamma_\mathrm{B}}(\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y)-\mathrm{bi}_K(\cdot;\overline{\mathrm{Bi}},y^K))\,u_{\mathcal{N},K}(\mu_K)\,e_{\mathcal{N},K}(\mu)\ .\tag{6-III.39}$$

Furthermore, the left-hand side (LHS) of (6-III.39) is bounded below by

$$\begin{aligned}\mathrm{LHS}&\geq a_C(e_{\mathcal{N},K}(\mu),e_{\mathcal{N},K}(\mu);(\kappa,\overline{\mathrm{Bi}}))\\&\geq\alpha_\mathrm{LB}\|e_{\mathcal{N},K}(\mu)\|_X^2\ ,\end{aligned}\tag{6-III.40}$$

and the right-hand side (RHS) of (6-III.39) is bounded above by

$$\begin{aligned}|\text{RHS}| &\le \overline{\text{Bi}}\tau_K \|u_{\mathcal{N},K}(\mu_K)\|_{L^2(\Gamma_\text{B})} \|e_{\mathcal{N},K}\|_{L^2(\Gamma_\text{B})} \\ &\le \overline{\text{Bi}}\tau_K\gamma_{\mathcal{N}} \|u_{\mathcal{N},K}(\mu_K)\|_X \|e_{\mathcal{N},K}(\mu)\|_X \\ &\le \overline{\text{Bi}}\tau_K\gamma_{\mathcal{N}} (\|u_{N,K}(\mu_K)\|_X + \Delta_{N,K}(\mu_K)) \|e_{\mathcal{N},K}(\mu)\|_X \ . \end{aligned} \tag{6-III.41}$$

The desired result (6-III.38) follows directly from (6-III.39)– (6-III.41). □

6-III-C-c Error Bounds for the Expected Value and Variance

Using the notations introduced in (6-III.35) and (6-III.36) we have, from the triangle inequality,

$$|s_{\mathcal{N}}(\mu) - s_{N,K}(\mu_K)| \le \Delta^o_{N,K}(\mu_K) := \Delta^s_{N,K}(\mu_K) + \Delta^t_{N,K}(\mu_K) \ . \tag{6-III.42}$$

Thus we obtain the error bound for the error in the expected value $\mathbb{P}$-a.s. as

$$|E_M[s_{\mathcal{N}}(\mathcal{M})] - E_M[s_{N,K}(\mathcal{M}_K)]| \le \Delta^o_E[s_{N,K}(\mathcal{M}_K)] := \Delta^s_E[s_{N,K}(\mathcal{M}_K)] + \Delta^t_E[s_{N,K}(\mathcal{M}_K)] \ , \tag{6-III.43}$$

using M i.i.d. (truncated) copies $(\mathcal{M}^m)_{1\le m\le M}$ of $\mathcal{M}$, and the following random variables :

$$\Delta^s_E[s_{N,K}(\mathcal{M}_K)] \equiv \frac{1}{M}\sum_{m=1}^M \Delta^s_{N,K}(\mathcal{M}^m_K) \ , \qquad \Delta^t_E[s_{N,K}(\mathcal{M}_K)] \equiv \frac{1}{M}\sum_{m=1}^M \Delta^t_{N,K}(\mathcal{M}^m_K) \ . \tag{6-III.44}$$

The error bound (6-III.43) consists of the RB estimate (6-III.35) and the KL truncation estimate (6-III.36). The two estimates depend on both N and K but in different ways : the former will decrease rapidly with increasing N and typically increase with increasing K, while the latter will decrease rapidly with increasing K.

For the error bound in the variance, we introduce a function of $\mu \in \Lambda^\mu$

$$s^\pm_{N,K}(\mu_K) := s_{N,K}(\mu_K) \pm \Delta^o_{N,K}(\mu_K) \ , \tag{6-III.45}$$

a random variable that is a sum of MC estimators :

$$E^\pm_M[s_{N,K}(\mathcal{M}_K)] := E_M[s_{N,K}(\mathcal{M}_K)] \pm \Delta^o_E[s_{N,K}(\mathcal{M}_K)] \ , \tag{6-III.46}$$

and random variables parametrized by $\mu_K \in \Lambda^\mu$

$$\begin{aligned} A_{N,K}(\mathcal{M}_K;\mu_K) &:= E^+_M[s_{N,K}(\mathcal{M}_K)] - s^-_{N,K}(\mu_K) \ , \\ B_{N,K}(\mathcal{M}_K;\mu_K) &:= E^-_M[s_{N,K}(\mathcal{M}_K)] - s^+_{N,K}(\mu_K) \ , \\ C_{N,K}(\mathcal{M}_K;\mu_K) &:= \begin{cases} 0 \quad \text{if } [s^-_{N,K}(\mu_K), s^+_{N,K}(\mu_K)] \cap \left[E^-_M[s_{N,K}(\mathcal{M}_K)], E^+_M[s_{N,K}(\mathcal{M}_K)]\right] \ne \emptyset \\ \min\{|A_{N,K}(\mathcal{M}_K;\mu_K)|, |B_{N,K}(\mathcal{M}_K;\mu_K)|\} \quad \text{otherwise} \end{cases} \ , \\ D_{N,K}(\mathcal{M}_K;\mu_K) &:= \max\{|A_{N,K}(\mathcal{M}_K;\mu_K)|, |B_{N,K}(\mathcal{M}_K;\mu_K)|\} \ . \end{aligned} \tag{6-III.47}$$

We thus have $\mathbb{P}$-a.s.

$$C^2_{N,K}(\mathcal{M}_K;\mu_K) \le (E_M[s_{\mathcal{N}}(\mathcal{M}_K)] - s_{\mathcal{N}}(\mu_K))^2 \le D^2_{N,K}(\mathcal{M}_K;\mu_K) \ , \tag{6-III.48}$$

and hence after summation, also $\mathbb{P}$-a.s.

$$V^{\text{LB}}_M[s_{N,K}(\mathcal{M}_K)] \le V_M[s_{\mathcal{N}}(\mathcal{M}_K)] \le V^{UB}_M[s_{N,K}(\mathcal{M}_K)] \ , \tag{6-III.49}$$

where we have used the MC estimators

$$V^{\text{LB}}_M[s_{N,K}(\mathcal{M}_K)] := \frac{1}{M-1}\sum_{m=1}^M C^2_{N,K}(\mathcal{M}_K;\mathcal{M}^m_K) \ , \quad V^{\text{UB}}_M[s_{N,K}] := \frac{1}{M-1}\sum_{m=1}^M D^2_{N,K}(\mathcal{M}_K;\mathcal{M}^m_K) \ , \tag{6-III.50}$$

with the same collection $\{\mathcal{M}^m_K\}$ as in the MC estimators (6-III.47).

Thus we obtain $\mathbb{P}$-a.s. a bound for the error in the variance as

$$|V_M[s_{\mathcal{N}}(\mathcal{M}_K)] - V_M[s_{N,K}](\mathcal{M}_K)| \le \Delta^o_V[s_{N,K}(\mathcal{M}_K)] \tag{6-III.51}$$

with

$$\Delta_V^o[s_{N,K}(\mathcal{M}_K)] \equiv \max\Big\{ |V_M[s_{N,K}(\mathcal{M}_K)] - V_M^{\rm UB}[s_{N,K}(\mathcal{M}_K)]|, \\ |V_M[s_{N,K}(\mathcal{M}_K)] - V_M^{\rm LB}[s_{N,K}(\mathcal{M}_K)]| \Big\}. \quad \text{(6-III.52)}$$

This variance error bound also includes both an RB contribution and a KL truncation contribution.

Finally, although it is not our main goal, we point out that without consideration of the KL truncation effect we may also obtain the error bounds (at fixed K)

$$\begin{aligned} |E_M[s_{\mathcal{N},K}(\mathcal{M}_K)] - E_M[s_{N,K}(\mathcal{M}_K)]| &\le \Delta_E^s[s_{N,K}(\mathcal{M}_K)] \ , \\ |V_M[s_{\mathcal{N},K}(\mathcal{M}_K)] - V_M[s_{N,K}(\mathcal{M}_K)]| &\le \Delta_V^s[s_{N,K}(\mathcal{M}_K)] \ . \end{aligned} \quad \text{(6-III.53)}$$

Here $\Delta_E^s[s_{N,K}(\mathcal{M}_K)]$ is given by (6-III.44), and $\Delta_V^s[s_{N,K}(\mathcal{M}_K)]$ is defined in the same way as $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ but replacing $\Delta_{N,K}^o(\mu_K)$ with $\Delta_{N,K}^s(\mu_K)$ in (6-III.45) and $\Delta_E^o[s_{N,K}]$ with $\Delta_E^s[s_{N,K}]$ in (6-III.46). We introduce the contribution due to the KL truncation to the variance error bound (6-III.51) as

$$\Delta_V^t[s_{N,K}(\mathcal{M}_K)] \equiv \Delta_V^o[s_{N,K}(\mathcal{M}_K)] - \Delta_V^s[s_{N,K}(\mathcal{M}_K)]. \quad \text{(6-III.54)}$$

6-III-D Offline-Online Computational Approach

6-III-D-a Construction-Evaluation Decomposition

The system (6-III.26) comprises N linear algebraic equations in N unknowns. However, its formation involves entities $\zeta_n, 1 \le n \le N$, associated with the $\mathcal{N}$-dimensional FE approximation space. If we must invoke FE fields in order to form the system *for each new value of* μ, the marginal cost per input-output evaluation $\mu \to s_{N,K}(\mu_K)$ will remain unacceptably large. Fortunately, we can compute this output very efficiently by constructing Offline-Online procedures [NVP05b, PRV$^+$02, RHP08], as we now discuss.

First, we note that the bilinear form a_K as introduced in (6-III.21) can be expressed as the following "affine" decomposition

$$a(w,v;\mu_K) = \sum_{k=1}^{K+3} \Theta_k(\mu_K) a_k(w,v), \quad \forall w,v \in X \ . \quad \text{(6-III.55)}$$

Here $\Theta_1(\mu_K) = 1$, $\Theta_2(\mu_K) = \kappa$, $\Theta_3(\mu_K) = \overline{\rm Bi}$, and $\Theta_{3+k}(\mu_K) = \overline{\rm Bi}\, y_k$, $1 \le k \le K$, are parameter-*dependent* functions, and $a_1(w,v) = \int_{\mathcal{D}_1} \nabla w \cdot \nabla v$, $a_2(w,v) = \int_{\mathcal{D}_2} \nabla w \cdot \nabla v$, $a_3(w,v) = \int_{\Gamma_B} G(x) wv$, and $a_{3+k}(w,v) = \int_{\Gamma_B} \Phi_k(\cdot) wv$, $1 \le k \le K$, are parameter-*independent* bilinear forms. Note the crucial role of the "separable" (in ω and x) form of the KL expansion is ensuring an affine representation; the affine representation is, in turn, crucial to the Offline-Online strategy.

We next express $u_{N,K}(\mu_K) = \sum_{m=1}^N c_{N,K,m}(\mu_K)\zeta_m$, choose $v = \zeta_n$, $1 \le n \le N$, and invoke the affine representation (6-III.55) to write the system (6-III.26) as

$$\sum_{m=1}^N \left(\sum_{k=1}^{K+3} \Theta_k(\mu_K) a_k(\zeta_m, \zeta_n) \right) c_{N,K,m}(\mu_K) = f(\zeta_n), \qquad 1 \le n \le N \ , \quad \text{(6-III.56)}$$

and subsequently evaluate our RB output as

$$s_{N,K}(\mu_K) = \sum_{n=1}^N c_{N,K,n}(\mu_K) f(\zeta_n) \ . \quad \text{(6-III.57)}$$

We observe that the quantities $a_k(\zeta_m, \zeta_n)$ and $f(\zeta_n)$ are independent of μ and thus can be pre-computed in a Construction-Evaluation decomposition.

In the Construction phase, we form and store the $f(\zeta_n)$ and $a_k(\zeta_m, \zeta_n)$, $1 \le n, m \le N_{\max}$, $1 \le k \le K+3$. In the Evaluation phase, we first perform the sum $\sum_{k=1}^{K+3} \Theta_k(\mu_K) a_k(\zeta_m, \zeta_n)$, we next solve the resulting $N \times N$ system (6-III.56) to obtain the $c_{N,K,n}(\mu_K)$, $1 \le n \le N$, and finally we evaluate the output (6-III.57). The operation count for the Evaluation phase is $O((K+3)N^2)$ to perform the sum, $O(N^3)$ to invert (6-III.56), and finally $O(N)$ to effect the inner product (6-III.57); the storage for the Evaluation phase (the data archived in the Construction

phase) is only $O(N_{\max}+(K+3)N_{\max}^2)$. The Evaluation cost (operation cost and storage) — and hence marginal cost and also asymptotic average cost — to evaluate $\mu \rightarrow s_{N,K}(\mu_K)$ is thus independent of $\mathcal{N}$. The implications are twofold : first, if N and K are indeed small, we shall achieve very fast response in many-query contexts (in which the initial Offline investment is eventually "forgotten") ; second, we may choose $\mathcal{N}$ very conservatively — to effectively eliminate the error between the exact and FE predictions — without adversely affecting the Evaluation (marginal) cost.

The Construction-Evaluation for the error bounds is a bit more involved. To begin, we note from standard duality arguments that $\|r(\cdot;\mu_K)\|_{X'_{\mathcal{N}}} = \|\mathcal{R}_{N,K}(\mu_K)\|_X$; here $\mathcal{R}_{N,K}(\mu_K) \in X_{\mathcal{N}}$ satisfies $(\mathcal{R}_{N,K}(\mu_K),v)_X = r(v;\mu_K)$, $\forall v \in X_{\mathcal{N}}$, where $r(v;\mu_K) \equiv f(v) - a(u_N(\mu),v;\mu_K)$, $\forall v \in X_{\mathcal{N}}$, is the residual introduced earlier. We can thus express (6-III.34) and (6-III.35) as

$$\Delta_{N,K}(\mu_K) = \frac{\|\mathcal{R}_{N,K}(\mu_K)\|_X}{\alpha_{\rm LB}}, \quad \text{and} \quad \Delta^s_{N,K}(\mu_K) = \frac{\|\mathcal{R}_{N,K}(\mu_K)\|^2_X}{\alpha_{\rm LB}} \ . \tag{6-III.58}$$

There are two components to the error bounds : the dual norm of the residual, $\|\mathcal{R}_{N,K}(\mu_K)\|_X$, and our lower bound for the coercivity constant, $\alpha_{\rm LB}$. The Construction-Evaluation decomposition for the coercivity constant lower bound is based on the Successive Constraint Method (SCM) described in detail in [CHMR08, HRSP07a, RHP08]. We focus here on the Construction-Evaluation decomposition for the dual norm of the residual and express our residual $r(v;\mu_K)$ in terms of (6-III.55)

$$(\mathcal{R}_{N,K}(\mu),v)_X = f(v) - \sum_{k=1}^{K+3}\sum_{n=1}^{N} \Theta_k(\mu) c_{N,K\,n}(\mu) a_k(\zeta_n,v), \quad \forall v \in X \ ,$$

and hence obtain by linear superposition

$$\mathcal{R}_{N,K}(\mu_K) = z_0 + \sum_{k=1}^{K+3}\sum_{n=1}^{N} \Theta_k(\mu_K) c_{N,K,n}(\mu_K) z^k_n \ ,$$

where $(z_0,v)_X = f(v)$, and $(z^k_n,v)_X = -a_k(\zeta_n,v)$, $\forall v \in X_{\mathcal{N}}$, $1 \le n \le N$, $1 \le k \le K+3$, thus

$$\|\mathcal{R}_{N,K}\|^2_X = (z_0,z_0)_X + 2\sum_{k,n=1}^{K+3,N} \Theta_k(\mu_K) c_{N,K\,n}(\mu_K)(z^k_n,z_0)_X + \sum_{k,k',n,n'=1}^{K+3,K+3,N,N} \Theta_k(\mu_K) c_{N,K,n}(\mu_K) \Theta_{k'}(\mu_K) c_{N,K,n'}(\mu_K)(z^k_n,z^{k'}_{n'})_X \ . \tag{6-III.59}$$

Since the $(\cdot,\cdot)_X$ inner products are independent of μ, we can pre-compute these quantities in the Construction-Evaluation decomposition.

In the Construction phase — parameter independent, and performed only once — we find z_0, z^k_n, $1 \le k \le K+3$, $1 \le n \le N$, and then form and store the inner products $(z_0,z_0)_X$, $(z^k_n,z_0)_X$, $1 \le k \le K+3$, $1 \le n \le N$, and $(z^k_n,z^{k'}_{n'})_X$, $1 \le k,k' \le K+3$, $1 \le n,n' \le N$. Then, in the Evaluation phase — given any desired value of μ_K — we simply evaluate (6-III.58) from the summation (6-III.59) and the SCM evaluation for $\alpha_{\rm LB}$ at cost $O((K+3)^2N^2)$. The crucial point, again, is that the cost and storage in the Evaluation phase — the *marginal* cost for each new value of μ — is independent of $\mathcal{N}$: thus we can not only evaluate our output prediction but also our rigorous output error bound very rapidly in the many-query (or real-time) context.

Finally, the error bound $\Delta^t_{N,K}(\mu_K)$ of (6-III.36) requires additional quantities : τ_K, $\gamma_{\mathcal{N}}$, $\|f\|_{X'_{\mathcal{N}}}$, and $\|u_{N,K}(\mu_K)\|_X$. Note the first three quantities are independent of μ : τ_K can be pre-computed for any $1 \le K \le \mathcal{K}$ from the expansion (6-II.20) ; $\gamma_{\mathcal{N}}$ can be pre-computed from the eigenvalue problem (6-III.3) ; and finally $\|f\|_{X'_{\mathcal{N}}}$ can be pre-computed (by duality) as a standard FE Poisson problem. We note further that

$$\|u_{N,K}(\mu_K)\|^2_X = \sum_{n,n'=1}^{N,N} c_{N,K,n}(\mu_K) c_{N,K,n'}(\mu_K)(\zeta_n,\zeta_{n'})_X \ , \tag{6-III.60}$$

which readily admits a Construction-Evaluation decomposition ; clearly, the Evaluation-phase summation (6-III.60) requires only $O(N^2)$ operations. In summary, in the Evaluation phase, we can evaluate $s_{N,K}(\mu_K)$, $\Delta^s_{N,K}(\mu_K)$, $\Delta^t_{N,K}(\mu_K)$, and $\Delta^o_{N,K}(\mu_K)$ at total cost $O(N^3+(K+3)^2N^2)$ operations.

6-III-D-b Greedy Sampling

Finally, we turn to the construction of our reduced basis ζ_n, $1 \le n \le N_{\max}$: we pursue a very simple but also very effective Greedy procedure [RHP08]. To initiate the Greedy procedure we specify a very large (exhaustive) "train" sample of n_{train} points in Λ^μ, Ξ_{train}, a maximum RB dimension $N_{\max}$, and an initial (say, random) sample $S_1 = \{\mu^1\}$ and associated RB space X_1. (In actual practice, we typically specify an error tolerance-*cum*-stopping criterion which then implicitly determines $N_{\max}$.) We specify $K = \mathcal{K}$ (in practice, finite) for the Greedy algorithm described below.

Then, for $N = 1, \ldots, N_{\max}$: Step (1) Find $\mu^{N+1} = \arg\max_{\mu \in \Xi_{\text{train}}} \Delta_{N,K}(\mu)$; Step (2) Update $S_{N+1} = S_N \cup \mu^{N+1}$ and $X_{N+1} = X_N + \text{span}\{u_{\mathcal{N},K}(\mu^{N+1})\}$. The heuristic is simple : we append to our sample the point μ^{N+1} which is least well represented by the space X_N (as predicted by the error bound associated with our RB Galerkin approximation). In practice, the basis must be orthogonalized with respect to the $(\cdot,\cdot)_X$ inner product ; the algebraic system then inherits the conditioning properties of the underlying partial differential equation. Note that the Greedy automatically generates *hierarchical* spaces X_N, $1 \le N \le N_{\max}$, which is computationally very advantageous.

The important point to note from the computational perspective is that the operation count for a few $N_{\max} \ll \mathcal{N}^k$ steps of the Greedy algorithm (using truncations at order $K = \mathcal{K} \ll \mathcal{N}^k$) is $O(\mathcal{N}^k + n_{\text{train}})$ and *not* $O(\mathcal{N}^k n_{\text{train}})$ (where $O(\mathcal{N}^k)$ is the complexity for numerically solving *one* system of size $\mathcal{N} \times \mathcal{N}$) — and hence much less expensive than classical approaches such as the KL (here Proper Orthogonal Decomposition, or POD) expansion for the sample $(u_{\mathcal{N},K}(\mu))_{\mu \in \Xi_{\text{train}}}$. The reason is simple : In Step (1), to calculate $\Delta_{N,K}(\mu)$ over Ξ_{train}, we invoke the Construction-Evaluation decomposition to obtain (per Greedy cycle) an operation count of $O(NK\mathcal{N}^k) + n_{\text{train}} O(K^2 N^2)$. (Of course, much of the computational economies are due not to the Greedy itself, but rather to the accommodation within the Greedy of the inexpensive error bounds.) As a result, we can take n_{train} very large — often 10^4 or larger — particularly important for the high — $K + P_\varrho$ — dimensional parameter domains encountered in the SPDE context (here P_ϱ is dimension of the deterministic parameter ϱ). Furthermore, extensive numerical results for a wide variety of problems indicate that the Greedy RB space X_N is typically as good as more global (and provably optimal) approaches such as the POD [RHP08]. (Of course, the latter result is norm dependent : the Greedy prefers $L^\infty(\Xi_{\text{train}})$, whereas the POD expansion is optimal in $L^2(\Xi_{\text{train}})$.)

6-III-D-c Offline-Online Stages

Finally, we delineate Offline and Online stages. The Offline stage comprises the Greedy sampling strategy, and thus appeals to both the Construction and Evaluation phases. The Online stage includes all subsequent evaluations of the RB output and output error bound for many-query computations : it involves only the Evaluation phase, and hence will be extremely rapid.

We now discuss the implications for the MC sums required for the evaluation of our statistical outputs — the focus of the current paper. In particular, it is clear the *total* operation count — Offline and Online — to evaluate $E_M[s_{N,K}(\mathcal{M}_K)]$, $V_M[s_{N,K}(\mathcal{M}_K)]$, $\Delta^o_E[s_{N,K}(\mathcal{M}_K)]$ and $\Delta^o_V[s_{N,K}(\mathcal{M}_K)]$ for J different values of $\varrho = (\kappa, \overline{\text{Bi}})$ scales as

$$W_{\text{Offline}}(N_{\max}, \mathcal{K}, \mathcal{N}) + W_{\text{Online}}(J, M, N, K) \text{ where}$$

$$W_{\text{Offline}}(N, K, \mathcal{N}) = O(NK\mathcal{N}^k) + n_{\text{train}} O(K^2 N^2) \text{ and } W_{\text{Online}}(J, M, N, K) = JM \times O(N^3 + K^2 N^2) \ .$$

Thus as either $M \to \infty$ or $J \to \infty$ and in particular as $J, M \to \infty$ — many evaluations of our statistical output — $W_{\text{Offline}} \ll W_{\text{Online}}$. We further note that if $N, K \ll \mathcal{N}$ then $W_{\text{Online}} \ll W_{\text{FE}} \equiv JM(O(\mathcal{N}^k))$, where W_{FE} is the operation count for standard FE evaluation of the MC sums. Hence the interest in the RB approach. In addition, here are two final observations. First, a "con" : as we consider less smooth covariance functions with less rapidly decaying spectra not only — for a fixed desired accuracy — will K increase, but also N will increase (due to the more extended domain Λ^y_K). Clearly for sufficiently non-smooth covariances the RB approach will no longer be competitive. Second, a "pro" : the *a posteriori* error bounds will permit us to choose N and K minimally — for minimum computational effort — without sacrificing accuracy and certainty.

6-III-E Numerical Results

In this section, we present numerical results for the model problem described in Section 6-III-A. We consider a *homogeneous random input field* with :

– a uniform mean, thus $G(x) \equiv 1$,

– and a finite-rank covariance kernel $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)$ that coincides with the first $\boxed{\mathcal{K}=25}$ terms in the KL expansion of $(\overline{\mathrm{Bi}}\Upsilon)^2 e^{-\frac{(x-y)^2}{\delta^2}}$.

The "additional" deterministic parameter $\varrho=(\kappa,\overline{\mathrm{Bi}})$ shall take value in the range $\Lambda^\varrho=[0.1,10]\times[0.1,1]$. For the "truth" FE approximation, we use a regular mesh with quadratic elements and $\mathcal{N}=6{,}882$ degrees of freedom.

First, we choose $\boxed{\delta=0.5}$ (recall that the length of Γ_{B} is 4, and hence δ is reasonably "small") — we shall subsequently consider even smaller δ. We calculate the eigenvalues and eigenvectors of $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)$ using the standard (Matlab®) Arpack routines. We present in Figure 6.2 the eigenvalues λ_k as a function of k; we observe that the eigenvalues decay exponentially with respect to k^2, which is in good agreement with theoretical bounds [ST06]. Then, to satisfy our assumption (6-II.19), we set $\boxed{\tau_0=\frac{1}{2}}$ which yields the requirement $\Upsilon\le\Upsilon_{\max}\equiv 0.058$. In the following numerical example, we choose $\boxed{\Upsilon=\Upsilon_{\max}\equiv 0.058}$.

We first report results for the case $\boxed{\kappa=2.0}$ and $\boxed{\overline{\mathrm{Bi}}=0.5}$. We show in Figure 6.3 four realizations $\big(\mathrm{bi}(x;\overline{\mathrm{Bi}},y_i^K)\big)_{1\le i\le 4}$ of the Biot number, and in Figure 6.4 the corresponding temperature fields $u_{\mathcal{N},K}(\mu_K^i)$ (where $K=\mathcal{K}$).

RB approximation : We present in Figure 6.5 the five leading basis functions $(\zeta_n)_{n=1,2,\ldots,5}$ obtained by pursuing the Greedy sampling procedure over a training set Ξ_{train} of $n_{\mathrm{train}}=10{,}000$ parameter points randomly selected with uniform law in the parameter space Λ^μ. Note $n_{\mathrm{train}}=10{,}000$ is arguably adequate given the rapid decay of the eigenvalues. In any event, our *a posteriori* error bounds will certify (in the Online stage) the accuracy of our RB predictions. The Greedy procedure terminates when a maximum number of basis functions $N_{\max}=18$ is reached, while the maximum error bound $\Delta_{N,K,\max}=\max_{\mu\in\Xi_{\mathrm{train}}}\Delta_{N,K}(\mu)$ is less than 5×10^{-3}.

Statistical outputs : We present in Figure 6.6 the expected value and variance as a function of M, obtained for $N=10$ and $K=20$ (note that we do not need to repeat the Offline stage for different M.) We next choose $M=10{,}000$ for our Monte-Carlo sums. We show in Table 6.1 the expected value and associated error bound for the integrated temperature at the bottom surface of the fin as a function of $N(\le N_{\max})$ and $K(\le\mathcal{K})$. Table 6.2 displays the corresponding variance and associated error bound. Figures 6.8(a) and 6.8(b) show the error bounds for the expected value and variance, respectively.

(N,K)-variations : We observe that the error bounds $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ and $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ depend on N and K in a strongly coupled manner : for a fixed value of K the error bounds initially decrease with increasing N and then level off for N large; when the error bounds no longer improve with increasing N, increasing K further reduces the error. This behavior of the error bounds is expected since the accuracy of our predictions is limited by both the RB error bound $\Delta_{N,K}^s(\mu)$ and the KL truncation error bound $\Delta_{N,K}^t(\mu)$: the former decreases rapidly with increasing N only while the latter decreases rapidly with increasing K only. We note that the KL truncation error bounds, $\Delta_E^t[s_{N,K}(\mathcal{M}_K)]$ and $\Delta_V^t[s_{N,K}(\mathcal{M}_K)]$, dominate the RB error bounds $\Delta_E^s[s_{N,K}(\mathcal{M}_K)]$ and $\Delta_V^s[s_{N,K}(\mathcal{M}_K)]$ respectively, as shown in Figures 6.9 and 6.10.

Reduction efficiency : The expectation and variance error bounds (and the actual errors) decrease very rapidly as both N and K increase (such a rapid convergence is expected because the solution is very smooth with respect to the Biot number Bi and also because the eigenvalues decay rapidly). For $N=10$ and $K=20$ the error bounds for the expected value and variance are 3.94×10^{-3} (corresponding to a relative error of 0.1%) and 8.32×10^{-4} (corresponding to a relative error of 20%), respectively, while the RB computational savings (including both Offline and Online effort) relative to the FE method is more than a factor of $\frac{1}{45}$. In the limit $J\to\infty$ of many $(\kappa,\overline{\mathrm{Bi}})$-queries, or $M\to\infty$ for better accuracy in the MC evaluations, the RB savings will approach $\frac{1}{200}$ — which reflects just the Online effort. The $(N=10,K=20)$-statistical results can be obtained Online in only 70 seconds (for a given $(\kappa,\overline{\mathrm{Bi}})$) on a Pentium IV 1.73 GHz; it would take roughly 4 hours for the FE method to perform the same calculation.

We see that for $\kappa=2.0$ and $\overline{\mathrm{Bi}}=0.5$, the standard deviation of the integrated temperature is less than 2% of the expected integrated temperature; we can conclude that, for this value of κ and $\overline{\mathrm{Bi}}$, uncertainties in Bi are not too important to "device performance." However, for larger κ and small $\overline{\mathrm{Bi}}$ we expect more sensitivity : we find that for $\kappa=10$ and $\overline{\mathrm{Bi}}=0.1$ the standard deviation of the integrated temperature is now 6% of the expected integrated temperature — and hence of engineering relevance. It is also possible to calculate the empirical cumulative distribution function to both assess the range and likelihood of "tails."

$(\kappa,\overline{\mathrm{Bi}})$-variations : We show in Figure 6.7 the expected value of the integrated temperature at the bottom surface of the heat sink as a function of κ and $\overline{\mathrm{Bi}}$. The statistical outputs, which are obtained for $N=10$, $K=20$ and $J=15\times15=225$ grid points in the parameter space, are plotted in Figure 6.7(a) for $M=5{,}000$ and in Figure 6.7(b) for $M=10{,}000$. The maximum relative error in the expectation over the 225 parameter grid points is 9.4×10^{-4}. (The results in Figure 6.7(a) and 6.7(b) each require $J=225$ evaluations of the empirical

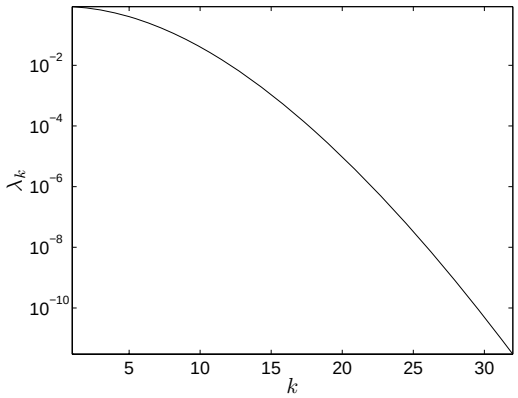


FIG. 6.2 – Eigenvalues λ_k as functions of k.

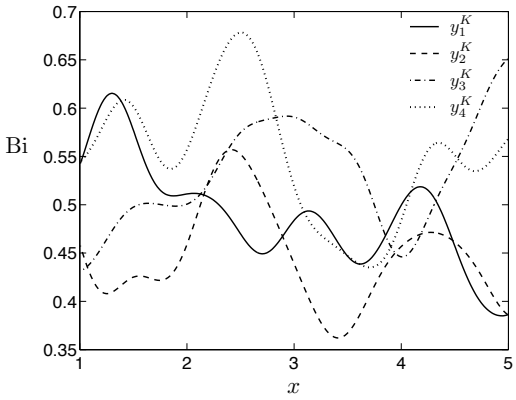
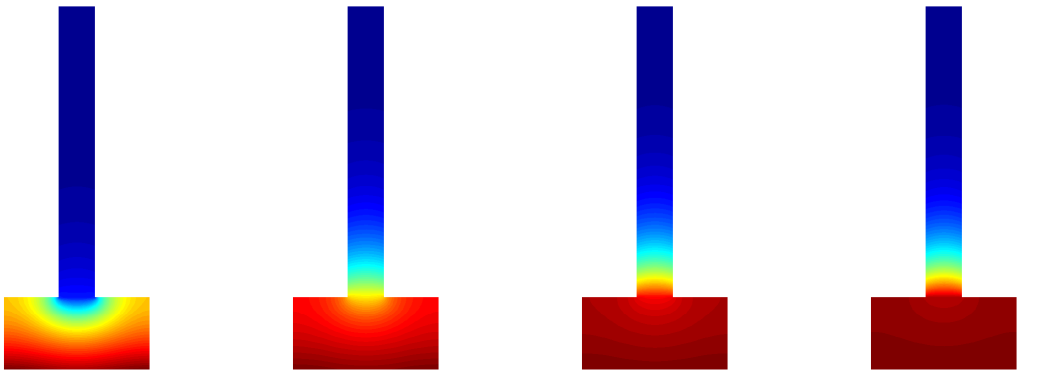


FIG. 6.3 – Four realizations of the Biot number $x \to \mathrm{bi}(x;\overline{\mathrm{Bi}}=0.5,y_i^K)$ – $1 \le i \le 4$ –.



(a) $\mu_1 = (0.1, 0.5, y_1^K)$

(b) $\mu_1 = (1.0, 0.5, y_2^K)$

(c) $\mu_1 = (4, 0.5, y_3^K)$

(d) $\mu_1 = (10, 0.5, y_4^K)$

FIG. 6.4 – The temperature field $u_{\mathcal{N},K}(\mu_i)$ for four different realizations $\mu_i = (\kappa_i, \overline{\mathrm{Bi}}=0.5, y_i)$ – $1 \le i \le 4$ – when $K = \mathcal{K}$, corresponding to the four realizations of $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}}=0.5,y_i^K)$ in Figure 6.3.

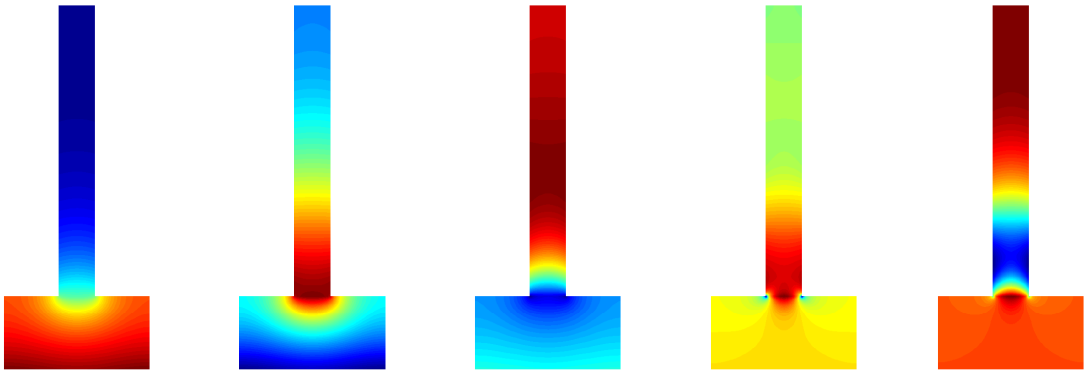
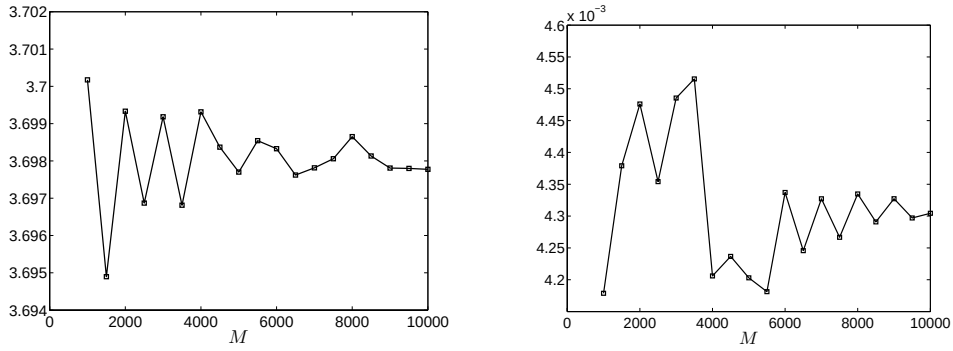
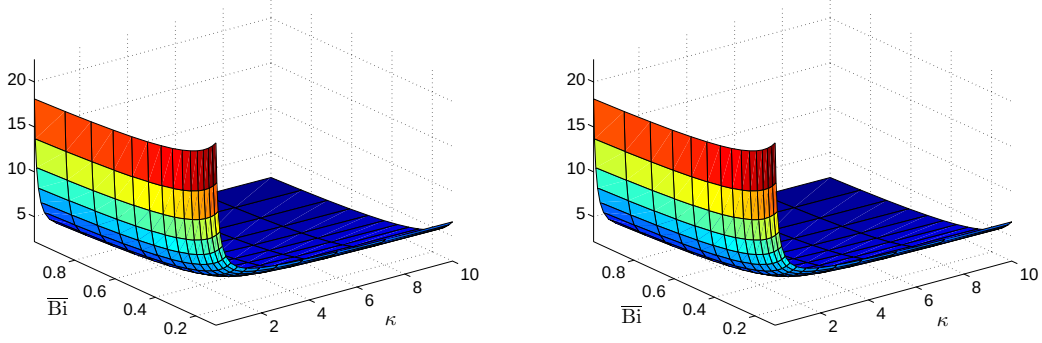


FIG. 6.5 – The five leading RB basis functions $(\zeta_n)_{n=1,2,\ldots,5}$, ordered from left to right and top to bottom as successively chosen (and orthonormalized) by the Greedy sampling procedure.



(a) Expected value (b) Variance

FIG. 6.6 – Outputs $E_M[s_{N,K}(\mathcal{M}_K)]$, $V_M[s_{N,K}(\mathcal{M}_K)]$ as functions of M, with $\varrho=(2.0,0.5)$.



(a) $M=5{,}000$ (b) $M=10{,}000$

FIG. 6.7 – Expected value of the integrated temperature at the bottom surface of the fin as a function of κ and $\overline{\text{Bi}}$ over $\Lambda^{\varrho}\equiv[0.1,10]\times[0.1,1]$.

N	$K=5$		$K=10$		$K=15$		$K=20$	
	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$
2	3.2602	4.74×10^{0}	3.2599	2.23×10^{0}	3.2600	1.59×10^{0}	3.2600	1.51×10^{0}
4	3.6920	2.20×10^{0}	3.6947	5.08×10^{-1}	3.6941	7.18×10^{-2}	3.6942	1.60×10^{-2}
6	3.6972	2.09×10^{0}	3.6974	4.76×10^{-1}	3.6979	5.80×10^{-2}	3.6966	4.54×10^{-3}
8	3.6981	2.09×10^{0}	3.6975	4.74×10^{-1}	3.6969	5.77×10^{-2}	3.6986	4.33×10^{-3}
10	3.6974	2.08×10^{0}	3.6977	4.71×10^{-1}	3.6976	5.69×10^{-2}	3.6978	3.94×10^{-3}
12	3.6973	2.07×10^{0}	3.6976	4.70×10^{-1}	3.6981	5.68×10^{-2}	3.6976	3.90×10^{-3}
14	3.6975	2.07×10^{0}	3.6974	4.70×10^{-1}	3.6977	5.68×10^{-2}	3.6978	3.89×10^{-3}

TAB. 6.1 – Expected value $E_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ for different values of the RB dimension N and of the KL truncation order K with $\varrho=(\kappa=2.0,\overline{\text{Bi}}=0.5)$.

N	$K=5$		$K=10$		$K=15$		$K=20$	
	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$v_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$
2	0.0039	9.38×10^{-1}	0.0041	4.38×10^{-1}	0.0041	3.23×10^{-1}	0.0041	3.00×10^{-1}
4	0.0039	4.54×10^{-1}	0.0045	1.11×10^{-1}	0.0045	1.56×10^{-2}	0.0045	3.52×10^{-3}
6	0.0037	4.05×10^{-1}	0.0043	1.02×10^{-1}	0.0043	1.23×10^{-2}	0.0043	9.89×10^{-4}
8	0.0037	4.05×10^{-1}	0.0043	1.08×10^{-1}	0.0043	1.26×10^{-2}	0.0043	9.09×10^{-4}
10	0.0038	4.16×10^{-1}	0.0043	9.72×10^{-2}	0.0043	1.24×10^{-2}	0.0043	8.32×10^{-4}
12	0.0038	4.16×10^{-1}	0.0043	9.72×10^{-2}	0.0043	1.24×10^{-2}	0.0043	8.36×10^{-4}
14	0.0038	4.12×10^{-1}	0.0043	9.72×10^{-2}	0.0043	1.23×10^{-2}	0.0043	8.46×10^{-4}

TAB. 6.2 – Variance $V_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ for different values of the RB dimension N and of the KL truncation order K with $\varrho=(\kappa=2.0,\overline{\text{Bi}}=0.5)$.

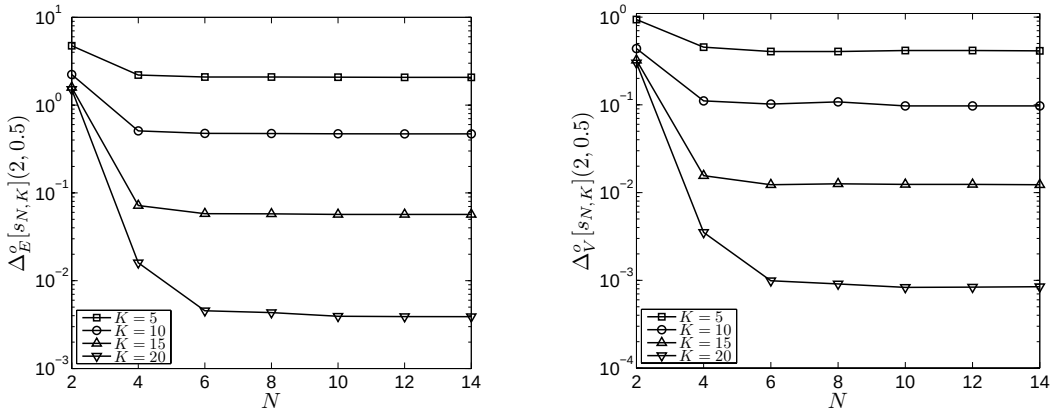


FIG. 6.8 – (a) $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ and (b) $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ as functions of N and K ; $\varrho = (2.0, 0.5)$.

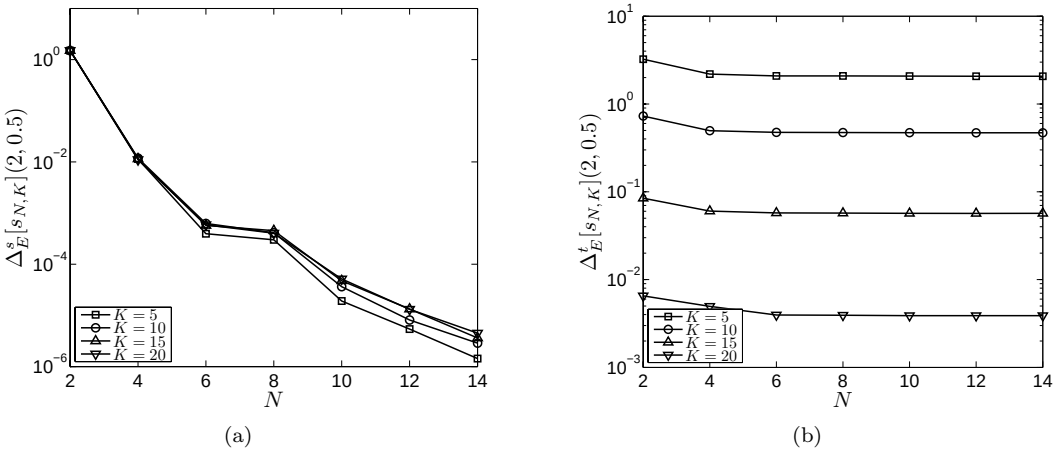


FIG. 6.9 – (a) $\Delta_E^s[s_{N,K}(\mathcal{M}_K)]$ and (b) $\Delta_E^t[s_{N,K}(\mathcal{M}_K)]$ as functions of N and K ; $\varrho = (2.0, 0.5)$.

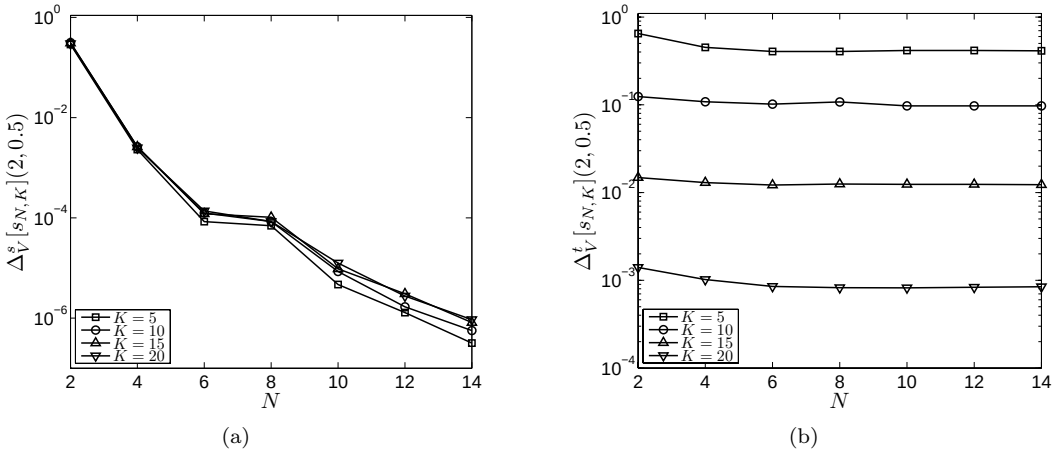


FIG. 6.10 – (a) $\Delta_V^s[s_{N,K}(\mathcal{M}_K)]$ and (b) $\Delta_V^t[s_{N,K}(\mathcal{M}_K)]$ as functions of N and K ; $\varrho = (2.0, 0.5)$.

estimations for the expectation and the variance.)

Next, we consider another finite-rank covariance kernel $\mathbf{Cov}_{\mathbb{P}}(\mathrm{Bi})(x,y)$ that coincides with the first $\boxed{\mathcal{K}=60}$ terms in the KL expansion of $(\overline{\mathrm{Bi}}\Upsilon)^2 e^{-\frac{(x-y)^2}{\delta^2}}$ for a smaller correlation length $\boxed{\delta=0.2}$. We present in Figure 6.11 the eigenvalues λ_k as a function of k. We see that the eigenvalues decay at a slower rate than the previous case (shown in Figure 6.2). We then obtain from (6-II.19) the requirement $\Upsilon_{\max}\equiv 0.074$; in our numerical examples we choose $\boxed{\Upsilon=\Upsilon_{\max}=0.074}$. Figure 6.12(a) shows four random realizations of the Biot number $\mathrm{Bi}(x,y)$ (these four random realizations vary more rapidly in space than the earlier instances of Figure 6.3). We then pursue the greedy sampling procedure which yields $N_{\max}=32$ for the same accuracy of 5×10^{-3} in the maximal error bound as in the case $\delta=0.5$. It is not surprising from the Figures 6.11 and 6.12(a) that the RB method needs larger $N_{\max}$ as the correlation length δ decreases.

For $\kappa=2.0$ and $\overline{\mathrm{Bi}}=0.5$ again, we show in Table 6.3 the expected value and associated error bound for the integrated temperature at the bottom surface of the heat sink as a function of N and K.[7] Table 6.4 displays the corresponding variance and associated error bound. Figure 6.13 shows the error bounds for the expected value and variance. We see that while the convergence pattern is similar to that of the previous case ($\delta=0.5$), we need to use larger N and K to obtain the same accuracy for $\delta=0.2$.

Nevertheless, the reduction in computational time is still quite significant : for $N=10$ and $K=45$ (for which the ratio $\Delta_E[s_{N,K}(\mathcal{M}_K)]/E_M[s_{N,K}(\mathcal{M}_K)]$ is $\mathbb{P}$-a.s. less than 0.01 at $\varrho=(2.0,0.5)$) the Online RB evaluation is still more than 50 times faster than the FE evaluation. Obviously, when the correlation length decreases further and further, the RB approach will no longer offer significant economies or may even become more expensive than the FE method; note however that, in three spatial dimensions, the RB method can "afford" a smaller correlation length since the FE truth will be considerably more expensive.

Finally, in the latter case of a correlation length $\delta=0.2$, we also consider $\boxed{\Upsilon=0.3>\Upsilon_{\max}}$ which yields a much larger domain Λ^y for the random parameter y^K. We note however that $\Upsilon=0.3$ does not satisfy the well-posedness requirement (6-II.19). As a result, $\mathrm{bi}(\cdot;\overline{\mathrm{Bi}},y^K)$ might be negative over the physical boundary $x\in[1,5]$ for some y^K. In such case, we simply ignore all possible values of y^K at which $\min_{x\in[1,5]}\mathrm{bi}(x;\overline{\mathrm{Bi}},y^K)\le 0$, in the Offline stage as well as Online computation. (This should not introduce a significant bias in the SLLN limit providing only very few realizations are rejected, which is indeed the case here with a rejection rate of approximately 1/100). We show in Figure 6.12(b) four random realizations of the Biot number $\mathrm{Bi}(x,Y)$; these four random realizations have much larger amplitudes than the instances of Figure 6.12(a). We then pursue the greedy sampling procedure for $K=60$ (*a priori* determined) to construct the nested basis sets X_N, $1\le N\le N_{\max}$; we obtain $N_{\max}=45$ — it is not surprising from Figure 6.12(b) that the RB method needs larger $N_{\max}$ as Υ increases.

For $\kappa=2.0$ and $\overline{\mathrm{Bi}}=0.5$ again, we further present in Table 6.5 the expected value and associated error bound for the integrated temperature at the bottom surface of the heat sink as a function of N and K. The expected values are now slightly larger than those shown in Table 6.3. Table 6.6 displays the corresponding variance and associated error bound. As expected, the variances are much larger than those shown in Table 6.4. More specifically, the standard deviation of the integrated temperature is approximately 7.2% of the expected integrated temperature, while the standard deviation of the integrated temperature is only 1.7% of the expected integrated temperature in the earlier results (see Table 6.3 and 6.4).

For $\boxed{\kappa=10}$ and $\boxed{\overline{\mathrm{Bi}}=0.1}$, we find that the standard deviation of the integrated temperature is 14.6% of the expected integrated temperature, which consequently defines much more stringent conditions. The reduction in computational time is still significant : for $N=10$ and $K=45$ (for which the ratio $\Delta_E[s_{N,K}(\mathcal{M}_K)]/E_M[s_{N,K}(\mathcal{M}_K)]$ is $\mathbb{P}$-a.s. less than 0.02 at $\varrho=(2.0,0.5)$ and thus slightly larger than that of the previous case) the Online RB evaluation is more than 50 times faster than the FE evaluation. These results demonstrate that the RB error bound is inexpensive and accurate even for a significant variation in the random variables y^K.

6-IV Conclusions

In this article we have developed the theoretical framework (error bounds) for, and numerically demonstrated the attractiveness of, an RB approach for the rapid and reliable computation of expectations of linear functionals of variational solutions to a BVP with ω-x "separable" random parameter fields. The *a posteriori* error bounds

[7] The values for $\delta=0.2$ are very similar to the values for $\delta=0.5$ for the same reason that the variance is in general small : the output is relatively insensitive to Bi fluctuations.

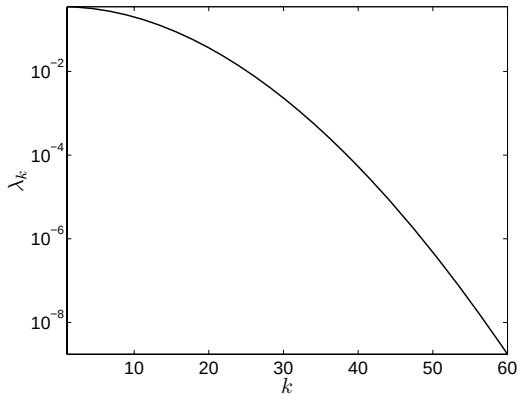
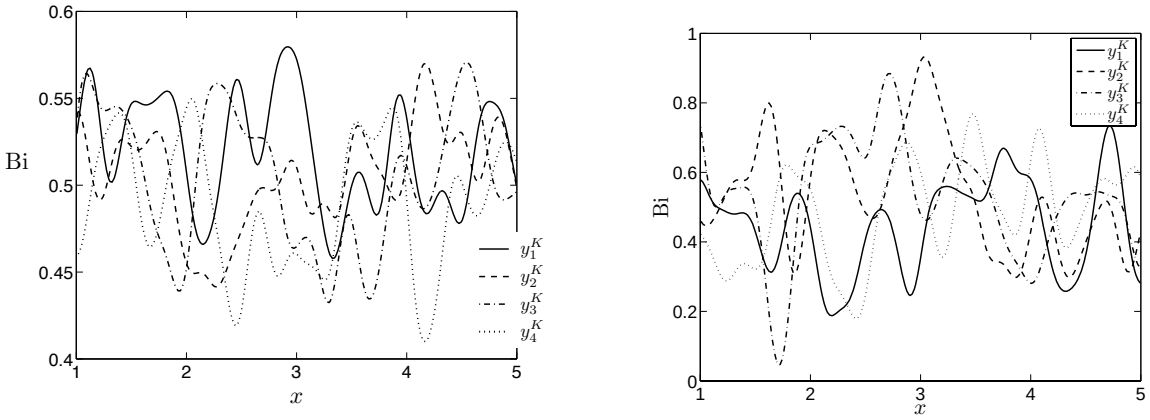
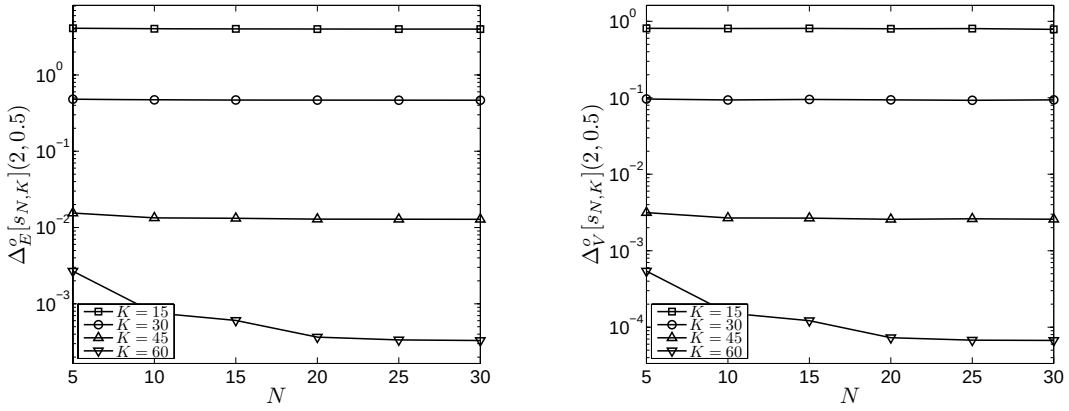


FIG. 6.11 – Eigenvalues λ_k as functions of k for the correlation length $\delta = 0.2$.



(a) $\Upsilon = 0.074$ (b) $\Upsilon = 0.3$

FIG. 6.12 – Four realizations of the Biot number $x \to \mathrm{bi}(x; \overline{\mathrm{Bi}}, y_i^K)$ – $1 \leq i \leq 4$ – for $\delta = 0.2$.



(a) (b)

FIG. 6.13 – (a) $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ and (b) $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ as functions of N and K with $\delta = 0.2$, $\Upsilon = 0.074$ and $\varrho = (2.0, 0.5)$.

N	$K=15$		$K=30$		$K=45$		$K=60$	
	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$
5	3.6975	4.09×10^{0}	3.6970	4.80×10^{-1}	3.6960	1.55×10^{-2}	3.6960	2.68×10^{-3}
10	3.6975	4.03×10^{0}	3.6973	4.71×10^{-1}	3.6979	1.34×10^{-2}	3.6963	7.62×10^{-4}
15	3.6973	4.02×10^{0}	3.6978	4.70×10^{-1}	3.6970	1.32×10^{-2}	3.6977	6.05×10^{-4}
20	3.6980	4.00×10^{0}	3.6980	4.67×10^{-1}	3.6973	1.29×10^{-2}	3.6980	3.65×10^{-4}
25	3.6969	3.99×10^{0}	3.6977	4.66×10^{-1}	3.6972	1.28×10^{-2}	3.6981	3.36×10^{-4}
30	3.6968	3.99×10^{0}	3.6975	4.66×10^{-1}	3.6972	1.28×10^{-2}	3.6975	3.30×10^{-4}

TAB. 6.3 – Expected value $E_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ for different values of N and K with $\delta=0.2$, $\Upsilon=0.074$ and $\varrho=(2.0,0.5)$.

N	$K=15$		$K=30$		$K=45$		$K=60$	
	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$v_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$
5	0.0038	8.09×10^{-1}	0.0039	9.64×10^{-2}	0.0039	3.15×10^{-3}	0.0038	5.41×10^{-4}
10	0.0039	8.04×10^{-1}	0.0039	9.36×10^{-2}	0.0039	2.68×10^{-3}	0.0039	1.53×10^{-4}
15	0.0040	8.07×10^{-1}	0.0039	9.50×10^{-2}	0.0040	2.67×10^{-3}	0.0039	1.21×10^{-4}
20	0.0039	7.99×10^{-1}	0.0039	9.39×10^{-2}	0.0040	2.57×10^{-3}	0.0039	7.28×10^{-5}
25	0.0039	8.02×10^{-1}	0.0039	9.28×10^{-2}	0.0040	2.62×10^{-3}	0.0040	6.76×10^{-5}
30	0.0039	7.84×10^{-1}	0.0040	9.39×10^{-2}	0.0040	2.58×10^{-3}	0.0040	6.71×10^{-5}

TAB. 6.4 – Variance $V_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ for different values of N and K with $\delta=0.2$, $\Upsilon=0.074$ and $\varrho=(2.0,0.5)$.

N	$K=15$		$K=30$		$K=45$		$K=60$	
	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$	$E_M[s_{N,K}]$	$\Delta_E^o[s_{N,K}]$
5	3.7230	1.82×10^{1}	3.7229	2.17×10^{0}	3.7215	1.02×10^{-1}	3.7239	4.25×10^{-2}
10	3.7312	1.70×10^{1}	3.7389	2.00×10^{0}	3.7273	6.00×10^{-2}	3.7299	5.65×10^{-3}
15	3.7341	1.67×10^{1}	3.7345	1.97×10^{0}	3.7287	5.59×10^{-2}	3.7311	2.53×10^{-3}
20	3.7327	1.66×10^{1}	3.7338	1.94×10^{0}	3.7328	5.40×10^{-2}	3.7351	1.08×10^{-3}
25	3.7323	1.65×10^{1}	3.7342	1.93×10^{0}	3.7350	5.33×10^{-2}	3.7364	6.73×10^{-4}
30	3.7322	1.64×10^{1}	3.7399	1.93×10^{0}	3.7385	5.30×10^{-2}	3.7370	5.20×10^{-4}

TAB. 6.5 – Expected value $E_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_E^o[s_{N,K}(\mathcal{M}_K)]$ for different values of N and K with $\delta=0.2$, $\Upsilon=0.3$ and $\varrho=(2.0,0.5)$.

N	$K=15$		$K=30$		$K=45$		$K=60$	
	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$v_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$	$V_M[s_{N,K}]$	$\Delta_V^o[s_{N,K}]$
5	0.0721	1.54×10^{1}	0.0716	1.85×10^{0}	0.0744	8.90×10^{-2}	0.0718	3.72×10^{-2}
10	0.0738	1.46×10^{1}	0.0764	1.78×10^{0}	0.0743	5.25×10^{-2}	0.0738	5.03×10^{-3}
15	0.0717	1.43×10^{1}	0.0734	1.68×10^{0}	0.0735	4.81×10^{-2}	0.0744	2.25×10^{-3}
20	0.0705	1.41×10^{1}	0.0737	1.69×10^{0}	0.0725	4.61×10^{-2}	0.0728	9.48×10^{-4}
25	0.0699	1.38×10^{1}	0.0699	1.62×10^{0}	0.0723	4.56×10^{-2}	0.0732	5.83×10^{-4}
30	0.0755	1.44×10^{1}	0.0757	1.68×10^{0}	0.0722	4.64×10^{-2}	0.0723	4.43×10^{-4}

TAB. 6.6 – Variance $V_M[s_{N,K}(\mathcal{M}_K)]$ and error bound $\Delta_V^o[s_{N,K}(\mathcal{M}_K)]$ for different values of N and K with $\delta=0.2$, $\Upsilon=0.3$ and $\varrho=(2.0,0.5)$.

certify the quality of the approximation and quantify the effects of both the FE → RB reduction for the BVP and the KL truncation in the random field expansion. The method also permits the study of the parametric dependence of the outputs with respect to other (deterministic) parameters entering the problem.

Future developments may include :

(a) test problems in which the random input field multiplies the solution field not only on the boundary but also over the entire domain (*e.g.* random diffusivity coefficient κ),

(b) more general variates (and sampling procedures) in the KL expansion of the input field,

(c) inputs developed with expansions other than KL (not necessarily decoupling $\mathcal{D}$ and Ω, and thus requiring empirical interpolation [BNMP04, GMNP07]),

(d) more general statistical outputs (that remain sufficiently smooth functionals of the random solution field — continuous in $L^p_{\mathbb{P}}(\Omega, H^1(\mathcal{D}))$), and

(e) application of the RB approach to Ω-weak/$\mathcal{D}$-weak collocation formulations [BNT07, NTW08].

But from our first results, it is arguably already interesting to apply an RB approach within many of the Ω-strong/$\mathcal{D}$-weak formulations in view of the simplicity of the implementation, the considerable reduction in computational time, and the availability of rigorous error bounds (suitably generalized, in particular as regards the contribution of the KL truncation and associated continuity constants).

We end this paper by pointing out that the RB methods and associated *a posteriori* error estimation have been developed for several classes of parametrized PDEs including linear coercive/noncoercive elliptic problems [BNMP04, CHMR08, HRSP07a, PRV$^+$02, RHP08, SVH$^+$06], linear elasticity [HP07b], eigenvalue problems [MMO$^+$00], linear parabolic problems [GP05, HO08], Boltzmann equations [PR07a], nonlinear elliptic and parabolic problems [GMNP07], and incompressible Navier-Stokes equations [NRHP09, NVP05b, VP05]. It appears that the extension to other classes of SPDEs beyond the particular linear elliptic SPDE discussed in this paper can be achieved by combining the current RB approach with those of the previous work. We consider to pursue this line of development in future work.

Acknowledgement 3. *We thank Gianluigi Rozza for helpful discussions.*

Chapitre 7

Une méthode de réduction de variance pour des équations différentielles stochastiques paramétrées en utilisant le paradigme des bases réduites

Ce chapitre est la reproduction de [BL09a]. Les résultats ont été obtenus en collaboration avec T. Lelièvre.

Nous y développons une approche par bases réduites en vue de calculer efficacement des espérances paramétrées pour un grand nombre de valeurs du paramètre, en utilisant des variables de contrôle comme méthode de réduction de variance. Deux algorithmes y sont proposés pour calculer rapidement, *en ligne*, des variables de contrôles pour les espérances de fonctionnelles de processus stochastiques d'Itô (des solutions d'équations différentielles stochastiques paramétrées). Dans chaque algorithme, une base réduite de variables de contrôle est précalculée *hors ligne*, selon une procédure dite *gloutonne*, qui minimise la variance au sein d'un échantillon d'espérances paramétrées représentant la sortie. Des résultats numériques dans des cas d'application pratiques illustrent l'efficacité de la méthode (la calibration de la volatilité en pricing d'options, et l'évolution d'un champ de vecteurs solutions d'une équation de Langevin paramètrée en théorie cinétique).

A Variance Reduction Method for Parametrized Stochastic Differential Equations using the Reduced Basis Paradigm

Sébastien Boyaval[a,b], Tony Lelièvre[a,b]

[a]Université Paris-Est, CERMICS (Ecole des ponts ParisTech, 6-8 avenue Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cedex 2, France).

[b]INRIA, MICMAC project team (Domaine de Voluceau, BP. 105, Rocquencourt, 78153 Le Chesnay Cedex, France).

In this chapter, we develop a reduced-basis approach for the efficient computation of parametrized expected values, for *a large number* of parameter values, using the control variate method to reduce the variance. Two algorithms are proposed to compute *online*, through a cheap reduced-basis approximation, the control variates for the computation of a large number of expectations of a functional of a parametrized Itô stochastic process (solution to a parametrized stochastic differential equation). For each algorithm, a reduced basis of control variates is pre-computed *offline*, following a so-called *greedy* procedure, which minimizes the variance among a trial sample of the output parametrized expectations. Numerical results in situations relevant to practical applications (calibration of volatility in option pricing, and parameter-driven evolution of a vector field following a Langevin equation from kinetic theory) illustrate the efficiency of the method.

Keywords : Variance Reduction, Stochastic Differential Equations, Reduced-Basis Methods.

7-I Introduction

This article develops a general variance reduction method for the *many-query* context where a large number of *Monte-Carlo* estimations of the expectation $\mathbf{E}\left(Z^\lambda\right)$ of a functional

$$Z^\lambda = g^\lambda(X_T^\lambda) - \int_0^T f^\lambda(s, X_s^\lambda)\, ds \tag{7-I.1}$$

of the solutions $\left(X_t^\lambda, t\in[0,T]\right)$ to the *stochastic differential equations* (SDEs) :

$$X_t^\lambda = x + \int_0^t b^\lambda(s, X_s^\lambda)\, ds + \int_0^t \sigma^\lambda(s, X_s^\lambda)\, dB_s \tag{7-I.2}$$

parametrized by $\lambda\in\Lambda$ have to be computed for many values of the parameter λ.

Such many-query contexts are encountered in finance for instance, where pricing options often necessitates to compute the price $\mathbf{E}\left(Z^\lambda\right)$ of an option with spot price X_t^λ at time t in order to *calibrate* the local volatility σ^λ as a function of a (multi-dimensional) parameter λ (that is minimize over λ, after many iterations of some optimization algorithm, the difference between observed statistical data with the model prediction). Another context for application is molecular simulation, for instance micro-macro models in rheology, where the mechanical properties of a flowing viscoelastic fluid are determined from the coupled evolution of a non-Newtonian stress tensor field $\mathbf{E}\left(Z^\lambda\right)$ due to the presence of many polymers with configuration X_t^λ in the fluid with instantaneous velocity gradient field λ. Typically, segregated numerical schemes are used : compute X_t^λ for a fixed field λ, and then compute λ for a fixed field $\mathbf{E}\left(Z^\lambda\right)$. Such tasks are known to be computationally demanding and the use of different *variance reduction* techniques to alleviate the cost of Monte-Carlo computations in those fields is very common (see [Aro04, MO95, OvdBH97, BP99] for instance).

In the following, we focus on one particular variance reduction strategy termed the *control variate* method [HDH64, New94, MT06]. More precisely, we propose new approaches in the context of the computation of $\mathbf{E}\left(Z^\lambda\right)$ for a large number of parameter values λ, with the control variate method. In these approaches, the control variates are computed through a *reduced-basis* method whose principle is related to the reduced-basis method [MMP01, MPT02a, PR07b, Boy08, BBM$^+$09] previously developed to efficiently solve parametrized partial differential equations (PDEs). Following the reduced-basis paradigm, a small-dimensional vector basis is first built *offline* to span a good linear approximation space for a large trial sample of the λ-parametrized control variates, and then used *online* to compute control variates at any parameter value. The offline computations are

typically expensive, but done once for all. Consequently, it is expected that the online computations (namely, approximations of $\mathbf{E}\left(Z^\lambda\right)$ for many values of λ) are very cheap, using the small-dimensional vector basis built offline for *efficiently* computing control variates online. Of course, such reduced-basis approaches can only be *efficient* insofar as :

1. online computations (of one output $\mathbf{E}\left(Z^\lambda\right)$ for one parameter value λ) are significantly cheaper using the reduced-basis approach than without, and
2. the amount of outputs $\mathbf{E}\left(Z^\lambda\right)$ to be computed online (for many different parameter values λ) is sufficient to compensate for the (expensive) offline computations (needed to build the reduced basis).

In this work, we will study numerically how the variance is reduced in two examples using control variates built with two different approaches.

The usual reduced-basis approach for parametrized PDEs also traditionally focuses on the certification of the reduction (in the parametrized solution manifold) by estimating *a posteriori* the error between approximations obtained before/after reduction for some *output* which is a functional of the PDE solution. Our reduced-basis approach for the parametrized control variate method can also be cast into a goal-oriented framework similar to the traditional reduced basis method. One can take the expectation $\mathbf{E}\left(Z^\lambda\right)$ as the reduced-basis output, while the empirically estimated variance $\mathrm{Var}_\mathrm{M}\left(Z^\lambda\right)$ serves as a computable (statistical) error indicator for the Monte-Carlo approximations $\mathrm{E}_\mathrm{M}\left(Z^\lambda\right)$ of $\mathbf{E}\left(Z^\lambda\right)$ *in the limit of large* M through the Central Limit Theorem (see error bound (7-II.4) in Section 7-II-A).

In the next Section 7-II, the variance reduction issue and the control variate method are introduced, as well as the principles of our reduced-basis approaches for the computation of parametrized control variates. The Section 7-III exposes details about the algorithms which are numerically applied to test problems in the last Section 7-IV.

The numerical simulations show good performance of the method for the two test problems corresponding to the applications mentionned above : a scalar SDE with (multi-dimensional) parametrized diffusion (corresponding to the calibration of a local volatility in option pricing), and a vector SDE with (multi-dimensional) parametrized drift (for the parameter-driven evolution of a vector field following a Langevin equation from kinetic theory). Using the control variate method with a 20-dimensional reduced basis of (precomputed) control variates, the variance is approximatively divided by a factor of 10^4 in the mean for large test samples of parameter in the applications we experiment here. As a consequence, our reduced-basis approaches allows to approximately divide the online computation time by a factor of 10^2, while maintaining the confidence intervals for the output expectation at the same value than without reduced basis.

This work intends to present a new numerical method and to demonstrate its interest on some relevant test cases. We do not have, for the moment, a theoretical understanding of the method. This is the subject of future works.

7-II The variance reduction issue and the control variate method

7-II-A Mathematical preliminaries and the variance reduction issue

Let $\left(B_t \in \mathbb{R}^d, t \in [0,T]\right)$ be a d-dimensional standard Brownian motion (where d is a positive integer) on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, endowed with a filtration $(\mathcal{F}_t, t \in [0,T])$. For any square-integrable random variables X, Y on that probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we respectively denote by $\mathbf{E}(X)$ and $\mathbf{Var}(X)$ the expected value and the variance of X with respect to the probability measure $\mathbb{P}$, and by $\mathbf{Cov}(X;Y)$ the covariance between X and Y.

For every $\lambda \in \Lambda$ (Λ being the set of parameter values), the Itô processes $\left(X_t^\lambda \in \mathbb{R}^d, t \in [0,T]\right)$ with deterministic initial condition $x \in \mathbb{R}^d$ are well defined as the solutions to the SDEs (7-I.2) under suitable assumptions on b^λ and σ^λ, for instance provided b^λ and σ^λ satisfy Lipschitz and growth conditions [KP00]. Let (X_t^λ) be solutions to the SDEs, and f^λ, g^λ be measurable functions such that Z^λ is a well-defined integrable random variable ($Z^\lambda \in L^1_\mathbb{P}(\Omega)$). Then, Kolmogorov's strong law of large numbers holds and, denoting by Z_m^λ ($m = 1, \ldots, M$) M independent copies of the random variables Z^λ (for all positive integer M), the output expectation $\mathbf{E}\left(Z^\lambda\right) = \int_\Omega Z^\lambda d\mathbb{P}$ can be approximated (almost surely) by Monte-Carlo estimations of the form :

$$\mathrm{E}_M\left(Z^\lambda\right) := \frac{1}{M} \sum_{m=1}^{M} Z_m^\lambda \xrightarrow[M \to \infty]{\mathbb{P}-a.s.} \mathbf{E}\left(Z^\lambda\right). \tag{7-II.1}$$

Furthermore, assume that the random variable Z^λ is square integrable ($Z^\lambda \in L^2_{\mathbb{P}}(\Omega)$) with variance $\mathbf{Var}\left(Z^\lambda\right)$, then an asymptotic error bound for the convergence occuring in (7-II.1) is given in probabilistic terms by the Central Limit Theorem as confidence intervals : for all $a>0$,

$$\mathbb{P}\left(\left|\mathrm{E}_M\left(Z^\lambda\right)-\mathbf{E}\left(Z^\lambda\right)\right| \leq a\sqrt{\frac{\mathbf{Var}\left(Z^\lambda\right)}{M}}\right) \underset{M\to\infty}{\longrightarrow} \int_{-a}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx. \tag{7-II.2}$$

In terms of the error bound (7-II.2), an approximation $\mathrm{E}_M\left(Z^\lambda\right)$ of the output $\mathbf{E}\left(Z^\lambda\right)$ is thus all the better, for a given M, as the variance $\mathbf{Var}\left(Z^\lambda\right)$ is small. In a many-query framework, the computation of approximations (7-II.1) for many outputs $\mathbf{E}\left(Z^\lambda\right)$ (corresponding to many queried values of the parameter $\lambda\in\Lambda$) would then be all the faster as the variance $\mathbf{Var}\left(Z^\lambda\right)$ for some $\lambda\in\Lambda$ could be decreased from some knowledge acquired from the $\lambda\in\Lambda$ computed beforehand. This typically defines a many-query setting with parametrized output suitable for a reduced-basis approach similar to the reduced-basis method developped in a deterministic setting for parametrized PDEs.

In addition, the convergence (7-II.1) controlled by the confidence intervals (7-II.2) can be easily observed using computable *a posteriori* estimators. Indeed, remember that since the random variable Z^λ has a finite second moment, then the strong law of large numbers also implies the following convergence :

$$\mathrm{Var}_M\left(Z^\lambda\right) := \mathrm{E}_M\left(\left(Z^\lambda-\mathrm{E}_M(Z^\lambda)\right)^2\right) \underset{M\to\infty}{\overset{\mathbb{P}-a.s.}{\longrightarrow}} \mathbf{Var}\left(Z^\lambda\right)\ . \tag{7-II.3}$$

Combining the Central Limit Theorem with Slutsky theorem for the couple of Monte-Carlo estimators $\left(\mathrm{E}_M\left(Z^\lambda\right),\mathrm{Var}_M\left(Z^\lambda\right)\right)$ (see for instance [GS92], exercise 7.2.(26)), we obtain a fully computable probabilistic (asymptotic) error bound for the Monte-Carlo approximation (7-II.1) of the output expectation : for all $a>0$,

$$\mathbb{P}\left(\left|\mathbf{E}\left(Z^\lambda\right)-\mathrm{E}_M\left(Z^\lambda\right)\right| \leq a\sqrt{\frac{\mathrm{Var}_M\left(Z^\lambda\right)}{M}}\right) \overset{M\to\infty}{\longrightarrow} \int_{-a}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx. \tag{7-II.4}$$

It is exactly the purpose of variance reduction techniques to reduce the so-called *statistical* error appearing in the Monte-Carlo estimation of the output expectation $\mathbf{E}\left(Z^\lambda\right)$ through the error bound (7-II.2). And this is usually achieved in practice by using the (*a posteriori*) estimation (7-II.4).

Remark 28 (SDE discretization and bias error in the output expectation)**.** *In practice, there is of course another source of error, coming from the time-discretizations of the SDE* (7-I.2) *and of the integral involved in the expression for* Z^λ.

In the following (for the numerical applications), we use the Euler-Maruyama numerical scheme with discretizations $0=t_0<t_1<\cdots<t_N=T$ *(*$N\in\mathbb{N}$*) of the time interval* $[0,T]$ *to approximate the Itô process* (X^λ_t) *:*

$$\begin{cases} \overline{X}^\lambda_n = \overline{X}^\lambda_{n-1} + |t_n-t_{n-1}|\ b^\lambda(t_{n-1},\overline{X}^\lambda_{n-1}) + \sqrt{|t_n-t_{n-1}|}\sigma^\lambda(t_{n-1},\overline{X}^\lambda_{n-1})G_{n-1}, \\ \overline{X}^\lambda_0 = x, \end{cases}$$

where $\{G_n,\ n=0,\ldots,N-1\}$ *is a collection of* N *independent* d*-dimensional normal centered Gaussian vectors. It is well-known that such a scheme if of weak order one, so that we have a bound for the bias due to the approximation of the output expectation* $\mathbf{E}\left(Z^\lambda\right)$ *by* $\mathbf{E}\left(\overline{Z}^\lambda\right)$ *(where* $\overline{Z}^\lambda$ *is a time-discrete approximation for* Z^λ *computed from* $(\overline{X}^\lambda_n)$ *with an appropriate discretization of the integral* $\int_0^T f^\lambda(s,X^\lambda_s)ds$*) :*

$$\left|\mathbf{E}\left(\overline{Z}^\lambda\right)-\mathbf{E}\left(Z^\lambda\right)\right| \underset{N\to\infty}{=} O\left(\max_{1\leq n\leq N}(|t_n-t_{n-1}|)\right).$$

The approximation of the output $\mathbf{E}\left(Z^\lambda\right)$ *by* $\mathrm{E_M}(\overline{Z}^\lambda)$ *thus contains two types of errors :*

- *first, a bias* $\mathbf{E}\left(Z^\lambda-\overline{Z}^\lambda\right)$ *due to discretization errors in the numerical integration of the SDE* (7-I.2) *and of the integral involved in* Z^λ*,*
- *second, a statistical error of order* $\sqrt{\mathbf{Var}\left(\overline{Z}^\lambda\right)/M}$ *in the empirical Monte-Carlo estimation* $\mathrm{E_M}(\overline{Z}^\lambda)$ *of the expectation* $\mathbf{E}\left(\overline{Z}^\lambda\right)$*.*

We focus here on the statistical error.

7-II-B Variance reduction with the control variate method

The idea of control variate methods for the Monte-Carlo evaluation of $\mathbf{E}\left(Z^\lambda\right)$ is to find a so-called *control variate* Y^λ (with $Y^\lambda \in L^2_{\mathbb{P}}(\Omega)$), and then to write :

$$\mathbf{E}\left(Z^\lambda\right) = \mathbf{E}\left(Z^\lambda - Y^\lambda\right) + \mathbf{E}\left(Y^\lambda\right),$$

where $\mathbf{E}\left(Y^\lambda\right)$ can be easily evaluated, while the expectation $\mathbf{E}\left(Z^\lambda - Y^\lambda\right)$ is approximated by Monte-Carlo estimations that have a smaller statistical error than direct Monte-Carlo estimations of $\mathbf{E}\left(Z^\lambda\right)$. In the following, we will consider control variates Y^λ such that $\mathbf{E}\left(Z^\lambda\right) = \mathbf{E}\left(Z^\lambda - Y^\lambda\right)$, equivalently

$$\mathbf{E}\left(Y^\lambda\right) = 0.$$

The control variate method will indeed be interesting if the statistical error of the Monte-Carlo estimations $\mathrm{E}_M(Z^\lambda - Y^\lambda)$ is significantly smaller than the statistical error of the Monte-Carlo estimations $\mathrm{E}_M(Z^\lambda)$. That is, considering the following error bound given by the Central Limit Theorem : for all $a > 0$,

$$\mathbb{P}\left(\left|\mathrm{E}_M\left(Z^\lambda - Y^\lambda\right) - \mathbf{E}\left(Z^\lambda\right)\right| \leq a\sqrt{\frac{\mathbf{Var}\left(Z^\lambda - Y^\lambda\right)}{M}}\right) \overset{M\to\infty}{\longrightarrow} \int_{-a}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx, \tag{7-II.5}$$

the Monte-Carlo estimations $\mathrm{E}_M(Z^\lambda - Y^\lambda)$ will indeed be more accurate approximations of the expectations $\mathbf{E}\left(Z^\lambda\right)$ than the Monte-Carlo estimations $\mathrm{E}_M(Z^\lambda)$ provided :

$$\mathbf{Var}\left(Z^\lambda\right) \geq \mathbf{Var}\left(Z^\lambda - Y^\lambda\right).$$

Clearly, the best possible control variate (in the sense of minimal variance) for a fixed parameter $\lambda \in \Lambda$ is :

$$Y^\lambda = Z^\lambda - \mathbf{E}\left(Z^\lambda\right), \tag{7-II.6}$$

since we then have $\mathbf{Var}\left(Z^\lambda - Y^\lambda\right) = 0$. Unfortunately, the result $\mathbf{E}\left(Z^\lambda\right)$ itself is necessary to compute Y^λ as $Z^\lambda - \mathbf{E}\left(Z^\lambda\right)$.

In the following, we will need another representation of the best possible control variate $Z^\lambda - \mathbf{E}\left(Z^\lambda\right)$. Under suitable assumptions on the coefficients b^λ and σ^λ (for well-posedness of the SDE), plus continuity and polynomial growth conditions on f^λ and g^λ, let us define $u^\lambda(t,y)$, for $(t,y) \in [0,T] \times \mathbb{R}^d$, as the unique solution $u^\lambda(t,y) \in C^1\left([0,T], C^2(\mathbb{R}^d)\right)$ to the backward Kolmogorov equation (7-II.7) satisfying the same polynomial growth assumptions at infinity than f^λ and g^λ (for instance, see Theorem 5.3 in [Fri75]) :

$$\begin{cases} \partial_t u^\lambda + b^\lambda(t,y) \cdot \nabla u^\lambda + \dfrac{1}{2}\sigma^\lambda(t,y)\sigma^\lambda(t,y)^T : \nabla^2 u^\lambda = f^\lambda(t,y)\ , \\ u^\lambda(T,y) = g^\lambda(y), \end{cases} \tag{7-II.7}$$

where the notation ∇u^λ means $\nabla_y u^\lambda(t,y)$ and $\sigma^\lambda(t,y)\sigma^\lambda(t,y)^T : \nabla^2 u^\lambda$ means $\sum_{i,j,k=1}^d \sigma^\lambda_{ik}(t,y)\sigma^\lambda_{jk}(t,y)\partial^2_{y_i,y_j} u^\lambda(t,y)$. Using Itô formula for $\left(u^\lambda(t,X^\lambda_t), t \in [0,T]\right)$ with u^λ solution to (7-II.7), we get the following integral representation of Z^λ (see also Appendix 7-VI-A for another link between the SDE (7-I.2) and the PDE (7-II.7), potentially useful to numerics) :

$$g^\lambda(X^\lambda_T) - \int_0^T f^\lambda(s, X^\lambda_s) ds = u^\lambda(0,x) + \int_0^T \nabla u^\lambda(s, X^\lambda_s) \cdot \sigma^\lambda(s, X^\lambda_s) dB_s. \tag{7-II.8}$$

Note that the left-hand side of (7-II.8) is Z^λ, and the right-hand side is the sum of a stochatic integral (with zero mean) plus a scalar $u^\lambda(0,x)$ (thus equal to the expected value $\mathbf{E}\left(Z^\lambda\right)$ of the left-hand side). Hence, the optimal control variate also writes :

$$Y^\lambda = Z^\lambda - \mathbf{E}\left(Z^\lambda\right) = \int_0^T \nabla u^\lambda(s, X^\lambda_s) \cdot \sigma^\lambda(s, X^\lambda_s) dB_s. \tag{7-II.9}$$

Of course, the formula (7-II.9) is again idealistic because, most often, numerically solving the PDE (7-II.7) is a very difficult task (especially in large dimension $d \geq 4$).

7-II-C Outline of the algorithms

Considering either (7-II.6) or (7-II.9), we propose two algorithms for the efficient online computation of the family of parametrized outputs $\{\mathbf{E}\big(Z^\lambda\big), \lambda \in \Lambda\}$, when the parameter λ can take any value in a given range Λ, using (for each $\lambda \in \Lambda$) a control variate built as a linear combination of objects precomputed offline.

More precisely, in *Algorithm 1*, we do the following :

- Compute *offline* an accurate approximation $\tilde{Y}^\lambda$ of Y^λ using (7-II.6), for a small set of selected parameters $\lambda \in \{\lambda_1, \ldots, \lambda_I\} \subset \Lambda$ (where $I \in \mathbb{N}_{>0}$).
- For any $\lambda \in \Lambda$, compute *online* a control variate for the Monte-Carlo estimation of $\mathbf{E}\big(Z^\lambda\big)$ as a linear combination of $\{\tilde{Y}^{\lambda_i}, i = 1, \ldots, I\}$:

$$\tilde{Y}_I^\lambda = \sum_{i=1}^{I} \mu_i^\lambda \tilde{Y}^{\lambda_i}.$$

And in *Algorithm 2*, we do the following :

- Compute *offline* an accurate approximation $\tilde{u}^\lambda$ of the solution u^λ to the Kolmogorov backward equation (7-II.7) for a small set of selected parameters $\lambda \in \{\lambda_1, \ldots, \lambda_I\} \subset \Lambda$.
- For any $\lambda \in \Lambda$, compute *online* a control variate for the Monte-Carlo computation of $\mathbf{E}\big(Z^\lambda\big)$, in view of (7-II.9), as a linear combination of $\int_0^T \nabla \tilde{u}^{\lambda_i}(s, X_s^\lambda) \cdot \sigma^\lambda(s, X_s^\lambda) dB_s$ (where $i = 1, \ldots, I$) :

$$\tilde{Y}_I^\lambda = \sum_{i=1}^{I} \mu_i^\lambda \int_0^T \nabla \tilde{u}^{\lambda_i}(s, X_s^\lambda) \cdot \sigma^\lambda(s, X_s^\lambda) dB_s. \tag{7-II.10}$$

For a fixed size I of the reduced-basis, being given a parameter λ, both algorithms compute the coefficients μ_i^λ, $i = 1, \ldots, I$, with a view to minimizing the variance of the random variable $Z^\lambda - \tilde{Y}_I^\lambda$ (in practice, the empirical variance $\mathrm{Var_M}(Z^\lambda - \tilde{Y}_I^\lambda)$).

For the moment being, we do not make further precise how we choose the set of parameters $\{\lambda_1, \ldots, \lambda_I\}$ offline. This will be done by the same *greedy procedure* for both algorithms, and will be the subject of the next section. Nevertheless, we would now like to make more precise how we build offline :

- in Algorithm 1, approximations $\{\tilde{Y}^{\lambda_i}, i = 1, \ldots, I\}$ for $\{Y^{\lambda_i}, i = 1, \ldots, I\}$, and
- in Algorithm 2, approximations $\{\nabla \tilde{u}^{\lambda_i}, i = 1, \ldots, I\}$ for $\{\nabla u^{\lambda_i}, i = 1, \ldots, I\}$,

assuming the parameters $\{\lambda_i, i = 1, \ldots, I\}$ have been selected.

For Algorithm 1, $\tilde{Y}^{\lambda_i}$ is built using the fact that it is possible to compute offline *accurate* Monte-Carlo approximations $\mathrm{E}_M(Z^{\lambda_i})$ of $\mathbf{E}\big(Z^{\lambda_i}\big)$ using a very large number $M = M_{\mathrm{large}}$ of copies of Z^{λ_i}, mutually independent and also independant of the copies of Z^λ used for the online Monte-Carlo estimation of $\mathbf{E}\big(Z^\lambda\big)$, $\lambda \neq \lambda_i$ (remember that the amount of offline computations is not meaningful in the case of a very large number of outputs to be computed online). The quantities $\mathrm{E}_{M_{\mathrm{large}}}(Z^{\lambda_i})$ are just real numbers that can be easily stored in memory at the end of the offline stage for re-use online to approximate the control variate $Y^{\lambda_i} = Z^{\lambda_i} - \mathbf{E}\big(Z^{\lambda_i}\big)$ through :

$$\tilde{Y}^{\lambda_i} = Z^{\lambda_i} - \mathrm{E}_{M_{\mathrm{large}}}(Z^{\lambda_i}). \tag{7-II.11}$$

For Algorithm 2, we compute approximations $\tilde{u}^{\lambda_i}$ as numerical solutions to the Kolmogorov backward equation (7-II.7). For example, in the numerical results of Section 7-IV, the PDE (7-II.7) is solved numerically with classical deterministic discretization methods (like finite differences in the calibration problem for instance).

Remark 29 (Algorithm 2 for stochastic processes with large dimension d)**.** *Most deterministic methods to solve a PDE (like the finite difference or finite elements methods) remain suitable only for $d \leq 3$. Beyond, one can for example resort to probabilistic discretizations : namely, a Feynman-Kac representation of the PDE solution, whose efficiency at effectively reducing the variance has already been shown in [New94]. We present this alternative probabilistic approximation in Appendix 7-VI-A, but we will not use it in the present numerical investigation.*

One crucial remark is that for both algorithms, in the online Monte-Carlo computations, the Brownian motions which are used to build the control variate (namely Z^{λ_i} in (7-II.11) for Algorithm 1, and the Brownian motion entering $\tilde{Y}_I^\lambda$ in (7-II.10) for Algorithm 2) are the same as those used for Z^λ.

Note last that, neglecting the approximation errors $\tilde{Y}^{\lambda_i} - Y^{\lambda_i}$ and $\tilde{u}^{\lambda_i} - u^{\lambda_i}$ in the reduced-basis elements computed offline, a comparison between Algorithms 1 and 2 is possible. Indeed, remembering the integral

representation :

$$Y^{\lambda_i} = \int_0^T \nabla u^{\lambda_i}(s, X_s^{\lambda_i}) \cdot \sigma^{\lambda_i}(s, X_s^{\lambda_i}) dB_s \,,$$

we see that the reduced-basis approximation of Algorithm 1 has the form :

$$Y_I^\lambda = \sum_{i=1}^I \mu_i^\lambda \int_0^T \nabla u^{\lambda_i}(s, X_s^{\lambda_i}) \cdot \sigma^{\lambda_i}(s, X_s^{\lambda_i}) dB_s \,,$$

while the reduced-basis approximation of Algorithm 2 has the form :

$$Y_I^\lambda = \sum_{i=1}^I \mu_i^\lambda \int_0^T \nabla u^{\lambda_i}(s, X_s^{\lambda}) \cdot \sigma^{\lambda}(s, X_s^{\lambda}) dB_s \,.$$

The residual variances $\mathbf{Var}\left(Y^\lambda - Y_I^\lambda\right)$ for Algorithms 1 and 2 then respectively read as :

$$\int_0^T \mathbf{E}\left(\left| \nabla u^\lambda \cdot \sigma^\lambda(s, X_s^\lambda) - \sum_{i=1}^I \mu_i^\lambda \nabla u^{\lambda_i} \cdot \sigma^{\lambda_i}(s, X_s^{\lambda_i}) \right|^2 \right) ds, \tag{7-II.12}$$

and :

$$\int_0^T \mathbf{E}\left(\left| \left(\nabla u^\lambda - \sum_{i=1}^I \mu_i^\lambda \nabla u^{\lambda_i} \right) \cdot \sigma^\lambda(s, X_s^\lambda) \right|^2 \right) ds. \tag{7-II.13}$$

The formulas (7-II.12) and (7-II.13) suggest that Algorithm 2 might be more robust than Algorithm 1 with respect to variations of λ. This will be illustrated by some numerical results in Section 7-IV.

7-III Practical variance reduction with approximate control variates

Let us now detail how to select parameters $\{\lambda_i \in \Lambda, i = 1, \ldots, I\}$ offline inside a large *a priori* chosen trial sample $\Lambda_{\text{trial}} \subset \Lambda$ of finite size, and how to effectively compute the coefficients $(\mu_i^\lambda)_{i=1,\ldots,I}$ in the linear combinations $\tilde{Y}_I^\lambda$ (see Section 7-III-C-b for details about practical choices of $\Lambda_{\text{trial}} \subset \Lambda$).

7-III-A Algorithm 1

Recall that some control variates Y^λ are approximated offline with a computationally expensive Monte-Carlo estimator using $M_{\text{large}} \gg 1$ independent copies of Z^λ :

$$\tilde{Y}^\lambda = Z^\lambda - E_{M_{\text{large}}}(Z^\lambda) \approx Y^\lambda, \tag{7-III.1}$$

for only a few parameters $\{\lambda_i, i = 1, \ldots, I\} \subset \Lambda_{\text{trial}}$ to be selected. The approximations $\tilde{Y}^{\lambda_i}$ are then used online to span a linear approximation space for the set of all control variates $\{Y^\lambda, \lambda \in \Lambda\}$, next linearly combined as $\tilde{Y}_I^\lambda$. For any $i = 1, \ldots, I$, we denote by $\tilde{Y}_i^\lambda$ (for any $\lambda \in \Lambda$) the reduced-basis approximation of Y^λ built as a linear combination of the first i selected random variables $\{\tilde{Y}^{\lambda_j}, j = 1, \ldots, i\}$:

$$\tilde{Y}_i^\lambda = \sum_{j=1}^i \mu_j^\lambda \tilde{Y}^{\lambda_j} \approx Y^\lambda \ , \tag{7-III.2}$$

where $(\mu_j^\lambda)_{j=1,\ldots,i} \in \mathbb{R}^i$ is a vector of coefficients to be computed for each λ (and each step i, but we omit to explicitly denote the dependence of each entry μ_j^λ, $j = 1, \ldots, i$, on i). The computation of the coefficients $(\mu_j^\lambda)_{j=1,\ldots,i}$ follows the same procedure offline (for each step $i = 1, \ldots, I-1$) during the reduced-basis construction as online (when $i = I$) : it is based on a variance minimization principle (see details in Section 7-III-A-b).

With a view to computing $\mathbf{E}\left(Z^\lambda\right)$ online through computationally *cheap* Monte-Carlo estimations $\mathrm{E}_{M_{\text{small}}}(Z^\lambda - \tilde{Y}_I^\lambda)$ using only a few M_{small} realizations for all $\lambda \in \Lambda$, we now explain how to select offline a subset $\{\lambda_i, i = 1, \ldots, I\} \subset \Lambda_{\text{trial}}$ in order to minimize $\mathbf{Var}\left(Z^\lambda - \tilde{Y}_I^\lambda\right)$ (or at least estimators for the corresponding statistical error).

Offline : select parameters $\{\lambda_i \in \Lambda_{\text{trial}}, i=1,\dots,I\}$ in $\Lambda_{\text{trial}} \subset \Lambda$ a large finite sample.
Selection under stopping criterium : maximal residual variance $\leq \varepsilon$.

Let $\lambda_1 \in \Lambda$ be already chosen,
Compute accurate approximation $\mathrm{E}_{M_{\text{large}}}(Z^{\lambda_1})$ of $\mathbf{E}\left(Z^{\lambda_1}\right)$.
Greedy procedure :
For step $i=1,\dots,I-1$ $(I>1)$:
For all $\lambda \in \Lambda_{\text{trial}}$, compute $\tilde{Y}_i^{\lambda}$ as (7-III.2) and (cheap) estimations :
$\epsilon_i(\lambda) := \mathrm{Var}_{M_{\text{small}}}\left(Z^{\lambda} - \tilde{Y}_i^{\lambda}\right)$ for $\mathbf{Var}\left(Z^{\lambda} - \tilde{Y}_i^{\lambda}\right)$.
Select $\lambda_{i+1} \in \underset{\lambda \in \Lambda_{\text{trial}} \setminus \{\lambda_j, j=1,\dots,i\}}{\operatorname{argmax}} \{\epsilon_i(\lambda)\}$.
If stopping criterium $\epsilon_i(\lambda_{i+1}) \leq \varepsilon$, Then Exit **Offline**.
Compute accurate approximation $\mathrm{E}_{M_{\text{large}}}(Z^{\lambda_{i+1}})$ of $\mathbf{E}\left(Z^{\lambda_{i+1}}\right)$.

FIG. 7.1 – Offline stage for Algorithm 1 : greedy procedure in metalanguage

7-III-A-a Offline stage : parameter selection

The parameters $\{\lambda_i, i=1,\dots,I\}$ are selected *incrementally* inside the trial sample Λ_{trial} following a *greedy procedure* (see Fig. 7.1). The incremental search between steps i and $i+1$ reads as follows. Assume that control variates $\{\tilde{Y}^{\lambda_j}, j=1,\dots,i\}$ have already been selected at the step i of the reduced basis construction (see Remark 32 for the choice of $\tilde{Y}^{\lambda_1}$). Then, $\tilde{Y}^{\lambda_{i+1}}$ is chosen following the principle of controlling the maximal residual variance inside the trial sample after the variance reduction using the first i selected random variables :

$$\lambda_{i+1} \in \underset{\lambda \in \Lambda_{\text{trial}} \setminus \{\lambda_j, j=1,\dots,i\}}{\operatorname{argmax}} \mathbf{Var}\left(Z^{\lambda} - \tilde{Y}_i^{\lambda}\right), \tag{7-III.3}$$

where the coefficients $(\mu_j^{\lambda})_{j=1,\dots,i}$ entering the linear combinations $\tilde{Y}_i^{\lambda}$ in (7-III.2) are computed, at each step i, like for $\tilde{Y}_I^{\lambda}$ in the online stage (see Section 7-III-A-b).

In practice, the variance in (7-III.3) is estimated by an empirical variance :

$$\mathbf{Var}\left(Z^{\lambda} - \tilde{Y}_i^{\lambda}\right) \simeq \mathrm{Var}_{\mathrm{M}_{\text{small}}}(Z^{\lambda} - \tilde{Y}_i^{\lambda}).$$

In our numerical experiments, we use the same number M_{small} of realizations for the offline computations (for all $\lambda \in \Lambda_{\text{trial}}$) as for the online computations, even though this is not necessary. Note that choosing a small number M_{small} of realizations for the offline computations is advantageous because the computational cost of the Monte-Carlo estimations in the greedy procedure is then cheap. This is useful since Λ_{trial} is very large, and at each step i, $\mathrm{Var}_{\mathrm{M}_{\text{small}}}(Z^{\lambda} - \tilde{Y}_i^{\lambda})$ has to be computed for all $\lambda \in \Lambda_{\text{trial}}$.

Remarkably, after each (offline) step i of the greedy procedure and for the next online stage when $i=I$, only a few real numbers should be stored in memory, namely the collection $\{\mathrm{E}_{M_{\text{large}}}(Z^{\lambda_j}), j=1,\dots,i\}$ along with the corresponding parameters $\{\lambda_j, j=1,\dots,i\}$ for the computation of the approximations (7-III.1).

Remark 30. *Another natural criterium for the parameter selection in the greedy procedure could be the maximal residual variance relatively to the output expectation*

$$\max_{\lambda \in \Lambda_{\text{trial}}} \frac{\mathbf{Var}\left(Z^{\lambda} - \tilde{Y}_i^{\lambda}\right)}{|\mathbf{E}(Z^{\lambda})|^2} \simeq \max_{\lambda \in \Lambda_{\text{trial}}} \frac{\mathrm{Var}_{\mathrm{M}_{\text{small}}}(Z^{\lambda} - \tilde{Y}_i^{\lambda})}{|\mathrm{E}_{\mathrm{M}_{\text{small}}}(Z^{\lambda})|^2}. \tag{7-III.4}$$

This is particularly relevant if the magnitude of the output $\mathbf{E}\left(Z^{\lambda}\right)$ is much more sensitive than that of $\mathbf{Var}\left(Z^{\lambda}\right)$ to the variations on λ. And it also proved useful for comparison and discrimination between Algorithms 1 and 2 in the calibration of a local parametrized volatility for the Black-Scholes equation (see Fig. 7.7).

7-III-A-b Online stage : reduced-basis approximation

To compute the coefficients $(\mu_j^{\lambda})_{j=1,\dots,i}$ in the linear combinations (7-III.2), both *online* for any $\lambda \in \Lambda$ when $i=I$ and *offline* for each $\lambda \in \Lambda_{\text{trial}}$ and each step i (see greedy procedure above), we solve a small-dimensional

least squares problem corresponding to the minimization of (estimators for) the variance of the random variable $Z^\lambda - \tilde{Y}_i^\lambda$.

More precisely, in the case $i=I$ (online stage) for instance, the I-dimensional vector $\mu^\lambda = (\mu_i^\lambda)_{1\leq i\leq I}$ is defined, for any $\lambda\in\Lambda$, as the unique global minimizer of the following strictly convex problem of variance minimization :

$$\mu^\lambda = \underset{\mu=(\mu_i)_{1\leq i\leq I}\in\mathbb{R}^I}{\operatorname{argmin}} \mathbf{Var}\left(Z^\lambda - \sum_{i=1}^I \mu_i \tilde{Y}^{\lambda_i}\right) , \tag{7-III.5}$$

or equivalently as the unique solution to the following linear system :

$$\sum_{j=1}^I \mathbf{Cov}\left(\tilde{Y}^{\lambda_i};\tilde{Y}^{\lambda_j}\right)\mu_j^\lambda = \mathbf{Cov}\left(\tilde{Y}^{\lambda_i};Z^\lambda\right), \forall i=1,\ldots,I. \tag{7-III.6}$$

Of course, in practice, we use the estimator (for $X,Y\in L^2_{\mathbb{P}}(\Omega)$ and $M\in\mathbb{N}_{>0}$) :

$$\mathrm{Cov}_{\mathrm{M}}(X;Y) := \frac{1}{M}\sum_{m=1}^M X_m Y_m - \left(\frac{1}{M}\sum_{m=1}^M X_m\right)\left(\frac{1}{M}\sum_{m=1}^M Y_m\right)$$

to evaluate the statistical quantities above. That is, defining a matrix $\mathbf{C}^{\mathrm{M_{small}}}$ with entries the following empirical Monte-Carlo estimators ($i,j\in\{1,\ldots,I\}$) :

$$\mathbf{C}_{i,j}^{\mathrm{M_{small}}} = \mathrm{Cov}_{\mathrm{M_{small}}}\left(\tilde{Y}^{\lambda_i};\tilde{Y}^{\lambda_j}\right),$$

and a vector $\mathbf{b}^{\mathrm{M_{small}}}$ with entries ($i\in\{1,\ldots,I\}$) $\mathbf{b}_i^{\mathrm{M_{small}}} = \mathrm{Cov}_{\mathrm{M_{small}}}\left(\tilde{Y}^{\lambda_i};Z^\lambda\right)$, the linear combinations (7-III.2) are computed using as coefficients the Monte-Carlo estimators which are entries of the following vector of $\mathbb{R}^I$:

$$\mu^{\mathrm{M_{small}}} = \left[\mathbf{C}^{\mathrm{M_{small}}}\right]^{-1}\mathbf{b}^{\mathrm{M_{small}}}. \tag{7-III.7}$$

The cost of one online computation for one parameter λ ranges as the computation of $\mathrm{M_{small}}$ (independent) realizations of the random variables $(Z^\lambda, Y^{\lambda_1},\ldots,Y^{\lambda_I})$, plus the Monte-Carlo estimators $\mathrm{E}_{M_{\mathrm{small}}}, \mathrm{Cov}_{M_{\mathrm{small}}}, \mathrm{Var}_{M_{\mathrm{small}}}$ and the computation of the solution $\mu^{\mathrm{M_{small}}}$ to the (small I-dimensional, but full) linear system (7-III.7).

In practice, one should be careful when computing (7-III.7), because the likely quasi-colinearity of some reduced-basis elements often induces ill-conditionning of the matrix $\mathbf{C}^{\mathrm{M_{small}}}$. Thus the QR or SVD algorithms [GvL96] should be preferred to a direct inversion of (7-III.6) with the Gaussian elimination or the Cholevsky decomposition. One important remark is that, once the reduced basis is built, the *same* (small I-dimensional) covariance matrix $\mathbf{C}^{\mathrm{M_{small}}}$ has to be inverted for *all* $\lambda\in\Lambda$, as soon as the same Brownian paths are used for each online evaluation. And the latter condition is easily satisfied in practice, simply by resetting the seed of the random number generator to the same value for each new online evaluation (that is for each new $\lambda\in\Lambda$).

Remark 31 (Final output approximations and bounds)**.** *It is a classical result that, taking first the limit* $\mathrm{M_{large}}\to\infty$ *then* $\mathrm{M_{small}}\to\infty$, $\mu^{\mathrm{M_{small}}} \underset{\mathrm{M_{small}},\mathrm{M_{large}}\to\infty}{\overset{\mathbb{P}-a.s.}{\longrightarrow}} \mu^\lambda$. *So, the variance is indeed (asymptotically) reduced to the minimum* $\mathbf{Var}\left(Z^\lambda - Y_I^\lambda\right)$ *in* (7-III.5)*, obtained with the optimal linear combination* Y_I^λ *of selected control variates* Y^{λ_i} *(without approximation). In addition, using Slutsky theorem twice successively for Monte-Carlo estimators of the coefficient vector* μ^λ *and of the variance* $\mathbf{Var}\left(Z^\lambda - Y_I^\lambda\right)$*, it also holds a computable version of the Central Limit Theorem, which is similar to* (7-II.4) *except that it uses Monte-Carlo estimations of* $Z^\lambda - \tilde{Y}_I^\lambda$ *instead of* Z^λ *to compute the confidence intervals (and with successive limits* $\mathrm{M_{large}}\to\infty$, $\mathrm{M_{small}}\to\infty$*). So our output approximations now read for all* $\lambda\in\Lambda$ *:*

$$\mathbf{E}\left(Z^\lambda\right) \simeq \mathrm{E}_{M_{\mathrm{small}}}\left(Z^\lambda - \sum_{i=1}^I \mu_i^{\mathrm{M_{small}}}\tilde{Y}^{\lambda_i}\right),$$

and asymptotic probabilistic error bounds are given by the confidence intervals (7-II.4)*.*

Offline : select parameters $\{\lambda_i \in \Lambda_{\text{trial}}, i=1,\dots,I\}$ in $\Lambda_{\text{trial}} \subset \Lambda$ a large finite sample.
Selection under stopping criterium : maximal residual variance $\leq \varepsilon$.

Let $\lambda_1 \in \Lambda$ be already chosen,
Compute approximation $\nabla\tilde{u}^{\lambda_1}$ of ∇u^{λ_1}.
Greedy procedure :
For step $i=1,\dots,I-1$ $(I>1)$:
For all $\lambda \in \Lambda_{\text{trial}}$, compute $\tilde{Y}_i^\lambda$ as (7-III.8) and estimations :
$\epsilon_i(\lambda) := \text{Var}_{M_{\text{small}}}\left(Z^\lambda - \tilde{Y}_i^\lambda\right)$ for $\mathbf{Var}\left(Z^\lambda - \tilde{Y}_i^\lambda\right)$.
Select $\lambda_{i+1} \in \underset{\lambda \in \Lambda_{\text{trial}} \setminus \{\lambda_j, j=1,\dots,i\}}{\text{argmax}} \{\epsilon_i(\lambda)\}$.
If stopping criterium $\epsilon_i(\lambda_{i+1}) \leq \varepsilon$, Then Exit **Offline**.
Compute approximation $\nabla\tilde{u}^{\lambda_{i+1}}$ of $\nabla u^{\lambda_{i+1}}$.

FIG. 7.2 – Offline stage for Algorithm 2 : greedy procedure in metalanguage

7-III-B Algorithm 2

In Algorithm 2, approximations $\nabla\tilde{u}^{\lambda_i}$ of the gradients ∇u^{λ_i} of the solutions u^{λ_i} to the backward Kolmogorov equation (7-II.7) are computed offline for only a few parameters $\{\lambda_i, i=1,\dots,I\} \subset \Lambda_{\text{trial}}$ to be selected. In comparison with Algorithm 1, approximations $(\nabla\tilde{u}^{\lambda_i})_{i=1,\dots,I}$ are now used online to span a linear approximation space for $\{\nabla u^\lambda,\ \lambda \in \Lambda\}$. At step i of the greedy procedure $(i=1,\dots,I)$, the reduced-basis approximations $\tilde{Y}_i^\lambda$ for the control variates Y^λ read (for all $\lambda \in \Lambda$) :

$$\tilde{Y}_i^\lambda = \sum_{j=1}^{i} \mu_j^\lambda \tilde{Y}_\lambda^{\lambda_j} \approx Y^\lambda, \tag{7-III.8}$$

$$\tilde{Y}_\lambda^{\lambda_j} = \int_0^T \nabla\tilde{u}^{\lambda_j}(s, X_s^\lambda) \cdot \sigma^\lambda(s, X_s^\lambda) dB_s\,. \tag{7-III.9}$$

where $(\mu_j^\lambda)_{j=1,\dots,i}$ are coefficients to be computed for each λ (again, the dependence of μ_j^λ on the step i is implicit). Again, the point is to explain, first, how to select parameters $\{\lambda_i, i=1,\dots,I\} \subset \Lambda_{\text{trial}}$ in the offline stage, and second, how to compute the coefficients $(\mu_j^\lambda)_{j=1,\dots,i}$ in each of the i-dimensional linear combinations $\tilde{Y}_i^\lambda$. Similarly to Algorithm 1, the parameters $\{\lambda_i, i=1,\dots,I\} \subset \Lambda_{\text{trial}}$ are selected offline following the greedy procedure, and, for any $i=1,\dots,I$, the coefficients $(\mu_j^\lambda)_{j=1,\dots,i}$ in the linear combinations offline and online are computed, following the same principle of minimizing the variance, by solving a least squares problem.

7-III-B-a Offline stage : parameter selection

The selection of parameters $\{\lambda_j, j=1,\dots,i\}$ from a trial sample Λ_{trial} follows a greedy procedure like in Algorithm 1 (see Fig. 7.2). In comparison with Algorithm 1, after i (offline) steps of the greedy procedure $(1 \leq i \leq I-1)$ and online $(i=I)$, note that discretizations of functions $(t,y) \to \nabla\tilde{u}^{\lambda_j}(t,y)$, $j=1,\dots,i+1$, are stored in memory to compute the stochastic integrals (7-III.8), which is possibly a huge amount of data.

7-III-B-b Online stage : reduced-basis approximation

Like in Algorithm 1, the coefficients $(\mu_j^\lambda)_{j=1,\dots,i}$ in the linear combination (7-III.8) are computed similarly *online* (and then $i=I$) for any $\lambda \in \Lambda$ and *offline* (when $1 \leq i \leq I-1$) for each $\lambda \in \Lambda_{\text{trial}}$ as minimizers of – a Monte Carlo discretization of – the least squares problem :

$$\min_{\mu \in \mathbb{R}^I} \mathbf{Var}\left(Z^\lambda - \sum_{i=1}^{I} \mu_i \tilde{Y}_\lambda^{\lambda_i}\right)\,, \tag{7-III.10}$$

where we recall that $\tilde{Y}_\lambda^{\lambda_i}$ are defined by (7-III.9). Note that contrary to the reduced-basis elements $\tilde{Y}^{\lambda_i}$ in Algorithm 1, the elements $\tilde{Y}_\lambda^{\lambda_i}$ in Algorithm 2 have to be recomputed for *each* queried parameter value $\lambda \in \Lambda$.

Again, in practice, the unique solution $(\mu_j^\lambda)_{j=1,\dots,i}$ to the variational problem (7-III.10) is equivalently the unique solution to the following linear system :

$$\sum_{j=1}^{I} \mathbf{Cov}\left(\tilde{Y}_\lambda^{\lambda_i};\tilde{Y}_\lambda^{\lambda_j}\right)\mu_j^\lambda = \mathbf{Cov}\left(\tilde{Y}_\lambda^{\lambda_i};Z^\lambda\right), \forall i=1,\dots,I, \tag{7-III.11}$$

and is in fact computed as the unique solution to the discrete minimization problem :

$$\mu^{\mathrm{M_{small}}} = \left[\mathbf{C}^{\mathrm{M_{small}}}\right]^{-1}\mathbf{b}^{\mathrm{M_{small}}}, \tag{7-III.12}$$

with $\mathbf{C}_{i,j}^{\mathrm{M_{small}}} = \mathrm{Cov}_{\mathrm{M_{small}}}\left(\tilde{Y}_\lambda^{\lambda_i};\tilde{Y}_\lambda^{\lambda_j}\right)$ and $\mathbf{b}_i^{\mathrm{M_{small}}} = \mathrm{Cov}_{\mathrm{M_{small}}}\left(\tilde{Y}_\lambda^{\lambda_i};Z^\lambda\right)$.

The cost of one computation online for one parameter λ is more expensive than that in Algorithm 1, and ranges as the computation of M_{small} independent realizations of Z^λ, *plus* the computation of I (discrete approximations of) the stochastic integrals (7-III.9), plus the Monte-Carlo estimators and the solution $\mu^{\mathrm{M_{small}}}$ to the (small I-dimensional, but full) linear system (7-III.12). In comparison to Algorithm 1, notice that the (discrete) covariance matrix $\mathbf{C}^{\mathrm{M_{small}}}$ to be inverted depends on λ, and thus cannot be treated offline once for all : it has to be recomputed for each $\lambda\in\Lambda$.

7-III-C General remarks about reduced-basis approaches

The success of our two reduced-basis approaches clearly depends on the variations of Z^λ with $\lambda\in\Lambda$. Unfortunately, we do not have yet a precise understanding of this, similarly to the PDE case [PR07b]. Our reduced-basis approaches have only been investigated numerically in relevant cases for application (see Section 7-IV). So we now provide some theoretical ground only for the *a priori* existence of a reduced basis, like in the PDE case [MPT02a], with tips for a practical use of the greedy selection procedure based on our numerical experience. Of course, it remains to show that the greedy procedure actually selects a good reduced basis.

7-III-C-a *A priori* existence of a reduced basis

Following the analyses [MPT02a, PR07b] for parametrized PDEs, we can prove the *a priori* existence of a reduced basis for some particular collections of parametrized control variates, under very restrictive assumptions on the structure of the parametrization.

Proposition 29. *Assume there exist collections of uncorrelated (parameter-independent) random variables with zero mean $Y_j\in L^2_{\mathbb{P}}(\Omega)$, $1\le j\le J$, and of positive $C^\infty(\mathbb{R})$ functions g_j, $1\le j\le J$, such that*

$$Y^\lambda = \sum_{j=1}^{J} g_j(\lambda)Y_j, \ \forall\lambda\in\Lambda, \tag{7-III.13}$$

and there exists a constant $C>0$ such that, for all parameter ranges $\Lambda=[\lambda_{\min},\lambda_{\max}]\subset\mathbb{R}$, there exists a C^∞ diffeomorphism τ_Λ defined on Λ satisfying :

$$\sup_{1\le j\le J}\sup_{\tilde{\lambda}\in\tau_\Lambda(\Lambda)} (g_j\circ\tau_\Lambda^{-1})^{(M)}(\tilde{\lambda}) \le M!C^M, \text{ for all } M\text{-derivatives of } g_j\circ\tau_\Lambda^{-1}. \tag{7-III.14}$$

Then, for all parameter ranges $\Lambda=[\lambda_{\min},\lambda_{\max}]\subset\mathbb{R}$, there exist constants $c_1,c_2>0$ independent of Λ and J such that, for all $N\in\mathbb{N}$,$N\ge N_0:=1+c_1(\tau_\Lambda(\lambda_{\max})-\tau_\Lambda(\lambda_{\min}))$, there exist N distinct parameter values $\lambda_n^N\in\Lambda$, $n=1,\dots,N$, (with $\lambda_n^N\le\lambda_{n+1}^N$ when $1\le n\le N-1$), sastisfying, with $\mathcal{Y}_N=\mathbf{Span}\left(Y^{\lambda_n^N}, n=1,\dots,N\right)$:

$$\inf_{Y_N^\lambda\in\mathcal{Y}_N}\mathbf{Var}\left(Z^\lambda-Y_N^\lambda\right) \le e^{-\frac{c_2}{N_0-1}(N-1)}\mathbf{Var}\left(Z^\lambda\right), \ \forall\lambda\in\Lambda. \tag{7-III.15}$$

One can always write Y^λ like (7-III.13) with uncorrelated random variables (using a Gram-Schmidt procedure) and with positive coefficients (at least on a range Λ where they do not vanish). But the assumption (7-III.14) is much more restrictive. The mapping τ_Λ for the parameter, which depends on the functions g_j, $j=1,\dots,J$, indeed tells us how the convergence depends on variations in the size of the parameter range Λ. See [MPT02a, PR07b] for an example of such functions g_j and τ_Λ, and Appendix 7-VI-B for a short proof inspired from [MPT02a, PR07b].

The Proposition 29 may cover a few interesting cases of application for the *a priori* existence theory. One example where the assumption (7-III.13) hold is the following. Consider an output $Z^\lambda = g(X_T^\lambda)$ with g a polynomial function, and :

$$X_t^\lambda = x + \int_0^t b^\lambda(s) X_s^\lambda \, ds + \int_0^t \sigma^\lambda(s) dB_s \,. \tag{7-III.16}$$

The optimal control variate Y^λ in such a case writes in the form (7-III.13) (to see this, one can first explicitly compute the reiterated (or multiple) Itô integrals in the polynomial expression of $g(X_T^\lambda)$ with Hermite polynomials [KS91]). Then, (7-III.14) may hold provided b^λ and σ^λ are smooth functions of $\lambda \in \Lambda$ (again, see [MPT02a, PR07b] for functions g_j satisfying (7-III.14)). But quite often, the reduced bases selected in practice by the *greedy* procedure are much better than $\mathcal{Y}_N$ (see [PR07b] for comparisons when λ is scalar).

7-III-C-b Requirements for efficient practical greedy selections

A comprehensive study would clearly need hypotheses about the regularity of Y^λ as a function of λ and about the discretization Λ_{trial} of Λ to show that the greedy procedure actually selects good reduced bases. We do not have precise results yet, but we would nevertheless like to provide the reader with conjectured requirements for the greedy procedure to work and help him as a potential user of our method.

Ideally, one would use the greedy selection procedure directly on $\{Y^\lambda, \lambda \in \Lambda\}$ for Algorithm 1 and on $\{\nabla u^\lambda, \lambda \in \Lambda\}$ for Algorithm 2. But in pratice, one has to resort to approximations only, $\{\tilde{Y}^\lambda, \lambda \in \Lambda\}$ for Algorithm 1 and $\{\nabla \tilde{u}^\lambda, \lambda \in \Lambda\}$ for Algorithm 2. So, following requirements on discretizations of parametrized PDEs in the classical reduced-basis method [PR07b], the stability of the reduced basis selected by the greedy procedure for parametrized control variates intuitively requires :

(H1) *For any required accuracy $\varepsilon > 0$, we assume the existence of approximations, $\tilde{Y}^\lambda$ for Y^λ in Algorithm 1 (resp. $\tilde{u}^\lambda$ for u^λ in Algorithm 2), such that the L^2-approximation error is uniformly bounded on Λ :*

$$\forall \lambda \in \Lambda, \mathbf{E}\left(|\tilde{Y}^\lambda - Y^\lambda|^2\right) \leq \varepsilon,$$

$$\left(\textit{resp.} \int_0^T \mathbf{E}\left(|\nabla \tilde{u}^\lambda - \nabla u^\lambda|^2 (X_t^\lambda)\right) dt \leq \varepsilon \text{ or } \|\nabla \tilde{u}^\lambda - \nabla u^\lambda\|_{L^2}^2 \leq \varepsilon \right).$$

Moreover, in practice, one can only manipulate finite nested samples of parameter Λ_{trial} instead of the full range Λ. So some representativity assumption about Λ_{trial} is also intuitively required for the greedy selection procedure to work on Λ :

(H2) *For any required accuracy $\varepsilon > 0$, we assume the existence of a sufficiently representative finite discrete subset $\Lambda_{\text{trial}} \subset \Lambda$ of parameters such that reduced bases built from Λ_{trial} are still good enough for Λ.*

Refering to Section 7-III-C-a, *good enough* reduced bases should satisfy exponential convergence like (7-III.15), with slowly deteriorating capabilities in terms of approximation when the size of the parameter range grows. Now, in absence of more precise result, intuitition has been necessary so far to choose good discretizations. The numerical results of Section 7-IV have been obtained with $\mathrm{M}_{\text{large}} = 100\,\mathrm{M}_{\text{small}}$ in Algorithm 1, and with a trial sample Λ_{trial} of 100 parameter values randomly chosen (with uniform distribution) in Λ.

In absence of theory for the greedy procedure, one could also think of using another parameter selection procedure in the offline stage. The interest of the greedy procedure is that it is cheap while effective in practice. In comparison, another natural reduced basis would be defined by the first I leading eigenvectors from the Principal Components Analysis (PCA) of the very large covariance matrix with entries $\mathbf{Cov}\left(Y^{\lambda_i}; Y^{\lambda_j}\right)_{(\lambda_i, \lambda_j) \in \Lambda_{\text{trial}} \times \Lambda_{\text{trial}}}$. The latter (known as the Proper Orthogonal Decomposition method) may yield similar variance reduction for most parameter values $\lambda \in \Lambda$ [PR07b], but would certainly require more computations during the offline stage.

Remark 32. *The choice of the first selected parameter λ_1 has not been precised yet. It is observed that most often, this choice does not impact the quality of the variance reduction. But to be more precise, we choose $\lambda_1 \in \Lambda_{\text{small trial}}$ such that Z^{λ_1} has maximal variance in a small initial sample $\Lambda_{\text{small trial}} \subset \Lambda$, for instance.*

7-IV Worked examples and numerical tests

The efficiency of our reduced-basis strategies for parametrized problems is now investigated numerically for two problems relevant to some applications.

Remark 33 (High-dimensional parameter)**.** *Although the maximal dimension in the parameter treated here is two, one can reasonably hope for our reduced-basis approach to remain feasible with moderately high-dimensions in the parameter range Λ, say twenty. Indeed, a careful mapping of a multi-dimensional parameter range may allow for an efficient sampling Λ_{trial} that makes a greedy procedure tractable and next yields a good reduced basis for Λ, as it was shown for the classical reduced-basis method with parametrized PDEs [Sen08, BBM+09].*

7-IV-A Scalar process with constant drift and parametrized diffusion

7-IV-A-a Calibration of the Black–Scholes model with local volatility

One typical computational problem in finance is the valuation of an option depending on a risky asset with value S_t at time $t\in[0,T]$. In the following we consider Vanilla European Call options with payoff $\phi(S_T;K)=\max(S_T-K,0)$, K being the exercise price (or strike) of the option at time $t=T$. By the no arbitrage principle for a portfolio mixing the risky asset of value S_t with a riskless asset of interest rate $r(t)$, the price (as a function of time) is a martingale given by a conditional expectation :

$$e^{-\int_t^T r(s)ds}\mathbf{E}(\phi(S_T)|\mathcal{F}_t) \tag{7-IV.1}$$

where, in the Black-Scholes model with local volatility, $S_t=S_t^\lambda$ is a stochastic process solving the Black-Scholes equation :

$$dS_t^\lambda = S_t^\lambda\left(r(t)\,dt+\sigma^\lambda(t,S_t^\lambda)\,dB_t\right)\qquad S_{t=0}^\lambda=S_0\,, \tag{7-IV.2}$$

and $(\mathcal{F}_t)$ is the natural filtration for the standard Brownian motion (B_t). For this model to be predictive the parameter λ in the (local) volatility σ^λ needs to be calibrated against observed data.

Calibration, like many numerical optimization procedures, defines a typical many-query context, where one has to compute many times the price (7-IV.1) of the option for a large number of parameter values until, for some optimal parameter value λ, a test of adequation with statistical data $P_{(K,\bar t_l)}$ observed on the market at times $\bar t_l\in[0,T]$, $l=0,\dots,\bar L$ is satisfied. For instance, a common way to proceed is to minimize in λ the quadratic quantity :

$$\mathcal{J}(\lambda)=\sum_{l=0}^{\bar L}\left|e^{-\int_{\bar t_l}^T r(s)\,ds}\mathbf{E}\left(\phi(S_T^\lambda;K)|\mathcal{F}_{\bar t_l}\right)-P_{(K,\bar t_l)}\right|^2,$$

most often regularized with some Tychonoff functional, using optimization algorithms like descent methods which indeed require many evaluations of the functional $\mathcal{J}(\lambda)$ for various λ. One could even consider the couple (K,T) as additional parameters to optimize the contract, but we do not consider such an extension here.

Note that the reduced-basis method for parameterized PDEs [MMP01, MPT02a, PR07b] has recently proved very efficient at treating a similar calibration problem [Pir08]. Our approach is different since we consider a probabilistic pricing numerical method.

In the following numerical results, we solve (7-IV.1) for many parameter values assuming that the interest rate r is a fixed given constant and the local volatility σ^λ has "hyperbolic" parametrization (7-IV.3) (used by practitionners in finance) :

$$\sigma^\lambda(t,S)=(\Gamma+1)\left(\frac{1}{C(0,S_0)}+\frac{\Gamma}{C(t,S)}\right)^{-1} \tag{7-IV.3}$$

where $C(t,S)=\frac12\left(\sqrt{C_A(t,S)^2+C_{\min}^2}+C_A(t,S)\right)$ with :

$$C_A(t,S)=a+\frac12\sqrt{(b-c)^2\log^2\left(\frac{S}{\alpha S_0e^{rt}}\right)+4a^2d^2}+\frac12(b+c)\log\left(\frac{S}{\alpha S_0e^{rt}}\right).$$

The local volatility σ^λ is thus parametrized with a 7-dimensional parameter $\lambda=(a,b,c,d,\alpha,\Gamma,C_{\min})$.

Our reduced-basis approach aims at building a vector space in order to approximate the family of random variables :

$$\left\{Y^\lambda:=e^{-rT}\max(S_T^\lambda-K,0)-e^{-rT}\mathbf{E}\left(\max(S_T^\lambda-K,0)\right),\lambda\in\Lambda\right\},$$

which are optimal control variates for the computation of the expectation of $e^{-rT}\max(S_T^\lambda-K,0)$. In Algorithm 2, we also use the fact that

$$Y^\lambda=\int_0^T\partial_S u^\lambda(t,S_t^\lambda)\sigma^\lambda(t,S_t^\lambda)S_t^\lambda dB_t\,, \tag{7-IV.4}$$

where the function $u^\lambda(t,S)$ solves for $(t,S)\in[0,T)\times(0,\infty)$:

$$\partial_t u^\lambda(t,S)+rS\partial_S u^\lambda(t,S)+\frac{\sigma^\lambda(t,S)^2S^2}{2}\partial_{SS}u^\lambda(t,S)=0, \tag{7-IV.5}$$

with final condition $u^\lambda(T,S)=e^{-rT}\max(S-K,0)$. Note the absence of boundary condition at $S=0$ because the advection and diffusion terms are zero at $S=0$. The backward Kolmogorov equation (7-IV.5) is *numerically* solved using finite differences [AP05]. More precisely, after a change of variable $u^\lambda(t,S)=e^{-rt}C^\lambda(t,S)$, equation (7-IV.4) rewrites :

$$Y^\lambda=\int_0^T e^{-rt}\partial_S C^\lambda(t,S_t^\lambda)\sigma^\lambda(t,S_t^\lambda)S_t^\lambda dB_t, \tag{7-IV.6}$$

where $C^\lambda(t,S)$ solves the classical Black-Scholes PDE :

$$\partial_t C^\lambda(t,S)-rC^\lambda(t,S)+rS\partial_S C^\lambda(t,S)+\frac{\sigma^\lambda(t,S)^2S^2}{2}\partial_{SS}C^\lambda(t,S)=0, \tag{7-IV.7}$$

with the final condition $C^\lambda(T,S)=\max(S-K,0)$. In the case of a low-dimensional variable S_t (like one-dimensional here), one can use a finite differences method of order 2 (with Crank-Nicholson discretization in time) to compute approximations $\tilde{C}^\lambda_{l,j}\simeq C^\lambda(t_l,x_j)$, $l=0,\dots,L$, $j=0,\dots,J$ on a grid for the truncated domain $[0,T]\times[0,3K]\subset[0,T]\times[0,\infty)$, with $L=100$ steps in time and $J=300$ steps in space of constant sizes (and with Dirichlet boundary condition $\tilde{C}^\lambda_{l,J+1}=(3-e^{-r(T-t_l)})K,\forall l=0,\dots,N$ at the truncated boundary). An approximation $\tilde{C}^\lambda(t,S)$ of $C^\lambda(t,S)$ at any $(t,S)\in[0,T]\times[0,3K]$ is readily reconstructed as a linear interpolation on tiles $(t,S)\in[t_l,t_{l+1}]\times[S_j,S_{j+1}]$.

7-IV-A-b Numerical results

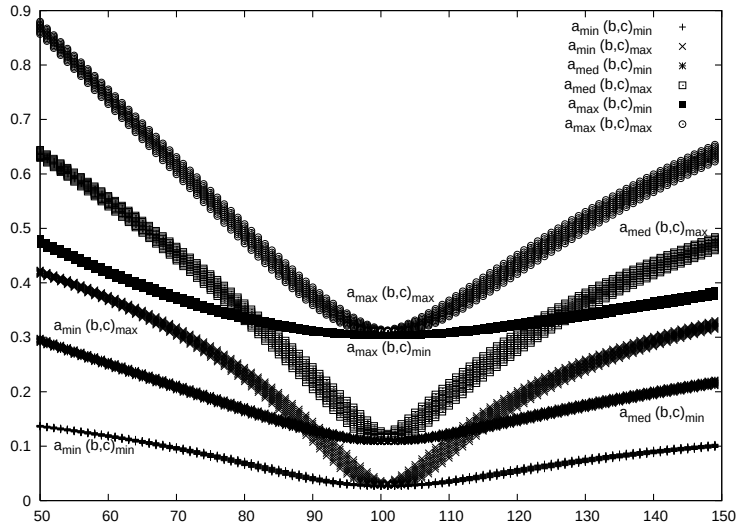


FIG. 7.3 – Variations of the "hyperbolic" local volatility function $\sigma^\lambda(t,S)$ with respect to $S\in[50,150]$. Six families of curves are shown (as time t evolves in $[0,1]$) for extremal and mid- values of the parameter $(a,b=c)$ in $[-.05,.15]\times\{b=c\in[.5,1.5]\}$: $(\min(a),\min(b=c))$, $(\min(a),\max(b=c))$, $(\mathrm{med}(a):=.5\min(a)+.5\max(a),\min(b=c))$, $(\mathrm{med}(a):=.5\min(a)+.5\max(a),\max(b=c))$, $(\max(a),\min(b=c))$, $(\max(a),\max(b=c))$. Each family of curves shows the time variations of $S\to\sigma^\lambda(t,S)$ for $t\in\{.1\times k|k=0,\dots,10\}$).

The Euler-Maruyama scheme with $N=10^2$ time steps of constant size $\Delta t=\frac{T}{N}=10^{-2}$ is used to compute one realization of a pay-off $\max(\tilde{S}^\lambda_N-K,0)$, for a strike $K=100$ at final time $t_N=T=1$ when the initial price is $\tilde{S}^\lambda_0=90$ and the interest rate $r=0.04$. Then, (a large number of) expectations $\mathbf{E}\Big(\max(\tilde{S}^\lambda_N-K,0)\Big)$ are approximated through Monte-Carlo evaluations $E_{\mathrm{M_{small}}}\Big(\max(\tilde{S}^\lambda_N-K,0)\Big)$ with $\mathrm{M_{small}}=10^3$ realizations, when the local volatility parameter $\lambda=(a,b,c,d,\alpha,\Gamma,C_{\min})$ assumes many values in the two-dimensional range

$\Lambda = [-.05,.15] \times \{b = c \in [.5,1.5]\} \times \{1.\} \times \{1.1\} \times \{5\} \times \{.05\}$ (variations of the function $\sigma^\lambda(t,S)$ with λ are shown in Fig. 7.3).

We build reduced bases of different sizes $I = 1,\dots,20$ from the same sample Λ_{trial} of size $|\Lambda_{\text{trial}}| = 100$, either with Algorithm 1 (Fig. 7.4 and 7.6) using approximate control variates computed with $\mathrm{M}_{\text{large}} = 100\mathrm{M}_{\text{small}}$ evaluations :

$$\tilde{Y}_I^\lambda = \sum_{i=1}^{I} \mu_i^{\mathrm{M}_{\text{small}}} \tilde{Y}^{\lambda_i} = \sum_{i=1}^{I} \mu_i^{\mathrm{M}_{\text{small}}} \left(\max(\tilde{S}_N^{\lambda_i} - K, 0) - E_{\mathrm{M}_{\text{large}}} \left(\max(\tilde{S}_N^{\lambda_i} - K, 0) \right) \right),$$

or with Algorithm 2 (Fig. 7.5 and 7.6) using approximate control variates :

$$\tilde{Y}_I^\lambda = \sum_{i=1}^{I} \mu_i^{\mathrm{M}_{\text{small}}} \left(\sum_{n=0}^{N-1} e^{-rt_n} \partial_S \tilde{C}^{\lambda_i}(t_n, \tilde{S}_n^\lambda) \sigma^\lambda(t_n, \tilde{S}_n^\lambda) \sqrt{|t_{n+1} - t_n|} G_n \right)$$

computed as first-order discretizations of the Itô stochastic integral (7-IV.6) using the finite-difference approximation of the solution to the backward Kolmogorov equation. We always start the greedy selection procedure by choosing λ_1 such that $\tilde{Y}^{\lambda_1}$ has the maximal correlation with other members in $\Lambda_{\text{small trial}}$, a small prior sample of 10 parameter values chosen randomly with uniform law in Λ, see Remark 32.

We show in Fig. 7.4 and 7.5 the absolute variance after variance reduction :

$$\mathrm{Var}_{\mathrm{M}_{\text{small}}} \left(\max(\tilde{S}_N^\lambda - K, 0) - \tilde{Y}_I^\lambda \right), \tag{7-IV.8}$$

and in Fig. 7.6 the relative variance after variance reduction :

$$\frac{\mathrm{Var}_{\mathrm{M}_{\text{small}}} \left(\max(\tilde{S}_N^\lambda - K, 0) - \tilde{Y}_I^\lambda \right)}{\mathrm{E}_{\mathrm{M}_{\text{small}}} \left(\max(\tilde{S}_N^\lambda - K, 0) - \tilde{Y}_I^\lambda \right)^2}. \tag{7-IV.9}$$

In each figure, the maximum, the minimum and the mean of one of the two residual variance above is shown, either within the offline sample deprived of the selected parameter values $\Lambda_{\text{trial}} \setminus \{\lambda_i, i = 1,\dots,I\}$, or within an online uniformly distributed sample test $\Lambda_{\text{test}} \subset \Lambda$ of size $|\Lambda_{\text{test}}| = 10|\Lambda_{\text{trial}}|$.

It seems that Algorithm 1 slighlty outperfoms Algorithm 2 with a sufficiently large reduced basis, comparing the (online) decrease rates for either the relative variance or the absolute variance. Yet, one should also notice that, with very small-dimensional reduced basis, the Algorithm 2 yields very rapidly good variance reduction. Comparing the decrease rates of the variance in offline and online samples tells us how good was the (randomly uniformly distributed here) choice of Λ_{trial}. The Algorithm 2 seems more robust than the Algorithm 1 for reproducing ("extrapolating") offline results from a sample Λ_{trial} in the whole range Λ. So, comparing the first results for Algorithms 1 and 2, it is not clear which algorithm performs the best variance reduction for a given size of the reduced basis.

Now, in Fig. 7.7 and 7.8, we show the online (absolute and relative) variance for a new sample test of parameters $\Lambda_{\text{testwide}}$ uniformly distributed in $\Lambda_{\text{wide}} = [-.15,.25] \times \{b = c \in]0,2[\} \times \{1.\} \times \{1.1\} \times \{5\} \times \{.05\}$, which is twice larger than $\Lambda = [-.05,.15] \times \{b = c \in [.5,1.5]\} \times \{1.\} \times \{1.1\} \times \{5\} \times \{.05\}$ where the training sample Λ_{trial} of the offline stage is nested : the quality of the variance reduction compared to that for a narrower sample test Λ_{test} seems to decrease faster for Algorithm 1 than for Algorithm 2. So Algorithm 2 definitely seems more robust with respect to the variations in λ than Algorithm 1. This observation is even further increased if we use the relative variance (7-IV.9) instead of the absolute variance (7-IV.8), as shown by the results in Fig. 7.7 and 7.8.

7-IV-B Vector processes with constant diffusion and parametrized drift

7-IV-B-a Molecular simulation of dumbbells in polymeric fluids

In rheology of polymeric viscoelastic fluids, the long polymer molecules responsible for the viscoelastic behaviour can be modelled through kinetic theories of statistical physics as Rouse chains, that is as chains of Brownian beads connected by springs. We concentrate on the most simple of those models, namely "dumbbells" (two beads connected by one spring) diluted in a Newtonian fluid.

Kinetic models consist in adding to the usual velocity and pressure fields $(\boldsymbol{u}, p)$ describing the (macroscopic) state of the Newtonian solvent, a field of dumbbells represented by their end-to-end vector $\boldsymbol{X}_t(\underline{x})$ at time t and

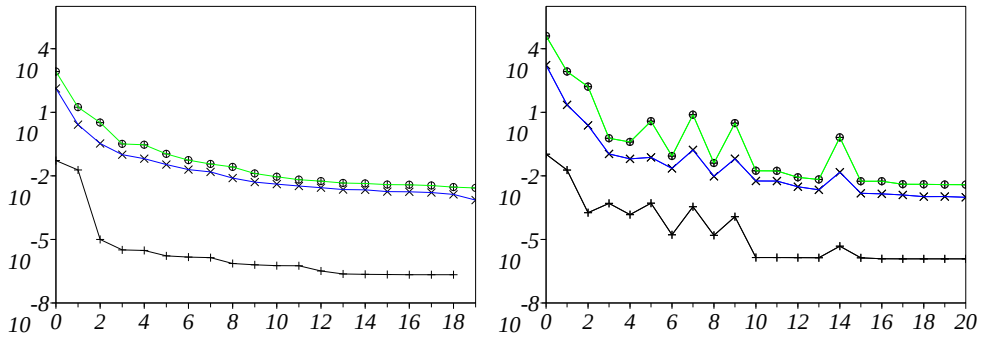


FIG. 7.4 – Algorithm 1 for Black–Scholes model with local "hyperbolic" volatility : Minimum +, mean × and maximum ∘ of the absolute variance (7-IV.8) in samples of parameters (left : offline sample $\Lambda_{\text{trial}} \setminus \{\lambda_i, i = 1,\dots,I\}$; right : online sample Λ_{test}) with respect to the size I of the reduced basis.

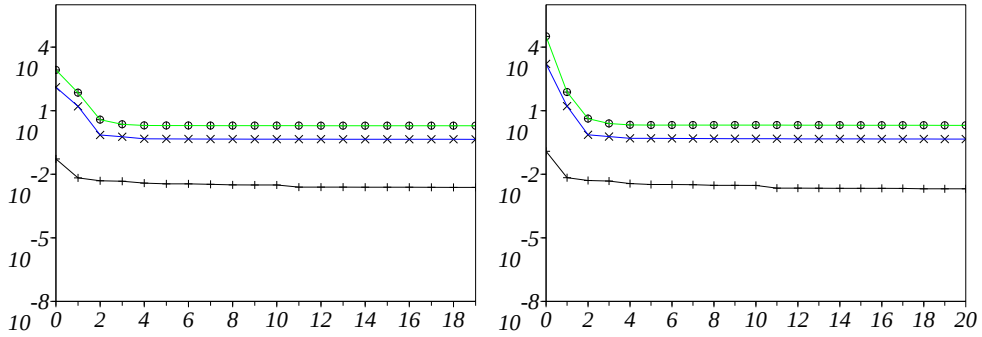


FIG. 7.5 – Algorithm 2 for Black–Scholes model with local "hyperbolic" volatility : Minimum +, mean × and maximum ∘ of the absolute variance (7-IV.8) in samples of parameters (left : offline sample $\Lambda_{\text{trial}} \setminus \{\lambda_i, i = 1,\dots,I\}$; right : online sample Λ_{test}) with respect to the size I of the reduced basis.

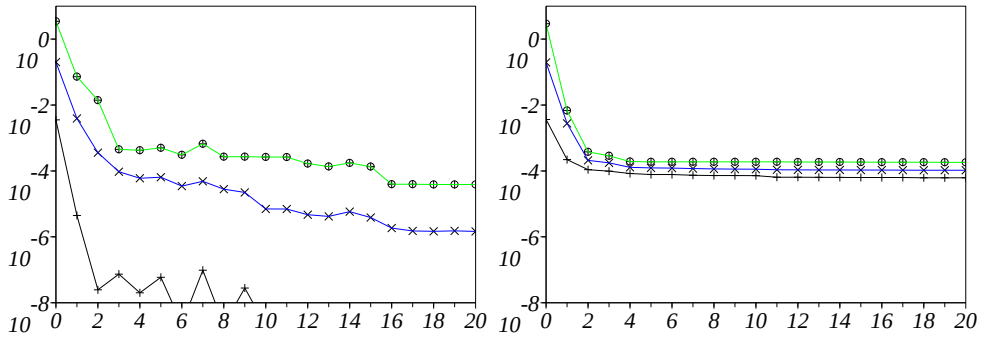


FIG. 7.6 – Algorithm 1 (left) and 2 (right) for Black–Scholes model with local "hyperbolic" volatility : Minimum +, mean × and maximum ∘ of the relative variance (7-IV.9) in a sample test (online) Λ_{test} of parameters with respect to the size I of the reduced basis.

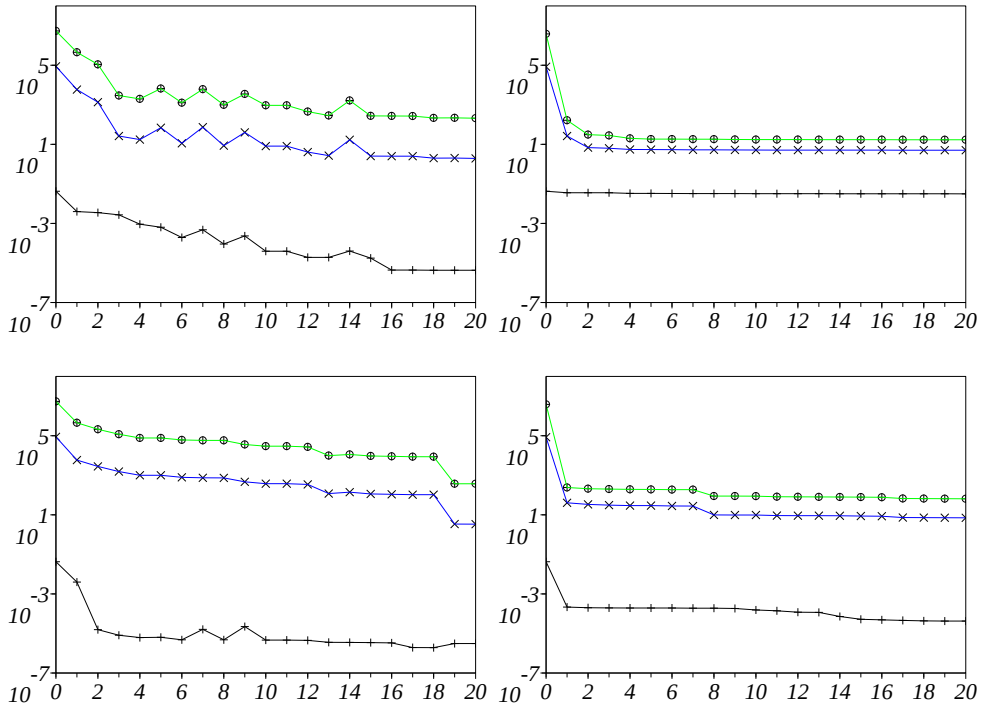


FIG. 7.7 – Algorithm 1 (left) and 2 (right) for Black–Scholes model with local "hyperbolic" volatility : Minimum $+$, mean $\times$ and maximum $\circ$ of the (online) absolute variance (7-IV.8) in a sample test $\Lambda_{\text{testwide}}$ of parameters with respect to the size I of the reduced basis. Greedy selection with absolute variance (7-IV.8) (top) and relative variance (7-IV.9) (bottom).

position $\underline{x}$ in the fluid. Vector stochastic processes $(\boldsymbol{X}_t(\underline{x}))$ encode the time evolution of the orientation and the stretch of the dumbbells (the idealized configuration of a polymer molecule) for each position $\underline{x}\in\mathcal{D}$ in a macroscopic domain $\mathcal{D}$ where the fluid flows. To compute the flow of a viscoelastic fluid with such multiscale dumbbell models [BL09b], segregated algorithms are used that iteratively, on successive time steps with duration T :

- first evolve the velocity and pressure fields $(\boldsymbol{u},p)$ of the Newtonian solvent under a fixed extra (polymeric) stress tensor field $\boldsymbol{\tau}$(typically following Navier-Stokes'equations), and
- then evolve the (probability distribution of the) polymer configurations vector field $(\boldsymbol{X}_t(\underline{x}))$ surrounded by the newly computed fixed velocity field $\boldsymbol{u}$.

The physics of kinetic models is based on a scale separation between the polymer molecules and the surrounding Newtonian fluid solvent. On the one side, the polymer configurations are directly influenced by the (local) velocity and pressure of the Newtonian solvent in which they are diluted. Reciprocally, on the other side, one needs to compute at every $\underline{x}\in\mathcal{D}$ the extra (polymeric) stress, given the Kramers formula :

$$\boldsymbol{\tau}(T,\underline{x})=\mathbf{E}(\boldsymbol{X}_T(\underline{x})\otimes\boldsymbol{F}(\boldsymbol{X}_T(\underline{x})))\,,$$

after one evolution step $t\in[0,T]$ over which the polymer configurations have evolved (remember that here $[0,T]$ should be understood as a timestep). The vector valued process $\boldsymbol{X}_t(\underline{x})$ in $\mathbb{R}^d$ ($d=2$ or 3) solves a Langevin equation at every physical point $\underline{x}\in\mathcal{D}$ (Eulerian description) :

$$d\boldsymbol{X}_t+\boldsymbol{u}\cdot\nabla_{\underline{x}}\boldsymbol{X}_t dt=\big((\nabla_{\underline{x}}\boldsymbol{u})\,\boldsymbol{X}_t-\boldsymbol{F}(\boldsymbol{X}_t)\big)\,dt+d\boldsymbol{B}_t\,.$$

This Langevin equation describes the evolution of polymers at each $\underline{x}\in\mathcal{D}$, under an advection $\boldsymbol{u}\cdot\nabla_{\underline{x}}\boldsymbol{X}_t$, a hydrodynamic force $(\nabla_{\underline{x}}\boldsymbol{u})\boldsymbol{X}_t$, Brownian collisions $(\boldsymbol{B}_t)$ with the solvent molecules, and an entropic force $\boldsymbol{F}(\boldsymbol{X}_t)$ specific to the polymer molecules. Typically, this entropic force reads either $\boldsymbol{F}(\boldsymbol{X})=\boldsymbol{X}$ (for Hookean dumbbells), or $\boldsymbol{F}(\boldsymbol{X})=\frac{\boldsymbol{X}}{1-|\boldsymbol{X}|^2/b}$ (for Finitely-Extensible Nonlinear Elastic or FENE dumbells, to model the finite extensibility of polymers : $|\boldsymbol{X}|<\sqrt{b}$).

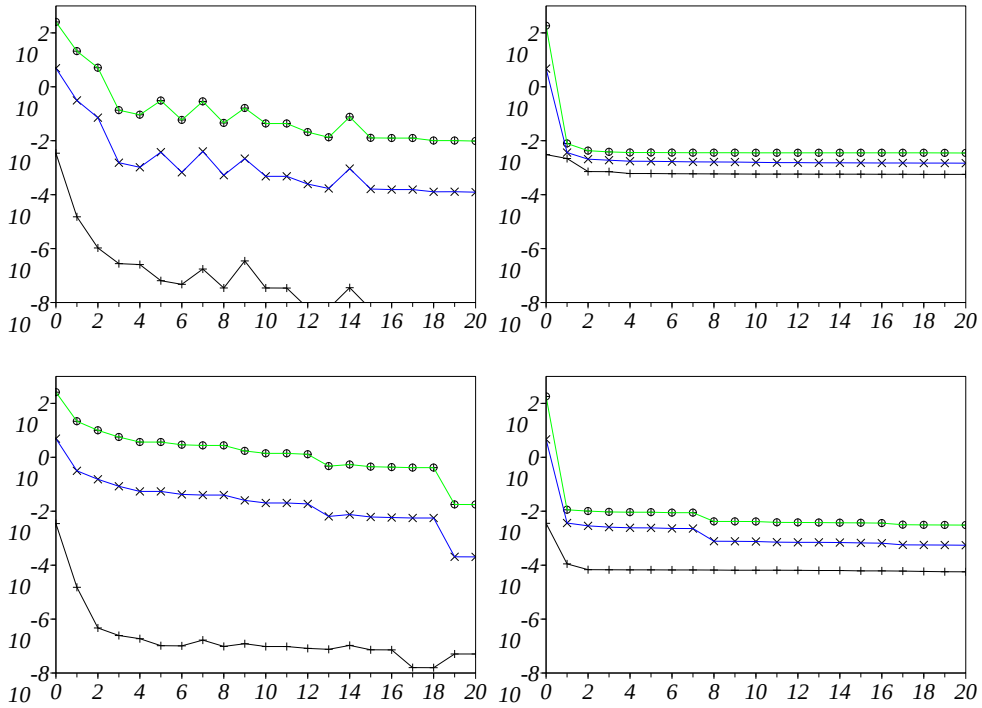


FIG. 7.8 – Algorithm 1 (left) and 2 (right) for Black–Scholes model with local "hyperbolic" volatility : Minimum $+$, mean $\times$ and maximum $\circ$ of the (online) relative variance (7-IV.9) in a sample test $\Lambda_{\text{testwide}}$ of parameters with respect to the size I of the reduced basis. Greedy selection with absolute variance (7-IV.8) (top) and relative variance (7-IV.9) (bottom).

In the following, we do not consider the advection term $\boldsymbol{u}\cdot\nabla_{\underline{x}}\boldsymbol{X}_t$ (which can be handled through integration of the characteristics in a semi-Lagrangian framework, for instance), and we concentrate on solving the parametrized SDE :

$$d\boldsymbol{X}_t = \Big(\underline{\underline{\lambda}}\boldsymbol{X}_t - \boldsymbol{F}(\boldsymbol{X}_t)\Big)\,dt + d\boldsymbol{B}_t\,, \tag{7-IV.10}$$

on a time slab $[0,T]$, with a fixed matrix $\underline{\underline{\lambda}}(\underline{x}) = \nabla_x \boldsymbol{u}(\underline{x})$. We also assume, as usual for viscoelastic fluids, that the velocity field is incompressible (that is $\mathrm{tr}(\underline{\underline{\lambda}}) = 0$), hence the parameter $\underline{\underline{\lambda}}$ is only (d^2-1)-dimensional.

This is a typical many-query context where the Langevin equation (7-IV.10) has to be computed many times at each (discretized) position $\underline{x}\in\mathcal{D}$, for each value of the $d\times d$-dimensional parameter $\underline{\underline{\lambda}}$ (since $\nabla_{\underline{x}}\boldsymbol{u}(\underline{x})$ depends on the position $\underline{x}$). Furthermore, the computation of the time-evolution of the flow defines a *very demanding* many-query context where the latter has to be done iteratively over numerous time steps of duration T between which the tensor field $\underline{\underline{\lambda}}(\underline{x})$ is evolved through a macroscopic equation for the velocity field $\boldsymbol{u}$.

Remark 34 (Initial Condition of the SDE as additional parameter)**.** *Let $T_0 = 0$ and $T_{n+1} = (n+1)T$. Segregated numerical schemes for kinetic models of polymeric fluids as described above simulate* (7-IV.10) *on successive time slabs $[T_n, T_{n+1}]$, for $n\in\mathbb{N}$. More precisely, on each time slab $[T_n, T_{n+1}]$, one has to compute*

$$\begin{aligned}\boldsymbol{\tau}(T_{n+1}) &= \mathbf{E}\big(\boldsymbol{X}_{T_{n+1}}\otimes\boldsymbol{F}(\boldsymbol{X}_{T_{n+1}})\big)\\ &= \mathbf{E}\big(\mathbf{E}\big(\boldsymbol{X}_{T_{n+1}}\otimes\boldsymbol{F}(\boldsymbol{X}_{T_{n+1}})|\boldsymbol{X}_{T_n}\big)\big)\end{aligned} \tag{7-IV.11}$$

at a fixed position $\underline{x}\in\mathcal{D}$. In practice, (7-IV.11) *can be approximated through*

$$\boldsymbol{\tau}(T_{n+1}) \simeq \frac{1}{R}\sum_{r=1}^{R}\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{X}^{r,m}_{T_{n+1}}\otimes\boldsymbol{F}(\boldsymbol{X}^{r,m}_{T_{n+1}}), \tag{7-IV.12}$$

after simulating MR processes $(\boldsymbol{X}^{r,m}_t)_{t\in[T_n,T_{n+1}]}$ driven by MR independent Brownian motions for a given set

of R different initial conditions, typically :

$$\boldsymbol{X}_{T_n^+}^{r,m} = \boldsymbol{X}_{T_n^-}^{r,1},\ r=1,\dots,R,\ m=1,\dots,M,$$

or any $\boldsymbol{X}_{T_n^-}^{r,m_0}$ (with $1 \le m_0 \le M$) given by the computation at final time of the previous time slab $[T_{n-1},T_n]$. In view of (7-IV.12)*, for a fixed r, the computation of $\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{X}_{T_{n+1}}^{r,m} \otimes \boldsymbol{F}(\boldsymbol{X}_{T_{n+1}}^{r,m})$ using the algorithms presented above requires a modification of the methods to the case when the initial condition of the SDE assumes many values.*

To adapt Algorithm 1 to the context of SDEs with many different initial conditions, one should consider reduced bases for control variates which depend on the joint-parameter (λ,x) where x is the initial condition of the SDE. And variations on the joint-parameter (λ,x) can be simply recast into the framework of SDEs with fixed initial condition used for presentation of Algorithm 1 after the change of variable $\hat{X}_t^{\lambda,x} = X_t^\lambda - x$, that is using the family of SDEs with fixed initial condition $\hat{X}^{\lambda,x}{}_0 = 0$:

$$d\hat{X}_t^{\lambda,x} = \hat{b}^{\lambda,x}(t,\hat{X}_t^{\lambda,x})dt + \hat{\sigma}^{\lambda,x}(t,\hat{X}_t^{\lambda,x})dB_t, \tag{7-IV.13}$$

where $\hat{b}^{\lambda,x}(t,X) = b^\lambda(t,X+x)$, $\hat{\sigma}^{\lambda,x}(t,X) = \sigma^\lambda(t,X+x)$, for all t, X and x. Then, with $\hat{g}^{\lambda,x}(X) = g^\lambda(X+x)$ and $\hat{f}^{\lambda,x}(t,X) = f^\lambda(t,X+x)$, the output is the expectation of

$$\hat{Z}^{\lambda,x} = \hat{g}^{\lambda,x}(\hat{X}_T^{\lambda,x}) - \int_0^T \hat{f}^{\lambda,x}(s,\hat{X}_s^{\lambda,x})\,ds.$$

And the corresponding "ideal" control variate reads $\hat{Y}^{\lambda,x} = \hat{Z}^{\lambda,x} - \mathbf{E}\left(\hat{Z}^{\lambda,x}\right)$.

In Algorithm 2, note that u^λ solution to (7-II.7) *does not depend on the initial condition used for the SDE. So, once parameters λ_i ($i=1,\dots,I$) have been selected offline, Algorithm 2 applies similarly for SDEs with one fixed, or many different, initial conditions. Though, the offline selection of parameters λ_i using SDEs with many different initial conditions should consider a larger trial sample than for one fixed initial condition. Indeed, the selection criterium in the greedy algorithm does depend on the initial condition of the SDE. So, defining a trial sample of initial conditions $\Lambda_{\rm IC}$, the following selection should be performed at step i in Fig. 7.2 :*

$$\text{Select } \lambda_{i+1} \in \underset{\lambda\in\Lambda_{\rm trial}\setminus\{\lambda_j, j=1,\dots,i\}}{\operatorname{argmax}} \max_{x\in\Lambda_{\rm IC}} \operatorname{Var}_{\mathrm{M}_{\rm small}}(Z^{\lambda,x} - \tilde{Y}_i^{\lambda,x}),$$

where $Z^{\lambda,x}$ and $\tilde{Y}_i^{\lambda,x}$, defined like Z^λ and $\tilde{Y}_i^\lambda$, depend on x because the stochastic process (X_t^λ) depends on $X_0^\lambda = x$.

It might be useful to build different reduced bases, one for each cell of a partition of the set of the initial condition. In summary, both algorithms can be extended to SDEs with variable initial condition, at the price of increasing the dimension of the parameter (see also Remark 33).

Remark 35 (Multi-dimensional output)**.** *Clearly, the full output $\boldsymbol{\tau}$ in the problem described above is three-dimensional (it is a symmetric matrix). So our reduced-basis approach such as presented so far would need three different reduced bases, one for each scalar output. Though, one could alternatively consider the construction of only one reduced basis for the three outputs, which may be advantageous, see [Boy08] for one example of such a construction.*

Note that it is difficult to compute accurate approximations of the solution to the backward Kolmogorov equation (7-II.7) in the FENE case, because of the nonlinear explosive term. It is tractable in some situations, see [LC03, CL04] for instance, though at the price of computational difficulties we did not want to deal with in this first work on our new variance reduction approach. On the contrary, the backward Kolmogorov equation (7-II.7) can be solved exactly in the case of Hookean dumbells. Hence we have approximated here u^λ in Algorithm 2 by the numerical solution $\tilde{u}^\lambda$ to the backward Kolmogorov equation (7-II.7) for Hookean dumbells, whatever the type of dumbbells used for the molecular simulation (Hookean or FENE).

We would like to mention the recent work [KP09] where the classical reduced-basis method for parameterized PDEs has been used in the FENE case (solving the FENE Fokker-Planck by dedicated deterministic methods). Our approach is different since we consider a stochastic discretization.

7-IV-B-b Numerical results

The SDE (7-I.2) for FENE dumbbells (when $d=2$) is discretized with the Euler-Maruyama scheme using $N=100$ iterations with a constant time step of $\Delta t=10^{-2}$ starting from a (deterministic) initial condition $\boldsymbol{X}_0=(1,1)$, with reflecting boundary conditions at the boundary of the ball with radius $\sqrt{b}$.

The number of realizations used for the Monte-Carlo evaluations, and the sizes of the (offline) trial sample Λ_{trial} and (online) test sample Λ_{test} for the three-dimensional matrix parameter $\underline{\underline{\lambda}}$ with entries $(\lambda_{11}=-\lambda_{22},\lambda_{12},\lambda_{21})$, are kept similar to the previous Section 7-IV-A. Samples Λ_{trial} and Λ_{test} for the parameter $\underline{\underline{\lambda}}$ are uniformly distributed in a cubic range $\Lambda=[-1,1]^3$. We will also make use of an enlarged (online) test sample $\Lambda_{\text{testwide}}$, uniformly distributed in the range $[-2,2]^3$.

When $b=9$, the variance reduction online with Algorithm 1 is again very interesting, of about 4 orders of magnitude with $I=20$ basis functions, whatever the criterium used for the selection (we only show the absolute variance, in Fig. 7.9). But when $b=4$, the reflecting boundary conditions are more often active, and the maximum online variance reduction slightly degrades (see Fig. 7.10).

We first tested our variance reduction with Algorithm 2 for Hookean dumbells and it appeared to work well ; but such a model is considered too simple generally. Then using the solution to the Kolmogorov backward equation for Hookean dumbells as $\tilde{u}^\lambda$ in Algorithm 2 for FENE dumbbells still yields good variance reduction while the boundary is not touched (see Fig. 7.11) ; when $b=4$ and many reflections at the boundary occur, the variance is hardly reduced. Again Algorithm 2 seems to be slightly more robust than Algorithm 1 in terms of extrapolation, that is when the (online) test sample is "enlarged" (see Fig.7.12 with $b=16$ and a sample test (online) $\Lambda_{\text{testwide}}$).

7-V Conclusion and perspectives

We have demonstrated the feasibility of a reduced-basis approach to compute control variates for the expectation of functionals of a parameterized Itô stochastic process. We have also tested the efficiency of such an approach with two possible algorithms, in two simple test cases where either the drift or the diffusion of scalar $(d=1)$, and vector $(d=2)$, Itô processes are parametrized, using 2- or 3-dimensional parameters.

Algorithm 2 is less generic than Algorithm 1 ; it is basically restricted to low-dimensional stochastic processes (X_t) since :

- one needs to solve (possibly high-dimensional) PDEs (offline), and
- discrete approximations of the PDEs solutions on a grid have to be kept in memory (which is possibly a huge amount of data).

Yet, Algorithm 2 seems more robust to variations in the parameter.

From a theoretical viewpoint, it remains to better understand the convergence of reduced-basis approximations for parametrized control variates depending on the parametrization (and on the dimension of the parameter in particular), on the reduced-basis construction (following a greedy procedure) and on an adequate discretization choice (including the computation of approximate control variates and the choice of a trial sample Λ_{trial}).

Acknowledgement 4. *We thank Claude Le Bris, Yvon Maday and Anthony T. Patera for fruitful discussions.*

7-VI Appendix to the Chapter 7

7-VI-A Appendix A. Algorithm 2 in a higher-dimensional setting ($d\geq 4$)

The solution $u^\lambda(t,y)$ to (7-II.7) can be computed at any $(t,y)\in[0,T]\times\mathbb{R}^d$ by the martingale representation theorem [KS91] :

$$g^\lambda(X_T^\lambda)-\int_t^T f^\lambda(s,X_s^\lambda)ds=u^\lambda(t,X_t^\lambda)+\int_t^T \nabla u^\lambda(s,X_s^\lambda)\cdot\sigma^\lambda(s,X_s^\lambda)dB_s, \tag{7-VI.1}$$

obtained by an Itô formula similar to (7-II.8). This gives the following Feynman-Kac formula for $u^\lambda(t,x)$, which can consequently be computed at any $(t,y)\in[0,T]\times\mathbb{R}^d$ through Monte-Carlo evaluations :

$$u^\lambda(t,y)=\mathbf{E}\left(g^\lambda(X_T^{\lambda,t,y})-\int_t^T f^\lambda(s,X_s^{\lambda,t,y})ds\right), \tag{7-VI.2}$$

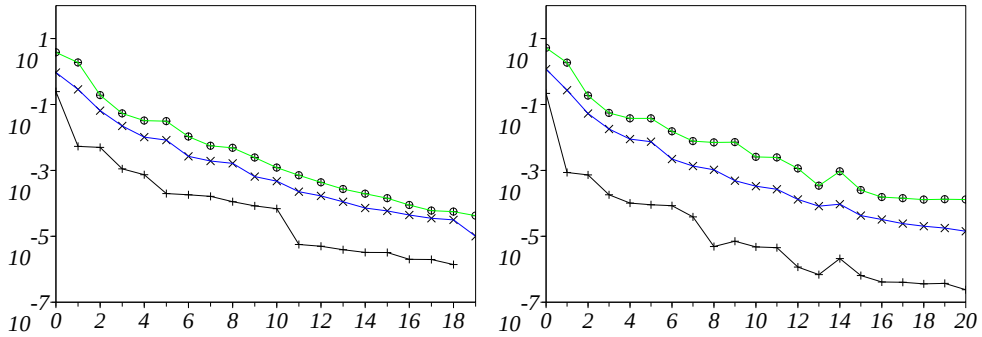


FIG. 7.9 – Algorithm 1 for FENE model with $b=9$: Minimum $+$, mean $\times$ and maximum $\circ$ of the absolute variance (7-IV.8) in samples of parameters (left : offline sample $\Lambda_{\text{trial}} \setminus \{\lambda_i, i=1,\ldots,I\}$; right : online sample Λ_{test}) with respect to the size I of the reduced basis.

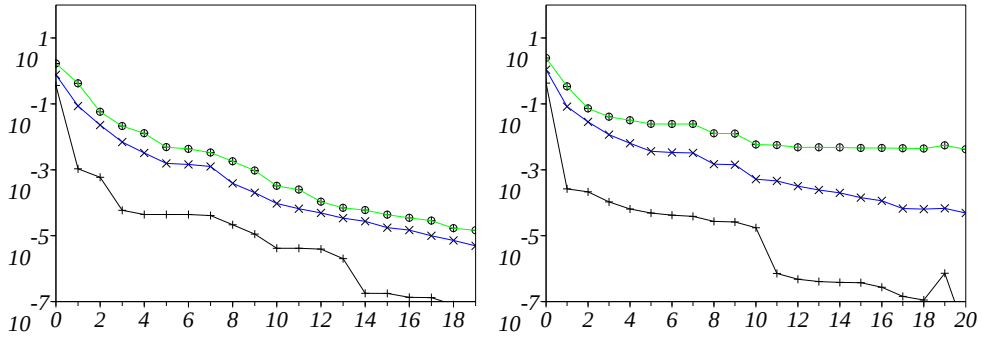


FIG. 7.10 – Algorithm 1 for FENE model with $b=4$: Minimum $+$, mean $\times$ and maximum $\circ$ of the absolute variance (7-IV.8) in samples of parameters (left : offline sample $\Lambda_{\text{trial}} \setminus \{\lambda_i, i=1,\ldots,I\}$; right : online sample Λ_{test}) with respect to the size I of the reduced basis.

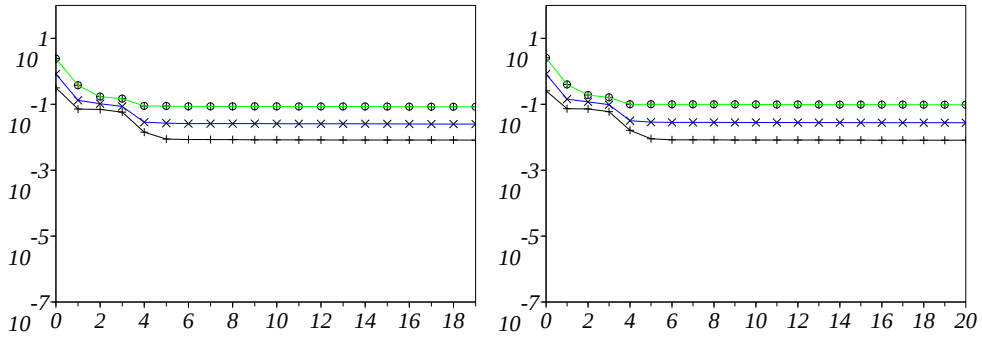


FIG. 7.11 – Algorithm 2 for FENE model with $b=9$: Minimum $+$, mean $\times$ and maximum $\circ$ of the absolute variance (7-IV.8) in samples of parameters (left : offline sample $\Lambda_{\text{trial}} \setminus \{\lambda_i, i=1,\ldots,I\}$; right : online sample Λ_{test}) with respect to the size I of the reduced basis.

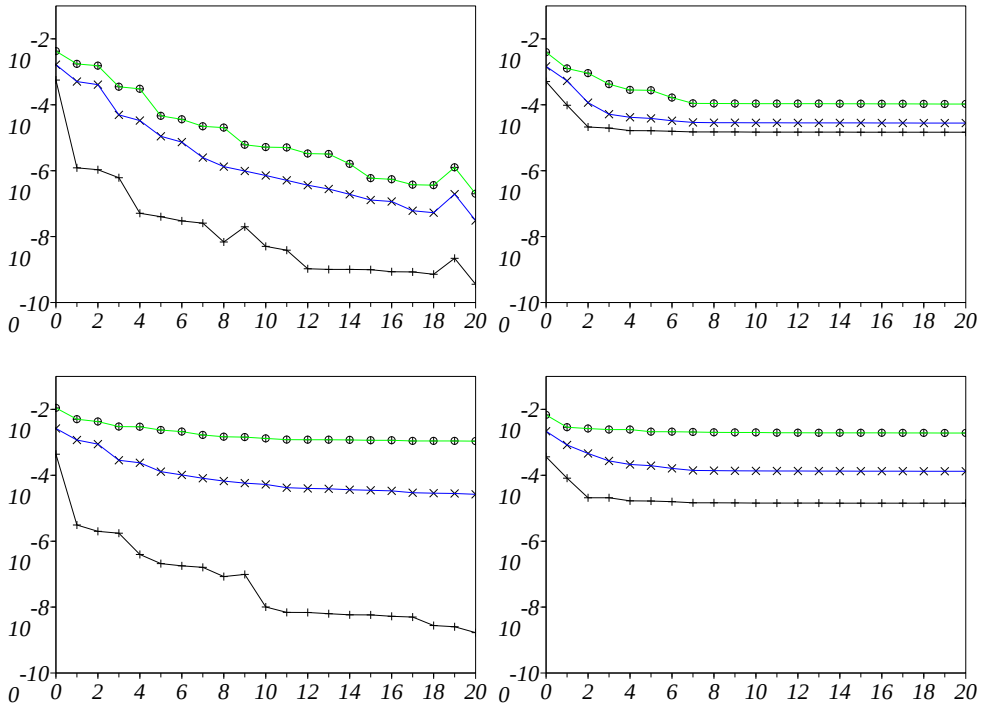


FIG. 7.12 – Algorithm 1 (left) and 2 (right) for FENE model with $b=16$: Minimum $+$, mean $\times$ and maximum $\circ$ of the relative variance (7-IV.9) in online test for samples Λ_{test} (top) and $\Lambda_{\text{testwide}}$ (bottom) of parameters, with respect to the size I of the reduced basis.

where $(X_t^{\lambda,t_0,y})_{t_0\le t\le T}$ is the solution to (7-I.2) with initial condition $X_{t_0}^{\lambda,t_0,y}=y$. Differentiating (7-VI.2) (provided f^λ and g^λ are differentiable), we even directly get a Feynman-Kac formula for $\nabla u^\lambda(t,y)$:

$$\nabla u^\lambda(t,y)=\mathbf{E}\left(\Phi_T^{\lambda,t,y}\cdot\nabla g^\lambda(X_T^{\lambda,t,y})-\int_t^T\Phi_s^{\lambda,t,y}\cdot\nabla f^\lambda(s,X_s^{\lambda,t,y})ds\right) \tag{7-VI.3}$$

where the stochastic processes $\left(\Phi_s^{\lambda,t,y},s\in[t,T]\right)$ in $\mathbb{R}^{d\times d}$ satisfy the first-order variation of the SDE (7-I.2) with respect to the initial condition, that is $\Phi_s^{\lambda,t,y}=\nabla_y X_s^{\lambda,t,y}$ for any $s\in[t,T]$:

$$\Phi_s^{\lambda,t,y}=\boldsymbol{I}_d+\int_t^s\Phi_{s'}^{\lambda,t,y}\cdot\nabla b^\lambda(s',X_{s'}^{\lambda,t,y})ds'+\int_t^s\Phi_{s'}^{\lambda,t,y}\cdot\nabla\sigma^\lambda(s',X_{s'}^{\lambda,t,y})dB_{s'}, \tag{7-VI.4}$$

where $\boldsymbol{I}_d$ denotes the $d\times d$ identity matrix (see [New94] for a more general and rigorous presentation of this Feynman-Kac formula in terms of the Malliavin gradient). The stochastic integral (7-II.9) can then be computed for each realization of (B_t), after discretizing $\left(\Phi_s^{\lambda,t,y},s\in[t,T]\right)$.

Discrete approximations of the Feynman-Kac formula (7-VI.3) have already been used succesfully in the context of computing control variates for the reduction of variance, in [New94] for instance. Note that this numerical strategy to compute $\nabla u^\lambda$from a Feynman-Kac formula requires a lot of computations. Yet, most often, the computation time of the functions $(t,y)\to\nabla u^\lambda(t,y)$ would not be a major issue in a reduced-basis approach, since this would be done *offline* (that is, in a pre-computation step, once for all) for only a few selected values of the parameter λ. What is nevertheless necessary for the reduced-basis approach to work is the possibility to store the big amount of data corresponding to a discretization of $\nabla u^\lambda(t,y)$ on a grid for the variable $(t,y)\in[0,T]\times\mathbb{R}^d$, (the parameter λ then assuming only a few values in Λ — of order 10 in our numerical experiments —), and to have rapid access to those data in the *online* stage (where control variates are computed for any $\lambda\in\Lambda$ using those precomputed data).

7-VI-B Appendix B. Proof of Proposition 29

First note that, since $\mathbf{E}\left(Y^\lambda\right)=0$, then $\mathbf{Var}\left(Z^\lambda\right)=\mathbf{Var}\left(Y^\lambda\right)$.

So, for all $\lambda\in\Lambda$ and for every linear combination of $Y^{\lambda_n^N}$, $n=1,\dots,N$:

$$Y_N^\lambda=\sum_{n=1}^N a_n(\lambda)\sum_{j=1}^J g_j(\lambda_n^N)Y_j$$

(with any choice $a_n(\lambda)\in\mathbb{R}$, $\lambda_n^N\in\Lambda$, $n=1,\dots,N$), there holds (recall that the Y_j, $j=1,\dots,J$, are uncorrelated) :

$$\begin{aligned}\mathbf{Var}\left(Z^\lambda-Y_N^\lambda\right)=\mathbf{Var}\left(Y^\lambda-Y_N^\lambda\right)&=\int_\Omega\left|\sum_{j=1}^J\left(g_j(\lambda)-\sum_{n=1}^N a_n(\lambda)g_j(\lambda_n^N)\right)Y_j\right|^2 d\mathbb{P}\\&\le\left(\sum_{j=1}^J|g_j(\lambda)|^2\,\mathbf{Var}\left(Y_j\right)\right)\sup_{1\le j\le J}\frac{\left|g_j(\lambda)-\sum_{n=1}^N a_n(\lambda)g_j(\lambda_n^N)\right|^2}{|g_j(\lambda)|^2}.\end{aligned}\tag{7-VI.5}$$

To get (7-III.15), we now explain how to choose the N coefficients $a_n(\lambda)$, $1\le n\le N$, for each $\lambda\in\Lambda$ when $\lambda_n^N\in\Lambda$, $n=1,\dots,N$ is given, and then how to choose those N parameter values $\lambda_n^N\in\Lambda$, $n=1,\dots,N$.

Assume the N parameter values $\lambda_n^N\in\Lambda$, $n=1,\dots,N$, are given, with $\lambda_1^N=\lambda_{\min}$, $\lambda_N^N=\lambda_{\max}$ and $\lambda_n^N\le\lambda_{n+1}^N$, $n=1,\dots,N-1$. Then, for a given $M\in\{2,\dots,N\}$ (to be determined later on) and for all $\lambda\in\Lambda$, it is possible to choose $1\le M_0(\lambda)\le N+1-M$ such that $\lambda_{M_0(\lambda)}^N\le\lambda\le\lambda_{M_0(\lambda)+M-1}^N$. Only the M coefficients corresponding to the M contiguous parameter values above are taken non zero, such that $\forall\lambda\in\Lambda$:

$$a_m(\lambda)\neq0\Leftrightarrow M_0(\lambda)\le m\le M_0(\lambda)+M-1,$$

and are more specifically chosen as $a_m(\lambda)=P_m^\lambda(\tau_\Lambda(\lambda))$ where P_m^λ are polynomials of degree $M-1$, such that, for all $M_0(\lambda)\le m,k\le M_0(\lambda)+M-1$, $P_m^\lambda(\tau_\Lambda(\lambda_k))=\delta_{mk}$. The polynomial function P_m^λ is the Lagrange interpolant defined on $[\tau_\Lambda(\lambda_{M_0(\lambda)}),\tau_\Lambda(\lambda_{M_0(\lambda)+M-1})]$, taking value 1 at $\tau_\Lambda(\lambda_m)$ and 0 at $\tau_\Lambda(\lambda_k)$, $k\neq m$. We will also need a function $d(\lambda)=|\tau(\lambda_{M_0(\lambda)})-\tau(\lambda_{M_0(\lambda)+M-1})|$. Using a Taylor-Lagrange formula for $g_j\circ\tau^{-1}$, we have (for some $0\le\eta\le1$) :

$$g_j(\lambda)-\sum_{n=1}^N a_n(\lambda)g_j(\lambda_n)=\frac{d(\lambda)^M}{M!}(g_j\circ\tau^{-1})^{(M)}\left(\eta\tau(\lambda_{M_0(\lambda)}^N)+(1-\eta)\tau(\lambda_{M_0(\lambda)+M-1}^N)\right).$$

Then, using (7-III.14) and the fact that $\mathbf{Var}\left(Z^\lambda\right)=\sum_{j=1}^J|g_j(\lambda)|^2\,\mathbf{Var}\left(Y_j\right)$, there exists a constant $C>0$ (independent of Λ and J) such that :

$$\mathbf{Var}\left(Z^\lambda-Y_N^\lambda\right)\le\mathbf{Var}\left(Z^\lambda\right)\left(C\,d(\lambda)\right)^{2M},\ \forall\lambda\in\Lambda.\tag{7-VI.6}$$

Finally, to get the result, we now choose a τ_Λ-equidistributed parameter sample :

$$\tau_\Lambda(\lambda_n^N)=\tau_\Lambda(\lambda_{\min})+\frac{n-1}{N-1}\left(\tau_\Lambda(\lambda_{\max})-\tau_\Lambda(\lambda_{\min})\right),\ n=1,\dots,N.$$

Then, $d(\lambda)=\frac{M-1}{N-1}\left(\tau_\Lambda(\lambda_{\max})-\tau_\Lambda(\lambda_{\min})\right)$ does not depend on λ. Minimizing $(Cd)^d$ as a function of $d\in(0,\frac{1}{C})$, we choose $d(\lambda)=\frac{1}{eC}$, and the choice $M=1+\lfloor\frac{1}{eC}\frac{N-1}{\tau_\Lambda(\lambda_{\max})-\tau_\Lambda(\lambda_{\min})}\rfloor$ (where $\lfloor x\rfloor$ denotes the integer part of a real number $x\in\mathbb{R}$) finishes the proof provided $N\ge N_0\equiv1+\lfloor Ce\left(\tau_\Lambda(\lambda_{\max})-\tau_\Lambda(\lambda_{\min})\right)\rfloor$. □

7-VII Addendum to chapter 7

We present here some unpublished results about a RB approach of the Fokker-Planck equation for Hookean dumbbells (1-IV.3), starting with a simple non-coercive parabolic equation as a toy model.

As mentionned in the previous sections of this chapter, such an approach may be an alternative to our variance-reduced Monte-Carlo method for the fast computation of functionals of a parametrized stochastic process (for many values of the parameter). Though, in its present formulation, the RB approach of the Fokker-Planck equation seems adequate only for the low-dimensional problems (like dumbbells rather than long Rouse chains of beads, in the field of micro-macro models for non-Newtonian fluids).

Note that the results presented here have been partly improved elsewhere by other authors since they were obtained in 2006. On the one hand, one should now take into account some improvements of the RB greedy algorithm for parabolic problems, following ideas first introduced in [HO08] for instance. Moreover, on the other hand, the problem of real-time simulations using the Fokker-Planck equation for dumbbells, for which we try here to develop a general RB approach, has been treated recently for the specific computation of the optical anisotropy of FENE-dumbbell polymer liquids under extensional flows, using dedicated estimates [KP09].

7-VII-A RB approach of a non-coercive parabolic equation as a model for a Fokker-Planck equation

7-VII-A-a Mathematical setting of the problem

As a toy-model for the equation (1-IV.3), we study a simple Cauchy-Dirichlet problem for $\phi(t,x)$, $(t,x)\in[0,T)\times(0,1)$, solution to a one-dimensional convection-diffusion equation

$$\partial_t\phi(t,x)=-\sigma(t)\partial_x(g(x)\phi(t,x))+\partial_x\left(\beta(x)\partial_x\phi(t,x)\right),\quad \phi(t=0,x)=\phi_0(x),\ \forall(t,x)\in[0,T)\times(0,1) \tag{7-VII.1}$$

supplied with homogeneous boundary conditions $\phi(t,0)=\phi(t,1)=0$. The initial condition ϕ_0 is chosen centered at $\frac{1}{2}$, and such that $\int_\Omega\phi_0(x)dx=1$. This is a parabolic problem parameterized by $\mu(t,x)=(\sigma(t),g(x),\beta(x))$, and we first tried to apply a RB approach similar to that in [GP05]. Note that, when β is constant and $g(x)=H(x-\frac{1}{2})$ with some constant H, (7-VII.1) has the form of the Fokker-Planck equation for a (truncated) Hookean dumbbell model in a given Couette flow with shear $\sigma(t)$, where finite-extensibility is enforced by truncating the domain of definition of the dumbbell polymers (or changing the definition of $\beta(x)$ and $g(x)$, this is simply the usual Hookean dumbbell model when $\mathbb{R}$ is mapped, for instance by the function $\mathrm{ath}(2x-1)$ where $x\in]0,1[$). We next define parametrized bilinear forms as, $\forall(u,v)\in H_0^1(\Omega)\times H_0^1(\Omega)$:

$$m(u,v)=\int_\Omega uv \quad\text{and}\quad P_{\mu(t)}(u,v)=\sigma(t)C_g(u,v)+D_\beta(u,v),\ \text{where}$$

$$D_\beta(u,v)=\int_\Omega\beta(x)\partial_x u\partial_x v \quad\text{and}\quad C_g(u,v)=C_g^a(u,v)+C_g^s(u,v)=\frac{1}{2}\int_\Omega g(x)\big(v\partial_x u-u\partial_x v\big)+\frac{1}{2}\int_\Omega uv\partial_x g(x).$$

A solution $\phi(t,\cdot)\in H_0^1(\Omega)$ to (7-VII.1) with time-derivative $\partial_t\phi(t,\cdot)\in H_0^1(\Omega)$ satisfies (al least in some distributional sense) for all $t\in(0,T)$:

$$m(\partial_t\phi,v)+P_{\mu(t)}(\phi,v)=m(f,v),\ \forall v\in H_0^1(\Omega). \tag{7-VII.2}$$

The configuration set Ω is a Lipschitz open bounded domain, a Poincaré-Friedrichs inequality then holds for some constant $\mathcal{P}=\mathcal{P}(\Omega)>0$ depending only on Ω :

$$\mathcal{P}\int_\Omega\phi^2\le\int_\Omega(\partial_x\phi)^2,\ \forall\phi\in H_0^1(\Omega). \tag{7-VII.3}$$

For any $t\in]0,T[$, it is natural to assume

$$\beta\ge\inf_{x\in\Omega}\beta>0, \tag{7-VII.4}$$

then the bilinear form $P_{\mu(t)}$ satisfies the following Gärding inequality with any ρ such that $0<\rho<\inf_{x\in\Omega}\beta$:

$$\eta_P(\rho)\int_\Omega\phi^2+\rho\int_\Omega(\partial_x\phi)^2\le P_{\mu(t)}(\phi,\phi)\ ,\ \forall\phi\in H_0^1(\Omega)\ , \tag{7-VII.5}$$

where we have used a Gärding constant $\eta_P(\rho) \leq \sigma(t) \inf_{x\in\Omega} \partial_x g + (\inf_{x\in\Omega} \beta - \rho)\mathcal{P}$. Then, assuming

$$\partial_x g \geq \inf_{x\in\Omega} \partial_x g > 0 \ , \tag{7-VII.6}$$

we have $\eta_P(\rho) > 0$ provided

$$\sigma(t) > -\frac{\left(\inf\limits_{x\in\Omega} \beta - \rho\right)\mathcal{P}}{\inf\limits_{x\in\Omega} \partial_x g} > -\frac{\inf\limits_{x\in\Omega} \beta}{\inf\limits_{x\in\Omega} \partial_x g}\mathcal{P}, \tag{7-VII.7}$$

which ensures well-posedness. Though, the variational problem (7-VII.2) may be non-coercive in the usual energy-norm.

7-VII-A-b Discretization and RB treatment

Time is discretized with a Euler backward scheme using a constant time step $\Delta t > 0$. The time range is split into $K = T/\Delta t \in \mathbb{N}^\star$ steps of equal length. We write $t_k = k\Delta t$, $1 \leq k \leq K$, the sequence of times when $\phi(t,x)$ is approximated. For every $1 \leq k \leq K$, the approximations $\phi^k(x)$ of $\phi(t_k,x)$ are solutions to :

$$\begin{cases} \forall k \in \mathbb{N}, 0 \leq k \leq K, \text{ find } \phi^k \in H_0^1(\Omega) \text{ such that} \\ m(\phi^{k+1},v) + \Delta t P_{\mu_k}(\phi^{k+1},v) = m(\phi^k,v), \ \forall v(x) \in H_0^1(\Omega), \end{cases} \tag{7-VII.8}$$

where the sequence $(\mu(t_k))_k$ is the corresponding discretization of the time-dependent parameter $\mu(t)$. As a matter of fact, our RB approach will consider the sequence of parameters $\tilde{\mu}_k = (\mu(t_k), t_k)$. We take $\phi^0(x) = \phi_0(x)$.

We compute Galerkin approximations of $\phi^k \in H_0^1(\Omega)$ solution to (7-VII.8) using the same finite-dimensional linear subspace $W \subset H_0^1(\Omega)$ for all $1 \leq k \leq K$. We still write ϕ^k those (supposedly sufficiently fine) approximations in W, W typically being a linear space of high dimension. We would then like to build RB approximations $\phi_N^k \in W_N$ of ϕ^k, with W_N a small N-dimensional subspace of W built as the span of well-chosen snapshots ϕ^k for well selected parameters $\tilde{\mu}_k$. As usual in RB techniques, we need good *a posteriori* estimators.

Let us define $R_N^k(\cdot;\tilde{\mu}_k)$ as a residual linear form in $H_0^1(\Omega)$:

$$R_N^k(v;\tilde{\mu}_k) = -\frac{m(\phi_N^k - \phi_N^{k-1}, v)}{\Delta t} - P_{\mu_k}(\phi_N^k, v), \ \forall v \in H_0^1(\Omega).$$

We call residual error the usual dual norm in $H^{-1}(\Omega)$ of this continuous linear form on $H_0^1(\Omega)$:

$$\epsilon_N^k(\tilde{\mu}_k) = \sup_{v\in H_0^1(\Omega)} \frac{R_N^k(v;\tilde{\mu}_k)}{\|v\|_{H_0^1(\Omega)}} \ .$$

The RB approximation error $e_N^k = \phi^k - \phi_N^k$ satisfies

$$m(e_N^{k+1},v) + \Delta t P_{\mu_k}(e_N^{k+1},v) = m(e_N^k,v) + \Delta t R_N(v;\tilde{\mu}_{k+1}), \ \forall v \in H_0^1(\Omega). \tag{7-VII.9}$$

Then, using $v = e_N^{k+1}$ and any couple of non-zero real numbers (ρ_1,ρ_2), on noting the continuity of the linear form $R_N^{k+1}(\cdot;\tilde{\mu}_{k+1})$ in $H_0^1(\Omega)$, we have :

$$R_N^{k+1}(e_N^{k+1};\tilde{\mu}_{k+1}) \leq \epsilon_N^{k+1}(\tilde{\mu}) \|e_N^{k+1}\|_{H_0^1(\Omega)} \tag{7-VII.10}$$

$$\leq \frac{1}{2\rho_1^2}\epsilon_N^{k+1}(\mu)^2 + \frac{\rho_1^2}{2}\left(m(e_N^{k+1},e_N^{k+1}) + D_1(e_N^{k+1},e_N^{k+1})\right) \tag{7-VII.11}$$

and the Cauchy-Schwarz and Young inequalities yield :

$$m(e_N^k,e_N^{k+1}) \leq m(e_N^k,e_N^k)^{1/2} m(e_N^{k+1},e_N^{k+1})^{1/2} \tag{7-VII.12}$$

$$\leq \frac{1}{2\rho_2^2} m(e_N^k,e_N^k) + \frac{\rho_2^2}{2} m(e_N^{k+1},e_N^{k+1}) \ . \tag{7-VII.13}$$

So, thanks to the Gärding inequality (7-VII.5), we finally get the error estimate :

$$\left(1 - \frac{\rho_2^2}{2} + \eta_P\left(\tilde{\mu}_k, \frac{\rho_1^2\left(\mathcal{P}^{-1}+1\right)}{2}\right)\Delta t\right)\|e_N^{k+1}\|_{L^2(\Omega)}^2 \leq \frac{1}{2\rho_2^2}\|e_N^k\|_{L^2(\Omega)}^2 + \frac{\Delta t}{2\rho_1^2}\epsilon_N^{k+1}(\tilde{\mu})^2 \ . \tag{7-VII.14}$$

A discrete version of Gronwall's lemma finally gives the following estimation :

$$\boxed{\|e_N^k\|^2_{L^2(\Omega)} \leq \sum_{i=1}^{k} \frac{\Delta t}{\rho_1^2 \rho_2^{2(k-i)} \left(2-\rho_2^2+2\eta_P\left(\frac{\rho_1^2(\mathcal{P}^{-1}+1)}{2}\right)\Delta t\right)^{k-i+1}} \epsilon_N^i(\tilde{\mu}_i)^2} \tag{7-VII.15}$$

provided there exists $\rho_1>0$ such that $1+\eta_P\left(\frac{\rho_1^2(\mathcal{P}^{-1}+1)}{2}\right)\Delta t>0$. (Note that this latter requirement – for (7-VII.15) to be true with some given $\Delta t>0$ – is less stringent than the Gärding assumption (7-VII.7), and it can always be satisfied as soon as $\Delta t>0$ is small enough : $\frac{-1}{\Delta t} \leq \sigma(t)\inf_{x\in\Omega}\partial_x g + \left(\inf_{x\in\Omega}\beta - \frac{\rho_1^2(\mathcal{P}^{-1}+1)}{2}\right)\mathcal{P}$.) We will next use the error estimate (7-VII.15) for a RB treatment of (7-VII.8) parametrized by $\tilde{\mu}$. The standard RB greedy algorithm is slightly modified in order to take into account the fact that the error is accumulated along time. In practice, we use the RB greedy algorithm proposed in [GP05].

Note that error estimations for parabolic equations were first developped in [Gre05] for the coercive case with an energy-norm. Using the continuity of $R_N^{k+1}(v;\mu)$ and the *coercivity* of $P_{\tilde{\mu}_k}$,

$$\alpha_P(\mu)\|v\|^2_{H_0^1(\Omega)} \leq P_{\tilde{\mu}_k}(v,v),$$

one gets there, for $0\leq k\leq K-1$ and $\rho\in\mathbb{R}$:

$$\|e_N^{k+1}\|^2_{L^2(\Omega)} - \|e_N^k\|^2_{L^2(\Omega)} + \Delta t\left(2-\frac{1}{\alpha_P(\mu)\rho^2}\right)P_{\tilde{\mu}_k}(e_N^{k+1},e_N^{k+1}) \leq \Delta t\rho^2\epsilon_N^{k+1}(\mu)^2,$$

The previous inequalities with $\alpha_P(\mu)\rho^2=1$ can then be summed to obtain the following estimation in energy-norm, $\forall k, 1\leq k\leq K$:

$$|||e_N(t_k)|||^2 = \|e_N^k\|^2_{L^2(\Omega)} + \Delta t\sum_{j=1}^{k} P^{j-1}_{\tilde{\mu}_k}(e_N^j,e_N^j) \leq \frac{\Delta t}{\alpha_P(\mu)}\sum_{i=1}^{k}\epsilon_N^i(\mu)^2 \ . \tag{7-VII.16}$$

So, our estimator (7-VII.15) is an extension of (7-VII.16), which is only valid in the coercive case. (By the way, in the coercive case, our L^2-norm error estimations could also be derived similarly as above when choosing $2\alpha_P(\mu)\rho^2=1$, thus :

$$\|e_N^k\|^2_{L^2(\Omega)} \leq \frac{\Delta t}{2\alpha_P(\mu)}\sum_{i=1}^{k}\epsilon_N^i(\mu)^2, \tag{7-VII.17}$$

which is the same as injecting $(\rho_1^2,\rho_2^2)=(2\alpha_P,1)$ in (7-VII.14).) See also [HO08] for a generalization of [GP05] (in the frame of finite volume discretizations).

7-VII-A-c Numerical results

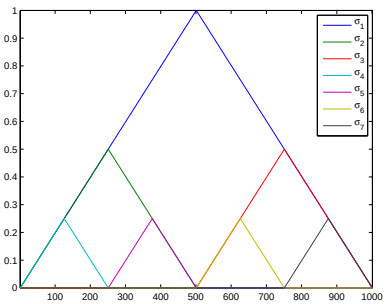


FIG. 7.13 – The hierarchical functions used to parametrize σ in (7-VII.18).

We compute a reduced basis for the collection of Galerkin approximations of the solutions to (7-VII.8) using regular $\mathbb{P}_1$ finite elements, when the parameter $\tilde{\mu}$ has the following prescribed variations : $\Delta t=10^{-3}$, $g(x)=x-\frac{1}{2}$

and $\beta = 10^{-3}$ are fixed, while

$$\sigma(t_k) = \mu_{\sigma_l}\sigma_l(t_k) + \sum_{i=1}^{7} \mu_{\sigma_i}\sigma_i(t_k) + \mu_{\sigma_r}\sigma_r(t_k) \ , \forall 0 \leq k \leq N \tag{7-VII.18}$$

describes the 9-dimensional range spanned by $\sigma_l(t_k) = 1 - t_k$, $\sigma_r(t_k) = t_k$ and the hierarchical hat functions σ_i (see Figure 7.13) :

$$\sigma_1(t) = 2t\mathbf{1}_{[0;.5]} + (1-2t)\mathbf{1}_{[.5;1]}\,,$$
$$\sigma_2(t) = 2t\mathbf{1}_{[0;.25]} + (.5-2t)\mathbf{1}_{[.25;.5]} \ \sigma_3(t) = (2t-1)\mathbf{1}_{[.5;.75]} + (1.5-2t)\mathbf{1}_{[.75;1]} \ldots$$

when $\mu_\sigma = (\mu_{\sigma_l}; \mu_{\sigma_i, i=1\cdots 7}; \mu_{\sigma_r}) \in [0;1]^9$. A training sample of 10^2 parameter values is chosen randomly in $[0;1]^9$.

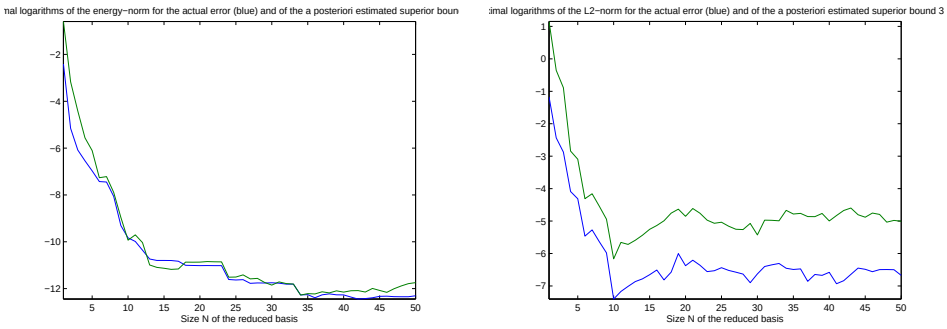


FIG. 7.14 – Maxima of the RB approximation error (blue) and the *a posteriori* estimation (green) observed offline during the greedy selection procedure within the training sample, using either the energy-norm estilmator (left) or the L^2-norm (right), with respect to the size of the reduced basis.

In this non-coercive case, the numerical results show that our greedy selection procedure yields some reduction, though much less than in the coercive case. In fact, using blindly the energy estimator shows that we miss some reduction possibilities (see Fig. 7.14), although the energy estimator (a H^1 norm) is not rigorous in the non-coercive case (contrary to our L^2 error estimator). Similar conclusions have been drawn in [HO08] (for the RB method applied to Finite-Volume schemes), though the use of a slightly different greedy selection (first proposed there) might slightly change this (somewhat unsatisfactory) conclusion. (See also [NRHP09] about the new selection procedure for parabolic problems : the time is nolonger merged with μ into a generalized parameter $\tilde{\mu}$, the standard greedy is actually used only to select iteratively μ, based on the error estimation at final time of the simulation, and at each step of the greedy, the reduced basis is enriched with the first POD-modes of the family of snapshots-in-time computed for the selected μ.)

Yet, our result for a toy-model is surely an indication that it will be difficult in a RB approach of the Fokker-Planck equation (1-IV.3) for dumbbells to control the (nonlinear) upper-convective term. (In fact, this is the same difficulty than for the existence theory, where an estimate misses to control that term.)

7-VII-B RB approach of the Fokker-Planck equation for dumbbells

7-VII-B-a Mathematical setting of the problem

We focus on the efficient computation of Galerkin approximations for the solutions to (variational formulations of) (1-IV.3) in Couette flows.

In a Couette flow, the velocity reads $\boldsymbol{u} = u(t,y)\mathbf{e_x}$, where $\mathbf{e_x}$ is a unit vector in the direction of shear and y the coordinate in the transverse direction. A natural cartesian frame of coordinates for the subset Ω of $\mathbb{R}^2$ is thus $(\mathbf{e_x}, \mathbf{e_y})$, and the spatial domain then reduces to $\Omega = \{\boldsymbol{x} = (x,y) | x = 0, y \in]0,1[\}$ (a one-dimensional range transverse to the shearing plates, respectively located at $y = 0$ and $y = 1$). Next, we specify our Couette flows using the boundary conditions $u(t,0) = V(t)$ and $u(t,1) = 0$ for the velocity, plus the initial condition $u(0,y) \equiv u_{t_0} \in \mathbb{R}$ (start up from a steady flow $\boldsymbol{\nabla u} = 0$).

The configuration space $\mathcal{D}$ for dumbbells is also some subset of $\mathbb{R}^2$. Given another cartesian frame of coordinates, the dumbbell configuration and the entropic force will then be denoted as vectors of two coordinates

$\boldsymbol{X}=(P,Q)$ and $\boldsymbol{F}=(F_P,F_Q)$. Our output for Couette flows, the extra-diagonal term $\tau\mathbf{e}_{\mathbf{x}}\times\mathbf{e}_{\mathbf{y}}$ of the extra-stress, thus reads :

$$\tau(t,y)=\frac{\varepsilon}{\mathrm{Wi}}\int_{\mathcal{D}}PF_Q(P,Q)\psi(t,y,P,Q)dPdQ=\frac{\varepsilon}{\mathrm{Wi}}\int_{\mathcal{D}}QF_P(P,Q)\psi(t,y,P,Q)dPdQ.$$

The entropic force $\boldsymbol{F}(\boldsymbol{X})=\nabla U(\boldsymbol{X})$ derives from a potential energy function $U(\boldsymbol{X})$ linked with the distribution of configurations $\boldsymbol{X}$ for dumbbells at thermodynamic equlibrium. We will consider $U(\boldsymbol{X})=H\boldsymbol{X}^2/2$ for Hookean dumbbells, with a view to extending next the results to $U(\boldsymbol{X})=-(Hb/2)\ln(1-|\boldsymbol{X}|^2/b)$ for FENE dumbbells. (Note that the potential energy U is a positive convex function in both cases, for $\boldsymbol{X}\in\mathcal{D}$, either with $\mathcal{D}=\mathbb{R}^2$ in the Hookean dumbbell case or with $\mathcal{D}=B(0,\sqrt{b})$ in the FENE case.) So we naturally supply (1-IV.3) with the initial condition $\psi(0,\boldsymbol{x},\boldsymbol{X})=\psi_0(\boldsymbol{x},\boldsymbol{X})$, where $\psi_0(\boldsymbol{x},\boldsymbol{X})$ is the density of the thermodynamically equilibrated state :

$$0=-\operatorname{div}_{\boldsymbol{X}}(-\boldsymbol{F}(\boldsymbol{X})\psi_0)+\Delta_{\boldsymbol{X}}\psi_0=\operatorname{div}_{\boldsymbol{X}}(\boldsymbol{F}(\boldsymbol{X})\psi_0+\nabla_{\boldsymbol{X}}\psi_0),$$

which implies $\psi_0(\boldsymbol{x},\boldsymbol{X})\propto e^{-U(\boldsymbol{X})}$ for all $\boldsymbol{x}\in\Omega$.

Note that the initial condition above for Hookean dumbbells is Gaussian, which then implies $\psi(t,\boldsymbol{x},\boldsymbol{X})$ remains a Gaussian density for all $(t,\boldsymbol{x})\in[0,T[\times\Omega$. It is thus interesting to rewrite the Fokker-Planck equation (1-IV.3) for the variable $\Psi=\frac{\psi}{\psi_0}$ instead of ψ. In particular, on noting $\boldsymbol{F}(\boldsymbol{X})=-\frac{\nabla_{\boldsymbol{X}}\psi_0}{\psi_0}$, the term $-\operatorname{div}_{\boldsymbol{X}}(-\boldsymbol{F}(\boldsymbol{X})\psi)+\Delta_{\boldsymbol{X}}\psi$ rewrites as a purely dissipative term :

$$\frac{\partial\Psi}{\partial t}\psi_0=\operatorname{div}_{\boldsymbol{X}}\left(\left[(\boldsymbol{\nabla u})\boldsymbol{X}\Psi-\frac{1}{2\mathrm{Wi}}\nabla_{\boldsymbol{X}}\Psi\right]\psi_0\right),\tag{7-VII.19}$$

and the long-time equilibrium probability density for the flows with stationary state $\boldsymbol{\nabla u}=0$ is clearly given by the following invariant measure [JLBLO06] :

$$d\mathbf{w}:=\psi_0(\boldsymbol{x},\boldsymbol{X})d\boldsymbol{X}=\frac{e^{-U(\boldsymbol{X})}d\boldsymbol{X}}{\int_{\mathcal{D}}e^{-U(\boldsymbol{X})d\boldsymbol{X}}}.$$

We define new spaces of integrable functions :

Definition 3 (Hilbert spaces)**.** *The following spaces are two separable Hilbert spaces :*

$$W_v=H^1_0(\Omega),\qquad W_\Phi=H^1_{d\mathbf{w}}(\mathcal{D})=\overline{C^\infty_c(\overline{\mathcal{D}})}^{|||\cdot|||_m},\tag{7-VII.20}$$

when, for any $m\in\mathbb{R}^+_\star$, they are respectively endowed with the inner products :

$$[u;v]=\int_\Omega\nabla u\cdot\nabla v,\ \forall u,v\in W_v,\quad(\Psi;\Phi)_m=\int_{\mathcal{D}}\left(m^2\Psi\Phi+\nabla\Psi\cdot\nabla\Phi\right)d\mathbf{w},\ \forall\Phi,\Psi\in W_\Phi,\tag{7-VII.21}$$

with Hilbertian norms :

$$\|u\|=[u;u]^{\frac12},\ \forall u,v\in W_v,\quad|||\Psi|||_m=(\Psi,\Psi)_m^{\frac12},\ \forall\Phi,\Psi\in W_\Phi.\tag{7-VII.22}$$

(Note also that for all $\Psi\in L^2_{d\mathbf{w}}(\mathcal{D})$, where square-integrability is with respect to the Borel measure $d\mathbf{w}=\psi_0(\boldsymbol{X})d\boldsymbol{X}$ for $\boldsymbol{X}\in\mathcal{D}$, we have $\Psi\sqrt{\psi_0}\in L^2(\mathcal{D})$, which implies $\psi=\Psi\psi_0\in L^1(\mathcal{D})$, on account of the Cauchy-Schwarz inequality and $\int_{\mathcal{D}}\psi_0=1$:

$$\int_{\mathcal{D}}|\psi|=\int_{\mathcal{D}}\left(|\Psi|\sqrt{\psi_0}\right)\sqrt{\psi_0}\le\|\Psi\|_{L^2_{d\mathbf{w}}(\mathcal{D})},$$

the latter being of course necessary for $\psi=\Psi\psi_0$ to remain a probability density function.)

Let us finally denote $W_u(V(t),0)=\left\{v\in L^2([0,T),H^1(]0,1[),v(0)=V(t),v(1)=0\right\}$ the space of velocities (which is not a vector space). In the following, we consider weak solutions $(u,\Psi)\in W_u(V(t),0)\times W_\Phi$ to the Cauchy-Dirichlet problem in the domain $(t,y,\boldsymbol{X})\in[0,T[\times\Omega\times\mathcal{D}$ such that :

1. $u(0,y)\equiv u_{t_0}\in\mathbb{R}$, in some sense, for all $(y,\boldsymbol{X})\in]0,1[\times\mathcal{D}$
2. $\psi(0,y,\boldsymbol{X})=\psi_0(y,\boldsymbol{X})$, in some sense, for all $(y,\boldsymbol{X})\in]0,1[\times\mathcal{D}$
3. and for all $t\in[0,T[$, we have in some sense $u(t,0)=V(t)$, $u(t,1)=0$ and

$$\left\{\begin{array}{c} \mathrm{Re}\int_\Omega \frac{\partial u}{\partial t}v = -(1-\varepsilon)\int_\Omega \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} - \int_\Omega \tau \frac{\partial v}{\partial y},\ \forall v \in W_v, \\ \tau = \frac{\varepsilon}{2\mathrm{Wi}}\int_{\mathcal{D}} \Psi \boldsymbol{X}^\dagger \cdot \nabla_{\boldsymbol{X}} \psi_0, \\ \int_{\mathcal{D}} \frac{\partial \Psi}{\partial t}\Phi d\mathbf{w} = \int_{\mathcal{D}} \Psi \frac{\partial u}{\partial y} Q \frac{\partial \Phi}{\partial P} d\mathbf{w} - \frac{1}{2\mathrm{Wi}}\int_{\mathcal{D}} \nabla_{\boldsymbol{X}}\Psi \cdot \nabla_{\boldsymbol{X}}\Phi d\mathbf{w},\ \forall \Phi \in W_\Phi, \end{array}\right. \tag{7-VII.23}$$

where we have used the notation $\boldsymbol{X}^\dagger = (Q,P)$ in (7-VII.23).

The solutions of (7-VII.23) always satisfy the normalization condition $\int_{\mathcal{D}}\psi \equiv 1$, since $\Phi \equiv 1 \in W_\Phi$ implies

$$\frac{d}{dt}\int_{\mathcal{D}} \Psi d\mathbf{w} = \int_{\mathcal{D}} \frac{\partial}{\partial t}\Psi d\mathbf{w} = 0.$$

Moreover, since "$(\boldsymbol{u}, \Psi(P,Q))$ solution of (7-VII.23)" $\Rightarrow$ "$(\boldsymbol{u}, \Psi(-P,-Q))$ solution of (7-VII.23)", the following symmetry property is conserved for all time $t \in [0,T[$, $\forall (\boldsymbol{x},\boldsymbol{X}) \in \Omega \times \mathcal{D}$:

$$\Psi(t=0,\boldsymbol{x},P,Q) = \Psi(t=0,\boldsymbol{x},-P,-Q). \tag{7-VII.24}$$

Last, every classical solution $(u,\Psi) \in C^1([0,T[,C^2(\Omega)) \times C^1([0,T[,C^2(\Omega \times \mathcal{D}))$ is clearly a weak solution.

Remark 36. *In the present study, the advection term in* (1-IV.3) *disappears because of the special geometrical configuration of a Couette flow. Yet, in more general geometries than a Couette flow, this term remains and may cause difficulties, because it is non-linear. Fortunately, in a first approach, we can always assume that this advection term can be handled with a characteristic method if the velocity field is sufficiently regular, and the Couette flow is still a good testcase then.*

Before proceeding to the numerical approximation of solutions to the system (7-VII.23), in particular RB approximations for the probability density functional of Hookean dumbbells, solution to (1-IV.3), let us try to grasp the mathematical nature of the problem. One term couples the Fokker-Planck equation with the macroscopic state of the fluid, given by u solution to a Navier-Stokes equation, this is $\Psi(\boldsymbol{\nabla u X})^T \nabla \Phi$, the sign of which we do not know in advance (since it depends on the coupling). This term is likely to cause difficulties in the analysis of the coupled system (7-VII.23). Let us first see how to handle this term in the uncoupled Fokker-Planck equation.

For hookean dumbbells in a Couette flow, $(\boldsymbol{\nabla u X})^T \nabla$ reduces to $\frac{\partial u}{\partial y} Q \frac{\partial}{\partial P}$, and this terms reads :

$$\int_{\mathcal{D}} \Psi(\boldsymbol{\nabla u X})^T \nabla \Phi d\mathbf{w} = \frac{1}{2}\frac{\partial u}{\partial y}\int_{\mathcal{D}} Q\left(\Psi \frac{\partial \Phi}{\partial P} - \Phi \frac{\partial \Psi}{\partial P}\right) d\mathbf{w} + \frac{1}{2}\frac{\partial u}{\partial y}\int_{\mathcal{D}} PQ\Psi\Phi d\mathbf{w}. \tag{7-VII.25}$$

(Note that this coupling term is exactly equal to the output τ when $\Phi \equiv 1$.)

Let us bound this coupling term (independently of the flow configuration) in the Fokker Planck equation :

$$\int_{\mathcal{D}} \frac{\partial \Psi}{\partial t}\Phi d\mathbf{w} = \int_{\mathcal{D}} \Psi(\boldsymbol{\nabla u X})^T \nabla \Phi d\mathbf{w} - \frac{1}{2\mathrm{Wi}}\int_{\mathcal{D}} \nabla_{\boldsymbol{X}}\Psi \cdot \nabla_{\boldsymbol{X}}\Phi d\mathbf{w}, \forall \Phi \in W_\Phi. \tag{7-VII.26}$$

For the sake of simplicity we take $H = 1$ in the following.

Proposition 30 (Finite-time weak solution to uncoupled Fokker-Planck)**.** *If $\boldsymbol{u}$ is sufficiently regular and $|\boldsymbol{\nabla u}| \leq \frac{1}{4Wi}$, then there exists $T < \infty$ such that, for every $\Psi(0,P,Q) \in L^2_{d\mathbf{w}}(\mathcal{D}))$, there exists one, and only one, solution*

$$\Psi \in C([0,T], H^1_{d\mathbf{w}}(\mathcal{D})) \cap L^2(]0,T[, L^2_{d\mathbf{w}}(\mathcal{D})), \frac{\partial}{\partial t}\Psi \in L^2(]0,T[, L^2_{d\mathbf{w}}(\mathcal{D})),$$

that satisfies (7-VII.26) *for almost all $t \in [0,T]$. This solution satisfies the estimate, for all $0 < t < T$:*

$$e^{\frac{-t}{2|\boldsymbol{\nabla u}|}} \|\Psi(t)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \int_0^t \left(2|\boldsymbol{\nabla u}| - \frac{1}{2\,Wi}\right) e^{\frac{-s}{2|\boldsymbol{\nabla u}|}} \|\nabla \Psi(s)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} ds \leq \|\Psi(t=0)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}. \tag{7-VII.27}$$

Proof : We concentrate on the term parametrized by $\boldsymbol{\nabla u}$. Using Young inequalities with $\rho_1^2 > 0$, we have :

$$\int_{\mathcal{D}} \Psi(\boldsymbol{\nabla u X})^T \nabla \Phi d\mathbf{w} \leq \frac{\rho_1^2}{2}\int_{\mathcal{D}} \Psi^2 |\boldsymbol{\nabla u X}|^2 \psi_0 + \frac{1}{2\rho_1^2}\int_{\mathcal{D}} |\nabla \Phi|^2 \psi_0. \tag{7-VII.28}$$

Then, using $\nabla\psi_0 = -\boldsymbol{X}\psi_0$ and the Green theorem, we find with $\mathrm{tr}(\boldsymbol{\nabla u}^T\boldsymbol{\nabla u}) = |\boldsymbol{\nabla u}|^2$:

$$\int_{\mathcal{D}} \Psi^2 |\boldsymbol{\nabla u X}|^2 \psi_0 = \int_{\mathcal{D}} \mathrm{div}\left(\Psi^2 \boldsymbol{\nabla u}^T \boldsymbol{\nabla u X}\right)\psi_0 = |\boldsymbol{\nabla u}|^2 \|\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + 2\int_{\mathcal{D}} \Psi \boldsymbol{X}^T \boldsymbol{\nabla u}^T \boldsymbol{\nabla u} \nabla\Psi\psi_0. \quad \text{(7-VII.29)}$$

We next bound the second term in the right hand side of the previous inequality, using Young inequalities again, with $\rho_2^2 > 0$:

$$\int_{\mathcal{D}} \Psi \boldsymbol{X}^T \boldsymbol{\nabla u}^T \boldsymbol{\nabla u} \nabla\Psi\psi_0 \leq \frac{\rho_2^2}{2} |\boldsymbol{\nabla u}|^2 \|\nabla\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{1}{2\rho_2^2}\int_{\mathcal{D}} \Psi^2 |\boldsymbol{\nabla u X}|^2 \psi_0.$$

Collecting the two last results, we get, if $1 < \rho_2^2$:

$$\int_{\mathcal{D}} \Psi^2 |\boldsymbol{\nabla u X}|^2 \psi_0 \leq \frac{\rho_2^2}{\rho_2^2 - 1} |\boldsymbol{\nabla u}|^2 \left(\|\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \rho_2^2 \|\nabla\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}\right),$$

which we reintroduce in a bound for the coupling term :

$$\int_{\mathcal{D}} \Psi(\boldsymbol{\nabla u X})^T \nabla\Phi d\mathbf{w} \leq \frac{\rho_2^2\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 \left(\|\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \rho_2^2 \|\nabla\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}\right) + \frac{1}{2\rho_1^2}\|\nabla\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}. \quad \text{(7-VII.30)}$$

With $\Phi = \Psi$, the following *a priori* bound for a solution $\Psi \in H^1_{d\mathbf{w}}(\mathcal{D})$ holds :

$$\frac{1}{2}\frac{d}{dt}\|\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \leq \frac{\rho_2^2\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 \|\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \left(\frac{\rho_2^4\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 + \frac{1}{2\rho_1^2} - \frac{1}{2\mathrm{Wi}}\right)\|\nabla\Psi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}. \quad \text{(7-VII.31)}$$

The *a priori* bound (7-VII.31) allows to prove the existence of a solution

$$\Psi \in C([0,T], H^1_{d\mathbf{w}}(\mathcal{D})) \cap L^2(]0,T[, L^2_{d\mathbf{w}}(\mathcal{D})),$$

with $\frac{\partial}{\partial t}\Psi \in L^2(]0,T[, L^2_{d\mathbf{w}}(\mathcal{D}))$, for some $0 < T < \infty$, provided a Gärding inequality holds *a.e.* for $t \in [0,T]$ (Lions-Magenes theorem). Using the Gromwall lemma, we thus apply the Lions-Magenes theory with some $\rho_1^2, \rho_2^2 > 0$ provided :

$$\frac{\rho_2^2\rho_1^2}{2(\rho_2^2-1)} \geq 0 \quad \left(\frac{\rho_2^4\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 + \frac{1}{2\rho_1^2} - \frac{1}{2\mathrm{Wi}}\right) \leq 0,$$

which is true while $|\boldsymbol{\nabla u}|^2$ stays bounded (the minimum of the second term above is reached for $\rho_1^2 = \frac{1}{2|\boldsymbol{\nabla u}|}, \rho_2^2 = 2$, while $\frac{\rho_2^2\rho_1^2}{2(\rho_2^2-1)} = \frac{1}{|\boldsymbol{\nabla u}|}$). $\diamond$

We also have the following lemma :

Lemma 13 (Weighted Poincaré)**.** *[Bec89] For all $\Phi \in W_\Phi$ in the weighted Sobolev space $H^1_{d\mathbf{w}}(\mathcal{D})$, the following Poincaré-Wirtinger inequality holds :*

$$\|\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \leq \left(\int_{\mathcal{D}} \Phi d\mathbf{w}\right)^2 + \|\nabla\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}. \quad \text{(7-VII.32)}$$

Sketch of the proof : We first show the lemma for the Hermite polynomials, which form an Hilbertian basis of $L^2_{d\mathbf{w}}(\mathcal{D})$ and $H^1_{d\mathbf{w}}(\mathcal{D})$. We conclude by a compactness argument. $\diamond$

With $\Phi = \min(\Psi, 0) \equiv \Psi^-$ and $\int_{\mathcal{D}} \Phi d\mathbf{w} = 1$, the previous lemma implies in particular :

$$\frac{1}{2}\frac{d}{dt}\|\Psi^-\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \leq \frac{\rho_2^2\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 \left(\int_{\mathcal{D}} \Psi^- d\mathbf{w}\right)^2 + \left(\frac{\rho_2^2(\rho_2^2+1)\rho_1^2}{2(\rho_2^2-1)} |\boldsymbol{\nabla u}|^2 + \frac{1}{2\rho_1^2} - \frac{1}{2\mathrm{Wi}}\right)\|\nabla\Psi^-\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}. \quad \text{(7-VII.33)}$$

Yet note that this is still not sufficient to prove a maximum principle for the Fokker-Planck equation (and to ensure that ψ is indeed a pdf, with $\psi \in L^1(\mathcal{D}; \mathbb{R}^+)$ and $\int_{\mathcal{D}} \psi \equiv 1$).

7-VII-B-b Discretization and RB treatment of the problem

For the discretization W_u^h of the velocity space, we use $\mathbb{P}_1$ continuous functions on a regular mesh for $]0,1[$ made of $\mathcal{N}$ intervals of constant size $\Delta x = 1/\mathcal{N}$.

For the time discretization, we use the Euler implicit scheme, for instance with $\Psi_h^n \in W_\Psi^h$:

$$\int_{\mathcal{D}} \left(\frac{\Psi_h^{n+1} - \Psi_h^n}{\Delta t} \right) \Phi_h d\mathbf{w} = \frac{\partial u_h^n}{\partial y} \int_{\mathcal{D}} \Psi_h^{n+1} Q \frac{\partial}{\partial P} \Phi_h d\mathbf{w} - \frac{1}{2\mathrm{Wi}} \int_{\mathcal{D}} \nabla_{\boldsymbol{X}} \Psi_h^{n+1} \cdot \nabla_{\boldsymbol{X}} \Phi_h d\mathbf{w}, \ \forall \Phi \in W_\Phi^h, \tag{7-VII.34}$$

using a constant time step Δt and $u(t^n) \approx u_h^n$.

In the Hookean case, recall that the potential energy function writes :

$$U(\boldsymbol{X}) = H\boldsymbol{X}^2/2.$$

The domain $\mathcal{D}$ is then the whole space $\mathbb{R}^2$, and $d\mathbf{w}$ is a centered Gaussian measure. An Hilbertian basis for W_Ψ is given by the tensor products of Hermite polynomials.

For any $l, m \in \mathbb{N}$, the tensor products $\Psi_{l,m}(P,Q) = H_l(P) H_m(Q)$ of Hermite polynomials defined as

$$H_m(P) = (-1)^m e^{\frac{P^2}{2}} \frac{d^m}{dx^m} e^{\frac{-P^2}{2}},$$

are orthonormal for the tensor product space, where $L^2_{d\mathbf{w}}(\mathcal{D})$ is equipped with the inner product :

$$\int_{\mathbb{R}} H_l(P) H_m(P) e^{\frac{-P^2}{2}} dP = \sqrt{2\pi} \delta_{lm}.$$

We choose our Galerkin projection space $W_\Phi^h = W_\Psi^h$ as the d^2-dimensional vector space spanned by the tensor products of degree-d Hermite polynomials in P and Q.

The RB method needs an *a posteriori* error bound, which we now evaluate for $\Psi_h^n - \Psi^n$ at every time step n where Ψ^n is solution of the time-discrete weak formulation

$$\int_{\mathcal{D}} \left(\frac{\Psi^{n+1} - \Psi^n}{\Delta t} \right) \Phi d\mathbf{w} = \frac{\partial u^n}{\partial y} \int_{\mathcal{D}} \Psi^{n+1} Q \frac{\partial}{\partial P} \Phi d\mathbf{w} - \frac{1}{2\mathrm{Wi}} \int_{\mathcal{D}} \nabla_{\boldsymbol{X}} \Psi^{n+1} \cdot \nabla_{\boldsymbol{X}} \Phi d\mathbf{w}, \ \forall \Phi \in W_\Phi. \tag{7-VII.35}$$

Let us define the residual form associated with the space discretization :

$$R_h^n(\Phi) = \int_{\mathcal{D}} \left(\frac{\Psi_h^{n+1} - \Psi_h^n}{\Delta t} \right) \Phi d\mathbf{w} - \frac{\partial u_h^n}{\partial y} \int_{\mathcal{D}} \Psi_h^{n+1} Q \frac{\partial}{\partial P} \Phi d\mathbf{w} + \frac{1}{2\mathrm{Wi}} \int_{\mathcal{D}} \nabla_{\boldsymbol{X}} \Psi_h^{n+1} \cdot \nabla_{\boldsymbol{X}} \Phi d\mathbf{w}, \tag{7-VII.36}$$

for any $\Phi \in W_\Psi$. We also recall the definition of a norm for $\Phi \in W_\Psi$, for any $m > 0$. :

$$|||\Phi||| = \sqrt{m^2 \|\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \|\nabla \Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}},$$

This is well-defined for any $\Phi \in W_\Psi$ thanks to the weight Ψ_0 in the integral. (The parameters m will be precised later with a view to optimizing the error estimate if possible.)

Proposition 31 (Dual norm)**.** *The dual norm of the linear form R_h^n in $(W_\Psi, |||\cdot|||)$ is well defined : $|||R_h^n|||_\star = \sup_{\Phi \in W_\Psi^h, |||\Phi|||=1} R_h^n(\Phi)$, as a norm on the dual space $(W_\Psi^{h\star}, |||\cdot|||_\star)$ of $(W_\Psi, |||\cdot|||)$.*

Proof : Using the Cauchy-Schwarz and the Young inequalities with the Green theorem, we get the following bound like in the continuous case viewed in the previous section, with any $\rho_1^2 > 0, \rho_2^2 > 1$:

$$\int_{\mathcal{D}} \Psi_h^{n+1} Q \frac{\partial \Phi}{\partial P} d\mathbf{w} \leq \frac{\rho_2^2 \rho_1^2}{2(\rho_2^2 - 1)} \left(\|\Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \rho_2^2 \|\nabla \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \right) + \frac{1}{2\rho_1^2} \|\nabla \Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}, \tag{7-VII.37}$$

which can be bounded using the $|||\cdot|||$ norm for Φ if we can make use of the $|||\cdot|||$ norm for Ψ_h^{n+1} (for $m \geq 0$). □

Then we have :

Proposition 32 (Error estimate)**.** *For solutions $(\Psi^n)_{0 \leq n \leq} \in H^1_{d\mathbf{w}}(\mathcal{D})^{N+1}$ of the semi-discrete uncoupled Fokker-Planck equation* (7-VII.35)*, we have the following estimate :*

$$\begin{aligned} \left(1 - \frac{\rho_1^2}{2} - \frac{\Delta t}{2} \left(m^2 \rho_2^2 + \frac{1}{2} \left| \frac{\partial u_h^n}{\partial y} \right| \right) \right) \|\Psi^{n+1} - \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \\ + \frac{\Delta t}{2} \left(\frac{1}{Wi} - \rho_2^2 - 4 \left| \frac{\partial u_h^n}{\partial y} \right| \right) \|\nabla \left(\Psi^{n+1} - \Psi_h^{n+1} \right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \\ \leq \frac{1}{2\rho_1^2} \|\Psi^n - \Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{\Delta t}{2\rho_2^2} |||R_h^n|||^2_\star. \end{aligned} \tag{7-VII.38}$$

Proof : Taking $\Phi = \Psi^{n+1} - \Psi_h^{n+1}$ in the residual form and using (7-VII.25), (7-VII.35) and the Young inequalities, we get the following estimate with any $\rho_1^2 > 0, \rho_2^2 > 0$:

$$\left(1 - \frac{\rho_1^2}{2}\right) \|\Psi^{n+1} - \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} - \frac{\Delta t}{2} \frac{\partial u_h^n}{\partial y} \int_{\mathcal{D}} PQ \left(\Psi^{n+1} - \Psi_h^{n+1}\right)^2 \psi_0 dP dQ - \frac{\Delta t}{2} \rho_2^2 |||\Psi^{n+1} - \Psi_h^{n+1}|||^2$$
$$+ \frac{\Delta t}{2} \frac{1}{\mathrm{Wi}} \|\nabla \left(\Psi^{n+1} - \Psi_h^{n+1}\right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} \leq \frac{1}{2\rho_1^2} \|\Psi^n - \Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{\Delta t}{2\rho_2^2} |||R_h^n|||^2_\star. \quad \text{(7-VII.39)}$$

Introducing the bound derived in the proof of Proposition 31, we finally get with any $\rho_3^2 > 0, \rho_4^2 > 1$:

$$\left(1 - \frac{\rho_1^2}{2}\right) \|\Psi^{n+1} - \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} - \frac{\Delta t}{2} \rho_2^2 |||\Psi^{n+1} - \Psi_h^{n+1}|||^2 + \frac{\Delta t}{2} \frac{1}{\mathrm{Wi}} \|\nabla \left(\Psi^{n+1} - \Psi_h^{n+1}\right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}$$
$$\leq \frac{1}{2\rho_1^2} \|\Psi^n - \Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{\Delta t}{2\rho_2^2} |||R_h^n|||^2_\star + \frac{\Delta t}{2} \left|\frac{\partial u_h^n}{\partial y}\right| \left(\frac{\rho_4^2 \rho_3^2}{2(\rho_4^2 - 1)} \|\Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \left(\frac{\rho_4^4 \rho_3^2}{2(\rho_4^2 - 1)} + \frac{1}{2\rho_3^2}\right) \|\nabla \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}\right), \quad \text{(7-VII.40)}$$

where $\left|\frac{\partial u_h^n}{\partial y}\right| = \max\left(\frac{\partial u_h^n}{\partial y}, 0\right)$. Collecting the different terms, we see that we can derive two types of *a posteriori* error bounds, that is in $L^2_{d\mathbf{w}}(\mathcal{D})$ or in $H^1_{d\mathbf{w}}(\mathcal{D})$:

$$\left(1 - \frac{\rho_1^2}{2} - \frac{\Delta t}{2}\left(m^2 \rho_2^2 + \left|\frac{\partial u_h^n}{\partial y}\right| \frac{\rho_4^2 \rho_3^2}{2(\rho_4^2 - 1)}\right)\right) \|\Psi^{n+1} - \Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}$$
$$+ \frac{\Delta t}{2}\left(\frac{1}{\mathrm{Wi}} - \rho_2^2 - \left|\frac{\partial u_h^n}{\partial y}\right| \left(\frac{\rho_4^4 \rho_3^2}{2(\rho_4^2 - 1)} + \frac{1}{2\rho_3^2}\right)\right) \|\nabla \left(\Psi^{n+1} - \Psi_h^{n+1}\right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}$$
$$\leq \frac{1}{2\rho_1^2} \|\Psi^n - \Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{\Delta t}{2\rho_2^2} |||R_h^n|||^2_\star. \quad \text{(7-VII.41)}$$

But problems occur on account of the coupling term, when the factor $\frac{\partial u_h^n}{\partial y}$ is too large, since for all $\rho_4^2 > 1$:

$$\frac{\rho_4^4}{(\rho_4^2 - 1)} \geq 4.$$

Remark 37. *Note that in the FENE case, this term can be handled more easily thanks to following inequality :*

$$\int_{\mathcal{D}} \Psi_h^{n+1} Q \frac{\partial \Phi}{\partial P} d\mathbf{w} \leq b \frac{\rho_4^2}{2} \|\Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \frac{1}{2\rho_4^2} \|\nabla \Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}, \quad \text{(7-VII.42)}$$

obtained after invoking the finite extensibility of $|\boldsymbol{X}| \leq \sqrt{b}$ *(*$\mathcal{D}$ *is bounded).*

Let us choose $\rho_4^2 = 2$, so that $\frac{\rho_4^4}{(\rho_4^2 - 1)} = 4$. We conclude by choosing $\rho_3^2 = \frac{1}{2}$ to minimize : $4\rho_3^2 + \frac{1}{\rho_3^2} \geq 4$. □

An *a posteriori* estimator for the L^2 error $\|\Psi_h^n - \Psi^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}$ can now be obtained when $\frac{\partial u_h^n}{\partial y}$ is not too large :

Proposition 33. *Assume that, for* $0 \leq k \leq n$*, the following condition is satisfied :*

$$\left|\frac{\partial u_h^k}{\partial y}\right| \leq \min\left(\frac{1}{4\,Wi}, \frac{4}{\Delta t}\right) \quad \text{(7-VII.43)}$$

(which we can also rewrite with a changing time step $\left|t^{k+1} - t^k\right|$ *instead of* Δt*). Then, provided the initial error is zero* $\|\Psi^0 - \Psi_h^0\|_{L^2_{d\mathbf{w}}(\mathcal{D})} = 0$*, we have the a posteriori error bound :*

$$\|\Psi^{n+1} - \Psi_h^{n+1}\|_{L^2_{d\mathbf{w}}(\mathcal{D})} \leq \Delta t \sum_{k=0}^{n} \left(\prod_{k'=k+1}^{n} \left(1 - \frac{\Delta t}{4} \left|\frac{\partial u_h^{k'}}{\partial y}\right|\right)^{-1}\right) m_k |||R_h^k|||_{m_k,\star}, \quad \text{(7-VII.44)}$$

where the norm $|||\cdot|||^k_\star$ *is defined at each step* $0 \leq k \leq n$ *as the dual norm of the following primal norm for* $\Phi \in W_\Phi$*,*

$$|||\Phi|||_{m_k} = \sqrt{m_k^2 \|\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})} + \|\nabla \Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}},$$

with $m_k^2 \Delta t = \dfrac{1 - \frac{\Delta t}{4}\left|\frac{\partial u_h^k}{\partial y}\right|}{\frac{1}{Wi} - 4\left|\frac{\partial u_h^k}{\partial y}\right|}$.

Proof : We start with the error estimate (7-VII.41) obtained above. To optimally turn this error estimate into an *a posteriori* superior bound, we would like to minimize the function

$$f(x,y)=\frac{ax+by}{xy(c-x-y)}$$

for $a,b,c>0$ when (x,y) belongs to the domain defined by $0<x, 0<y\leq d, x+y<c$. This corresponds to the minimization of the right-hand side of the error estimate (7-VII.41) when

- $a=\|\Psi^n-\Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}$,
- $b=m^2\Delta t^2|||R_h^n|||^2_\star$,
- $c=2\left(1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|\right)$,
- $d=m^2\Delta t\left(\frac{1}{\mathrm{Wi}}-4\left|\frac{\partial u_h^n}{\partial y}\right|\right)$,
- $x=\rho_1^2$,
- $y=m^2\Delta t\rho_2^2$,

under the condition : $\left|\frac{\partial u_h^n}{\partial y}\right|\leq\min\left(\frac{1}{4\mathrm{Wi}},\frac{4}{\Delta t}\right)$.

A simple analysis of the zeros of the derivative of f shows that f has only one admissible local extremum (x_0,y_0) in the domain of study when $y_0\leq d$:

$$x_0=\frac{c}{2}\frac{\sqrt{b}}{\sqrt{a}+\sqrt{b}}\quad y_0=\frac{c}{2}\frac{\sqrt{a}}{\sqrt{a}+\sqrt{b}}.$$

Last, we know the sign of the derivatives, which cross zero only once in the domain of concern. Then, (x_0,y_0) can only be an absolute extremum on the domain of study, that is a global minimum since $f\to\infty$ on the boundaries[1]. We have $f(x_0,y_0)=\left(\frac{2(\sqrt{a}+\sqrt{b})}{c}\right)^2$. And when $y_0>d$, there is one global minima on the line $y=d$.

Thus we choose :

$$\begin{aligned}\rho_1^2&=\left(1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|\right)\frac{\|\Psi^n-\Psi_h^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}}{\|\Psi^n-\Psi_h^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}+m\Delta t|||R_h^n|||_\star}\\ m^2\Delta t\rho_2^2&=\left(1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|\right)\frac{m\Delta t|||R_h^n|||_\star}{\|\Psi^n-\Psi_h^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}+m\Delta t|||R_h^n|||_\star}\end{aligned}\tag{7-VII.45}$$

to get :

$$\|\Psi^{n+1}-\Psi_h^{n+1}\|_{L^2_{d\mathbf{w}}(\mathcal{D})}\leq\left(1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|\right)^{-1}\left(\|\Psi^n-\Psi_h^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}+m\Delta t|||R_h^n|||_\star\right),\tag{7-VII.46}$$

with m such that $y_0\leq d$, that is such that :

$$\left(1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|\right)m\Delta t|||R_h^n|||_\star\leq m^2\Delta t\left(\frac{1}{\mathrm{Wi}}-4\left|\frac{\partial u_h^n}{\partial y}\right|\right)\left(\|\Psi^n-\Psi_h^n\|_{L^2_{d\mathbf{w}}(\mathcal{D})}+m\Delta t|||R_h^n|||_\star\right).$$

Finally, with the choice $m^2\Delta t=\frac{1-\frac{\Delta t}{4}\left|\frac{\partial u_h^n}{\partial y}\right|}{\frac{1}{\mathrm{Wi}}-4\left|\frac{\partial u_h^n}{\partial y}\right|}$ minimizing the L^2 error bound, we get the result.□

Remark 38 (An error bound for FENE dumbbells)**.** *The FENE case is more favorable, since the finite extensibility implies :*

$$\int_{\mathcal{D}}\Psi_h^{n+1}Q\frac{\partial\Phi}{\partial P}d\mathbf{w}\leq b\frac{\rho_3^2}{2}\|\Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}+\frac{1}{2\rho_3^2}\|\nabla\Phi\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})},\tag{7-VII.47}$$

so that we have :

$$\begin{aligned}\left(1-\frac{\rho_1^2}{2}-\frac{\Delta t}{2}\left(m^2\rho_2^2+b\rho_3^2\left|\frac{\partial u_h^n}{\partial y}\right|\right)\right)&\|\Psi^{n+1}-\Psi_h^{n+1}\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}\\ &+\frac{\Delta t}{2}\left(\frac{1}{Wi}-\rho_2^2-\frac{1}{\rho_3^2}\left|\frac{\partial u_h^n}{\partial y}\right|\right)\|\nabla\left(\Psi^{n+1}-\Psi_h^{n+1}\right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}\\ &\leq\frac{1}{2\rho_1^2}\|\Psi^n-\Psi_h^n\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}+\frac{\Delta t}{2\rho_2^2}|||R_h^n|||^2_\star,\end{aligned}\tag{7-VII.48}$$

[1]We can also compute the hessian of f at (x_0,y_0). The determinant and the trace of this 2×2 matrix are positive, which shows that the two eigenvalues of the hessian are positive and that (x_0,y_0) is a local minimum for f.

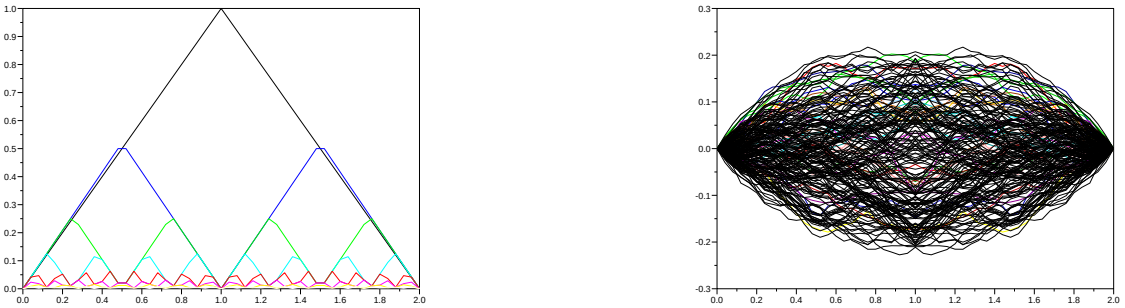


FIG. 7.15 – The eight first hierarchical basis (σ_i) (left), whic are used to generate 200 random linear combinations (right) as a training parameter sample of $\left(\frac{\partial u_h^n}{\partial y}\right)_n$ in the following experiment.

where one can always take ρ_3^2 *as large as desired to handle* $\|\nabla\left(\Psi^{n+1}-\Psi_h^{n+1}\right)\|^2_{L^2_{d\mathbf{w}}(\mathcal{D})}$, *while* Δt *can still be chosen very small for* $\left(1-\frac{\rho_1^2}{2}-\frac{\Delta t}{2}\left(m^2\rho_2^2+b\rho_3^2\left|\frac{\partial u_h^n}{\partial y}\right|\right)\right)$ *to stay positive. It is then possible to derive an a posteriori error bound for any Weissenberg number Wi.*

We will next use the RB method with the *a posteriori* error estimator (7-VII.44) and the same RB greedy algorithm than in [GP05].

7-VII-B-c Numerical results

We now build a reduced-basis for the collection of (uncoupled) Fokker-Planck equations where the (time-dependent) velocity gradient is the parameter $\mu=(\partial_y u_h^n)_n$, with a view to speed up the simulation of the system of (possibly coupled) equations . We take $\mathrm{Wi}=1$, a constant time step with 50 iterations until $T=2$, and a Hermite polynomial basis of dimension $11^2=121$ (for the fine approximations).

A training sample of parameters is generated using the same hierarchical "cereal box" training method as in Section 7-VII-B-c, that is with linear combinations of hierarchical basis elements, using a set of uniformly distributed random coefficients within a bounded range. At each time step, we write :

$$\mu(t^k)=\sum_{i=1}^{8}\mu_{\sigma_i}\sigma_i(t^k)\ ,\forall 0\le k\le N\,, \tag{7-VII.49}$$

using the hierarchical hat functions (Fig. 7.15) : $\sigma_i(t)=1-2^{-i}\left|\left(2^i\frac{t}{T}[2]\right)-1\right|$, $\forall t\in[0,T]$.

We recall that we use the RB greedy algorithm of [GP05], which is as such. First, we include the initial equilibrated state $\Psi^0\equiv 1$ in the reduced basis. Then, we increase the reduced basis by adding the solution $\Psi_h^n(\mu)$, among all $(\Psi_h^n(\mu))_{n,\mu}$, for which the increase of the relative error

$$\Delta_h^n(\mu)=\frac{\|\Psi^n(\mu)-\Psi_h^n(\mu)\|_{L^2_{d\mathbf{w}}(\mathcal{D})}}{\|\Psi_h^n(\mu)\|_{L^2_{d\mathbf{w}}(\mathcal{D})}}-\frac{\|\Psi^{n-1}(\mu)-\Psi_h^{n-1}(\mu)\|_{L^2_{d\mathbf{w}}(\mathcal{D})}}{\|\Psi_h^{n-1}(\mu)\|_{L^2_{d\mathbf{w}}(\mathcal{D})}}$$

is maximal.

In the offline stage, we observe that the maximal increment of relative error $\Delta_h^n(\mu)$ decreases very fast with the size of the reduced basis, see Fig. 7.16.

We also recall that at each step of the greedy algorithm, the reduced-basis approximation and the corresponding *a posteriori* error estimation for $(\Psi_h^n(\mu))_{n,\mu}$ are computed for every n and for every μ in the training sample through the reduced-basis, using the error bound derived in the previous section. So, it is only for the selected parameter value that we compute the "exact" solution. Though, we also need to compute some coefficient m_k at each iteration $(k=1,\dots,K)$ and for each parameter in the training sample in order to evaluate the error bound, which induces additional computations compared to the standard RB method (possibly a burden for online certification of the RB approximation error).

We next test the efficiency of the resulting reduced basis in some online stage using a test sample of parameters. For this, we compute the rate of convergence of the RB relative error within a sample of test parameter

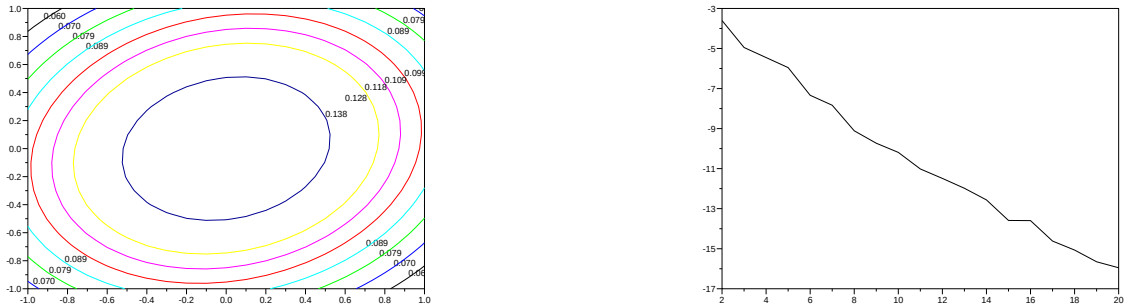


FIG. 7.16 – Sketch of a solution in the plane (P,Q) (left) and maximum increment of the relative error during the training (greedy selection) *wrt* the size of the reduced basis (right).

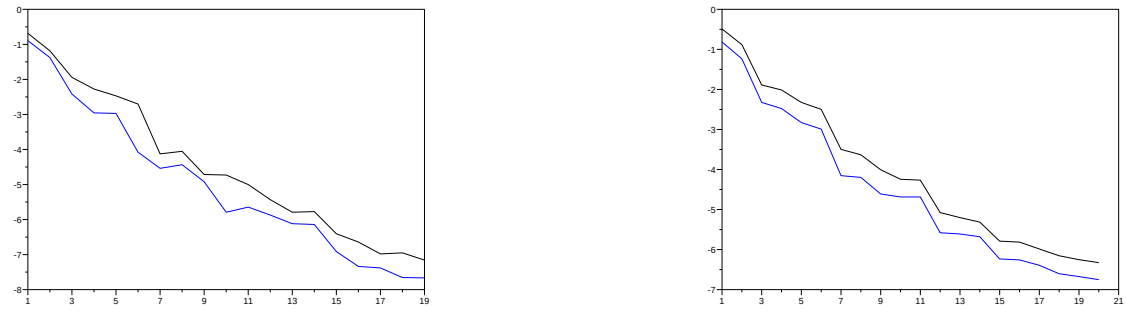


FIG. 7.17 – In log-scale, maximal (relative) RB approximation error (in blue), with its *a posteriori* estimation (in black), for a training sample of parameter values during the offline stage (left), and for a test sample of same size 200 (right), with respect to the size of the reduced basis.

(Fig. 7.17), which we compare with its *a posteriori* estimation (Fig. 7.18). The effectivity of the *a posteriori* error estimation in L^2 norm is shown in Fig. 7.18 : it is not too bad here, hence the good reduction yielded by the RB method. Yet, one should not that we have strongly limited the range of velocity gradients, and it is not clear whether this could be useful in a more realistic computation.

7-VII-B-d A RB approach for the coupled Fokker-Planck / Navier-Stokes system

With a view to using our RB method in a (real-time) simulation of the coupled system of equations (Fokker-Planck + Navier-Stokes), we propose a new training method. We recall that we limit to configurations using pure shear flows of the Couette type.

For the coupled case, note that our *a posteriori* error estimation is not rigorous anymore, since the coupling of approximation errors in each equation should be taken into account (as time evolves), but for the moment being and as a first approach, we will build a reduced basis for the solutions to the Fokker-Planck equation using the same *a posteriori* error estimate as in the uncoupled case.

In the offline stage, we propose to train the Fokker-Planck equation using the velocity gradient field generated by a macroscopic model in the same flow configuration. Here, we use the Oldroyd-B equation. This training method can be compared to the training method used in the previous section (sometimes termed "cereal box" in reference to the hierarchy in the ranges of successively introduced dimensions for the parameter), for a given simulation of the coupled system : of course, it proves better since the training sample is more adapted (see Fig. 7.19).

In fact, it seems that the convergence of RB approximations is very fast (when used for a coupled simulation), as shown by Fig. 7.19 : it is exponential and drops down to zero (machine) with a very small-dimensional basis. This is partly due to the fact that, although our error estimator is not rigorous in the coupled case, it is yet still effective here (see Fig. 7.20).

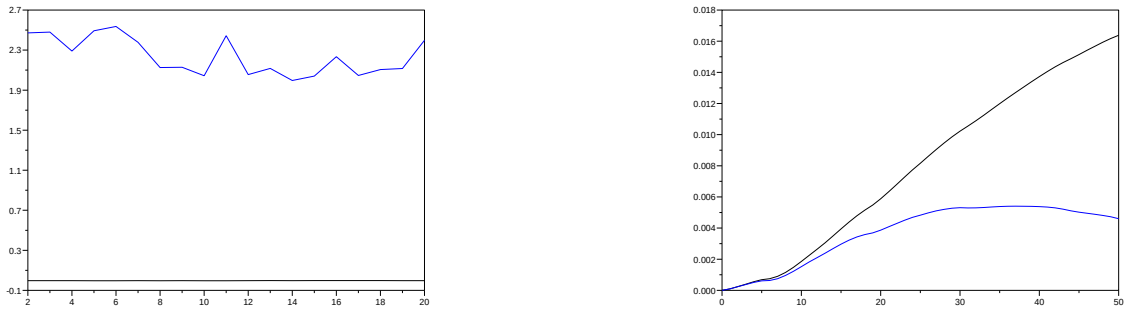


FIG. 7.18 – Maximum and minimum of the effectivity of the *a posteriori* estimation for a test sample (left, in log-scale) with respect to the size of the reduced basis. And time evolution of the actual error with its error bound for one particular parameter μ value (right).

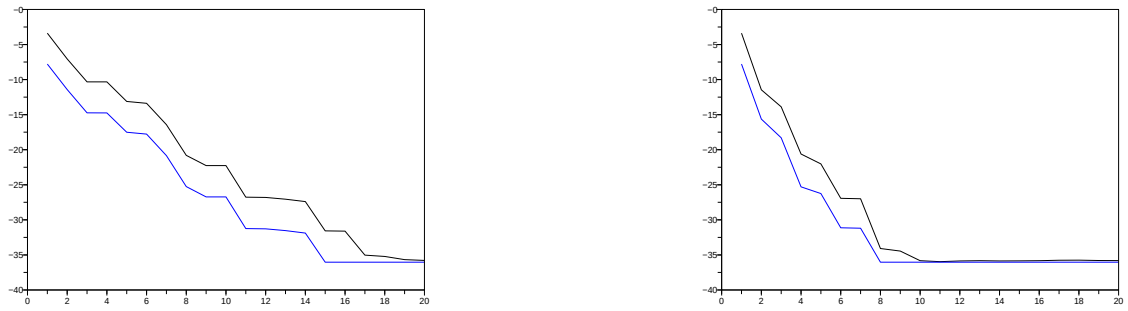


FIG. 7.19 – RB approximation errors in (squared) $L^\infty([0,T],L^2)$ norm using a log-scale, for the outputs $\boldsymbol{u}$ (blue) and $\operatorname{div}\boldsymbol{\tau}$ (black) in a coupled simulation, with respect to the size of thr reduced basis. On the left : "cereal box" training method with 100 parameter "histories"; on the right : "macro" training method with the "true" solution of the coupled simulation.

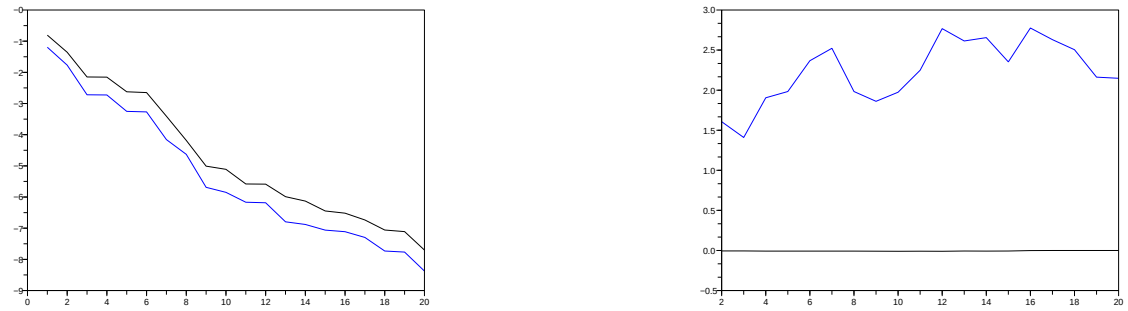


FIG. 7.20 – Left : maximal relative error for RB approximations of the pdf ψ in L^2 norm (blue), and their *a posteriori* error estimations (black), using a log-scale, with respect to the size of the reduced basis (in the offline stage using the new training method). Right : corresponding effectivity of the error estimator (minimum in blue and maximum in black, for the whole family of pdf generated in a coupled simulation).

But furthermore, this is a very special situation that we tested here. Not only because the velocity gradient fields that this method generates is exactly the same as the one that should be computed by a (Fokker-Planck + Navier-Stokes) coupled simulation (since the Oldroyd-B equation can be derived exactly from a Hookean dumbbell model – this is not the case anymore if we use the FENE dumbbell model trained on the Oldroyd-B equation for instance, or even better, the FENE-P model –). But also because, in the Hookean case, the output τ for the RB-approximated pdf (solution to the Fokker-Planck equation) is computed exactly when using $\Phi = QF_P(P,Q) = HPQ$ as a test function in the semi-discrete Fokker-Planck equation, as soon as $PQ \in W_\Psi^h$. It is consequently not very meaningful to test the RB method in this case, because there is actually no approximation error due to the Fokker-Planck equation (which is exactly computable) once the RB approximation space is large enough to contain these basis functions (as it can be shown in Fig. 7.19). In other words, the (incredibly fast) success of the RB method for the coupled simulation of Fig. 7.19 does not take into account the fact that it is much easier to obtain the output τ than the L^2 norm for Ψ.

For the coupled case, the point would be to test the RB method for FENE dumbbells, for instance. Then, computation of the entries of the Kramers matrix (which are integrals of $\boldsymbol{X}\boldsymbol{X}^T U'(\boldsymbol{X})$, with weight the pdf) would be more challenging, since the Fokker-Planck equation is not exactly solvable then, so that the reduction effort would be more meaningful (a theoretically infinite number of basis functions is generally needed for RB approximations to become exact). To this aim, one needs *a posteriori* estimations for the FENE Fokker-Planck equation. This has been done partly in the recent work [KP09], using another output functional (the optical anisotropy) for the solutions to the FENE Fokker-Planck equation in another type of flow configurations (extensional flows). But the need for computational reduction in the computation of (high-dimensional, parametrized) Fokker-Planck equation is still a challenge to a vast field of applications (see for example [SST08, LBLM08], which are two recent works in this direction).

Bibliographie

[AB05] G. Allaire and R. Brizzi. A multiscale finite element method for numerical homogenization. *SIAM MMS*, 4(3) :790–812, 2005.

[All92] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal*, 23(6) :1482–1518, 1992.

[AO00] M. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley-Interscience, 2000.

[AP05] Y. Achdou and O. Pironneau. *Computational Methods for Option Pricing (Frontiers in Applied Mathematics 30)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[AQ92] D. N. Arnold and J. Qin. Quadratic velocity/linear pressure Stokes elements. In R. Vichnevetsky, D. Knight, and G. Richter, editors, *Advances in Computer Methods for Partial Differential Equations*, volume VII, pages 28–34. IMACS, 1992.

[Aro04] B. Arouna. Robbins-monroe algorithms and variance reduction in finance. *The Journal of Computational Finance*, 7(2) :35–62, 2004.

[ASB78] B. O. Almroth, P. Stern, and F. A. Brogan. Automatic choice of global shape functions in structural analysis. *AIAA Journal*, 16 :525–528, 1978.

[BB09a] J. W. Barrett and S. Boyaval. Existence and approximation of a (regularized) FENE-P model. (in preparation), 2009.

[BB09b] J. W. Barrett and S. Boyaval. Existence and approximation of a (regularized) Oldroyd-B model. *(preprint submitted for publication http ://fr.arxiv.org/abs/0907.4066)*, 2009.

[BBM+09] S. Boyaval, C. Le Bris, Y. Maday, N.C. Nguyen, and A.T. Patera. A reduced basis approach for variational problems with stochastic parameters : Application to heat conduction with variable robin coefficient. *Computer Methods in Applied Mechanics and Engineering*, 198(41–44) :3187–3206, 2009.

[BCAH87a] R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager. *Dynamics of polymeric liquids*, volume 1 : Fluid Mechanics. John Wiley & Sons, New York, 1987.

[BCAH87b] R. B. Bird, C. F. Curtiss, R. C. Armstrong, and O. Hassager. *Dynamics of Polymeric Liquids*, volume 2 : Kinetic Theory. John Wiley & Sons, New York, 1987.

[BCP06] A. Bonito, P. Clément, and M. Picasso. Finite element analysis of a simplified stochastic hookean dumbbells model arising from viscoelastic flows. *ESAIM : Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 40(4) :785–814, 2006.

[BE94] A. N. Beris and B. J. Edwards. *Thermodynamics of flowing systems with internal microstructure*. Oxford University Press, New York, 1994.

[Bec89] W. Beckner. A generalized poincaré inequality for gaussian measures. *Proc. Amer. Math. Soc.*, 105(2) :397–400, 1989.

[BF91] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, Berlin, 1991.

[BFT93] M. Behr, L.P. Franca, and T.E. Tezduyar. Stabilized finite element methods for the velocity-pressure-stress formulation of incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 104(31–48), 1993.

[BGL06] J. Burkardt, M. D. Gunzburger, and H. C. Lee. POD and CVT-based reduced order modeling of Navier-Stokes flows. *Comp. Meth. Applied Mech.*, 196 :337–355, 2006.

[BH82] A. N. Brooks and T. J. R. Hughes. Streamline upwind / petrov-galerkin formulations for convection dominated flows with particular emphasis on the incompressible navier-stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32 :199–259, 1982.

[BHW89] H.A. Barnes, J.F. Hutton, and K.F.R.S. Walters. *An introduction to rheology*. Elsevier Science Publisher, Amsterdam, The Netherlands, 1st edition, 1989.

[BJDM85] F. Brezzi, Jr. J. Douglas, and L.D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47 :217–235, 1985.

[BJDM86] F. Brezzi, Jr. J. Douglas, and L.D. Marini. Recent results on mixed finite element methods for second order elliptic problems. In A.V. Balakrishnan, A.A. Dorodnitsyn, and J.L. Lions, editors, *Vistas in Applied Mathematics : Numerical Analysis, Atmospheric Sciences, Immunology*, pages 25–43, 1986.

[BL04] C. Le Bris and P.-L. Lions. Renormalized solutions of some transport equations with partially $w^{1,1}$ velocities and applications. *Annali di Matematica pura ed applicata*, 183 :97–130, 2004.

[BL09a] S. Boyaval and T. Lelièvre. A variance reduction method for parametrized stochastic differential equations using the reduced basis paradigm. In Pingwen Zhang, editor, *Accepted for publication in Communication in Mathematical Sciences*, volume Special Issue "Mathematical Issues on Complex Fluids", 2009. ARXIV :0906.3600.

[BL09b] C. Le Bris and T. Lelièvre. Multiscale modelling of complex fluids : A mathematical initiation. In B. Engquist, P. Lötstedt, and O. Runborg, editors, *Multiscale Modeling and Simulation in Science*, volume 66 of *Lecture Notes in Computational Science and Engineering*, pages 49–138. Springer, 2009. Proceedings of the Summer School in Stockholm, June 07.

[BLL06] X. Blanc, C. Le Bris, and P.L. Lions. Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques. *C.R. Acad. Sci. Paris Ser. I*, 343(11–12) :717–724, 2006.

[BLM09] S. Boyaval, T. Lelièvre, and C. Mangoubi. Free-energy-dissipative schemes for the Oldroyd-B model. *ESAIM : Mathematical Modelling and Numerical Analysis*, 43(3) :523–561, may 2009.

[BLP78] A. Bensoussan, J. L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 5 of *Studies in Mathematics and its applications*. North-Holland Publisher Company, 1978.

[BM97] J. Baranger and A. Machmoum. Existence of approximate solutions and error bounds for viscoelastic fluid flow : Characteristics method. *Comput. Methods Appl. Mech. Engrg.*, 148 :39–52, 1997.

[BNMP04] M. Barrault, N. C. Nguyen, Y. Maday, and A. T. Patera. An "empirical interpolation" method : Application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris, Série I.*, 339 :667–672, 2004.

[BNT07] I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 45(3) :1005–1034, 2007.

[Boy08] S. Boyaval. Reduced-basis approach for homogenization beyond the periodic setting. *SIAM Multiscale Modeling & Simulation*, 7(1) :466–494, 2008.

[Boy09] S. Boyaval. *Mathematical modeling and simulation for material science*. PhD thesis, Université Paris Est, 2009. In preparation.

[BP84] F. Brezzi and J. Pitkäranta. On the stabilization of finite element approximations of the Stokes equations. In W. Hackbusch, editor, *Efficient Solution of Elliptic System*, pages 11–19. Vieweg & Sohn, Braunschweig, 1984. Notes on Numerical Fluid Mechanics, vol. 10.

[BP99] J. Bonvin and M. Picasso. Variance reduction methods for CONNFFESSIT-like simulations. *J. Non-Newtonian Fluid Mech.*, 84 :191–215, 1999.

[BP04] A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. I.H. Poincaré*, 40 :153–165, 2004.

[BPL06] A. Bonito, M. Picasso, and M. Laso. Numerical simulation of 3d viscoelastic flows with free surfaces. *J. Comput. Phys.*, 215(2) :691–716, 2006.

[BPP08] M. Bajaj, M. Pasquali, and J. R. Prakash. Coil-stretch transition and the breakdown of computations for viscoelastic fluid flow around a confined cylinder. *J. Rheol.*, 52 :197–223, 2008.

[BPS01] J. Bonvin, M. Picasso, and R. Stenberg. GLS and EVSS methods for a three fields Stokes problem arising from viscoelastic flows. *Comp. Meth. Appl. Mech. Eng.*, 190 :3893–3914, 2001.

[BR01] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 37 :1–225, 2001.

[Bri94] M. Briane. Homogenization of a non periodic material. *J. Math. Pures Appl.*, 73(1) :47–66, 1994.

[BS92a] J. Baranger and D. Sandri. Finite element approximation of viscoelastic fluid flow. *Numer. Math.*, 63 :13–27, 1992.

[BS92b] J. Baranger and D. Sandri. A formulation of stokes's problem and the linear elasticity equations suggested by the oldroyd model for viscoelastic flow. *RAIRO - Modélisation mathématique et analyse numérique*, 26(2) :331–345, 1992.

[BS07] J. W. Barrett and E. Süli. Existence of global weak solutions to some regularized kinetic models of dilute polymers. *Multiscale Model. Simul.*, 6 :506–546, 2007.

[BS08] J. W. Barrett and E. Süli. Existence of global weak solutions to dumbbell models for dilute polymers with microscopic cut-off. *Math. Models Methods Appl. Sci.*, 18 :935–971, 2008.

[BS09] J. W. Barrett and E. Süli. Finite element approximation of kinetic dilute polymer models with microscopic cut-off. (submitted for publication), 2009.

[BSS05] J. W. Barrett, Ch. Schwab, and E. Süli. Existence of global weak solutions for some polymeric flow models. *Math. Models Methods Appl. Sci.*, 15 :939–983, 2005.

[BTZ05] I. Babuška, R. Tempone, and G. Zouraris. Solving elliptic boundary value problems with uncertain coefficients by the finite element method : the stochastic formulation. *Comput. Meth. Appl. Mech. Eng.*, 194 :1251–1294, 2005.

[CDD] A. Cohen, R. A. Devore, and W. Dahmen. Personal communication.

[CDW84] M. J. Crochet, A. R. Davies, and K. Walters. *Numerical Simulation of Non-Newtonian Flow*. Elsevier, London, 1984.

[CGRdB04] D.B. Chung, M.A. Gutiérrez, J.J.C. Remmers, and R. de Borst. Stochastic finite element modelling of fibre-metal laminates. In *45th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics & Materials Conference*, April 2004.

[CHMR08] Y. Chen, J. Hesthaven, Y. Maday, and J. Rodríguez. A monotonic evaluation of lower bounds for inf-sup stability constants in the frame of reduced basis approximations. *C.R. Math. Acad. Sci. Paris*, 2008.

[Cia78] Ph. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.

[CL04] C. Chauvière and A. Lozinski. Simulation of dilute polymer solutions using a Fokker-Planck equation. *Computers and Fluids*, 33 :687–696, 2004.

[CO01] C. Chauvière and R. G. Owens. A new spectral element method for the reliable computation of viscoelastic flow. *CMAME*, 190(31) :3999–4018, 2001.

[Cod98] R. Codina. Comparison of some finite element methods for solving the diffusion-convection-reaction equation. *Comp. Meth. Appl. Mech. Engrg.*, 156 :185–210, 1998.

[CR73] M. Crouzeix and P. A. Raviart. Conforming and non-conforming finite element methods for solving the stationary Stokes equations. *RAIRO Anal. Numer.*, 3 :33–75, 1973.

[DBO01] M.K. Deb, I.M. Babuška, and J.T. Oden. Solution of stochastic partial differential equations using galerkin finite element techniques. *Comp. Meth. Appl. Mech. Engrg.*, 190(48) :6359–6372, 2001.

[DE98] M. Doï and S.F. Edwards. *The Theory of Polymer Dynamics*. Oxford Science, 1998.

[Dep08] S. Deparis. Reduced basis error bound computation of parameter-dependent Navier-Stokes equations by the natural norm approach. *SIAM Journal of Numerical Analysis*, 46(4) :2039–2067, 2008.

[Dev93] R. A. Devore. Constructive approximation. *Acta Numerica*, 7 :51–150, 1993.

[DGRH07] A. Doostan, R. G. Ghanem, and J. Red-Horse. Stochastic model reduction for chaos representations. *Computer Methods in Applied Mechanics and Engineering*, 196(37-40) :3951–3966, 2007.

[DNP+04] B.J. Debusschere, H.N. Najm, P.P. Pebay, O.M. Knio, R.G. Ghanem, and O.P. Le Maître. Numerical challenges in the use of polynomial chaos representations for stochastic processes. *SIAM Journal on Scientific Computing*, 26(2) :698–719, 2004.

[EEL+07] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. The heterogeneous multiscale method : A review. *Commun. Comput. Phys.*, 2 :367–450, 2007.

[EG04] A. Ern and J. L. Guermond. *Theory and Practice of Finite Elements*. Springer Verlag, New York, 2004.

[ELZ04] W. E, T. J. Li, and P.-W. Zhang. Well-posedness for the dumbbell model of polymeric fluids. *Comm. Math. Phys.*, 248 :409–427, 2004.

[Emm08] E. Emmrich. Convergence of a time discretization for a class of non-newtonian fluid flow. *Commun. Math. Sci.*, 6(4) :827–843, 2008.

[EPR] J.L. Eftang, A.T. Patera, and E.M. Rønquist. Personal communication. HP-type reduced basis method for parametrized partial differential equations.

[FCGO02] E. Fernández-Cara, F. Guillèn, and R. R. Ortega. Mathematical modeling and analysis of viscoelastic fluids of the oldroyd kind. In P.G. Ciarlet et al., editor, *Handbook of numerical analysis VIII : Solution of equations in $\mathbb{R}^n$ (Part 4). Techniques of scientific computing (Part 4). Numerical methods of fluids (Part 2).*, pages 543–661. Elsevier North-Holland, Amsterdam, 2002.

[FF89] M. Fortin and A. Fortin. A new approach for the fem simulation of viscoelastic flows. *Journal of Non-Newtonian Fluid Mechanics*, 32 :295–310, 1989.

[FGP97] M. Fortin, R. Guénette, and R. Pierre. Numerical analysis of the modified evss method. *Comput. Methods Appl. Mech. Engrg.*, 143 :79–95, 1997.

[FHKK07] A. Fattal, O. H. Hald, G. Katriel, and R. Kupferman. Global stability of equilibrium manifolds, and "peaking" behavior in quadratic differential systems related to viscoelastic models. *J. Non-Newtonian Fluid Mech.*, 144 :30–41, 2007.

[FK04] R. Fattal and R. Kupferman. Constitutive laws for the matrix-logarithm of the conformation tensor. *Journal of Non-Newtonian Fluid Mechanics*, 123(2–3) :281–285, 2004.

[FK05] R. Fattal and R. Kupferman. Time-dependent simulation of visco-elastic flows at high weissenberg number using the log-conformation representation. *J. Non-Newtonian Fluid Mech.*, 126 :23–27, 2005.

[FR83] J. P. Fink and W. C. Rheinboldt. On the error behavior of the reduced basis technique for nonlinear finite element approximations. *Z. Angew. Math. Mech.*, 63(1) :21–28, 1983.

[Fri75] A. Friedman. *Stochastic differential equations and applications, Vol. 1.* Academic Press (New York ; London ; Toronto), 1975.

[FS91] L. Franca and R. Stenberg. Error analysis of some gls methods for elasticity equations. *SIAM J. Numer. Anal.*, 28 :1680–1697, 1991.

[FST05] P. Frauenfelder, C. Schwab, and R.A. Todor. Deterministic fem for elliptic problems with stochastic coefficients. *Comput. Meth. Appl. Mech. Eng.*, 194 :205–228, 2005.

[GD98] R. Ghanem and S. Dham. Stochastic finite element analysis for multiphase flow in heterogeneous porous media. *Tranp. Porous Media*, 32(239), 1998.

[Glo06a] A. Gloria. An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *Multiscale Modeling and Simulation*, 5(3) :996–1043, 2006.

[Glo06b] A. Gloria. A direct approach to numerical homogenization in nonlinear elasticity. *NHM*, 1(1) :109–141, 2006. *Erratum*, pp. 503-514.

[GMNP07] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *M2AN (Math. Model. Numer. Anal.)*, 41(2) :575–605, 2007. (doi : 10.1051/m2an :2007031).

[GO09] A. Gloria and F. Otto. An optimal variance estimate in stochastic homogenization of discrete elliptic equations. *(preprint submitted for publication http ://hal.archives-ouvertes.fr/hal-00383953/en/)*, 2009.

[GP05] M. A. Grepl and A. T. Patera. *A Posteriori* error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. *M2AN (Math. Model. Numer. Anal.)*, 39(1) :157–181, 2005.

[Gre05] M. Grepl. *Reduced-Basis Approximations and A Posteriori Error Estimation for Parabolic Partial Differential Equations*. PhD thesis, Massachusetts Institute of Technology, May 2005.

[Grm98] M. Grmela. Letter to the editor : Comment on "thermodynamics of viscoelastic fluids : The temperature equation" [j. rheol. [bold 42], 999–1019 (1998)]. *Journal of Rheology*, 42(6) :1565–1567, 1998.

[GS90] C. Guillopé and J.-C. Saut. Existence results for the flow of viscoelastic fluids with a differential constitutive law. *Nonlinear Analysis, Theory, Methods & Appl.*, 15 :849–869, 1990.

[GS91] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements : A Spectral Approach*. Springer Verlag, New York, 1991. Revised First Dover Edition, 2003.

[GS92] G. Grimmett and D. Stirzaker. *Probability and Random Processes, second ed.* Oxford Science Publications, Oxford, 1992.

[GSD07] R. Ghanem, G. Saad, and A. Doostan. Efficient solution of stochastic systems : Application to the embankment dam problem. *Structural Safety*, 29 :138–251, 2007.

[Gue99] J.L. Guermond. Stabilization of galerkin approximations of transport equations by subgrid modeling. *Math. Model. Numer. Anal.*, 33(6) :1293–1316, 1999.

[GvL96] G. Golub and C. van Loan. *Matrix computations, third edition*. The Johns Hopkins University Press, London, 1996.

[HDH64] J. Hammersley and eds. D. Handscomb. *Monte Carlo Methods*. Chapman and Hall Ltd, London, 1964.

[HF87] T. J. R. Hughes and L. P. Franca. A new finite element formulation for CFD : VII the Stokes problem with various well-posed boundary conditions : Symmetric formulations that converge for all velocity/pressure spaces. *Comp. Meth. App. Mech. Eng.*, 65 :85–96, 1987.

[HFK05] M. A. Hulsen, R. Fattal, and R. Kupferman. Flow of viscoelastic fluids past a cylinder at high weissenberg number : Stabilized simulations using matrix logarithms. *Journal of Non-Newtonian Fluid Mechanics*, 127(1) :27–39, 2005.

[HKY+09] D.B.P. Huynh, D.J. Knezevic, Y.Chen, J.S. Hesthaven, and A. T. Patera. A natural-norm successive constraint method for inf-sup lower bounds. *Computer Methods in Applied Mechanics and Engineering*, 2009. submitted, preprint version available on URL http ://augustine.mit.edu/methodology/papers/.

[HL07] D. Hu and T. Lelièvre. New entropy estimates for the Oldroyd-B model, and related models. *Commun. Math. Sci.*, 5(4) :906–916, 2007.

[HO08] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *Mathematical Modelling and Numerical Analysis (M2AN)*, 42(3) :277–302, 2008. (doi : 10.1051/m2an :2008001).

[How09] J. S. Howell. Computation of viscoelastic fluid flows using continuation methods. *J. Comput. Appl. Math.*, 225(1) :187–201, 2009.

[HP07a] J. Hao and T.-W. Pan. Simulation for high Weissenberg number viscoelastic flow by a finite element method. *Applied Mathematics Letters*, 20 :988–993, 2007.

[HP07b] D. B. P. Huynh and A. T. Patera. Reduced-basis approximation and *a posteriori* error estimation for stress intensity factors. *Int. J. Num. Meth. Eng.*, 72(10) :1219–1259, 2007. (doi : 10.1002/nme.2090).

[HR82] J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier–Stokes problem I : Regularity of solutions and second-order error estimates for spatial discretization. *SIAM J. Numer. Anal.*, 19 :275–311, 1982.

[HRSP07a] D. B. P. Huynh, G. Rozza, S. Sen, and A. T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *C. R. Acad. Sci. Paris, Analyse Numérique*, 345(8) :473–478, 2007. (doi : 10.1016/j.crma.2007.09.019).

[HRSP07b] D.B.P. Huynh, G. Rozza, S. Sen, and A.T. Patera. A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *C. R. Math. Acad. Sci. Paris*, 345 :473–478, 2007.

[Hul90] M. A. Hulsen. A sufficient condition for a positive definite configuration tensor in differential models. *J. Non-Newtonian Fluid Mech.*, 38 :93–100, 1990.

[HW97] T. Y. Hou and X. H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134 :169–189, 1997.

[Isr92] J. N. Israelachvili. *Intermolecular and Surface Forces, Second Edition : With Applications to Colloidal and Biological Systems (Colloid Science)*. Academic Press, 1992.

[JG04] M. Jardak and R. Ghanem. Spectral stochastic homogenization for of divergence-type pdes. *Computer Methods in Applied Mechanics and Engineering*, 193 :429–447, 2004.

[JKO94] V. Jikov, S. Kozlov, and O. Oleinik. *Homogenization of differential operators and integral functionals*. Springer, Berlin, 1994.

[JLBL04] B. Jourdain, C. Le Bris, and T. Lelièvre. On a variance reduction technique for micro-macro simulations of polymeric fluids. *J. Non-Newtonian Fluid Mech.*, 122 :91–106, 2004.

[JLBLO06] B. Jourdain, C. Le Bris, T. Lelièvre, and F. Otto. Long-time asymptotics of a multiscale model for polymeric fluid flows. *Archive for Rational Mechanics and Analysis*, 181(1) :97–148, 2006.

[JLL04] B. Jourdain, T. Lelièvre, and C. Le Bris. Existence of solution for a micro-macro model of polymeric fluid : the FENE model. *J. Funct. Anal.*, 209 :162–193, 2004.

[Jos90] D. D. Joseph. *Fluid dynamics of viscoelastic liquids*. Applied Mat. Springer Verlag, 1990.

[Jou09] B. Jourdain. *to appear*, chapter Adaptive variance reduction techniques in finance. Radon Series Comp. Appl. Math 8. De Gruyter, 2009.

[JZ08] B. Jin and J. Zou. Inversion of robin coefficient by a spectral stochastic finite element approach. *Journal of Computational Physics*, 227(6) :3282–3306, 2008.

[Kar46] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae*, 37, 1946.

[Kei92] R. A. Keiller. Numerical instability of time-dependent flows. *J. Non-Newtonian Fluid Mech.*, 43 :229–246, 1992.

[Keu90] R. Keunings. Fundamentals of computer modeling for polymer processing. In C.L. Tucker, editor, *Simulation of viscoelastic fluid flow*, pages 402–470. C. Hanser Verlag, 1990.

[Keu00] R. Keunings. A survey of computational rheology. In D. M.Binding et al., editor, *Proc. 13th Int. Congr. on Rheology*, pages 7–14, 2000. British Society of Rheology.

[KL95] Y. Kwon and A. V. Leonov. Stability constraints in the formulation of viscoelastic constitutive equations. *J. Non-Newtonian Fluid Mech.*, 58(1) :25–46, 1995.

[KM05] A. Keese and H. G. Matthies. Hierarchical parallelisation for the solution of stochastic finite element equations. *Computers & Structures*, 83(14) :1033–1047, 2005.

[KMT08] R. Kupferman, C. Mangoubi, and E. Titi. A Beale-Kato-Majda breakdown criterion for an Oldroyd-B fluid in the creeping flow regime. *Comm. Math. Sci.*, 6 :235–256, 2008.

[KP00] P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 2000.

[KP09] D.J. Knezevic and A.T. Patera. A certified reduced basis method for the fokker-planck equation of dilute polymeric fluids : Fene dumbbells in extensional flow. submitted to SIAM Journal of Scientific Computing, 2009.

[KS91] I. Karatzas and S.E. Shreve. *Brownian Motion and Stochastic Calculus*. SpringerVerlag, 1991.

[KS92] N. Kechkar and D. Silvester. Analysis of locally stabilized mixed finite element methods for the Stokes problem. *Mathematics of Computation*, 58(197) :1–10, 1992.

[KS09a] D. Knezevic and E. Süli. A heterogeneous alternating-direction method for a micro-macro dilute polymeric fluid model. *M2AN : Mathematical Modeling and Numerical Analysis*, 2009. Published Online.

[KS09b] D. Knezevic and E. Süli. Spectral Galerkin approximation of Fokker–Planck equations with unbounded drift. *M2AN : Mathematical Modeling and Numerical Analysis*, 43(3), 2009.

[KV02] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM J. Num. Analysis*, 40(2) :492–515, 2002.

[Kwo04] Y. Kwon. Finite element analysis of planar 4 :1 contraction flow with the tensor-logarithmic formulation of differential constitutive equations. *Korea-Australia Rheology Journal*, 16(4) :183–191, 2004.

[LB05] Claude Le Bris. *Systèmes multi-échelles : Modélisation & simulation*, volume 47 of *Mathématiques & applications*. Springer, smai edition, 2005.

[LBLM08] C. Le Bris, T. Lelievre, and Y. Maday. Results and questions on a nonlinear approximation approach for solving high-dimensional partial differential equations. *(preprint submitted for publication http ://hal.inria.fr/inria-00336911/en/)*, 2008.

[LC03] A. Lozinski and C. Chauvière. A fast solver for Fokker-Planck equations applied to viscoelastic flows calculations : 2D FENE model. *J. Comput. Phys.*, 189 :607–625, 2003.

[Lel04] T. Lelièvre. *Problèmes mathématiques et numériques posés par la simulation d'écoulement de fluides polymériques*. PhD thesis, Ecole Nationale des Ponts et Chaussées, 2004. Available at `http ://cermics.enpc.fr/ lelievre/rapports/these.pdf`. In French.

[Leo92] A. I. Leonov. Analysis of simple constitutive equations for viscoelastic liquids. *J. non-newton. fluid mech.*, 42(3) :323–350, 1992.

[LL02] J. H. Lienhard IV and J. H. Lienhard V. *A Heat Transfer Textbook*. Phlogiston Press, Cambridge, Mass., 2002.

[LLPW00] J L Lions, D. Lukkassen, L E Persson, and P. Wall. Reiterated homogenization of monotone operators. *C. R. Acad. Sci. Paris*, 330 :675–680, 2000.

[LLZ05] F.-H. Lin, C. Liu, and P. W. Zhang. On hydrodynamics of viscoelastic fluids. *Comm. Pure Appl. Math.*, 58(11) :1437–1471, 2005.

[LLZ08] Z. Lei, C. Liu, and Y. Zhou. Global solutions for incompressible viscoelastic fluids. *Archive for Rational Mechanics and Analysis*, 188(3) :371–398, 2008.

[LM95] H. Le Meur. Existence locale de solutions des équations d'un fluide viscoélastique avec frontière libre. *C. R. Acad. Sci. Paris Sér. I Math.*, 320(1) :125–130, 1995.

[LM00] P.-L. Lions and N. Masmoudi. Global solutions for some Oldroyd models of non-Newtonian flows. *Chin. Ann. Math. Ser. B*, 21 :131–146, 2000.

[LM07] P.-L. Lions and N. Masmoudi. Global existence of weak solutions to some micro-macro models. *C. R. Math. Acad. Sci. Paris*, 345 :15–20, 2007.

[LO03] A. Lozinski and R. G. Owens. An energy estimate for the Oldroyd-B model : theory and applications. *J. Non-Newtonian Fluid Mech.*, 112 :161–176, 2003.

[Loè78] M. Loève. *Probability Theory*, volume I-II. Springer, New York, 1978.

[Loz03] A. Lozinski. *Spectral methods for kinetic theory models of viscoelastic fluids*. PhD thesis, EPFL, 2003.

[LR74] P. Lesaint and P.A. Raviart. *Mathematical Aspects of Finite Elements in Partial Differential Equations*, chapter On a Finite Element Method for Solving the Neutron Transport Equation, pages 89–123. Academic Press, (c. de boor ed.) edition, 1974.

[LX06] Y. J. Lee and J. Xu. New formulations, positivity preserving discretizations and stability analysis for non-Newtonian flow models. *Comput. Methods Appl. Mech. Engrg.*, 195 :1180–1206, 2006.

[LZ07] T. Li and P.-W. Zhang. Mathematical analysis of multi-scale models of complex fluids. *Commun. Math. Sci.*, 5 :1–51, 2007.

[LZZ04] T. Li, H. Zhang, and P.-W. Zhang. Local existence for the dumbbell model of polymeric fuids. *Comm. Partial Differential Equations*, 29 :903–923, 2004.

[Mad06] Y. Maday. Reduced–basis method for the rapid and reliable solution of partial differential equations. In *Proceedings of International Conference of Mathematicians, Madrid*. European Mathematical Society Eds., 2006.

[Mas08] N. Masmoudi. Well posedness of the FENE dumbbell model of polymeric flows. *Comm. Pure Appl. Math.*, 61 :1685–1714, 2008.

[MBS00] A M Matache, I. Babuska, and C. Schwab. Generalized p-FEM in homogenization. *Numerische Mathematik*, 86 :319–375, 2000.

[MC87] J.M. Marchal and M.J. Crochet. A new mixed finite element for calculating viscoelastic flow. *J. Non-Newtonian Fluid Mech*, 26 :77–114, 1987.

[MHK09] C. Mangoubi, M. A. Hulsen, and R. Kupferman. Numerical stability of the method of brownian configuration fields. *Journal of Non-Newtonian Fluid Mechanics*, 157(3) :188–196, 2009.

[MHZ05] L. Mathelin, M.Y. Hussaini, and T.A. Zang. Stochastic approaches to uncertainty quantification in cfd simulations. *Numerical Algorithms*, 38 :209–236, 2005.

[MK05] H.G. Matthies and A. Keese. Galerkin methods for linear and nonlinear elliptic stochastic pdes. *Comput. Meth. Appl. Mech. Eng.*, 194 :1295–1331, 2005.

[MMO+00] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D.V. Rovas. Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems. *C. R. Acad. Sci. Paris, Série I*, 331(2) :153–158, 2000.

[MMP01] L. Machiels, Y. Maday, and A. T. Patera. Output bounds for reduced-order approximations of elliptic partial differential equations. *Comp. Meth. Appl. Mech. Engrg.*, 190(26-27) :3413–3426, 2001.

[MNK08] P. Surya Mohan, P. B. Nair, and A. J. Keane. Multi-element stochastic reduced basis methods. *Computer Methods in Applied Mechanics and Engineering*, 197(17-18) :1495–1506, 2008.

[MNPP09] Y. Maday, N. C. Nguyen, A. T. Patera, and G. Pau. A general, multipurpose interpolation procedure : the magic points. *Communications on Pure and Applied Analysis*, 8(1) :383–404, 2009.

[MO95] M. Melchior and H.C. Öttinger. Variance reduced simulations of stochastic differential equations. *J. Chem. Phys.*, 103 :9506–9509, 1995.

[MPS88] K. W. Morton, A. Priestley, and E. Süli. Convergence analysis of the Lagrange-Galerkin method with non-exact integration. *M2AN*, 22(4) :625–653, 1988.

[MPT02a] Y. Maday, A. T. Patera, and G. Turinici. *A Priori* convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. *Journal of Scientific Computing*, 17(1-4) :437–446, 2002.

[MPT02b] Y. Maday, A.T. Patera, and G. Turinici. Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. *C. R. Acad. Sci. Paris, Ser. I*, 335(3) :289–294, 2002.

[MT86] R. Mneimne and F. Testard. *Introduction a la théorie des groupes de Lie classiques.* Hermann, 1986.

[MT06] G.N. Milstein and M.V. Tretyakov. Practical variance reduction via regression for simulating diffusions. Technical Report MA-06-19, School of Mathematics and Computer Science, University of Leicester, 2006.

[New94] N. J. Newton. Variance reduction for simulated diffusions. *SIAM J. Appl. Math.*, 54(6) :1780–1805, 1994.

[New96] A. J. Newman. Model reduction via the Karhunen-Loeve expansion part i : an exposition. Technical Report 96-322, Institute for System Research, University of Maryland, 1996.

[NK02] P. B. Nair and A. J. Keane. Stochastic reduced basis methods. *AIAA Journal*, 40(8) :1653–1664, 2002.

[Nou07] A. Nouy. A generalized spectral decomposition technique to solve a class of linear stochastic partial differential equations. *Computer Methods in Applied Mechanics and Engineering*, 196(45-48) :4521–4537, 2007.

[Nou08] A. Nouy. Generalized spectral decomposition method for solving stochastic finite element equations : Invariant subspace problem and dedicated algorithms. *Computer Methods in Applied Mechanics and Engineering*, 197(51-52) :4718–4736, 2008.

[NP80] A. K. Noor and J. M. Peters. Reduced basis technique for nonlinear analysis of structures. *AIAA Journal*, 18(4) :455–462, 1980.

[NRHP09] N. C. Nguyen, G. Rozza, D. B. P. Huynh, and A. T Patera. Reduced basis approximation and a posteriori error estimation for parametrized parabolic pdes; application to real-time bayesian parameter estimation. In Biegler, Biros, Ghattas, Heinkenschloss, Keyes, Mallick, Tenorio, van Bloemen Waanders, and Willcox, editors, *Computational Methods for Large Scale Inverse Problems and Uncertainty Quantification*, John Wiley & Sons, UK, 2009.

[NRP09] N. C. Nguyen, G. Rozza, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for the time-dependent viscous burgers equation. *Calcolo*, (46) :157–185, 2009.

[NTW08] F. Nobile, R. Tempone, and C.G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46 :2309–2345, 2008.

[NVP05a] N. C. Nguyen, K. Veroy, and A. T. Patera. Certified real-time solution of parametrized partial differential equations. In S. Yip, editor, *Handbook of Materials Modeling*, pages 1523–1558. Springer, 2005.

[NVP05b] N. C. Nguyen, K. Veroy, and A.T. Patera. *Certified real-time solution of parametrized partial differential equations*, pages 1523–1558. Springer, 2005. (S. Yip, editor).

[Ö96] H. C. Öttinger. *Stochastic Processes in Polymeric Fluids*. Springer Verlag, Berlin - Heidelberg, 1996.

[Ö05] H. C. Öttinger. *Beyond equilibrium thermodynamics*. John Wiley, New Jersey, 2005.

[Oks03] B. Oksendal. *Stochastic Differential Equations. An Introduction with Applications.* Springer-Verlag, 2003.

[OP02] R. G. Owens and T. N. Philips. *Computational rheology*. Imperial College Press / World Scientific, 2002.

[OSW99] M. Ostoja-Starzewski and X. Wang. Stochastic finite elements as a bridge between random material microstructure and global response. *Computer Methods in Applied Mechanics and Engineering*, 168(1-4) :35–49, 1999.

[OvdBH97] H. C. Öttinger, B. H. A. A. van den Brule, and M. A. Hulsen. Brownian configuration fields and variance reduced CONNFFESSIT. *J. Non-Newtonian Fluid Mech.*, 70(30) :255 – 261, 1997.

[PG00] M. F. Pellissetti and R. G. Ghanem. Iterative solution of systems of linear equations arising in the context of stochastic finite elements. *Adv. Eng. Softw.*, 31(8-9) :607–616, 2000.

[Pin85] A. Pinkus. *n-Widths in Approximation Theory*. Springer, 1985.

[Pir82] O. Pironneau. On the transport-diffusion algorithm and its application to the Navier-Stokes equations. *Numer. Math.*, 3 :309–332, 1982.

[Pir08] O. Pironneau. Calibration of options on a reduced basis. *Journal of Computational and Applied Mathematics*, In Press, Corrected Proof :–, 2008.

[Por85] T. A. Porsching. Estimation of the error in the reduced basis method solution of nonlinear equations. *Mathematics of Computation*, 45(172) :487–496, 1985.

[PP05] G. Pagès and J. Printems. Functional quantization for numerics with an application to option pricing. *Monte Carlo Methods and Appl.*, 11(4) :407–446, 2005.

[PR07a] A. T. Patera and E. M. Rønquist. Reduced basis approximations and *a posteriori* error estimation for a Boltzmann model. *Computer Methods in Applied Mechanics and Engineering*, 196 :2925–2942, 2007.

[PR07b] A. T. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. Copyright MIT, 2006–2007. To appear in MIT Pappalardo Monographs in Mechanical Engineering.

[Pro08] Andreas Prohl. Convergent finite element discretizations of the nonstationary incompressible magnetohydrodynamics system. *ESAIM : M2AN*, 42(6) :1065–1087, nov 2008.

[PRV+02] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. T. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations : Reduced-basis output bounds methods. *Journal of Fluids Engineering*, 124(1) :70–80, 2002.

[Qua09] A. Quarteroni. *Numerical Models for Differential Problems*, volume 2 of *Modeling, Simulation and Applications*. Springer, a. quarteroni et al. edition, 2009.

[Ren00] M. Renardy. *Mathematical Analysis of Viscoelastic Flows*, volume 73 of *CBMS-NSF Conference Series in Applied Mathematics*. SIAM, 2000.

[RH88] J. M. Rallison and E. J. Hinch. Do we understand the physics in the constitutive equation ? *J. Non-Newtonian Fluid Mech.*, 29 :37–55, 1988.

[RHP08] G. Rozza, D.B.P. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations — application to transport and continuum mechanics. *Arch. Comput. Methods Eng.*, 15(3) :229–275, 2008.

[S88] E. Süli. Convergence and nonlinear stability of the Lagrange-Galerkin method for the Navier-Stokes equations. *Numer. Math.*, 53 :459–483, 1988.

[San99] D. Sandri. Non integrable extra stress tensor solution for a flow in a bounded domain of an Oldroyd fluid. *Acta Mech.*, 135(1–2) :95–99, 1999.

[SB95] R Sureshkumar and A N Beris. Effect of artificial stress diffusivity on the stability of numerical calculations and the flow dynamics of time-dependent viscoelastic flows. *Journal of Non-Newtonian Fluid Mech*, 60 :53–80, 1995.

[SBL06] B. Sudret, M. Berveiller, and M. Lemaire. A stochastic finite element procedure for moment and reliability analysis. *Rev. Eur. Méca. Num*, 15(7-8) :825–866, 2006.

[Sch06] J. D. Schieber. Generalized Brownian configuration field for Fokker–Planck equations including center-of-mass diffusion. *J. Non-Newtonian Fluid Mech.*, 135 :179–181, 2006.

[Sen08] S. Sen. Reduced-basis approximation and a posteriori error estimation for many-parameter heat conduction problems. *Numerical Heat Transfer, Part B : Fundamentals*, 54(5), 2008.

[SF73a] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, 1973.

[SF73b] W.G. Strang and G.J. Fix. *An Analysis of the Finite Element Method*. Wellesley-Cambridge Press, 1973.

[SG04] C. Soize and R. Ghanem. Physical systems with random uncertainties : Chaos representations with arbitrary probability measure. *SIAM Journal on Scientific Computing*, 26(2) :395–410, 2004.

[Sim87] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Math. Pura. Appl.*, 146 :65–96, 1987.

[SNK06] S. K. Sachdeva, P. B. Nair, and A. J. Keane. Hybridization of stochastic reduced basis methods with polynomial chaos expansions. *Probabilistic Engineering Mechanics*, 21 :182–192, 2006.

[SR94] T. Sato and M. S. Richardson. Numerical simulation method for viscoelastic flows with free surfaces - fringe element generation method. *International Journal for Numerical Methods in Fluids*, 19(7) :555–574, 1994.

[SST08] Ch. Schwab, E. Süli, and R.-A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. *Math. Models Methods Appl. Sci.*, 42 :777–820, 2008.

[ST06] C. Schwab and R.A. Todor. Karhunen-loève approximation of random fields by generalized fast multipole methods. *Journal of Computational Physics*, 217(1) :100–122, 2006.

[SV85] L. R. Scott and M. Vogelius. Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials. *RAIRO Model. Math. Anal. Numer.*, 19 :111–143, 1985.

[SVH+06] S. Sen, K. Veroy, D. B. P. Huynh, S. Deparis, N. C. Nguyen, and A. T. Patera. "Natural norm" *a posteriori* error estimators for reduced basis approximations. *Journal of Computational Physics*, 217 :37–62, 2006.

[Tem66] R. Temam. Sur l'approximation des équations de Navier-Stokes. *C. R. Acad. Sc. Paris, Série A*, 262 :219–221, 1966.

[Tem84] R. Temam. *Navier–Stokes Equations. Theory and Numerical Analysis (Third Edition)*, volume 2 of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, 1984.

[TS07a] B. Thomases and M. Shelley. Emergence of singular structures in Oldroyd-B fluids. *Phys. Fluids*, 19(12) :103103.1–103103.12, 2007.

[TS07b] R.A. Todor and C. Schwab. Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA Journal of Numerical Analysis*, 27 :232–261, 2007.

[VJ05] G. Venkiteswaran and M. Junk. Quasi-monte carlo algorithms for diffusion equations in high dimensions. *Math. Comput. Simul.*, 68(1) :23–41, 2005.

[V.N08] V.N.Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3 :33–107, 2008.

[VP05] K. Veroy and A. T. Patera. Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations ; Rigorous reduced-basis *a posteriori* error bounds. *International Journal for Numerical Methods in Fluids*, 47 :773–788, 2005.

[VPRP03a] K. Veroy, C. Prud'homme, D. V. Rovas, and A. T. Patera. *A Posteriori* error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*, 2003. Paper 2003-3847.

[VPRP03b] K. Veroy, C. Prud'homme, D.V. Rovas, and A. T. Patera. A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations. In *16th AIAA Computational Fluid Dynamics Conference*, June 2003.

[VRP02] K. Veroy, D. Rovas, and A. T. Patera. *A Posteriori* error estimation for reduced-basis approximation of parametrized elliptic coercive partial differential equations : "convex inverse" bound conditioners. *ESAIM : Control, Optimization and Calculus of Variations*, 8 :1007–1028, 2002.

[WH98a] P. Wapperom and M. A. Hulsen. Response to "comment on : 'thermodynamics of viscoelastic fluids : The temperature equation' " [j. rheol. [bold 42], 1565–1567 (1998)]. *Journal of Rheology*, 42(6) :1569–1570, 1998.

[WH98b] P. Wapperom and M. A. Hulsen. Thermodynamics of viscoelastic fluids : the temperature equation. *J. Rheol.*, 42(5) :999–1019, 1998.

[Wie38] N. Wiener. The homogeneous chaos. *Am. J. Math.*, 60 :897–936, 1938.

[WK05] X. Wan and G.E. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *Journal of Computational Physics*, 209 :617–642, 2005.

[WK09] X. Wan and G.E. Karniadakis. Error control in multi-element generalized polynomial chaos method for elliptic problems with random coefficients. *Communications in Computational Physics*, 5(2–4) :793–820, 2009.

[WKL00] P. Wapperom, R. Keunings, and V. Legat. The backward-tracking lagrangian particle method for transient viscoelastic flows. *J. Non-Newtonian Fluid Mech.*, 91 :273–295, 2000.

[XH05] D. Xiu and J.S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3) :1118–1139, 2005.

[XK02] D. Xiu and G.E. Karniadakis. The Wiener–Askey polynomial chaos for stochastic differential equations. *Journal on Scientific Computing*, 24 :619–644, 2002.

[Xu07] X. Frank Xu. A multiscale stochastic finite element method on elliptic problems involving uncertainties. *Computer Methods in Applied Mechanics and Engineering*, 196(25-28) :2723–2736, 2007.

[YH07] J. Yvonnet and Q. C. He. The reduced model multiscale method (r3m) for the non-linear homogenization of hyperelastic media at finite strains. *Journal of Computational Physics*, 223 :341–368, April 2007.