Reduction of complex models for simulation and estimation Application to cardiac modelling

Asven Gariah

université paris 6 – pierre et marie curie
école doctorale de sciences mathématiques de paris centre

mathématiques appliquées

Réduction de modèles complexes pour la simulation et l'estimation Application à la modélisation cardiaque

Asven Gariah

Thèse de doctorat en mathématiques,
soutenue le 9 novembre 2011.

Président	M. Yvon Maday
Rapporteur	M. Claudio Canuto
Examinateurs	M. Martin Grepl
	M. Cyril Touzé
Directeur	M. Dominique Chapelle
Co-directeur	M. Jacques Sainte-Marie

inria paris–rocquencourt – projet macs
domaine de voluceau, bp 105, 78153 le chesnay

octobre 2008 – septembre 2011

Asven Gariah

Réduction de modèles complexes pour la simulation et l'estimation
Application à la modélisation cardiaque

Thèse présentée en vue d'obtenir le grade de docteur de l'Universtié Pierre et Marie Curie en mathématiques,

préparée au sein du projet Macs à Inria Paris–Rocquencourt sous la direction de MM. Dominique Chapelle et Jacques Sainte-Marie,

et soutenue publiquement le 9 novembre 2011 à l'Université Pierre et Marie Curie.

UPMC SORBONNE UNIVERSITÉS

informatiques mathématiques Inria

Résumé

Ce mémoire analyse et valide des applications possibles de méthodes de *réduction de modèle* pour la simulation directe, et la résolution de problèmes inverses d'estimation de paramètres sur des modèles complexes. Il se concentre sur la *réduction par* proper orthogonal decomposition *(POD)*, et ses extensions.

On démontre d'abord de nouvelles estimations a priori pour *l'erreur de réduction* sur des problèmes abstraits types (paraboliques et hyperboliques, linéaires ou avec non-linéarités lipschitziennes), validées dans de nombreux cas non linéaires. On évite notamment le problème de contrôle des termes d'ordre élevé par l'exploitation d'une suite spécifique de normes de projecteurs.

Puis, pour couvrir les systèmes dépendant de paramètres, et par des résultats d'interpolation, on adapte la méthode précédente en *réduction par multi-POD*. On étend aussi, au prix d'un terme additif, les estimations a priori précédentes pour *l'erreur maximum de réduction* sur une plage paramétrique donnée. On illustre la puissance de la méthode sur le système électrophysiologique de FitzHugh–Nagumo, fortement sensible aux variations paramétriques.

On valide enfin numériquement les versions réduites, toujours avec la réduction par multi-POD, de problèmes d'estimation de paramètres : de type variationnel avec le système de FitzHugh–Nagumo, et de type séquentiel (filtrage « kalmanien ») avec un modèle mécanique de cœur (multi-échelles, 3D, grandes déformations). En particulier, la méthode présente une efficacité et une robustesse similaires à celles obtenues pour les problèmes directs.

Mots-clés : analyse numérique, réduction de modèle par POD, approximation de Galerkin, estimation de paramètres, FitzHugh–Nagumo, modélisation cardiaque.

Abstract

Reduction of complex models for simulation and estimation
Application to cardiac modelling

This report analyzes and validates possible applications of some model reduction methods for direct simulations, and the solving of inverse problems of parameter estimation on complex models. It focuses on the *reduction by proper orthogonal decomposition (POD)*, and its extensions.

We start by proving new a priori estimates for the *reduction error* on typical abstract problems (parabolic and hyperboic, linear or with Lipschitz-continuous nonlinearities), also validated in various nonlinear cases. In

particular, we avoid the issue of controlling the high-order terms by using a specific sequence of projector norms.

Then, in order to tackle parameter-dependent systems, and using some interpolation results, we adapt the previous method in a *multi-POD reduction* strategy. We also extend the previous estimates for the *maximum reduction error* over a given parameter range, at the cost of an additive term. We illustrate the power of the method on the electrophysiology FitzHugh–Nagumo system, known to be highly parameter-sensitive.

Finally, we numerically validate the reduced versions, still with the multi-POD reduction, of some parameter estimation problems : of variational kind with the FitzHugh–Nagumo system, and of sequential kind (Kalmanian filtering) with a mechanical model of a heart (multi-scale, 3D, large displacements). In particular, we exhibit similar efficiency and robustness of the method as with direct problems.

Keywords: numerical analysis, reduced-order modelling with POD, Galerkin approximation, parameter estimation, FitzHugh–Nagumo, cardiac modelling.

Table des matières

Chapitre 1

Introduction

La puissance croissante des outils informatiques de simulation à disposition des ingénieurs, à la fois en termes de capacité de stockage et de rapidité d'exécution, repousse sans cesse la limite des calculs possibles. Cet avantage technologique accompagne alors une plus grande facilité d'analyse de problèmes mathématiques appliqués, en particulier pour ceux qui reposent sur une solution de système d'équations aux dérivées partielles. En effet, de telles solutions peuvent être discrétisées plus finement, et à supposer qu'elles passent convenablement à la limite avec les pas de discrétisation, être évaluées de manière plus précise, plus rapidement.

Cependant, l'utilité d'améliorer de telle façon un calcul numérique soulève deux questions.

D'une part, de tels problèmes appliqués formulent une modélisation particulière de systèmes physiques. Ils représentent une vision idéale, à la fois synthétique et suffisamment complète, pour reproduire les phénomènes observés. La course à la précision numérique dans le cadre d'un système complexe est donc illusoire en soi.

D'autre part, du point de vue de l'analyse numérique, l'amélioration réelle d'un calcul se mesure plus au rapport entre la précision gagnée et l'augmentation du temps pour l'exécuter, plutôt qu'à la durée seule. Or, pour un nombre de degrés de liberté N_h d'un système, comme on le développe plus loin, le coût numérique évolue au moins en N_h^2 pour une simulation directe. Pire, une méthode d'optimisation pour retrouver la valeur d'un paramètre inconnu du système d'équations, et ainsi résoudre un exemple de formulation de problème inverse, peut nécessiter un grand nombre d'évaluations de telles simulations. Ainsi, à la limite, pour un gain de précision de plus en plus faible, par exemple pour obtenir un nouveau chiffre significatif à chaque fois, le calcul demande une durée de plus en plus longue, et au final inappropriée.

Puisqu'une telle démarche de raffinement n'évite pas une grande accumulation des informations à traiter, on peut envisager au contraire de *simplifier* les modèles, c'est-à-dire de se diriger vers une description où le nombre de degrés de liberté raisonnablement borné, serait jugé suffisant et fiable. C'est l'objet des méthodes de *réduction de modèle.* Celles-ci

concentrent un intérêt particulier depuis une trentaine d'années. On en présente ci-dessous deux grandes familles, chacune déployée dans une littérature appliquée abondante : la réduction par base réduite *(reduced basis)*, et celle par analyse en composante principale *(POD*, pour *proper orthogonal decomposition*, entre autres dénominations), qu'on choisit de développer et analyser théoriquement, ainsi qu'étendre et appliquer numériquement, dans ce rapport. Ces méthodes introduisent en particulier des nombres très réduits de degrés de liberté, et se montrent en effet beaucoup plus puissantes en ce sens que des décompositions plus classiques telles que l'analyse modale peut en fournir.

On commence par présenter la motivation concrète, et étudiée dans ce rapport, pour appliquer ces méthodes de réduction à un modèle dynamique et électromécanique de cœur. On insiste sur le jeu très complexe qui intervient dans ce modèle entre les déplacements, les vitesses, les variables internes liées à la loi de comportement et les paramètres, ainsi que sur l'application concrète, appuyée sur des données médicales réelles, qui en est réalisée. Notre étude repose sur un modèle développé dans l'équipe Macs (Inria Rocquencourt) dans le cadre du projet de recherche CardioSense3D (Inria, prix ARTS 2007, voir [SMCCS06]) et du projet européen de développement industriel euHeart.

En revenant sur un terrain plus abstrait, on dresse ensuite un état de l'art des méthodes de réduction précitées, en mettant l'accent sur la méthode retenue par POD.

Enfin, on annonce l'organisation du rapport en détaillant les contributions. Notamment, on intègre sous forme légèrement remaniée un article en voie de publication dans le journal *ESAIM : M²AN (Modélisation mathématique et analyse numérique)*.

1.1 Objectif d'application à un modèle cardiaque

Les développements industriels autour de modalités d'imagerie médicale de pointe, telles que l'IRM ou les ultrasons, ont permis des avancées significatives pour le *diagnostic* médical, à travers par exemple l'IRM ou l'échographie. Le traitement de données physiologiques toujours plus précises à disposition des médecins affine donc la compréhension, la détection, et ainsi le *pronostic* de certaines maladies. À ce titre, l'étude pour la simulation du système cardio-vasculaire et la prise en compte numérique de ses cas pathologiques (anévrisme, ischémie), présente un double intérêt.

Premièrement, elle intervient dans un contexte où les maladies cardio-vasculaires constituent un problème de santé majeur. En effet, l'Organisation Mondiale de la Santé précise qu'elles sont « une cause majeure d'incapacité et de décès prématurés dans le monde entier » [Wor07].

Deuxièmement, c'est un défi de modélisation biomathématique. En effet, les interactions physiques entre la dynamique du sang, les propriétés mécaniques et structurelles du tissu musculaire, ainsi que l'activité électrique, mis en jeu au cours des cycles cardiaques, restent très complexes à décrire. Elles couplent de plus nécessairement des considérations entre les échelles moléculaire et macroscopique. On donne une description plus précise des couches élémentaires du modèle aux chapitres 3 et 6.

En articulant ces interactions comme un couplage de sous-modèles, elles introduisent de nombreux paramètres, que les simulations directes exigent de déterminer. Or, certains paramètres tels que le champ de contractilité du muscle cardiaque, étudié au chapitre 6 et propriété intrinsèque du sous-modèle mécanique, ne sont pas directement mesurables. De même, lorsqu'on cherche les plages de variations des variables de vitesse, de pression ou de tension électrique, les seules mesures dont on dispose se limitent dans le meilleur des cas à des moyennes spatiales. Par exemple, en ce qui concerne le champ électrique tridimensionnel dans l'ensemble des fibres musculaires, on n'en mesure l'effet qu'en des points éloignés du cœur à travers les électrocardiogrammes.

Ainsi, d'un point de vue technique, la calibration de ces paramètres, étape incontournable pour aborder le modèle, pose alors de nombreuses difficultés. D'un point de vue théorique, toutes ces incertitudes forment un remarquable problème *d'estimation de paramètres*. Ce type de problème inverse présente en soi une complexité de modélisation et un coût numérique largement supérieurs aux problèmes directs.

1.2 Naissance et essor de deux méthodes majeures de réduction de modèle

Afin de présenter et comparer les deux familles majeures de méthodes de réduction de modèle identifiées dans la littérature, on donne un premier cadre mathématique, pour des raisons introductives, uniquement formel. Les variables entre autres mécaniques, électriques, thermodynamiques ou chimiques d'un modèle biomathématique sont régies par des équations aux dérivées partielles et non-linéaires, généralement paraboliques ou hyperboliques. Pour suivre par exemple le premier cas, étant donné un domaine spatial $\Omega \subset \mathbb{R}^d$, les équations prennent la forme

$$\frac{\partial u}{\partial t} - \operatorname{div}(D\nabla u) = f(t,u), \quad \text{dans } \Omega,$$

avec $D(x) \in \mathbb{R}^{d\times d}$ uniformément elliptique, ou (comme pour le principe des travaux virtuels en mécanique) sous forme variationnelle

$$\text{trouver} \quad u(t) \in V \quad \text{tel que}$$
$$\frac{\partial}{\partial t}(u(t),\, v) + a(u(t),\, v) = (f(t,\, u(t)),\, v), \quad \forall v \in V,$$

où V est un sous-espace de $H^1(\Omega)$ dépendant des conditions aux limites pour $u(t)$, et

$$(w,\, v) = \int_\Omega wv, \quad a(w,\, v) = \int_\Omega (\nabla w)^\top D \nabla v.$$

Pour la discrétisation spatiale, on s'intéresse uniquement à la méthode des éléments finis dans ce rapport. Celle-ci définit un sous-espace V_h de V engendré par un nombre fini N_h de fonctions de forme, et la solution semi-discrète $u_h(t)$ correspondante par l'approximation de Galerkin

$$\text{trouver} \quad u_h(t) \in V_h \quad \text{tel que}$$
$$\frac{\partial}{\partial t}(u_h(t), v_h) + a(u_h(t), v_h) = (f(t, u_h(t)), v_h), \quad \forall v_h \in V_h. \tag{1.1}$$

Les méthodes courantes de discrétisation temporelle (θ-méthode, schéma de Newmark), conduisent à résoudre à chaque pas de temps des systèmes non-linéaires de dimension finie et possiblement grande N_h, qui s'écrivent

$$\text{trouver} \quad U_h^{n+1} \in \mathbb{R}^{N_h} \quad \text{tel que} \quad G^n(U_h^{n+1}) = H^n, \tag{1.2}$$

où le vecteur $H^n \in \mathbb{R}^{N_h}$ et l'application non-linéaire $G^n : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ sont entièrement déterminés à l'instant t^n. À ce stade, on peut alors rapidement rendre explicite le résultat de coût en N_h^2 évoqué plus haut. On résout classiquement (1.2) par algorithme de Newton. Dans sa forme la plus simple, chaque sous-étape k de l'algorithme à l'itération n s'écrit comme le système non-linéaire

$$\text{trouver} \quad U_h^{n,k+1} \in \mathbb{R}^{N_h} \quad \text{tel que}$$
$$\mathrm{d}G^n(U^{n,k}) \cdot (U^{n,k+1} - U^{n,k}) = H^n - G^n(U^{n,k}).$$

Comme les éléments finis rendent la matrice $\mathrm{d}G(U^{n,k}) \in \mathbb{R}^{N_h\times N_h}$ creuse, la première phase d'assemblage coûte $O(N_h)$ opérations, et la seconde phase d'inversion $O(N_h^2)$ opérations.

Les méthodes de réduction par base réduite et par POD partagent l'idée de réduire ce coût par sous-approximation de Galerkin de (1.1) sur un sous-espace $V^l \subset V_h$ de dimension

$$l \ll N_h.$$

En revanche, les critères pour choisir un *espace de réduction* V^l adéquat, ainsi que ceux qui évaluent la qualité des réductions correspondants, diffèrent. On les présente ci-dessous.

1.2.1 Réduction de modèle par base réduite

La terminologie de base réduite apparaît initialement dans un article de Noor et Peters [NP80] pour l'analyse des grandes déflexions, non-linéaires, des matériaux composites utilisés dans l'industrie aérospatiale. La méthode pour choisir l'espace de réduction est alors très proche d'une analyse modale non-linéaire, puisqu'il est engendré par des vecteurs provenant d'une analyse de Rayleigh-Ritz. Plusieurs publications qui lui succèdent exploitent cette méthode de manière empirique.

Maday et Ronquist [MR02] clarifient mathématiquement la notion de base réduite pour une solution, dans un premier temps statique, de problème elliptique présentant une dépendance affine en des fonctions d'un paramètre. Un premier point fort de la méthode ressort du procédé en deux étapes *offline* et *online* distinctes, qui permet comme détaillé ci-dessous, au prix d'un coût de préparation considérable, une évaluation approchée rapide et contrôlable de la solution en des valeurs quelconques du paramètre. De plus, par la suite (voir par exemple [MPR02]), des estimateurs a posteriori *rigoureux*, c'est-à-dire qui forment des majorations serrées de l'erreur de réduction, et qui constituent un deuxième point fort, sont développés dans le même cadre. Si ces estimateurs évaluent la qualité de réduction de la méthode, ils permettent surtout, par un algorithme glouton, de l'améliorer de manière constructive en enrichissant convenablement la base réduite. Pour une revue détaillée, agrémentée de nombreux résultats numériques, sur les propriétés et les possibilités des bases réduites pour les équations elliptiques paramétriques, on renvoie à un article de Rozza et al. [RHP07].

Des extensions de la méthode ainsi formalisée et d'estimateurs a posteriori correspondants sont alors proposées. Pour les cas statiques non-linéaires, Barrault et al. [BMNP04] emploient une technique d'interpolation *(magic points)*, qui ramène les opérateurs à la forme de dépendance affine précédente. Pour les problèmes paraboliques linéaires discrétisés en temps, Grepl et Patera [GP05] développent un schéma réduit naturel, et traitent la variable temporelle comme un paramètre pour la méthode d'enrichissement de la base.

Pour illustrer la méthode en poursuivant l'exemple (1.1), on se place dans un cas statique linéaire ($f(t,u)$ devient $f \in V'$), et on se donne une partition

$$\overline{\Omega} = \bigcup_{k=1}^{p} \overline{\Omega}_k$$

du domaine spatial. On peut ainsi introduire une dépendance affine en paramètre sur la forme bilinéaire $a(\cdot,\cdot)$ de la manière suivante : en supposant que la matrice de diffusion $D(x)$ est constante et scalaire par sous-domaine

Ω_i, i.e.

$$D(x) = \Big(\sum_{k=1}^{p} \mathbf{1}_{\Omega_k}(x) D^{(k)} \Big) \mathrm{Id}_d,$$

formant un vecteur paramètre

$$D = (D^{(1)}, \ldots, D^{(p)}) \quad \in \mathcal{D} = (0, \infty)^p.$$

Le problème elliptique qui résulte, de solution $u(D)$, s'écrit

$$\begin{aligned} &\text{trouver} \quad u_h(D) \in V_h \quad \text{tel que} \\ &a(u_h(D), v_h; D) = (f, v_h), \quad \forall v_h \in V_h, \end{aligned} \tag{1.3}$$

où la forme bilinéaire se décompose comme

$$a(w, v, D) = \sum_{k=1}^{p} D^{(k)} a_k(w, v), \quad a_k(w, v) = \int_{\Omega_k} (\nabla w)^\top \nabla v.$$

La première phase *offline* calcule pour un certain nombre $l \ll N_h$, que la méthode ne peut pas déterminer a priori, une *base réduite* (en réalité une famille très probablement libre) de solutions non réduites

$$\mathcal{S}^l = (u_h(D_1), \ldots, u_h(D_l)),$$

où $D_1, \ldots, D_l \in \mathcal{D}$ sont des valeurs particulières et distinctes de vecteur paramètre. On définit alors l'espace de réduction V^l comme le sous-espace généré par $\mathcal{S}^l$, qui nécessite un coût de résolution de $O(l \times N_h^2)$ opérations, et un coût de stockage de $l \times N_h$ valeurs.

Pour tout $D \in \mathcal{D}$, l'approximation de la solution $u_h(D)$ du problème (1.3) est donc cherchée sous la forme

$$u^l(D) = \sum_{i=1}^{l} \alpha^{(i)}(D) u_h(D_i),$$

et suivant l'approximation de Galerkin de (1.3) sur V^l, i.e.

$$\begin{aligned} &\text{trouver} \quad u^l(D) \in V^l \quad \text{tel que} \\ &a(u^l(D), v^l; D) = (f, v^l), \quad \forall v^l \in V^l. \end{aligned}$$

Or, étant donné les hypothèses de dépendance affine, le calcul de $u^l(D)$ revient à très faible coût, puisque matriciellement, cela revient à résoudre le système

$$K^l(D)\, \alpha(D) = F^l,$$

où la matrice $K^l(D)$, pleine et de petite taille $l \times l$, et le vecteur force $F^l \in \mathbb{R}^l$ sont définis par

$$K^l(D) = \sum_{k=1}^{p} D^{(k)} K_k^l, \quad K_{k,ij}^l = a_k(u_h(D_j),\, u_h(D_i)) \quad (1 \le i, j \le l),$$
$$F^{l,(i)} = (f,\, u_h(D_i)) \quad (1 \le i \le l).$$

Ainsi, si on assemble puis stocke *offline* les matrices K_k^l, $1 \le k \le p$ et le vecteur force F^l, qui sont *indépendants* du paramètre, le coût *online* d'évaluation de l'approximation $u^l(D)$ pour $D \in \mathcal{D}$ quelconque se réduit à $O(l^3)$ opérations, sans avoir à revenir à la dimension N_h. Cela représente un gain considérable pour des valeurs raisonnables de l.

Pour garantir la qualité de l'approximation

$$R^l(D) = u(D) - u^l(D) \ll u(D),$$

on recourt à une estimation a posteriori [RHP07, Sec. 9]. Celle-ci découle de représentations de Riesz particulières sur la décomposition

$$a(R^l(D),\, v;\, D) = (f,\, v) - \sum_{i=1}^{l} \alpha^{(i)}(D) \sum_{k=1}^{p} D^{(k)} a_k(u(D_i),\, v),$$

et prend la forme d'une quantité scalaire $\Delta^l(D)$ vérifiant

$$\|R^l(D)\|_D = a(R^l(D),\, R^l(D);\, D)^{1/2} \le \Delta^l(D).$$

Comme précédemment et sous la condition de certains stockages pertinents *offline*, l'estimateur a posteriori $\Delta^l(D)$ devient de très faible coût à évaluer *online*. Ainsi, on peut enrichir la base $\mathcal{S}^l$ d'une nouvelle solution $u(D_{l+1})$ en minimisant, plutôt que l'erreur de réduction $R^l(D)$ qui requiert l'évaluation de $u(D)$, l'estimateur $\Delta^l(D)$, directement accessible, en recouvrant numériquement un certain sous-domaine de $\mathcal{D}$. Pour vérifier que $\Delta(D)$ est un estimateur rigoureux, une analyse d'efficacité, omise pour cette introduction, est aussi disponible dans ce cas linéaire.

Pour terminer, on formule quelques remarques sur la méthode (voir par exemple [RHP07]). Premièrement, on recommande d'utiliser plutôt une orthogonalisation de Gram-Schmidt de $\mathcal{S}^l$ afin de disposer d'un système de coordonnées plus performant numériquement. Enfin, la qualité des estimateurs a posteriori et des méthodes de construction hiérarchique de $\mathcal{S}^l$ dans les cas non linéaires reste en réalité très difficile à contrôler.

1.2.2 Réduction par *proper orthogonal decomposition*

On retrace d'abord l'apparition progressive de la méthode de réduction par transformée de Karhunen-Loève, aucune traduction française fidèle de "*proper orthogonal decomposition*" n'étant répandue, dans la littérature des mathématiques appliquées. On renvoie à la synthèse de Berkooz

et al. [BHL93] à ce sujet pour plus de détails. Ensuite, alors qu'on s'aperçoit que la base de décomposition correspondante convient pour une description d'ensemble, c'est-à-dire non locale, de l'information essentielle d'une solution, on explique en quel sens celle-ci est optimale. Enfin, on expose les estimations a priori existantes, et peu nombreuses, pour justifier l'emploi de la méthode de réduction associée.

Premières explorations expérimentales de la POD et de la méthode de réduction associée

Le théorème de Karhunen-Loève, ou résultat de *proper orthogonal decomposition* dans la littérature anglophone, généralise la méthode classique en statistiques d'analyse en composantes principales pour un processus aléatoire continu X_t, $t \in [0, T]$. Il détermine de manière constructive et spectrale une base hilbertienne $(e_i)_{i=1}^{\infty}$ de $L^2(0, T)$ sur laquelle le processus se décompose comme

$$X_t = \sum_{i=1}^{\infty} X^{(i)} e_i(t),$$

où les coefficients $X^{(i)}$ sont des variables aléatoires de covariances nulles deux à deux. Les vecteurs $e_i(t)$ constituent alors les *composantes principales* du signal X_t.

Alors que la communauté probabiliste démontre ce résultat à partir des années 1930–40 et indépendamment par plusieurs mathématiciens, initialement Kosambi [Kos43], puis Karhunen [Kar47], Loève (même période, voir cependant l'ouvrage de référence [Loè78]) entre autres, l'application de celui-ci pour l'étude des modèles physiques, et en particulier la mécanique des fluides, émerge dès la fin des années 1960.

Lumley [Lum67] introduit ce concept pour l'analyse des mesures régulières d'un fluide en régime turbulent non homogène. Il découvre alors que les composantes principales calculées, de même nature que les images discrètes recueillies pour les mesures, correspondent à des *structures cohérentes*, c'est-à-dire des formes spatiales remarquables qui réapparaissent cycliquement au cours du temps.

L'espace engendré par un nombre défini l de composantes principales *les plus significatives*, en un sens à préciser plus loin, paraît donc susceptible de contenir les informations physiques les plus pertinentes *à l'ordre l*. Avec l'attention que les systèmes dynamiques suscitent dans les années 1960, le procédé d'approximation de Galerkin d'équations sur cet espace particulier se multiplie dans les publications. Pour la littérature sur la turbulence, on cite par exemple [Sir87, AHLS89, LP96].

Enfin, pour appuyer le paradoxe évoqué au début de ce rapport, on insiste sur le fait que ce sont les solutions des équations de Navier–Stokes qui ont motivé cette démarche de réduction par POD. En effet, dans un

premier but d'analyse quantitative et d'observation, l'emploi de l'analyse en composantes principales a permis de *résumer l'essentiel* de l'information contenue dans les turbulences. Cette idée apparaît alors que les limites techniques interdisaient encore l'étude fine, par le détail, des solutions correspondantes, et très complexes, de Navier–Stokes à travers la simulation numérique.

Caractère non local et interprétation variationnelle des bases POD

En analyse numérique, on retient plutôt la dénomination de « bases POD » *(POD bases)* pour les composantes principales. Avant de préciser comment on interprète les bases POD comme des bases optimales, on discute la notion de degré de liberté sous deux angles opposés, l'un local et l'autre global.

Pour construire u_h dans l'exemple (1.1), les éléments finis adoptent une approche *locale* en utilisant une *base de fonctions de forme* aux supports majoritairement deux à deux disjoints. L'avantage réside dans l'apparition de matrices creuses dans l'écriture du schéma de résolution. Cependant, pour certaines solutions qui ressemblent qualitativement à celles de l'équation de la chaleur, par exemple celles de la version linéaire de (1.1), i.e.

$$\begin{gathered}\text{trouver} \quad u_h(t) \in V_h \quad \text{tel que} \\ \frac{\partial}{\partial t}(u_h(t), v_h) + a(u_h(t), v_h) = (f(t), v_h), \quad \forall v_h \in V_h, \end{gathered} \tag{1.4}$$

on peut désirer revenir à une autre description qui présente des qualités mathématiques complémentaires.

En effet, muni d'une hypothèse de compacité, on peut résoudre (1.4) par l'approche *globale* de décomposition spectrale. En prenant la base de modes propres (w_i) définie par

$$a(w_i, v) = \omega_i^2 (w_i, v), \quad \forall v \in V, \qquad a(w_i, w_j) = \delta_{ij},$$

on décompose la solution sous la forme $u_h(t) = \sum_i u_i(t) w_i$ et on vérifie les relations scalaires découplées des coefficients

$$\frac{\partial u_i}{\partial t}(t) + \omega_i^2 u_i(t) = f_i(t).$$

De ce point de vue analytique d'une part, les *modes propres* utilisés ont pour support, contrairement aux fonctions de forme des éléments finis, le domaine entier Ω. Toutefois, tandis que les modes propres correspondent à la base de Fourier dans le cas de domaines rectangulaires, leur calcul approché dans le cas général reste coûteux et peut nécessiter un assez grand nombre de modes pour une approximation convenable. Surtout, la définition de cette base ne présente aucune extension naturelle pour les problèmes linéaires.

D'un point de vue physique d'autre part, ces solutions suivent spatialement un mouvement d'étalement, qui tend à diffuser sur tout le domaine l'information contenue plutôt qu'à la confiner en zones d'accumulation. Intuitivement, en considérant des modes particuliers, distincts dans ce cas des modes propres, qui ne dépendent eux que de la géométrie de Ω, elles s'approchent plus efficacement en sommes finies de tels modes qu'en sommes finies de fonctions de formes localisées.

Pour formuler concrètement cette recherche de décomposition la mieux adaptée pour $u_h(t)$ sur un intervalle $[0, T]$, on introduit le critère variationnel de rang $l \geq 1$ arbitraire

$$\text{minimiser} \int_0^T \left\| u_h(t) - \sum_{i=1}^{l} \beta_i(t)\varphi_i \right\|^2 \mathrm{d}t,$$

sur les fonctions mesurables $\beta_i : [0, T] \to \mathbb{R}$ et les familles de vecteurs $(\varphi_i)_{i=1}^l$ de V_h.

Comme pour toute famille $(\beta_i, \varphi_i)_{i=1}^l$, l'orthonormalisation de Gram–Schmidt $(\tilde{\varphi}_i)$ sur (φ_i) fournit une nouvelle famille $(\tilde{\beta}_i, \tilde{\varphi}_i)_{i=1}^l$ telle que

$$\sum_i \tilde{\beta}_i(t)\tilde{\varphi}_i = \sum_i \beta_i(t)\varphi_i \,,$$

on peut restreindre pour $(\varphi_i)_{i=1}^l$ la recherche sur les familles *orthonormales* pour la norme associée à a. Puis, par le théorème de Pythagore, en reconnaissant les coefficients de projection orthogonale de $u_h(t)$ sur une telle famille,

$$\left\| u_h(t) - \sum_{i=1}^{l} \beta_i(t)\varphi_i \right\|^2 = \left\| u_h(t) - \sum_{i=1}^{l} a(u_h(t), \varphi_i)\varphi_i \right\|^2 + \sum_{i=1}^{l} \left| \beta_i - a(u_h(t), \varphi_i) \right|^2,$$

ce qui montre que pour tout candidat $(\beta_i, \varphi_i)_{i=1}^l$, la famille

$$\Big(a(u_h(t), \varphi_i), \varphi_i \Big)_{i=1}^l$$

est associée à une valeur encore inférieure. La recherche porte donc uniquement à présent sur les familles orthogonales $(\varphi_i)_{i=1}^l$. Pour terminer, on remarque que

$$v = \sum_{i=1}^{l} a(u_h(t), \varphi_i)\varphi_i$$

est la projection orthogonale de $u_h(t)$ sur $V^l = \mathrm{Vect}(\varphi_i)_{i=1}^l$. Or le projecteur orthogonal en question est entièrement déterminé par son image V^l, et notamment toute autre famille (φ_i') engendrant le même sous-espace V^l

fournit le même vecteur v. On peut donc au final paramétrer le problème précédent par les *projecteurs orthogonaux sur V de rang l*, ou de manière équivalente par les sous-espaces de rang l.

Ainsi, on appelle *projecteur POD* π^l toute solution du problème

$$\min_{\tilde{\pi}^l} \int_0^T \|u_h(t) - \tilde{\pi}^l u_h(t)\|^2 \, \mathrm{d}t,$$

et aussi *espace POD* l'image de ce projecteur, et *base POD* toute base de l'espace POD. Ce critère est en réalité défini pour toute fonction $L^2(0,T;V_h)$, quelle que soit l'équation qu'elle vérifie, et sans hypothèse particulière de linéarité.

On montre alors (comme rappelé au chapitre 2), et dans le cadre plus général de fonctions non discrétisées en espace, que la solution de ce problème *existe* et que *l'espace engendré par les l vecteurs propres dominants d'une décomposition de Karhunen-Loève, est solution.* La suite de valeurs propres (λ_i) correspondante, positive et décroissante vers 0, vérifie alors le résultat fondamental d'erreur de projection

$$\|u_h - \pi^l u_h\|_{L^2(0,T;V)} = \Big\{ \sum_{i>l} \lambda_i \Big\}^{1/2}. \tag{1.5}$$

L'avantage essentiel de ce résultat réside dans la décroissance du reste de la série des valeurs propres, numériquement toujours très rapide, et cela particulièrement pour les solutions d'équations paraboliques.

Enfin, pour compléter les commentaires formulés plus haut sur leur caractère non local, les bases POD peuvent bien entendu aussi être calculées pour des solutions ressemblant à celles de l'équation des ondes. Si le raisonnement précédent sur la diffusion spatiale de l'information ne tient plus, puisqu'elle est au contraire cette fois *propagée*, les bases POD correspondantes subissent alors une interprétation opposée. En effet, dans ce cas, elles convergent bien pour $T \to \infty$ vers les modes propres, comme énoncé en section 3.3.5. On verra aussi au cours du chapitre 5 avec les équations de FitzHugh–Nagumo, que les solutions propagatives présentent des difficultés de réduction.

Estimations d'erreurs de réduction existantes et robustesse des bases POD

Contrairement à la méthode de réduction par base réduite, même pour les cas élémentaires, la construction d'estimateurs a posteriori semble, en revanche, moins naturelle ici. On dénombre aussi peu d'estimations a priori dans la littérature.

Henri et Yvon [HY05] démontrent une estimation pour un problème parabolique continu et linéaire, i.e.

$$\begin{gathered}\text{trouver} \quad u(t) \in V \quad \text{tel que} \\ \frac{\partial}{\partial t}(u(t),\, v) + a(u(t),\, v) = (f(t),\, v), \quad \forall v \in V, \\ u(0) = u_0.\end{gathered}$$

En posant V^l l'espace POD associé à $u(t)$ sur $[0, T]$ et pour le produit scalaire a, la réduction $u^l(t)$ de $u(t)$ est définie par l'approximation de Galerkin

$$\begin{gathered}\text{trouver} \quad u^l(t) \in V^l \quad \text{tel que} \\ \frac{\partial}{\partial t}(u^l(t),\, v^l) + a(u^l(t),\, v^l) = (f(t),\, v^l), \quad \forall v^l \in V^l, \\ u^l(0) = u_0^l \in V^l.\end{gathered}$$

Ainsi [HY05, Th. 9], pour u_0^l convenablement choisi,

$$\|u - u^l\|_{L^2(0,T;V)} \leq C\Big(\Big\{\sum_{i>l} \lambda_i\Big\}^{1/2} + \Big\|\frac{\partial u}{\partial t} - \pi^l \frac{\partial u}{\partial t}\Big\|_{L^2(0,T;\, L^2(\Omega))}\Big). \tag{1.6}$$

Cependant, la POD n'étant pas adaptée pour la dérivée $\frac{\partial u}{\partial t}$, le deuxième terme au second membre n'est pas contrôlé. On relève, dans des cadres différents, plusieurs autres propositions d'estimations qui contiennent des termes analogues, et présentent donc le même inconvénient : Kunisch et Volkwein [KV02] pour un problème parabolique non-linéaire et discrétisé en temps, Hinze et Volkwein [HV08] pour un problème de contrôle optimal linéaire-quadratique, entre autres.

Pour contourner ce problème, Henri et Yvon (ibid.) utilisent une POD associée au couple $\Big(u(t), \frac{\partial u}{\partial t}(t)\Big)$, i.e. minimisent la somme des fonctionnelles associées aux deux éléments, ce qui a effectivement pour effet de supprimer le terme problématique. Toutefois, nous ne retenons pas ce procédé dans ce rapport. Comme raison mineure, on a vérifié, dans un premier temps exploratoire sur des problèmes paraboliques 1D non-linéaires, qu'il n'améliorait pas la réduction, ni la décroissance du reste $\sum_{i>l} \lambda_i$. La version totalement discrète de $\frac{\partial u}{\partial t}$, i.e. $\frac{U^{n+1}-U^n}{\Delta t}$ n'est en effet qu'une combinaison linéaire de la solution (U^n), et n'apporte pas davantage d'informations.

Surtout, on propose dans ce rapport une autre estimation, bien qu'incomplète. Celle-ci nous a semblé ne pas apparaître dans la littérature, et justifier *par une simple vérification numérique*, le procédé classique sans les dérivées temporelles.

L'idée a consisté à reprendre la preuve de l'estimation (1.6), et de faire jouer, plutôt que le découpage

$$u - u^l = (u - \pi^l u) + (\pi^l u - u^l), \tag{1.7}$$

celui qui utilise $\pi^l_{L^2}$, le projecteur de même image $V^l = \operatorname{Im} \pi^l$ et orthogonal cette fois pour le produit scalaire $(\cdot, \cdot)$ de $L^2(\Omega)$, i.e.

$$u - u^l = (u - \pi^l_{L^2} u) + (\pi^l_{L^2} u - u^l). \tag{1.8}$$

L'estimation correspondante devient, pour un choix convenable de u^l_0,

$$\|u - u^l\|_{L^2(0,T;V)} \leq C(1 + \rho_l)\|u - \pi^l u\|_{L^2(0,T;V)}, \tag{1.9}$$

où ρ_l est la norme d'opérateur

$$\rho_l = \sup_{v \in V \setminus \{0\}} \frac{\|\pi^l_{L^2} v\|}{\|v\|}. \tag{1.10}$$

On a vérifié par la suite numériquement le caractère borné de la suite $(\rho_l)_{l \geq 1}$ dans de nombreux cas étudiés. Pour argumenter cette hypothèse, on a aussi montré que pour certains jeux admissibles et particuliers de sous-espaces V^l, construits depuis des décompositions modales, la suite $(\rho_l)_{l \geq 1}$ correspondante est effectivement bornée.

On souligne le caractère théorique des estimations précédentes, qui étudient un cas *d'auto-réduction*, où la solution est réduite sur son propre espace POD, ce qui ne présente pas d'intérêt directement pratique. En effet, l'espace POD a d'une part nécessité un calcul complet de la solution, et d'autre part, est d'emblée optimal pour l'erreur de projection de celle-ci.

En vue d'applications concrètes, ou simplement moins maîtrisées par les estimations d'erreurs existantes, ces investigations analytiques nécessitent donc une extension, et n'échappent pas à la question de *robustesse* de la réduction par POD, qu'on peut poser sous plusieurs angles.

À des considérations de *dernière valeur propre multiple* près, on peut bien définir l'application

$$u \in \mathcal{U} \subset L^2(0, T; V) \quad \mapsto \quad \pi^l(u),$$

associant à toute élément de $\mathcal{U}$ son projecteur POD de rang l. On souhaite naturellement étudier la sensibilité correspondante du projecteur POD par rapport à la solution, résumée par $\frac{\partial \pi^l}{\partial u}$. On doit alors garder à l'esprit que l'ensemble décrit par les projecteurs orthogonaux de rang l est une variété nommée *grassmannienne*, et non un espace vectoriel.

Rathinam et Petzold [RP03] donnent un sens précis à la dérivée $\frac{\partial \pi^l}{\partial u}$ dans ce contexte de variétés, pour une solution d'EDO linéaire en dimension finie, et l'illustre par des exemples de sensibilité parfois très grande, et bien infinie dans le cas $\lambda_l = \lambda_{l+1}$ écarté plus haut. On n'a pas poursuivi cette piste, essentiellement théorique, et difficile à explorer dans un cadre non-linéaire, dans ce rapport. On en a cependant tiré une mise en garde,

puisqu'elle montre qu'une base POD calculée pour une solution particulière peut se révéler très inadaptée pour réduire les solutions voisines.

Dans un contexte paramétrique, i.e. $u(t;\theta)$, $\theta \in \Theta$, la question intervient plus clairement dans celle de la sensibilité du projecteur POD par rapport au paramètre, i.e. $\frac{\partial \pi^l}{\partial \theta}$. Amsallem et Farhat [AFo8] proposent à ce titre une méthode d'interpolation de bases POD qui permet de décrire la grassmannienne précisée plus haut, et d'approcher $\pi^l(\theta)$ par la donnée d'un certain nombre de projecteurs déjà calculés $\pi^l(\theta_i)$. C'est une méthode qui paraît difficile à analyser, et qui ne dispose pas d'estimation d'erreur couvrant des cas simples. On ne l'a pas non plus explorée pour ce rapport, et on a choisi de privilégier une piste différente.

En effet, on a décidé de traiter de telles solutions paramétriques par une stratégie plus simple, que nous appelons *multi-POD*. Cette méthode est l'objet du chapitre 4. On l'y teste avec succès sur *les équations de FitzHugh–Nagumo*, qui apparaissent dans les modèles électrophysiologiques de potentiel d'action. À l'inverse du point de vue de la sensibilité, qui cherche à étendre au maximum le domaine de validité (sur l'espace des paramètres) d'un projecteur POD particulier, on propose de délimiter une zone raisonnable de variabilité paramétrique et de définir une généralisation de projecteur POD dessus. Plus précisément, en posant $\mathcal{D} \subset \Theta$ une zone de la forme

$$\mathcal{D} = [a_1, b_1] \times \cdots \times [a_p, b_p],$$

on définit, par généralisation quadratique du problème POD *standard*, le problème *multi-POD de degré* 1 comme

$$\min_{\tilde{\pi}^l} \sum_{\substack{\alpha_i \in \{a_i, b_i\} \\ 1 \leq i \leq p}} \int_0^T \|u(t; \alpha_1, \ldots, \alpha_p)\|^2 \, \mathrm{d}t, \tag{1.11}$$

où la somme fait intervenir toutes les sommets du rectangle $\mathcal{D}$. On montre facilement que le projecteur minimiseur est celui d'une POD standard appliquée sur une concaténation en temps de chacune des trajectoires

$$u(\cdot; \alpha_1, \ldots, \alpha_p)$$

impliquées. Notons alors (μ_i) la suite de valeurs propres associée. On conserve alors un critère de la forme

$$\sum_{\substack{\alpha_i \in \{a_i, b_i\} \\ 1 \leq i \leq p}} \|u(\alpha_1, \ldots, \alpha_p)\|^2_{L^2(0,T;V)} = \left\{ \sum_{i>l} \mu_i \right\}^{1/2}.$$

Si la vitesse de convergence du reste $\sum_{i>l} \mu_i$ est généralement plus faible que celle associée à une POD standard, par exemple pour le paramètre barycentre de Θ, elle reste néanmoins très satisfaisante.

On généralise alors le résultat (1.9) en une estimation a priori pour l'erreur maximale de réduction par multi-POD sur le domaine $\mathcal{D}$. Pour cela, on utilise des résultats classiques d'interpolation dans la norme $C^0(\mathcal{D})$ correspondante, et à distinguer des résultats analogues avec les normes de Sobolev.

1.3 Organisation du rapport et contributions

La suite du rapport couvre cinq chapitres qui avancent, dans l'ensemble, des considérations les plus fondamentales vers les résultats les plus appliqués, suivis d'une démonstration d'analyse fonctionnelle en annexe. On s'est concentré jusqu'ici sur l'introduction des fondements et des pistes actuelles du domaine des méthodes de réduction. Néanmoins, les deux derniers chapitres, qui abordent deux formulations distinctes de problèmes inverses, contiennent également, pour clarifier les contextes d'application, des passages introductifs et classiques sur les méthodes correspondantes.

On annonce ci-dessous le plan de ces chapitres et les contributions qu'on a apportées.

Chapitre 2. Revue mathématique de la *proper orthogonal decomposition* pour un problème continu et abstrait

Le chapitre 2 démontre entièrement le résultat principal de la POD, énoncé en l'équation (1.5), cette fois dans le cadre général de fonctions continues temps à valeurs hilbertiennes. On a souhaité simplifié les démonstrations qui utilisent des *méthodes lagrangiennes* (par exemple [Volo1] ou [AH92]). L'idée principale est inspirée d'une manipulation algébrique sur les valeurs propres qui apparaît dans la courte démonstration d'une variante d'un théorème de Weyl [Fan49]. On note aussi, avec une idée similaire sans méthode lagrangienne, dans un cadre stochastique et de dimension finie, la courte démonstration de Dür [Dür98].

Chapitre 3. Approximation de Galerkin utilisant la *proper orthogonal decomposition* : nouvelles estimations d'erreurs et exemples illustratifs

Dans le chapitre 3, on démontre l'estimation a priori résumée plus haut par l'équation (1.9) pour une équation parabolique abstraite linéaire.

Ensuite, premièrement, on généralise celle-ci pour une équation parabolique sous-linéaire, c'est-à-dire avec un second membre lipschitzien en la solution. On s'appuie pour cela sur l'annexe A, dans laquelle on démontre le caractère bien posé du problème dans notre cadre variationnel. Deuxièmement et plus simplement, on transpose l'estimation pour une équation des ondes linéaire abstraite. On parvient en effet encore à perdre le terme

de projection des dérivées en temps les plus élevées par rapport aux estimations classiques.

Puis, pour clore la partie analytique, on exhibe deux cas de suites de sous-espaces de réduction (V^l) pour lesquels (ρ_l) (1.10) est effectivement bornée. On démontre d'abord le résultat pour le cas lié aux familles quasi-uniformes de discrétisation par éléments finis, et ensuite pour le cas lié aux sous-espaces modaux d'un opérateur elliptique.

On illustre ensuite la qualité des estimations a priori proposées par des résultats numériques simples 1D qui dépassent les non-linéarités du cadre théorique. Enfin, on implémente la réduction par POD pour le modèle de cœur. On vérifie notamment pour le champ de déplacement, pour une géométrie simplifiée, que l'erreur de réduction est encore numériquement contrôlée par l'erreur de projection POD.

> *Ce résultat partiellement théorique, consolidé par plusieurs vérifications numériques, notamment sur le complexe modèle de cœur [SMCCS06], fait l'objet d'un article en voie de publication : Chapelle, Gariah et Sainte-Marie "Galerkin approximation with proper orthogonal decomposition: new error estimates and illustrative examples", ESAIM:M2AN. Celui-ci constitue une adaptation du chapitre 3 sous forme autonome.*

Chapitre 4. Stratégie de réduction de modèle pour les problèmes paramétriques

Le chapitre 4 propose d'étendre les estimations a priori précédentes pour des solutions paramétriques avec une méthode *multi-POD*, qu'on a décrite à l'équation (1.11). En effet, les excellents résultats du chapitre 3 pour la réduction d'une solution *réduite sur elle-même*, c'est-à-dire sur le sous-espace qui lui est associé par la POD standard, forment le minimum requis pour se permettre d'appliquer une méthode de réduction POD plus pratique, empirique et, pour laquelle le projecteur POD devient théoriquement sous-optimal pour chacune des solutions.

On reprend le cas parabolique linéaire pour des raisons de simplicité. En s'inspirant des premiers développements pour la méthode de réduction par base réduite [MR02], on introduit dans l'opérateur elliptique un paramètre D qui varie dans un domaine rectangulaire $\mathcal{D}$. L'estimation se décompose alors en deux parties. La première partie estime l'erreur maximum de réduction

$$\|u - u^l\|_{C^0(\mathcal{D};L^2(0,T;V))}$$

par l'erreur maximale de projection POD

$$\|u - \pi^l u\|_{C^0(\mathcal{D};L^2(0,T;V))},$$

de manière totalement analogue au cas non paramétrique.

La deuxième utilise un résultat d'interpolation en norme $C^0(\mathcal{D})$. On a souhaité redémontrer ce résultat de Ciarlet et Raviart [CR72] pour faire ressortir qu'il repose essentiellement sur une propriété purement algébrique des polynômes de Lagrange. Cela permet d'estimer en retour cette erreur maximale de projection POD par la somme de l'erreur de projection multi-POD, c'est-à-dire la racine carrée de reste

$$\left\{ \sum_{i>l} \lambda_i \right\}^{1/2},$$

et d'un terme d'erreur d'interpolation, dépendant de la régularité de la solution par rapport à D.

Pour finir cette extension théorique, on dresse un bilan des avantages et inconvénients numériques de la méthode, en distinguant ses deux phases *offline* et *online*.

En visant ensuite à tester les limites de la réduction par multi-POD, on l'a appliquée numériquement, et en comparaison avec une réduction par POD standard, sur les équations de FitzHugh–Nagumo. En effet, d'une part, on illustre comment ces équations peuvent être hautement sensibles par rapport à leurs paramètres, notamment autour *d'effets de seuil*. D'autre part, elles présentent un caractère propagatif, connu pour être difficile à approcher par une somme de termes qui découplent les variables spatiale et temporelle. Par cohérence avec la partie théorique, on s'est intéressé à la variation du coefficient de diffusion uniquement. En se donnant une fenêtre de variation paramétrique, on a divisé l'étude en deux cas, l'un aux variations homogènes de la solution, et l'autre aux variations plus brutales.

Chapitre 5. Réduction d'une estimation variationnelle de paramètres sur un modèle électrophysiologique

Le chapitre 5 reprend les équations de FitzHugh–Nagumo. Comme on a efficacement réduit par multi-POD les simulations directes, tout du moins pour les plages paramétriques qui n'induisent pas d'effets de seuil, on attaque dans ce chapitre le problème inverse d'estimation de paramètres, avec une approche variationnelle (voir l'article fondateur [DT86]), c'est-à-dire formulée par une fonction coût à minimiser.

On commence par poser le problème mathématique précis pour estimer une condition initiale et un coefficient de diffusion inconnus, avec une observation spatialement distribuée en un petit nombre de points, et bruitée. Puisque ce rapport n'est pas centré sur des considérations probabilistes, on a souhaité détailler et justifier la modélisation classique des incertitudes sans passer par l'approche bayésienne. On a aussi restreint l'espace à explorer pour la condition initiale, pour concentrer le problème sur l'estimation paramétrique. On détaille ensuite, afin d'utiliser une méthode de

minimisation du premier ordre, le calcul du gradient de la fonction coût. On donne en parallèle les versions réduites par POD de la fonction coût et de son gradient, dont les expressions sont naturelles et immédiates.

Ensuite, partant d'un problème d'estimation de référence sans méthode de réduction, on étudie la *convergence en rang POD* des estimations réduites, à la fois pour la méthode multi-POD et la moins performante POD standard, et pour deux cas de solution exacte repris du chapitre 4.

Chapitre 6. Réduction d'une estimation séquentielle de paramètres sur un modèle mécanique de cœur

Enfin, le chapitre 6 emploie la méthode multi-POD sur un problème inverse utilisant des vraies données cliniques, notamment issues d'imagerie. On dispose d'un modèle complexe de cœur [SMCCS06], tridimensionnel, fondé sur la mécanique en grandes déformations et muni d'une loi de comportement qui retranscrit les processus à l'œuvre à l'échelle moléculaire. En appliquant une méthode particulière d'estimation par *filtrage* sur ce modèle, et en y intégrant pour seule information un jeu de données IRM d'un cœur ayant subi un infarctus, on cherche à retrouver la zone infarcie en estimant le champ de contractilité associé. Ph. Moireau (équipe Macs, Inria Rocquencourt, [40, 41, 39]) a apporté une contribution essentielle à ce chapitre.

Les techniques de filtrage qui généralisent le *filtre de Kalman* pour les cas non-linéaires définissent des *observateurs.* Ceux-ci sont définis comme les solutions des équations originales dans lesquelles on supprime les incertitudes, et ajoute au second membre un terme correctif particulier, mesurant la distance des données au modèle. Dans un esprit différent des techniques variationnelles, elles corrigent, par prise en compte successive des données recueillies, la trajectoire des observateurs au cours du temps de simulation.

Pour lier plus explicitement ce point de vue au chapitre 5, on présente rapidement le filtre de Kalman, écrit pour les systèmes linéaires, avec une construction par approche variationnelle dépendant du temps. On présente ensuite une extension courante de ce filtre pour les problèmes non-linéaires, par linéarisation de la dynamique et de l'opérateur d'observation, et qui reste très coûteuse pour des problèmes de grande taille. On fournit alors la version réduite par POD, qui s'écrit naturellement et rend cette méthode praticable.

On présente ensuite une méthode de filtrage par "*unscented transform*", qu'on utilise pour la validation numérique. Elle possède l'avantage d'éviter les linéarisations du filtre précédent en réalisant une interpolation. Elle consiste à propager les moyennes et covariances empiriques du système filtré, sur des directions spéciales à cet effet dans l'espace d'incertitude. On détaille le schéma en temps correspondant, écrit sous une forme de

prédiction-correction. La version réduite par POD est ensuite immédiate.

Puis, on décrit dans les grands lignes le protocole expérimental réalisé, les procédures d'adaptation du modèle pour les données recueillies, et la configuration numérique pour l'estimation des problèmes réduits par multi-POD. Enfin, on montre une étude de convergence en rang POD des résultats correspondants, et notamment des champs de contractilité estimés en temps final.

Références pour l'introduction

[AF08] D. Amsallem and C. Farhat. Interpolation method for the adaptation of reduced-order models to parameter changes and its application to aeroelasticity. *AIAA Journal*, 46 :1803–1813, 2008.

[AH92] A. N. Akansu and R. A. Haddad. *Multiresolution Signal Decomposition*. Academic Press, 1992.

[AHLS89] N. Aubry, Ph. Holmes, J. L. Lumley, and E. Stone. Application of dynamical system theory to coherent structures in the wall region. *Physica D : Nonlinear Phenomena*, 37(1–3) :1–10, 1989.

[BHL93] G. Berkooz, Ph. Holmes, and J. L. Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annu. Rev. Fluid Mech.*, 25 :539–575, 1993.

[BMNP04] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An "empirical interpolation" method : application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris, Ser. I*, 339, 2004.

[CR72] P. G. Ciarlet and P. A. Raviart. General Lagrange and Hermite interpolation in $\mathbb{R}$ with applications to finite element methods. *Arch. Rational Mech. Anal.*, 46 :177–199, 1972.

[DT86] F.-X. Le Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects. *Tellus*, 38 A :97–110, 1986.

[Dür98] A. Dür. On the optimality of the discrete Karhunen-Loève expansion. *SIAM J. Control. Optim.*, 36(6) :1937–1939, 1998.

[Fan49] K. Fan. On a theorem of Weyl concerning eigenvalues of linear transformations. I. *Proc. Nat. Acad. Sci. U. S. A.*, 35 :652–655, 1949.

[GP05] M. A. Grepl and A. T. Patera. A posteriori error bounds for reduced-order approximations of parametrized parabolic partial differential equations. *ESAIM : M2AN*, 39(1) :157–181, 2005.

[HV08] M. Hinze and S. Volkwein. Error estimates for abstract linear–quadratic optimal control problems using proper orthogonal decomposition. *Computational Optimization and Applications*, 39 :319–345, 2008.

[HY05] T. Henri and J.-P. Yvon. Convergence estimates of POD-Galerkin methods for parabolic problems. In *System Modeling and Optimization*, volume 166 of *IFIP International Federation for Information Processing*, pages 295–306. Springer Boston, 2005.

[Kar47] K. Karhunen. Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys.*, 37, 1947.

[Kos43] D. D. Kosambi. Statistics in function space. *J. Indian Math. Soc. (N.S.)*, 7 :76–88, 1943.

[KV02] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM J. Numer. Anal.*, 40(2) :492–515, 2002.

[Loè78] M. Loève. *Probability Theory II*. Number 46 in Graduate Texts in Mathematics. Springer, 1978.

[LP96] J. L. Lumley and B. Podvin. Dynamical systems theory and extra rates of strain in turbulent flows. *Experimental Thermal and Fluid Science*, 13(3) :180 – 189, 1996.

[Lum67] J. L. Lumley. The structure of inhomogeneous turbulence. In A. M. Yaglom and V. I. Tatarski, editors, *Atmospheric Turbulence and Wave Propagation*, pages 166–178. Nauka, 1967.

[MPR02] Y. Maday, A. T. Patera, and D. V. Rovas. A blackbox reduced-basis output bound method for noncoercive linear problems. In *Nonlinear Partial Differential Equations and their Applications – Collège de France Seminar Volume XIV*, volume 31 of *Studies in Mathematics and Its Applications*, pages 533 – 569. Elsevier, 2002.

[MR02] Y. Maday and E. M. Rønquist. A reduced-basis element method. *J. Sci. Comput.*, 17(1–4) :447–459, 2002.

[NP80] A. K. Noor and J. M. Peters. Reduced basis technique for nonlinear analysis of structures. *AIAA Journal*, 18(4) :455–462, 1980.

[RHP07] 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. *Arch. Comput. Methods Eng.*, 15(3) :1–47, 2007.

[RP03] M. Rathinam and L. R. Petzold. A new look at proper orthogonal decomposition. *SIAM J. Numer. Anal*, 41(5) :1893–1925, 2003.

[Sir87] L. Sirovich. Turbulence and the dynamics of coherent structures, part i–iii. *Quarterly of Applied Mathematics*, 45(3) :561–590, 1987.

[SMCCS06] J. Sainte-Marie, D. Chapelle, R. Cimrman, and M. Sorine. Modeling and estimation of the cardiac electromechanical activity. *Computers & Structures*, 84 :1743–1759, 2006.

[Vol01] S. Volkwein. Optimal control of a phase-field model using the proper orthogonal decomposition. *Z. Angew. Math. Mech.*, 81 :83–97, 2001.

[Wor07] World Health Organization. *Prevention of cardiovascular disease : pocket guidelines for assessment and management of cardiovascular risk – WHO/ISH cardiovascular risk prediction charts for the African Region*. WHO Press, 2007.

Chapter 2

Mathematical review of the abstract continuous proper orthogonal decomposition

Let V be a separable Hilbert space with scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. Let $z(t)$, $t\in[0,T]$, be a function with regularity

$$z\in L^2(0,T;V). \tag{2.1}$$

Performing the POD (time-continuous here) of rank l of z over $[0,T]$ means to find the orthogonal projector π^l of rank l solution of

$$\min_{\tilde{\pi}^l_V}\|z(t)-\tilde{\pi}^l z(t)\|_{L^2(0,T;V)}. \tag{2.2}$$

The integer l is called the *POD rank*. Thanks to (2.1), Problem (2.2) has a sense. It is solved as follows.

Let us introduce $\mathrm{Cov}:V\to V$ the *covariance operator* defined by

$$\mathrm{Cov}\,\varphi=\int_0^T\langle z(t),\varphi\rangle z(t)\,\mathrm{d}t.$$

We need the following property of Cov.

Proposition 1. *There exists a unique real sequence $(\lambda_i)_{i\in I}$, with I at most countable, such that*

$$\lambda_i>0, \tag{2.3}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_N \quad \textit{if finite} \quad (I=\{1,\,2,\,,\ldots,\,N\}), \tag{2.4}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_i\geq\ldots, \quad \lambda_i\underset{i\to\infty}{\longrightarrow}0 \quad \textit{if infinite} \quad (I=\mathbb{N}\setminus\{0\}), \tag{2.5}$$

and an orthonormal sequence $(\varphi_i)_{i\in I}$ of V of corresponding eigenvectors of the operator Cov*, in finite number for each non-null eigenvalue,*

$$\mathrm{Cov}\,\varphi_i=\lambda_i\varphi_i,\quad \forall i\in I,$$

such that $(\varphi_i)_{i\in I}$ *is total in the orthogonal complement of the kernel of* Cov *i.e.*

$$V = \operatorname{Ker}\operatorname{Cov} \overset{\perp}{\oplus} \overline{\operatorname{Span}\{\varphi_i,\, i \in I\}}. \tag{2.6}$$

In order to understand (2.6), it is helpful to characterize the kernel of Cov with respect to z. As the following proposition intimates, it is made of the vectors that are not concerned by the evolution of $z(t)$.

Proposition 2. *The kernel of* Cov *is made of the vectors that are orthogonal to* $z(t)$ *for almost every* $t \in [0,T]$*, i.e.*

$$\operatorname{Ker}\operatorname{Cov} = \{\varphi \;;\; \langle z(t), \varphi\rangle = 0 \text{ a.e. } t \in [0,T]\}.$$

Proof of Prop. 2. For the non-trivial inclusion, consider $(\operatorname{Cov}\varphi, \varphi)$. □

Then, we assert the classical result, that uses the notations of Prop. 1.

Proposition 3. *For all* $1 \le l \le \operatorname{Card} I$*, a solution* π^l *of Problem (2.2) is determined by*

$$\operatorname{Im}\pi^l = \operatorname{Span}(\varphi_1, \dots, \varphi_l).$$

Moreover,

$$\|z - \pi^l z\|_{L^2(0,T;V)} = \left\{ \sum_{i=l+1}^{\operatorname{Card} I} \lambda_i \right\}^{1/2}. \tag{2.7}$$

In various numerical experiments in the sequel, we verify that (λ_i) decreases very rapidly, typically at exponential rate. Hence, the approximation

$$z(t) \approx \sum_{i=1}^{l} \langle z(t), \varphi_i\rangle \varphi_i$$

converges at exponential rate w.r.t. l in $L^2(0,T;V)$.

We prove Props. 1 and 3 hereafter.

2.1 Diagonalisation of the covariance operator

Let us introduce $\widetilde{\operatorname{Cov}}$ the $L^2(0,T;\mathbb{R}) \to L^2(0,T;\mathbb{R})$ operator defined by

$$\widetilde{\operatorname{Cov}}\, v\,(s) = \int_0^T \langle z(t), z(s)\rangle v(t)\,\mathrm{d}t = \left\langle \int_0^T v(t) z(t)\,\mathrm{d}t,\, z(s) \right\rangle,$$

for a.e. $s \in [0,T]$. Since we assume (2.1), Cov and $\widetilde{\operatorname{Cov}}$ are well defined. We relate the properties of these two operators thanks to the two following lemmas.

Lemma 1. Cov *and* $\widetilde{\mathrm{Cov}}$ *share the same non-null eigenvalues, with identical multiplicities.*

Proof of Lemma 1. Assume $\lambda \neq 0$ is an eigenvalue of Cov of multiplicity $m(\lambda)$ (eventually ∞), i.e.

$$\mathrm{Cov}\,\varphi_k = \lambda\varphi_k,$$

with $(\varphi_k)_{k=1}^{m(\lambda)}$ an orthonormal family of V. We define the $L^2(0,T)$ functions v_k by

$$v_k(s) = \langle z(s), \varphi_k\rangle. \tag{2.8}$$

Then we verify on the one hand that

$$\begin{aligned}\widetilde{\mathrm{Cov}}\, v_k(s) &= \int_0^T \langle z(t), z(s)\rangle\langle z(t), \varphi_k\rangle \,\mathrm{d}t \\ &= \langle \mathrm{Cov}\,\varphi_k, z(s)\rangle \\ &= \lambda v_k(s),\end{aligned}$$

and on the other hand that $(v_k)_{k=1}^{m(\lambda)}$ is an orthogonal family of $L^2(0,T)$.

$$\begin{aligned}(v_k, v_j)_{L^2(0,T)} &= \int_0^T \langle z(t), \varphi_k\rangle\langle z(t), \varphi_j\rangle \,\mathrm{d}t \\ &= \langle \mathrm{Cov}\,\varphi_k, \varphi_j\rangle \\ &= \lambda\delta_{k,j}.\end{aligned} \tag{2.9}$$

This proves that λ is an eigenvalue of $\widetilde{\mathrm{Cov}}$ of multiplicity $\widetilde{m}(\lambda) \geq m(\lambda)$. [1]

Conversely, assume $\mu \neq 0$ is an eigenvalue of $\widetilde{\mathrm{Cov}}$ of multiplicity $\widetilde{m}(\mu)$, i.e.

$$\widetilde{\mathrm{Cov}}\, w_k = \mu w_k,$$

with $(w_k)_{k=1}^{\widetilde{m}(\mu)}$ an orthonormal family of $L^2(0,T)$. We define the V elements

$$\psi_k = \int_0^T w_k(t) z(t) \,\mathrm{d}t.$$

Then similarly we verify that μ is an eigenvalue of Cov and $(\varphi_k)_{k=1}^{\widetilde{m}(\lambda)}$ is an orthogonal family of eigenvectors associated to μ. Thus $m(\mu) \geq \widetilde{m}(\mu)$. This ends the proof. □

Lemma 2. $\widetilde{\mathrm{Cov}}$ *is compact.*

1. Note that the proof does not work for the null eigenvalue. If $\lambda = 0$, defining v_k as in (2.8) leads to $v_k = 0$ in $L^2(0,T)$, which can be seen in (2.9).

Proof of Lemma 2. We define

$$k(t,s) = \langle z(t), z(s) \rangle.$$

Then by Cauchy-Schwarz's inequality and Fubini's theorem,

$$k \in L^2([0,T]^2).$$

Thanks to this property, we show that $\widetilde{\mathrm{Cov}}$ is a Hilbert–Schmidt operator, which means that for some Hilbertian basis (e_i),

$$\sum_{i=1}^{\infty} \|\widetilde{\mathrm{Cov}}\, e_i\|^2_{L^2(0,T)} < \infty.$$

Indeed, let $(e_i)_{i=1}^{\infty}$ be a Hilbertian basis of $L^2(0,T)$. We verify the equality

$$\begin{aligned} (\widetilde{\mathrm{Cov}}\, e_i, e_j)_{L^2(0,T)} &= \int_0^T \int_0^T \langle z(t), z(s) \rangle e_i(t) e_j(s)\, \mathrm{d}t\mathrm{d}s \\ &= (k, e_i \otimes e_j)_{L^2([0,T]^2)}. \end{aligned}$$

But $(e_i \otimes e_j)_{i,j=1}^{\infty}$ is a Hilbertian basis of $L^2([0,T]^2)$, so we conclude by Parseval's equality

$$\begin{aligned} \sum_{i=1}^{\infty} \|\widetilde{\mathrm{Cov}}\, e_i\|^2_{L^2(0,T)} &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (\widetilde{\mathrm{Cov}}\, e_i, e_j)^2_{L^2(0,T)} \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} (k, e_i \otimes e_j)^2_{L^2([0,T]^2)} < \infty. \end{aligned}$$

Finally, we use the property that Hilbert–Schmidt operators are compact, see e.g. [3, Lemma 3.4.2]. □

We verify that $\widetilde{\mathrm{Cov}}$ is self-adjoint and positive. In conjunction with Lemma 2, we deduce that $\widetilde{\mathrm{Cov}}$ is diagonalisable on $L^2(0,T)$, i.e. that $L^2(0,T)$ admits a Hilbertian basis of eigenvectors of $\widetilde{\mathrm{Cov}}$. Moreover the non-null eigenvalues are positive and of finite multiplicity [7, Th. VI.11].

Therefore, there exists a unique sequence $(\lambda_i)_{i \in I}$, with I at most countable, of numbers λ_i that satisfies (2.3–2.5), and an orthonormal sequence $(v_i)_{i \in I}$ of $L^2(0,T)$ of corresponding eigenvectors of $\widetilde{\mathrm{Cov}}$, in finite number for each non-null eigenvalue,

$$\widetilde{\mathrm{Cov}}\, v_i = \lambda_i v_i, \quad \forall i \in I,$$

such that $(v_i)_{i \in I}$ is total in the orthogonal complement of the kernel of $\widetilde{\mathrm{Cov}}$ i.e.

$$L^2(0,T) = \mathrm{Ker}\, \widetilde{\mathrm{Cov}} \overset{\perp}{\oplus} \overline{\mathrm{Span}\{v_i,\, i \in I\}}.$$

We define, for all $i \in I$,

$$\varphi_i = \frac{1}{\lambda_i} \int_0^T v_i(t) z(t) \, \mathrm{d}t.$$

Then, following the proof of Lemma 1, $(\varphi_i)_{i \in I}$ is an orthonormal sequence of V of eigenvectors of Cov, with the same sequence of corresponding eigenvalues

$$\mathrm{Cov}\, \varphi_i = \lambda_i \varphi_i, \quad \forall i \in I.$$

Finally, we use the following lemma.

Lemma 3. *$(\varphi_i)_{i \in I}$ is total in the orthogonal complement of the kernel of* Cov.

Proof of Lemma 3. Let $W = \overline{\mathrm{Span}\{\varphi_i,\, i \in I\}}$. We shall prove that

$$W^\perp = \mathrm{Ker}\, \mathrm{Cov}. \tag{2.10}$$

Let $w \in W^\perp$. Then $\forall i \in I$,

$$\int_0^T v_i(t) \langle z(t), w \rangle \, \mathrm{d}t = \lambda_i \langle w, \varphi_i \rangle = 0.$$

Since $L^2(0,T)$ is decomposed as (2.1), it means

$$\langle z(\cdot), w \rangle \in \mathrm{Ker}\, \widetilde{\mathrm{Cov}},$$

that is to say

$$\left\langle \int_0^T (z(t), w)_V z(t) \, \mathrm{d}t,\, z(s) \right\rangle = 0 \quad \text{a.e. } s \in [0,T].$$

Multiplying by $(z(s), w)_V$ and integrating over $s \in [0,T]$, we get

$$\| \mathrm{Cov}\, w \|^2 = 0,$$

which proves $W^\perp \subset \mathrm{Ker}\, \mathrm{Cov}$.

Conversely, let $\psi \in \mathrm{Ker}\, \mathrm{Cov}$. Then

$$\int_0^T \langle z(t), \psi \rangle^2 \, \mathrm{d}t = \langle \mathrm{Cov}\, \psi, \psi \rangle = 0,$$

so that

$$\langle z(t), \psi \rangle = 0 \quad \text{a.e. } t \in [0,T].$$

Multiplying by $v_i(t)$ and integrating over $[0,T]$,

$$\langle \varphi_i, \psi \rangle = 0, \quad \forall i \in I,$$

which, by density, proves the inverse inclusion. Finally we proved (2.10), but as W is closed in V, it implies $W = (\mathrm{Ker}\, \mathrm{Cov})^\perp$. This ends the proof. □

Finally we proved (2.6), which ends the proof of Prop. 1.

2.2 Solution of the POD problem

Let $\tilde{\pi}^l$ be any orthogonal projector of V of rank l, and $(\psi_1, \dots, \psi_l)$ any orthonormal basis of $\operatorname{Im} \tilde{\pi}^l$. By Pythagoras' theorem,

$$\|z - \tilde{\pi}^l z\|^2_{L^2(0,T;V)} = \|z\|^2_{L^2(0,T;V)} - \underbrace{\|\tilde{\pi}^l z\|^2_{L^2(0,T;V)}}_{J(\tilde{\pi}^l)}, \tag{2.11}$$

with $J(\tilde{\pi}^l)$ bearing the following expressions

$$J(\tilde{\pi}^l) = \int_0^T \sum_{k=1}^{l} \langle z(t), \psi_k \rangle^2 \, \mathrm{d}t = \sum_{k=1}^{l} \langle \operatorname{Cov} \psi_k, \psi_k \rangle.$$

This transforms Problem (2.2) into the equivalent problem

$$\max_{\tilde{\pi}^l} J(\tilde{\pi}^l).$$

Naming π^l the orthogonal projector onto the first l vectors φ_i defined by Prop. 1, i.e.

$$\operatorname{Im} \pi^l = \operatorname{Span}\{\varphi_1, \dots, \varphi_l\},$$

we remark that J takes on it the value

$$J(\pi^l) = \sum_{i=1}^{l} \lambda_i.$$

It is then sufficient to show that for an arbitrary $\tilde{\pi}^l$

$$J(\tilde{\pi}^l) \leq \sum_{i=1}^{l} \lambda_i. \tag{2.12}$$

Using (2.6), each $\psi \in V$ has a unique decomposition

$$\psi = \widehat{\psi} + \overline{\psi}, \quad \text{with } \overline{\psi} = \sum_{i=1}^{\operatorname{Card} I} \langle \psi, \varphi_i \rangle \varphi_i, \tag{2.13}$$

such that $\widehat{\psi} \in \operatorname{Ker} \operatorname{Cov}$. Since Cov is self-adjoint and continuous, we remark

$$\langle \operatorname{Cov} \psi, \psi \rangle = \langle \operatorname{Cov} \overline{\psi}, \overline{\psi} \rangle = \sum_{i \in I} \lambda_i \langle \psi, \varphi_i \rangle^2.$$

Then J becomes

$$J(\tilde{\pi}^l) = \sum_{k=1}^{l} \underbrace{\sum_{i \in I} \lambda_i \langle \psi_k, \varphi_i \rangle^2}_{K(\psi_k)}.$$

Following the idea developped in [Fan49, Th. 1] we interpose the eigenvalue λ_l

$$\begin{aligned} K(\psi_k) &= \sum_{i=1}^{\operatorname{Card} I} \underbrace{(\lambda_i - \lambda_l)}_{\leq 0 \text{ when } i \geq l} \langle \psi_k, \varphi_i \rangle^2 + \lambda_l \underbrace{\sum_{i \in I} \langle \psi_k, \varphi_i \rangle^2}_{\leq \|\psi_k\|^2 = 1} \\ &\leq \sum_{i=1}^{l} (\lambda_i - \lambda_l) \langle \psi_k, \varphi_i \rangle^2 + \lambda_l. \end{aligned}$$

Taking the sum on k,

$$J(\tilde{\pi}^l) \leq \sum_{i=1}^{l} (\lambda_i - \lambda_l) \underbrace{\sum_{k=1}^{l} \langle \psi_k, \varphi_i \rangle^2}_{\leq \|\varphi_i\|^2 = 1} + l \lambda_l,$$

which directly leads to (2.12). This proves of the first part of Prop. 3.

Finally, in order to prove (2.7), according to (2.11) we need to show

$$\|z\|^2_{L^2(0,T;V)} = \sum_{i=1}^{\infty} \lambda_i.$$

A priori, following (2.13), z is decomposed as

$$z(t) = \hat{z}(t) + \sum_{i=1}^{\infty} \langle z(t), \varphi_i \rangle \varphi_i,$$

with $\hat{z}(t) \in \operatorname{Ker} \operatorname{Cov}$. Actually we show that $\hat{z} = 0$. Indeed, if $\operatorname{Ker} \operatorname{Cov} \neq \{0\}$ (otherwise the result is obvious), let $(\hat{\varphi}_j)_{j \in J}$ be a Hilbertian basis of it, J either finite or $\mathbb{N} \setminus \{0\}$. Then

$$\|\hat{z}\|^2_{L^2(0,T;V)} = \int_0^T \sum_{j \in J} \langle \hat{z}(t), \hat{\varphi}_j \rangle^2 \, \mathrm{d}t,$$

but remarking $\langle \hat{z}(t), \hat{\varphi}_j \rangle = \langle z(t), \hat{\varphi}_j \rangle$ because of (2.6) it becomes

$$\|\hat{z}\|^2_{L^2(0,T;V)} = \sum_{j \in J} \int_0^T \langle z(t), \hat{\varphi}_j \rangle^2 \, \mathrm{d}t = \sum_{j \in J} \langle \operatorname{Cov} \hat{\varphi}_j, \hat{\varphi}_j \rangle = 0.$$

Then (2.2) comes from Parseval's equality. This ends the proof of Prop. 3.

Chapter 3

Galerkin approximation with proper orthogonal decomposition

3.1 Introduction

In general, the simulation of partial differential equations resorts to discretization techniques such as finite differences, finite elements, or finite volumes. This typically results into discrete systems of large dimensions, hence the solution process can be rather costly, especially in situations when many computational iterations are required, as often occurs in design, control applications and inverse modeling.

In order to obtain reduced-order models, two main approaches are generally used. The first one consists in analyzing the dynamics operator of the system considered and retaining only the "most significant parts". *Modal Analysis* (linear or non-linear normal modes), but also the *Moment Matching Method* [16, 2] and *Balanced Truncation* [50, 51] belong to this first family. Unfortunately, for complex and large systems, these tools can be difficult to use in practice, since e.g. the eigenmodes are costly to obtain.

The second strategy is more data-oriented in the sense that it mainly uses snapshots of the system to perform its reduction. The *Reduced Basis* [37, 46, 43, 52, 47] and the *Proper Orthogonal Decomposition* (POD) [35, 32, 34, 22, 53] are two techniques belonging to this second family. This second approach consists in projecting the system onto subspaces of reduced sizes, albeit containing the major part of the expected dynamical solution. The aim is to obtain low-dimensional systems capturing the essence of the phenomena of interest.

Proper Orthogonal Decomposition, also known as Karhunen-Loève decomposition or principal component analysis, is a method initially introduced for analyzing multidimensional data. This method essentially provides an orthonormal basis for representing the given data in an optimal manner with respect to a quadratic criterion. The work in [Kos43] has been

pioneering in the development of the POD technique. In fluid mechanics, POD has been successfully used to access the coherent structures in turbulent flows [24], and it is now widely used in engineering in general.

Despite its relative simplicity of development and use, the POD technique has some limitations, since in particular it does not guarantee stability e.g. when parametric variations are considered [1]. Moreover, existing error estimates are expressed with respect to quantities which are not controlled in the construction of the POD basis [23]. This latter important issue is our primary concern here.

In this article, we propose new error estimates for the POD-based Galerkin approximation of the solutions of some classical and widely used PDE systems. First, we briefly recall the foundations of the POD decomposition. Then we derive the estimates for linear and non-linear parabolic equations, and also for linear hyperbolic systems. Finally, the theoretical results are confronted with numerical tests in various situations including a complex 3D biomechanical heart model.

3.2 Classical principles of POD reduction

Considering $z(x,t)$ the solution of a PDE problem, the POD-based reduced order modeling, or more simply POD reduction, consists in building a spatial Galerkin approximation $z^l(x,t)$ of $z(x,t)$ in the POD space $V^l = \mathrm{Span}(\varphi_1,\dots,\varphi_l)$. Then the key point is to be able to control the *reduction error*, namely

$$\|z - z^l\|_{L^2(0,T;V)}.$$

We tackle this problem in the following section.

3.3 New estimates for the POD reduction error

In this section, our objective is to derive POD-reduction error estimates bounded by approximation terms which can be conveniently controlled in the construction of the POD basis, namely, in particular without undue time derivatives.

For the sake of generality and homogeneity with the existing literature, we introduce the classical abstract mathematical framework. Nevertheless, to fix the ideas the reader can keep in mind that in the examples considered, the abstract spaces H and V will typically correspond to $L^2(\Omega)$ and $H_0^1(\Omega)$, respectively.

Let $(V, ((\cdot,\cdot)), \|\cdot\|)$ and $(H, (\cdot,\cdot), |\cdot|)$ be two separable Hilbert spaces with continuous and dense embedding $V \hookrightarrow H$, i.e.

$$|v| \le C_\Omega \|v\|, \quad \forall v \in V.$$

We choose H as the pivot space — namely, we perform the identification of H with its dual H' — and

$$V \hookrightarrow H \hookrightarrow V'.$$

Let a be a symmetric bilinear form on V, continuous, and coercive, namely,

$$a(v,w) \le C_a \|v\| \|w\|, \quad \forall v, w \in V,$$
$$a(v,v) \ge c_a \|v\|^2, \quad \forall v \in V.$$

Then a also defines a scalar product on V and we denote by $\|\cdot\|_a$ the associated norm.

We point out that in our estimations we use C to denote a generic positive constant, independent of all discretization parameters, and that may take different values at various occurrences, including in the same equation.

3.3.1 Galerkin estimates for linear parabolic problems with H-orthogonal projectors

We formally introduce the *abstract parabolic equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}(u(t),v) + a(u(t),v) = (f(t),v), \quad \forall v \in V, \tag{3.1}$$
$$u(0) = u_0. \tag{3.2}$$

Equation (3.1) is to be understood in the sense of distributions in time. We then have the following existence and uniqueness result [17, XVIII, §3.2, Th. 1, §3.3, Th. 2].

Proposition 4. *Assume $f \in L^2(0,T;H)$ and $u_0 \in H$. Then there exists a unique solution u of Eqs. (3.1)-(3.2) such that*

$$u \in L^2(0,T;V) \cap C([0,T];H), \quad \frac{\mathrm{d}u}{\mathrm{d}t} \in L^2(0,T;V').$$

Considering now a finite-dimensional subspace V^l, we formally introduce the spatial Galerkin approximation u^l of u

$$u^l(t) \in V^l, \tag{3.3}$$
$$\frac{\mathrm{d}}{\mathrm{d}t}(u^l(t),v^l) + a(u^l(t),v^l) = (f(t),v^l), \quad \forall v^l \in V^l, \tag{3.4}$$
$$u^l(0) = u_0^l. \tag{3.5}$$

Since V^l is a finite-dimensional space, it is easy to prove the following result [17, XVIII, §3.3.1, Lemma 1].

Proposition 5. *Assume $f \in L^2(0,T;H)$ and $u_0^l \in V^l$. Then there exists a unique solution u^l of Eqs. (3.4)-(3.5) such that*

$$u^l \in C([0,T];V^l), \quad \frac{\mathrm{d}u^l}{\mathrm{d}t} \in L^2(0,T;V^l).$$

Note that we have more regularity here than for the continuous solution only because we are considering a finite dimensional problem, and of course the corresponding estimates are not uniform with respect to the discretization.

Finally, let π_H^l and π_V^l respectively denote the H-orthogonal and V-orthogonal projectors of V onto the reduction space V^l. For all $v \in V$

$$|v - \pi_H^l v| = \inf_{v^l \in V^l} |v - v^l|,$$
$$\|v - \pi_V^l v\| = \inf_{v^l \in V^l} \|v - v^l\|.$$

With a view to estimating the *reduction error* $\|u - u^l\|_{L^2(0,T;V)}$, classical error estimates are of the form [17]

$$\begin{aligned}\|u - u^l\|_{L^2(0,T;V)} \leq C\Big(&\|u - \pi_V^l u\|_{L^2(0,T;V)} + \Big\|\frac{\partial}{\partial t}(u - \pi_V^l u)\Big\|_{L^2(0,T;V)} \\ &+ |u_0^l - \pi_V^l u_0|\Big).\end{aligned} \tag{3.6}$$

However, the POD criterion (2.2) does not provide a direct control on the time-derivative term in the right-hand side, and our objective is to circumvent this difficulty. To that end we use the *H-projection error*, still in the same $L^2(0,T;V)$-norm, i.e.

$$\|u - \pi_H^l u\|_{L^2(0,T;V)}.$$

and the following result holds.

Proposition 6. *For all $T > 0$,*

$$\|u - u^l\|_{L^2(0,T;V)} \leq C(|\pi_H^l u_0 - u_0^l| + \|u - \pi_H^l u\|_{L^2(0,T;V)}).$$

Proof. We split $u - u^l$ into two parts

$$u - u^l = p^l + q^l,$$

where $p^l = u - \pi_H^l u$ and $q^l = \pi_H^l u - u^l$. Since $q^l \in V^l$, and using the definition of u^l, q^l satisfies the variational equation

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(q^l(t), v^l) + a(q^l(t), v^l) = &-(f(t), v^l) + \frac{\mathrm{d}}{\mathrm{d}t}(\pi_H^l u(t), v^l) \\ &+ a(\pi_H^l u(t), v^l), \quad \forall v^l \in V^l.\end{aligned}$$

The projection π_H^l satisfies $(\pi_H^l u(t), v^l) = (u(t), v^l)$, so that using the definition of u we get

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} q^l(t), v^l\right) + a(q^l(t), v^l) = -a(p^l(t), v^l).$$

Taking $v^l = q^l(t)$, we then obtain the energy estimate

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\{|q^l|^2\}(t) + \|q^l(t)\|_a^2 = -a(p^l(t), q^l(t)),$$

which we now integrate on $[0, T]$ to obtain, combined with Young's inequality,

$$\int_0^T \|q^l(t)\|_a^2 \,\mathrm{d}t \leq |q^l(0)|^2 + \int_0^T \|p^l(t)\|_a^2 \,\mathrm{d}t.$$

This directly entails

$$\|q^l\|_{L^2(0,T;V)} \leq C\left(|q^l(0)|^2 + \|p^l\|_{L^2(0,T;V)}\right),$$

using the properties of the scalar product a, and the triangle inequality

$$\|u(t) - u^l(t)\| \leq \|p^l(t)\| + \|q^l(t)\|$$

concludes the proof. □

Therefore, via the introduction of π_H^l we avoid the time derivative appearing in the right-hand side of the more standard estimate (3.6). However, we now need to deal with the approximation term $\|u - \pi_H^l u\|_{L^2(0,T;V)}$, which is the topic of the next section.

3.3.2 Galerkin estimates for linear parabolic problems with V-orthogonal projectors

Note first that, since V^l is finite-dimensional, we have an inverse inequality of the form

$$\exists \alpha^l > 0, \quad \forall v^l \in V^l, \quad \|v^l\| \leq \alpha^l |v^l|.$$

Hence, π_H^l is continuous as an endomorphism of V, as can be seen by directly writing

$$\|\pi_H^l v\| \leq \alpha^l |\pi_H^l v| \leq \alpha^l |v| \leq C\alpha^l \|v\|.$$

However, as inverse inequality constants blow up when the dimension of the V_l subspace increases, we will obtain some better insight by using the V-projection as follows

$$\begin{aligned}
\|\pi_H^l v\| &\leq \|\pi_H^l v - \pi_V^l v\| + \|\pi_V^l v\| \\
&\leq \alpha^l |\pi_H^l v - \pi_V^l v| + \|v\| = \alpha^l |\pi_H^l (v - \pi_V^l v)| + \|v\| \\
&\leq \alpha^l |v - \pi_V^l v| + \|v\|.
\end{aligned}$$

Denoting by $\mathcal{L}(V)$ the space of V-endomorphisms and by $\mathcal{L}(V,H)$ the space of linear operators from V to H, this entails

$$\|\pi_H^l\|_{\mathcal{L}(V)} \leq 1 + \alpha^l \|\mathrm{Id} - \pi_V^l\|_{\mathcal{L}(V,H)},$$

where Id is the identity operator and the inverse inequality constant is now multiplied by a projection error term which can be conjectured to vanish in various cases when increasing l, since V is more regular than H. Hence, we can transform any estimate with $\|v - \pi_H^l v\|$ into an estimate with $\|v - \pi_V^l v\|$. Indeed, as π_H^l and π_V^l project onto the same subspace we have

$$\begin{aligned} v - \pi_H^l v &= (\mathrm{Id} - \pi_H^l)(v - \pi_V^l v) \\ &= (\mathrm{Id} - (\pi_H^l - \pi_V^l))(v - \pi_V^l v). \end{aligned}$$

Since $(\pi_H^l - \pi_V^l)v = \pi_H^l(v - \pi_V^l v)$ for all $v \in V$, we remark that

$$(\pi_H^l - \pi_V^l)v = \begin{cases} \pi_H^l v & \text{if } v \in (V^l)^\perp, \\ 0 & \text{if } v \in V^l, \end{cases}$$

denoting by $(V^l)^\perp$ the V-orthogonal complement of V^l. Hence,

$$\|\pi_H^l - \pi_V^l\|_{\mathcal{L}(V)} \leq \|\pi_H^l\|_{\mathcal{L}(V)}.$$

Defining

$$\rho_l = \|\pi_H^l\|_{\mathcal{L}(V)}, \quad \sigma_l = \|\pi_H^l - \pi_V^l\|_{\mathcal{L}(V)},$$

we can then convert the projector by writing

$$\begin{aligned} \|v - \pi_H^l v\| &\leq (1 + \sigma_l)\|v - \pi_V^l v\| && (3.7) \\ &\leq (1 + \rho_l)\|v - \pi_V^l v\| && (3.8) \end{aligned}$$

from which we directly infer the following estimate.

Corollary 1. *For all $T > 0$,*

$$\|u - u^l\|_{L^2(0,T;V)} \leq C\left(|\pi_H^l u_0 - u_0^l| + (1 + \sigma_l)\|u - \pi_V^l u\|_{L^2(0,T;V)}\right).$$

However, we do not formally assert that, for a general reduction space, neither ρ_l nor σ_l have a bounded behavior with respect to l. This behavior is likely to be dependent on the specific types of variational problem and Galerkin reduction considered, and can be numerical assessed when no analytical treatment is at hand.

Note that the last term in the right-hand side of (1) is the quantity that is in fact minimized in the construction of POD subspaces. Furthermore, for POD reduction subspaces the sequences (σ_l) and (ρ_l) remain bounded in a large class of situations, see Section 3.3.5 for some theoretical insight.

3.3.3 Extension to a non-linear parabolic equation

We formally introduce the *abstract non-linear parabolic equation*

$$\frac{\mathrm{d}}{\mathrm{d}t}(u(t),v)+a(u(t),v)=(f(t,u(t)),v),\quad \forall v\in V, \tag{3.9}$$

$$u(0)=u_0. \tag{3.10}$$

Unlike in Equation (3.1), f is some $[0,T]\times V\to V$ function. The general theory is very delicate. Especially the solution may explode in finite time. We provide the following proposition, where we assume that f is Lipschitz-continuous in the second variable. As proven in the appendix, this guarantees, for any $T>0$, the well-posedness of Equations (3.9)-(3.10) in the same spaces as in the linear case.

Proposition 7. *Assume $u_0\in H$, $f\in C([0,T]\times H;H)$, and that f is L-Lipschitz continuous in its second variable, i.e. that there exists a constant L such that*

$$\forall t\in[0,T],\quad \forall h_1,h_2\in H,\quad |f(t,h_1)-f(t,h_2)|\leq L|h_1-h_2|.$$

Assume also that the embedding $V\hookrightarrow H$ is compact. Then there exists a unique solution u of Eqs. (3.9)-(3.10) *such that*

$$u\in L^2(0,T;V)\cap C([0,T];H),\quad \frac{\mathrm{d}u}{\mathrm{d}t}\in L^2(0,T;V').$$

Let now u^l be the spatial Galerkin approximation of u in V^l

$$u^l(t)\in V^l, \tag{3.11}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(u^l(t),v^l)+a(u^l(t),v^l)=(f(t,u^l(t)),v^l),\quad \forall v^l\in V^l, \tag{3.12}$$

$$u^l(0)=u_0^l. \tag{3.13}$$

More simply, with the Peano existence theorem, we obtain the following result.

Proposition 8. *Assume $u_0^l\in V^l$, $f\in C([0,T]\times H;H)$ and f is L-Lipschitz continuous in its second variable. Then there exists a unique solution u^l of* (3.11)-(3.13) *such that*

$$u^l\in C^1([0,T];V^l).$$

The proof of this result can also be seen as contained in that of Proposition 7, proven in the appendix.

We now show the following result for the reduction error.

Proposition 9. *For all $T > 0$,*

$$\|u - u^l\|_{L^2(0,T;V)} \leq C_1(L,T)\Big(|\pi_H^l u_0 - u_0^l| + C_2(L)\|u - \pi_H^l u\|_{L^2(0,T;V)}\Big), \quad (3.14)$$

where, for all $L > 0$,

$$C_1(L,T) = Ce^{LT}, \quad C_2(L) = C(L+1).$$

In addition, we have

$$\begin{aligned}\|u - u^l\|_{L^2(0,T;V)} \leq C_1(L,T)\Big(&|\pi_H^l u_0 - u_0^l| \\ &+ (1+\sigma_l)C_2(L)\|u - \pi_V^l u\|_{L^2(0,T;V)}\Big).\end{aligned} \quad (3.15)$$

Moreover, under the condition $L < \frac{c_a}{C_\Omega^2}$, we have the improved constants

$$C_1(L,T) = C_1(L) = \frac{C}{\sqrt{\frac{c_a}{C_\Omega^2} - L}}, \quad C_2(L) = \frac{C}{\sqrt{\frac{c_a}{C_\Omega^2} - L}},$$

where C_1 is now independent of T.

Proof. We split $u - u^l$ into two parts

$$u - u^l = p^l + q^l,$$

where $p^l = u - \pi_H^l u$ and $q^l = \pi_H^l u - u^l$. Using the same property of π_H^l as in the proof of Prop. 6, we obtain

$$\left(\frac{\mathrm{d}q^l}{\mathrm{d}t}(t), v^l\right) + a(q^l(t), v^l) = -a(p^l(t), v^l) + (f(t,u(t)) - f(t,u^l(t)), v^l).$$

Taking $v^l = q^l(t)$, and integrating on $[0,t]$, $0 \leq t \leq T$, we obtain

$$\begin{aligned}\frac{1}{2}|q^l(t)|^2 + c_a\|q^l\|^2_{L^2(0,t;V)} &\leq \frac{1}{2}|q^l(0)|^2 + C_a\|p^l\|_{L^2(0,t;V)}\|q^l\|_{L^2(0,t;V)} \\ &\quad + L\int_0^t |u(s) - u^l(s)| \cdot |q^l(s)|\,\mathrm{d}s \\ &\leq \frac{1}{2}|q^l(0)|^2 + L\|q^l\|^2_{L^2(0,t;H)} \\ &\quad + (C_a + LC_\Omega^2)\|p^l\|_{L^2(0,t;V)}\|q^l\|_{L^2(0,t;V)}.\end{aligned} \quad (3.16)$$

Let us first assume $L < \frac{c_a}{C_\Omega^2}$. Hence, for $t = T$ and using the continuous embedding $V \hookrightarrow H$, Eq. (3.16) entails

$$(c_a - LC_\Omega^2)\|q^l\|^2_{L^2(0,T;V)} \leq \frac{1}{2}|q^l(0)|^2 + (C_a + c_a)\|p^l\|_{L^2(0,T;V)}\|q^l\|_{L^2(0,T;V)}.$$

Using Young's inequality, we can then conclude as in Prop. 6.

Let us now consider the general case for L. By Young's inequality on (3.16),

$$|q^l(t)|^2 + c_a\|q^l\|^2_{L^2(0,t;V)} \leq |q^l(0)|^2 + \frac{(C_a + LC_\Omega^2)^2}{c_a}\|p^l\|^2_{L^2(0,T;V)} + 2L\|q^l\|^2_{L^2(0,t;H)}. \tag{3.17}$$

Then, we use Gronwall's inequality for $t \mapsto |q^l(t)|^2$, which leads to

$$|q^l(t)|^2 \leq e^{2Lt}\left(|q^l(0)|^2 + \frac{(C_a + LC_\Omega^2)^2}{c_a}\|p^l\|^2_{L^2(0,T;V)}\right).$$

Finally, re-incorporating this estimate in (3.17) gives, for $t = T$,

$$\|q^l\|^2_{L^2(0,T;V)} \leq \frac{e^{2LT}}{c_a}\left(|q^l(0)|^2 + \frac{(C_a + LC_\Omega^2)^2}{c_a}\|p^l\|^2_{L^2(0,T;V)}\right),$$

and we conclude for (3.14) as in Prop. 6.

Of course (3.15) directly follows like in Corollary 1. □

3.3.4 Galerkin estimates for the wave-like equation

Let us now consider the wave-like equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(y(t), v) + a(y(t), v) = (f(t), v), \quad \forall v \in V, \tag{3.18}$$

$$y(0) = y_0, \quad \frac{\mathrm{d}y}{\mathrm{d}t}(0) = \dot{y}_0, \tag{3.19}$$

for which we have the classical existence and uniqueness results [17, XVIII, §5.5.2, Th. 1, §5.5.3, Th. 2].

Proposition 10. *We assume $f \in L^2(0,T;H)$, $y_0 \in V$ and $\dot{y}_0 \in H$. Then there exists a unique solution y of Eqs. (3.18)-(3.19) such that*

$$y \in C([0,T];V), \quad \frac{\mathrm{d}y}{\mathrm{d}t} \in C([0,T];H), \quad \frac{\mathrm{d}^2y}{\mathrm{d}t^2} \in L^2(0,T;V').$$

As in Section 3.3.1, we formally introduce the spatial Galerkin approximation y^l of y

$$y^l(t) \in V^l, \tag{3.20}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}(y^l(t), v^l) + a(y^l(t), v^l) = (f(t), v^l), \quad \forall v^l \in V^l, \tag{3.21}$$

$$y^l(0) = y_0^l, \quad \frac{\mathrm{d}y^l}{\mathrm{d}t}(0) = \dot{y}_0^l. \tag{3.22}$$

And the following holds [17, XVIII, §5.3.1, Lemma 2].

Proposition 11. *We assume $f \in L^2(0,T;H)$, $y_0 \in V^l$ and $\dot{y}_0 \in V^l$. Then there exists a unique solution y^l of Eqs. (3.20)-(3.22) such that*

$$y^l \in C^1([0,T];V^l), \quad \frac{\mathrm{d}^2 y^l}{\mathrm{d}t^2} \in L^2(0,T;V^l).$$

The error estimate between the solutions of Problems (3.18)-(3.19) and (3.20)-(3.22) is given by the following Proposition.

Proposition 12. *For all $T > 0$,*

$$\begin{aligned} &\|y - y^l\|_{L^2(0,T;V)} + \left\|\frac{\mathrm{d}}{\mathrm{d}t}(y - y^l)\right\|_{L^2(0,T;H)} \\ &\leq C\Bigg\{\sqrt{T}\left(\|y_0 - \pi_H^l y_0\| + \|\pi_H^l y_0 - y_0^l\| + |\pi_H^l \dot{y}_0 - \dot{y}_0^l|\right) \\ &\qquad + \|y - \pi_H^l y\|_{L^2(0,T;V)} + (1+T)\left\|\frac{\mathrm{d}}{\mathrm{d}t}\left(y - \pi_H^l y\right)\right\|_{L^2(0,T;V)}\Bigg\}, \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} &\|y - y^l\|_{L^2(0,T;V)} + \left\|\frac{\mathrm{d}}{\mathrm{d}t}(y - y^l)\right\|_{L^2(0,T;H)} \\ &\leq C\Bigg\{\sqrt{T}(\|y_0 - \pi_H^l y_0\| + \|\pi_H^l y_0 - y_0^l\| + |\pi_H^l \dot{y}_0 - \dot{y}_0^l|) \\ &\qquad + (1+\sigma_l)\Big(\|y - \pi_V^l y\|_{L^2(0,T;V)} + (1+T)\left\|\frac{\mathrm{d}}{\mathrm{d}t}(y - \pi_V^l y)\right\|_{L^2(0,T;V)}\Big)\Bigg\}. \end{aligned} \tag{3.24}$$

Proof. We split $y - y^l$ into two parts

$$y - y^l = p^l + q^l,$$

where $p^l = y - \pi_H^l y$ and $q^l = \pi_H^l y - y^l$. Since $q^l \in V^l$, and using the definition of y^l, q^l verifies the variational equation

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(q^l(t), v^l) + a(q^l(t), v^l) = -(f(t), v^l) + \frac{\mathrm{d}^2}{\mathrm{d}t^2}(\pi_H^l y(t), v^l) \\ + a(\pi_H^l y(t), v^l), \quad \forall v^l \in V^l. \end{aligned}$$

The projector π_H^l verifies $(\pi_H^l y(t), v^l) = (y(t), v^l)$, so that using the definition of y we get

$$\left(\frac{\mathrm{d}^2 q^l}{\mathrm{d}t^2}(t), v^l\right) + a(q^l(t), v^l) = -a(p^l(t), v^l).$$

We infer the energy balance by taking $v^l = \frac{\mathrm{d}q^l}{\mathrm{d}t}(t)$, viz.

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\{\left|\frac{\mathrm{d}q^l}{\mathrm{d}t}\right|^2 + \|q^l\|_a^2\right\}(t) = -a\left(p^l(t), \frac{\mathrm{d}q^l}{\mathrm{d}t}(t)\right).$$

Performing an integration by parts over time and using Young's inequality in the right-hand side, we have

$$\left|\frac{\mathrm{d}q^l}{\mathrm{d}t}(t)\right|^2 + (1-\eta)\|q^l(t)\|_a^2 \leq \left|\frac{\mathrm{d}q^l}{\mathrm{d}t}(0)\right|^2 + \|q^l(0)\|_a^2 + 2a(p^l(0), q^l(0)) + \frac{1}{\eta}\|p^l(t)\|_a^2 + \theta\|q^l\|_{L^2(0,T;a)}^2 + \frac{1}{\theta}\left\|\frac{\mathrm{d}p^l}{\mathrm{d}t}\right\|_{L^2(0,T;a)}^2.$$

By integration on $[0,T]$ again and taking $\eta = \frac{1}{4}$, $\theta = \frac{1}{4T}$, we get

$$\left\|\frac{\mathrm{d}q^l}{\mathrm{d}t}\right\|_{L^2(0,T;H)}^2 + \|q^l\|_{L^2(0,T;a)}^2 \leq C\left\{T\left(\|p^l(0)\|_a^2 + \|q^l(0)\|_a^2 + \left|\frac{\mathrm{d}q^l}{\mathrm{d}t}(0)\right|^2\right) + \|p^l\|_{L^2(0,T;a)}^2 + T^2\left\|\frac{\mathrm{d}p^l}{\mathrm{d}t}\right\|_{L^2(0,T;a)}^2\right\}.$$

Using the properties of the scalar product a, we get back to the $\|\cdot\|$ norm, and the triangular inequality

$$\|y(t) - y^l(t)\| \leq \|p^l(t)\| + \|q^l(t)\|$$

ends the proof for (3.23), whence (3.24) directly follows. □

3.3.5 Boundedness of (σ^l)

As already mentioned, we need some characterization of the behavior of the sequences $(\rho_l)_{l\geq 1}$ and $(\sigma_l)_{l\geq 1}$ in order for the above estimations to be meaningful. To provide some insight into this issue we give some examples of reduction subspaces for which these sequences can be proven to be bounded.

Let us start by showing this boundedness when the Galerkin subspace is given by finite element discretization procedures. To fix the ideas we consider a standard $\mathbf{P}_1$ discretization, but this result can be extended with ease to most other finite element procedures.

Proposition 13. *Let Ω be an open convex polyhedral subset of $\mathbb{R}^2$ or $\mathbb{R}^3$. Let $H = L^2(\Omega)$ and $V = H_0^1(\Omega)$. Let $(\mathcal{T}_h)_{h>0}$ be a quasi-uniform family of triangulations of Ω, and V_h the $\mathbf{P}_1$-Lagrange finite element subspace of V built on $\mathcal{T}_h$, with $\pi_{h,H}$ the H-orthogonal projector onto V_h. Then $(\pi_{h,H})_{h>0}$ is bounded in $\mathcal{L}(V)$, i.e.*

$$\forall h > 0, \quad \forall v \in V, \quad \|\pi_{h,H} v\| \leq C\|v\|. \tag{3.25}$$

Proof. Let us introduce a family of Clément interpolation operators $(\mathcal{C}_h)_{h>0}$ associated with $(\mathcal{T}_h)_{h>0}$ and uniformly bounded from V to V_h [13]. Since $(\mathcal{T}_h)_{h>0}$ is quasi-uniform, an inverse inequality holds [12, Th. 3.2.6], so that

$$\begin{aligned} \|\pi_{h,H} v\| &\leq \|\pi_{h,H} v - \mathcal{C}_h v\| + \|\mathcal{C}_h v\| \\ &\leq Ch^{-1} |\pi_{h,H} v - \mathcal{C}_h v| + \|\mathcal{C}_h v\|. \end{aligned}$$

Remark that, by the characterization of an orthogonal projector,

$$|\pi_{h,H} v - \mathcal{C}_h v| \leq |v - \pi_{h,H} v| + |v - \mathcal{C}_h v| \leq 2|v - \mathcal{C}_h v|. \tag{3.26}$$

Now we use the property

$$\forall h > 0, \quad \forall v \in V, |v - \mathcal{C}_h v| \leq Ch\|v\|, \tag{3.27}$$

and the boundedness of $(\mathcal{C}_h)_{h>0}$ in $\mathcal{L}(V)$ [13, Th. 2]. This shows our result. □

We now consider spectral analysis, namely, taking Galerkin subspaces provided by the eigenmodes of the bilinear form a. We thus assume the embedding $V \hookrightarrow H$ to be compact, which is satisfied when Ω is bounded, $H = L^2(\Omega)$ and $V = H_0^1(\Omega)$. Then there exists a Hilbertian basis of H, (w_i), characterized by

$$\begin{gathered} a(w_i, v) = \omega_i^2 (w_i, v), \\ 0 < \omega_1 \leq \omega_2 \leq \ldots, \quad \omega_i \underset{i \to \infty}{\longrightarrow} +\infty. \end{gathered}$$

Introducing $\tilde{w}_i = \frac{1}{\omega_i} w_i$, $(\tilde{w}_i)_{i \geq 1}$ is a Hilbertian basis of V for the scalar product associated with a.

Proposition 14. *Assuming $V^l = \mathrm{Span}(w_1, \ldots, w_l)$, the sequences (ρ_l) and (σ_l) are bounded.*

Proof. We remark

$$a(v, \tilde{w}_i)\tilde{w}_i = (v, w_i) w_i.$$

Summing this identity from 1 to l directly entails that $\pi_H^l = \pi_a^l$, where π_a^l is the a-orthogonal projector of V onto V^l. Moreover

$$c_a^{1/2} \|\pi_a^l v\| \leq \|\pi_a^l v\|_a \leq \|v\|_a \leq C_a^{1/2} \|v\|,$$

which leads to

$$\|\pi_a^l\|_{\mathcal{L}(V)} \leq \left(\frac{C_a}{c_a} \right)^{1/2}.$$

and the property $\pi_H^l = \pi_a^l$ concludes the proof. □

As a third example, we will consider the case of the POD subspaces arising from the analysis of the homogeneous wave-like equation. The following result is very straightforward to establish by decomposing the solution on the eigenmodes. We also refer to [18] for related discussions.

Proposition 15. *Let y be the solution of the homogeneous wave equation, namely, (3.18)-(3.19) with $f = 0$. Denoting by $(\varphi_i(T))_{i=1}^l$ the POD basis constructed with y over $[0,T]$, for $T \in [0,\infty)$,*

$$\varphi_i(T) \underset{T\to\infty}{\longrightarrow} \widetilde{w}_{\sigma(i)},$$

where σ describes a certain reordering determined by the initial conditions. Therefore

$$\rho(l,T) = \|\pi_H^l(T)\|_{\mathcal{L}(V)} \underset{T\to\infty}{\longrightarrow} C.$$

3.4 Numerical validations

In this section, we provide some numerical validations of the above error estimates for some examples of one-dimensional problems. As in the rest of the paper, we only consider the case of *self-reduction*, i.e. when the reduction space we use is the POD space generated from the trajectory of the reference solution u itself. In particular, we aim at assessing whether or not the sequences (σ_l) and (ρ_l) are bounded in several examples. Of course, since the reference solution is needed to compute the POD space, it is mostly a theoretical study on synthetic data. However, this is an important first step before tackling the practical situation of *parametric variations*, when a unique POD space is used to reduce a family of solutions. This issue will be addressed in forthcoming papers.

3.4.1 Discretization and corresponding reduction for parabolic problem

Here, we consider the reduction of (3.9)-(3.10) with the one-dimensional non-linear equation

$$\begin{aligned} \partial_t u - \partial_{xx}^2 u = f(t,u) \quad &\text{in } (0,T)\times(0,1), \\ u(t,0) = u(t,1) &= 0, \\ u(0,x) = u_0(x) \quad &\text{in } (0,1). \end{aligned}$$

where now f is simply a $[0,T]\times\mathbb{R} \to \mathbb{R}$ function. In the sequel, $H = L^2(0,1)$ and $V = H_0^1(0,1)$. We keep the notations $(\cdot,\cdot)$ and $a(\cdot,\cdot)$ for their respective scalar products.

Semi-discrete solution and reduced form

Let u_h be the $\mathbf{P}_1$ approximation of u on the regular mesh $(x_i)_{i=1}^{N_h}$

$$x_i = ih, \quad 1 \le i \le N_h, \quad h = \frac{1}{N_h + 1},$$

associated with the basis of shape functions $(e_i)_{i=1}^{N_h}$. The discrete solution u_h is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}(u_h(t), e_i) + a(u_h(t), e_i) = (f(t, u_h(t)), e_i), \quad 1 \le i \le N_h, \tag{3.28}$$

$$u_h(0, x) = u_{h,0}(x). \tag{3.29}$$

This discrete solution is the reference solution with which the reduced solutions will be compared. However, the POD basis $(\varphi_1, \dots, \varphi_l)$ itself will be constructed based on the fully-discrete solution u_h^n described below. Nevertheless, we emphasize that we do not consider time discretization issues in this paper, hence in our numerical trials we choose the time step "sufficiently small" for the discrete solution to be converged in time.

The corresponding reduced form u_h^l of u_h satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(u_h^l(t), \varphi_k) + a(u_h^l(t), \varphi_k) = (f(t, u_h^l(t)), \varphi_k), \quad 1 \le k \le l, \tag{3.30}$$

$$u_h^l(0, x) = u_{h,0}^l(x). \tag{3.31}$$

As before, we have local existence and uniqueness of the solutions u_h and u_h^l in the classical sense.

We also similarly introduce the $L^2(0,1)$-orthogonal and $H_0^1(0,1)$-orthogonal projectors $\pi_{L^2}^l$ and $\pi_{H_0^1}^l$ from V_h onto V^l, and the corresponding sequences

$$\rho_l = \|\pi_{L^2}^l\|_{\mathcal{L}(H_0^1)},$$
$$\sigma_l = \|\pi_{L^2}^l - \pi_{H_0^1}^l\|_{\mathcal{L}(H_0^1)},$$

that still verify

$$\sigma_l \le \rho_l.$$

We can then directly adapt Proposition 9.

Proposition 16. *Assume $u_{h,0} \in V_h$, $u_{h,0}^l \in V^l$, $f \in C([0,T] \times \mathbb{R}; \mathbb{R})$, and f is Lipschitz continuous in its second variable. Then there exists unique classical and global solutions u_h and u_h^l of Equations* (3.28)-(3.29) *and* (3.30)-(3.31)*, respectively. Moreover, for all $T > 0$,*

$$\begin{aligned} \|u_h - u_h^l\|_{L^2(0,T;H_0^1)} &\le C\left(\|\pi_{L^2}^l u_{h,0} - u_{h,0}^l\|_{L^2(\Omega)} + \|u_h - \pi_{L^2}^l u_h\|_{L^2(0,T;H_0^1)}\right) \\ &\le C\Big(\|\pi_{L^2}^l u_{h,0} - u_{h,0}^l\|_{L^2(\Omega)} \\ &\qquad + (1 + \sigma_l)\|u_h - \pi_{H_0^1}^l u_h\|_{L^2(0,T;H_0^1)}\Big). \end{aligned} \tag{3.32}$$

Note that – if we assume that the POD basis is constructed in a continuous-time discrete-space setting – the last term in this error estimate directly corresponds to the POD remainder, recall (2.7), hence it is perfectly controlled in the POD construction itself.

Full discretization

We use the classical θ-method as a time discretization scheme. In order to compute the reference solution u_h, we need the non-reduced mass matrix M and stiffness matrix K

$$M = [(e_j, e_i)]_{1\le i,j\le N_h}, \quad K = [a(e_j, e_i)]_{1\le i,j\le N_h},$$

and the reaction term application $F : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ of coefficient

$$(F(t,\beta))_i = \int_0^1 f\Big(t, \sum_{k=1}^{N_h} \beta_k e_k(x)\Big) e_i(x)\,\mathrm{d}x$$

Then the vector $U_h(t) \in \mathbb{R}^{N_h}$ concatenating the coordinates of $u_h(x,t)$ in $(e_i(x))_{i=1}^{N_h}$ satisfies

$$\begin{aligned} M\dot{U}_h(t) + KU_h(t) &= F(t, U_h(t)),\\ U_h(0) &= U_{h,0}. \end{aligned}$$

Next we apply a semi-implicit time scheme by θ-method

$$\frac{1}{\Delta t} M\left(U_h^{n+1} - U_h^n\right) + K\left(\theta U_h^{n+1} + (1-\theta)U_h^n\right) = \theta F(t^{n+1}, U_h^{n+1}) + (1-\theta)F(t^n, U_h^n),$$

leading to a non-linear problem in U_h^{n+1} once U_h^n is known, which we can solve for using a Newton algorithm.

For the reduced solutions u_h^l, we follow exactly the same path (spatial discretization $U_h^l(t)$, full discretization $(U_h^{l,n})$), except that we substitute the POD basis $(\varphi_i)_{i=1}^l$ for the finite element basis $(e_i)_{i=1}^{N_h}$. This gives the reduced mass and stiffness matrices M^l and K^l, and the reduced reaction term F^l. We emphasize that although these reduced matrices are of limited size, they are full. We call Φ^l the matrix

$$\Phi^l = [\varphi_1, \ldots, \varphi_l] \quad \in \mathbb{R}^{N_h \times l},$$

where vectors φ_i are expressed as column vectors of coordinates in $(e_i)_{i=1}^{N_h}$. Then we obtain the following relations between reduced and non-reduced

operators

$$M^l = (\Phi^l)^\top M \Phi^l,$$
$$K^l = (\Phi^l)^\top K \Phi^l,$$
$$F^l(t, \beta^l) = (\Phi^l)^\top F(t, \Phi^l \beta^l),$$
$$\mathrm{d}_{\beta^l} F^l(t, \beta^l) = (\Phi^l)^\top \mathrm{d}_\beta F(t, \Phi^l \beta^l) \Phi^l,$$

where $\mathrm{d}_{\beta^l} F^l$ and $\mathrm{d}_\beta F$ denote the differential quantities needed in the Newton algorithm computations.

3.4.2 Sharpness indicators for the new estimates

Here, we define the quantities that we need to check to ensure the *sharpness* of the new estimates. Note first that the POD eigenvalues λ_i typically decrease exponentially, hence the maximum POD rank to be considered is set as

$$\lambda_{l_{\max}+1} \leq 10^{-12} \lambda_1 \leq \lambda_{l_{\max}}. \tag{3.33}$$

in order to preserve sufficient K-orthogonality of the POD basis $(\varphi_i)_{i=1}^l$ when we perform the diagonalization of the covariance matrix. Indeed, since the covariance matrix is ill-conditioned, this orthogonality tends to rapidly deteriorate with l and we should preserve

$$\left\| (\Phi^l)^\top K \Phi^l - \mathrm{Id}_l \right\| \leq \varepsilon_{\mathrm{tol}}.$$

Summary of the estimation chain

In Prop. 16, we mainly handle three error terms:
– the *reduction error* $R(l)$

$$R(l) = \|u_h - u_h^l\|_{L^2(0,T;H_0^1)};$$

– the *L^2-projection error* $Q(l)$

$$Q(l) = \|u_h - \pi_{L^2}^l u_h\|_{L^2(0,T;H_0^1)};$$

– and the *H_0^1-projection error* $P(l)$

$$P(l) = \|u_h - \pi_{H_0^1}^l u_h\|_{L^2(0,T;H_0^1)}, \tag{3.34}$$

that coincides, in this situation of self-reduction, with the *POD remainder* $\varepsilon(l)$

$$\varepsilon(l) = \left\{ \sum_{i>l} \lambda_i \right\}^{1/2}.$$

Note that we do not need the reduced solution u_h^l to compute $Q(l)$ nor $P(l)$. Moreover, we point out that these quantities – except for $P(l)$ which can be obtained as a by-product of the covariance computation – are auxiliary quantities only computed to evaluate the reduction performance and accuracy.

If we prescribe the initial condition

$$u_{h,0}^l = \pi_{L^2}^l u_{h,0},$$

then the first term in the right-hand side of Eq. (3.32) vanishes. Thus we summarize the estimation chain by

$$R(l) \leq CQ(l) \leq C(1+\sigma_l)P(l).$$

Let us introduce the following *sharpness indicator*

$$\mathcal{S}_{\mathrm{Gal}}(l) = \frac{R(l)}{Q(l)},$$

which is clearly bounded under the assumptions of Prop. 16, but can be considered in a more general framework. By contrast, note that for the second inequality that only relies on (3.7), the bound

$$\frac{Q(l)}{(1+\sigma_l)P(l)} \leq 1$$

always holds.

Finally, we aim at numerically verifying, in various cases, that:

- the maximum POD rank $l_{\max}$ is reasonably limited compared to the number of degrees of freedom of the system

  $$l \ll N_h,$$

  for the POD subspace to accurately approximate the solution, recall (3.34);
- the quantity $\max_{1\leq l\leq l_{\max}} \rho_l$, that is an upper bound of $\max_{1\leq l\leq l_{\max}} \sigma_l$, remains small;
- the indicator $\mathcal{S}_{\mathrm{Gal}}(l)$, $1 \leq l \leq l_{\max}$, remains numerically bounded, especially in cases of strong non-linearities.

Computation of the (ρ_l) and (σ_l) sequences

In order to compute ρ_l, it is useful to manipulate the L^2-orthonormal basis $(\psi_k)_k$ that results from a Gram–Schmidt L^2-orthonormalization on the H_0^1-orthonormal POD basis (φ_k). Thus

$$V^l = \mathrm{Span}(\psi_1, \ldots, \psi_l),$$

with ψ_i independent of the POD rank l. Denoting by Ψ^l the matrix

$$\Psi^l = [\psi_1, \ldots, \psi_l],$$

where the ψ_i elements are expressed as column vectors of coordinates in $(e_i)_{i=1}^{N_h}$, we notice that $\pi_{L^2}^l$ has the following matrix in $(e_i)_{i=1}^{N_h}$

$$\Pi_{L^2}^l = \Psi^l (\Psi^l)^\top M \quad \in \mathbb{R}^{N_h \times N_h}.$$

We use the definition of ρ_l

$$\begin{aligned}\rho_l^2 &= \sup_{v \in V_h \setminus \{0\}} \frac{\int_0^1 ([\pi_{L^2}^l v]'(x))^2 \, \mathrm{d}x}{\int_0^1 v'(x)^2 \, \mathrm{d}x} \\ &= \sup_{\beta \in \mathbb{R}^{N_h} \setminus \{0\}} \frac{\beta^\top (\Pi_{L^2}^l)^\top K_h \Pi_{L^2}^l \beta}{\beta^\top K_h \beta}.\end{aligned}$$

Then ρ_l is the solution of the "largest K-eigenvalue" problem

$$\rho_l = \sup\left\{\omega \geq 0 \,|\, \exists \beta \in \mathbb{R}^{N_h} \setminus \{0\},\, (\pi_{L^2}^l)^\top K \pi_{L^2}^l \beta = \omega^2 K \beta\right\}.$$

Similarly, let Φ^l be the matrix

$$\Phi^l = [\varphi_1, \ldots, \varphi_l],$$

and $\widetilde{\Pi}_{L^2}^l$ be the matrix of the truncated projector $(\pi_{L^2}^l - \pi_{H_0^1}^l)$

$$\widetilde{\Pi}_{L^2}^l = \Psi^l (\Psi^l)^\top M - \Phi^l (\Phi^l)^\top K.$$

Then σ_l is the solution of the problem

$$\sigma_l = \sup\left\{\omega' \geq 0 \,|\, \exists \beta \in \mathbb{R}^{N_h} \setminus \{0\},\, (\widetilde{\Pi}_{L^2}^l)^\top K \widetilde{\Pi}_{L^2}^l \beta = \omega'^2 K \beta\right\}.$$

These properties will allow the numerical evaluation of the sequences.

3.4.3 Numerical experiments and validation for parabolic problems

We present three numerical cases of POD reduction on the generic discrete parabolic equation (3.30)-(3.31). The corresponding parameters are gathered in Table 3.1. In all these cases, we take $\theta = \frac{2}{3}$ in the θ-method for time discretization, and $N_h = 100$ for the spatial discretization.

Case	Case A	Case B	Case C
nb. timesteps	10^3	10^3	800
Δt	10^{-4}	10^{-4}	10^{-5}
$u_0(x)$	$\mathbf{1}_{[\frac{1}{3},\frac{2}{3}]}(x)$	$\frac{27}{4}x^2(1-x)$	$\frac{27}{4}x^2(1-x)$
$f(t,u)$	$9u$	$10u^2$	$100u^2$

Table 3.1: Cases of study for the reduction of parabolic equations

Case A: Lipschitz continuous reaction term

Case A satisfies the assumptions of Prop. 16 since f is linear with respect to u. We display the corresponding results in Figs. 3.1 and 3.2. In these figures the POD rank l varies from 1 to $l_{\max}$.

Figure 3.1, as an indication, shows the shape in space and time of the non-reduced solution u_h, and a numerical comparison between the indicators $P(l)$, $Q(l)$ and $R(l)$.

Figure 3.2 displays the sequence of POD constants ρ_l and their truncated versions σ_l, together with the sharpness indicator $\mathcal{S}_{\mathrm{Gal}}(l)$.

In this simple case, all our verifications are successful, namely,

- the POD remainder decreases at an exponential rate, and $l_{\max} = 10$;
- the POD constants ρ_l are of magnitude $O(1)$ and remain bounded with l. Also, the improvement provided by σ_l is limited;
- as expected with Prop. 16, the sharpness indicator is bounded and of small value, viz.

$$\mathcal{S}_{\mathrm{Gal}} \in [0.3,\, 0.8].$$

Cases B and C: super-linear reaction term

By contrast, the other two cases B and C are beyond the assumptions of Prop. 16 because we consider super-linear reaction terms. Moreover, while case B remains bounded, case C appears to explode in finite time, which is why we reduced the time range in this case while keeping a similar number of time steps, see Table 3.1. Even though our above error estimates do not hold in these cases, we can still compute the same numerical error quantities for illustrative purposes. Case B is reported on in Figs. 3.3 and 3.4, and case C in Figs. 3.5 and 3.6.

In fact, the maximum POD rank as well as the behavior and magnitude of the indicators ρ_l, σ_l, and $\mathcal{S}_{\mathrm{Gal}}$ reveal no significant difference compared to case A. Furthermore, the reduction error still decreases as fast as the POD remainder.

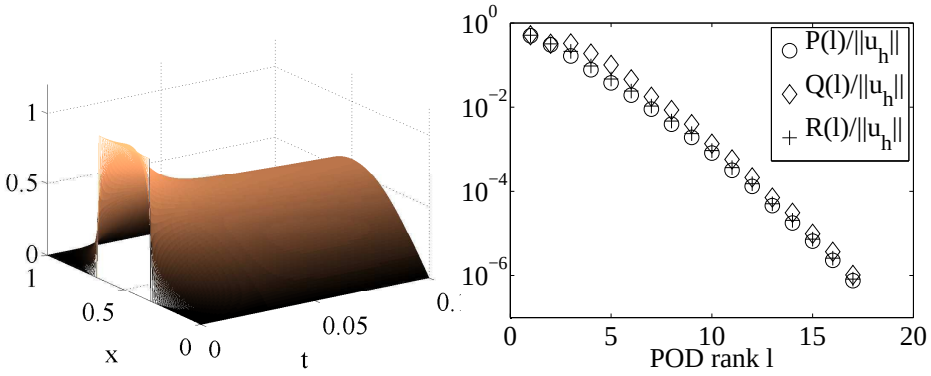

Figure 3.1: Case A. Left: full solution. Right: relative errors of H_0^1-projection, L^2-projection and reduction.

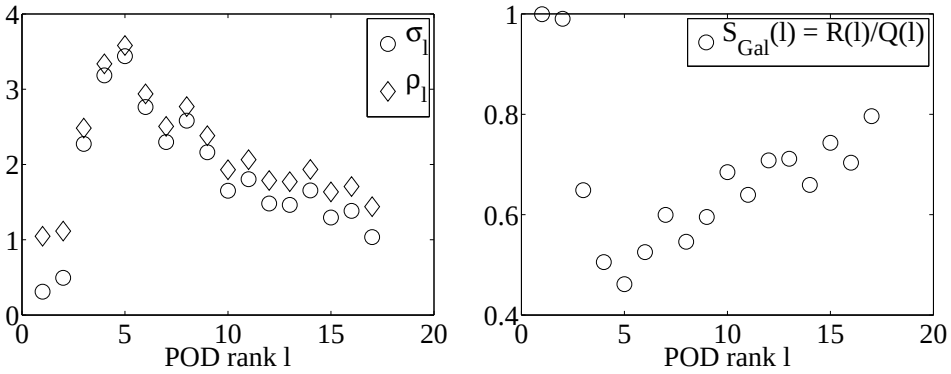

Figure 3.2: Case A: Left: POD sequences (ρ_l) and (σ_l). Right: sharpness indicator $(\mathcal{S}_{\mathrm{Gal}}(l))$.

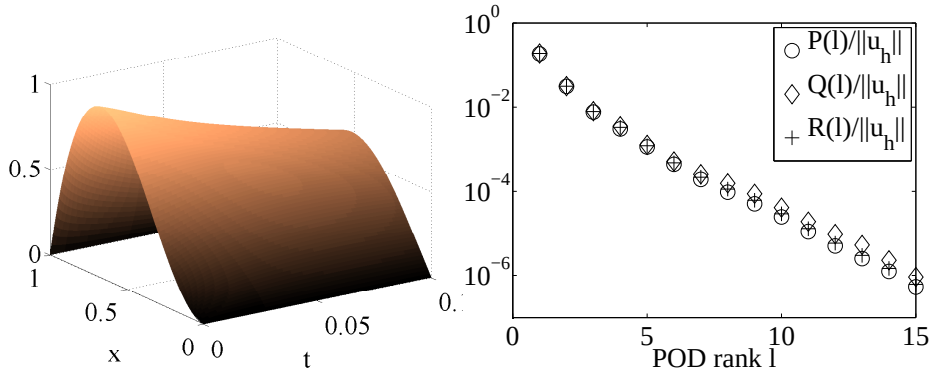

Figure 3.3: Case B. Left: full solution. Right: relative errors of H_0^1-projection, L^2-projection and reduction.

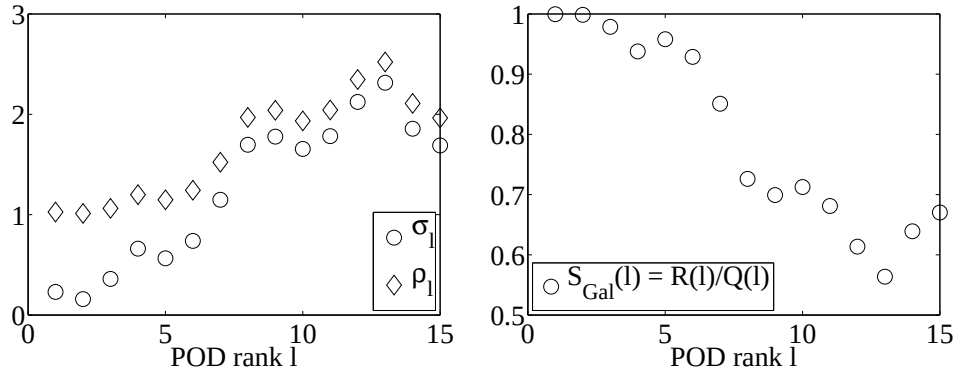

Figure 3.4: Case B: Left: POD sequences (ρ_l) and (σ_l). Right: sharpness indicator $(\mathcal{S}_{\mathrm{Gal}}(l))$.

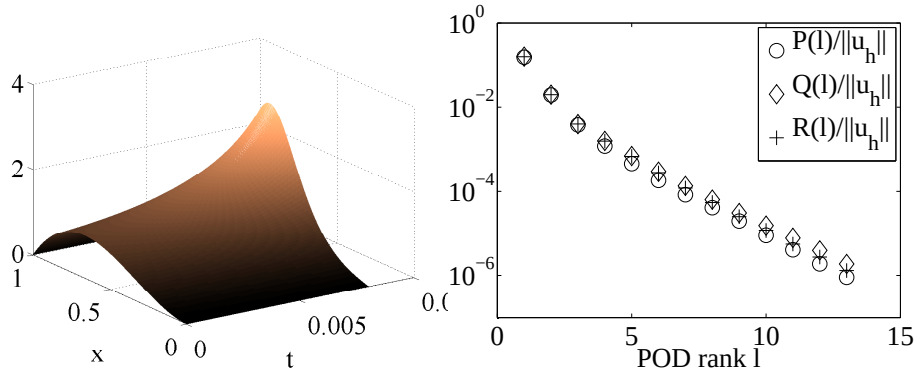

Figure 3.5: Case C. Left: full solution. Right: relative errors of H_0^1-projection, L^2-projection and reduction.

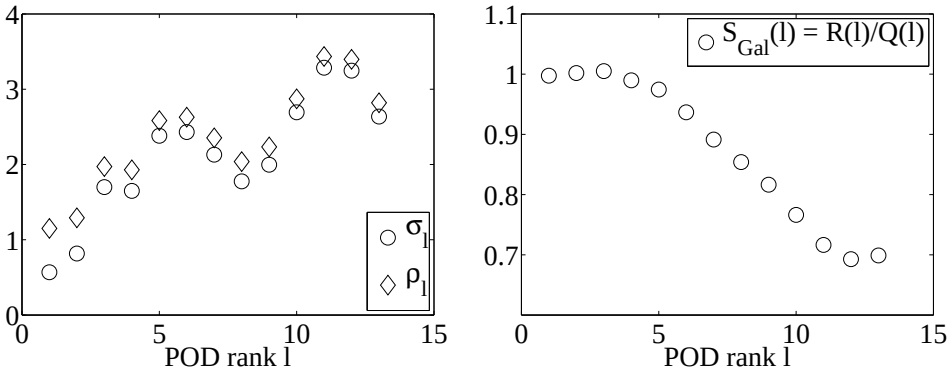

Figure 3.6: Case C: Left: POD sequences (ρ_l) and (σ_l). Right: sharpness indicator $(\mathcal{S}_{\mathrm{Gal}}(l))$.

3.4.4 Numerical assessment of the wave equation reduction

Considering now the 1D homogeneous wave equation,

$$\begin{aligned} \partial^2_{tt} y - c^2 \partial^2_{xx} y = 0 \quad &\text{in } (0,T)\times(0,1), \\ y(t,0) = y(t,1) &= 0, \\ y(0,x) = y_0(x) \quad &\text{in } (0,1), \\ \partial_t y(0,x) = \dot{y}_0(x) \quad &\text{in } (0,1). \end{aligned}$$

we report on the numerical values obtained for the various error terms. We discretize in space with finite elements on a regular mesh, and in time with a Newmark scheme according to the classical parameters $\beta = \frac{1}{4}$ and $\gamma = \frac{1}{2}$, see e.g. [44]. We take a regular cutoff function for $y_0(x)$, and $\dot{y}_0(x) = 0$. The corresponding results are shown in Figs. 3.7 and 3.8.

We verify that the POD basis is very close to a set of $H^1_0(0,1)$-eigenmodes $(\tilde{w}_i)$ of the Dirichlet Laplacian as substantiated in Section 3.3.5. This also explains why σ_l is much lower than ρ_l, since the L^2 and H^1_0 projectors onto eigenspaces coincide.

Note that the estimate of Prop. 15 contains the first-order time derivative $\frac{\partial}{\partial t}(u - \pi^l_{H^1_0} u)$ which is not controlled by the POD construction. Nevertheless, we observe from Figure 3.7 that the POD reduction is very effective, and indeed converges nearly-exponentially with the POD-rank.

3.5 Reduction of a complex system: a biomechanical heart model

In this section, we explore and validate the application of the analyzed POD-based reduction method on a three-dimensional continuum mechanics model, in *large displacements* and *large strains*, of a beating heart, coupled with an electrical model for its *activation*.

We start by briefly describing the major ingredients of this electromechanical model, in order to present the complexity of its multi-scale approach. A detailed discussion, with the elements that prove the physiological and thermomechanical consistance of the involved submodels, appears in [48]. Its discretization is described in [49], and a validation by confrontation with clinical data is given in [9]. Then, we present the result for the error terms defined in the previous sections for this direct simulation application. In constitutes the first step before adapting this study of a reduced-order heart model to the tackling of inverse problems of parameter estimation (see Chapter 6).

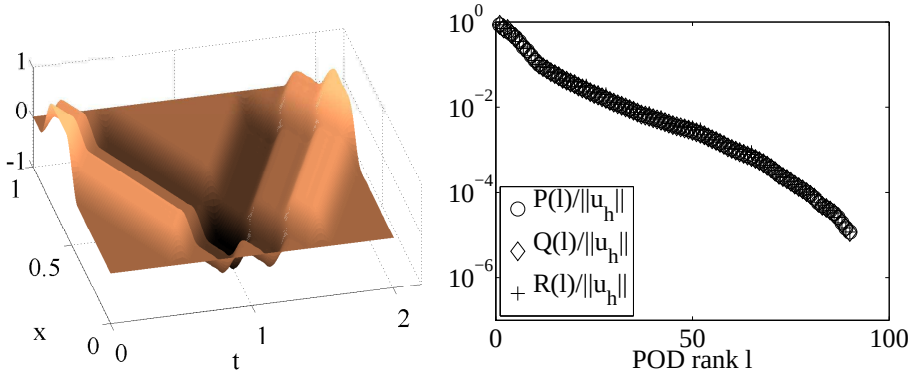

Figure 3.7: Homogeneous wave equation. Left: full solution. Right: relative errors of H_0^1-projection, L^2-projection and reduction.

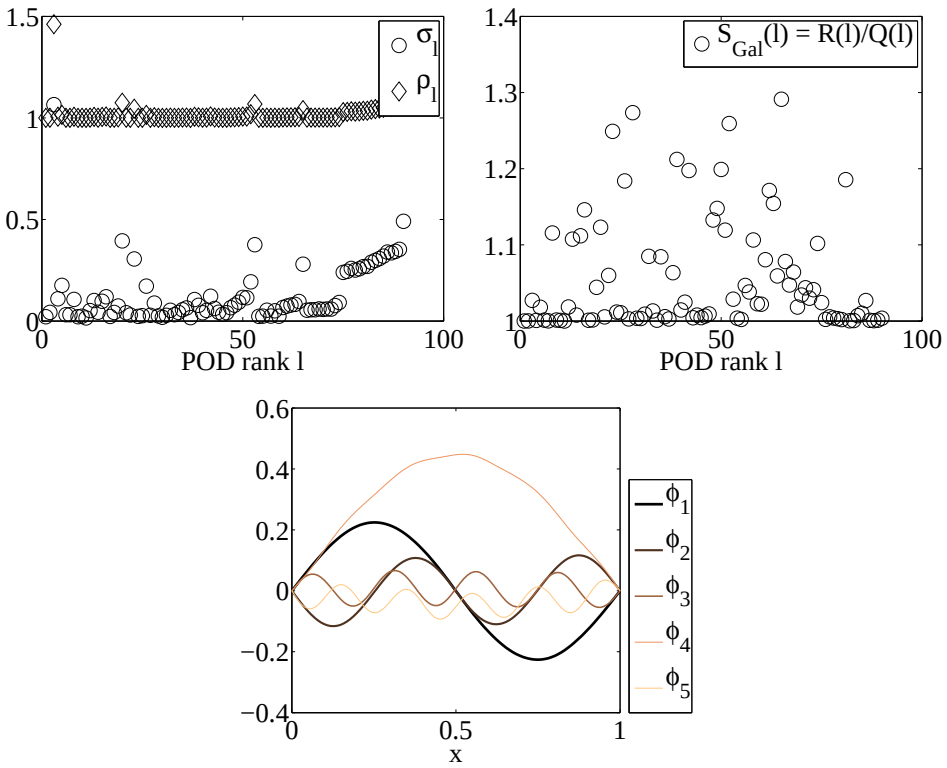

Figure 3.8: Homogeneous wave equation. Top-left: POD sequences (ρ_l) and (σ_l). Top-right: sharpness indicator $(\mathcal{S}_{\mathrm{Gal}}(l))$. Bottom: the first five POD modes.

3.5.1 Electromechanical heart model

The cardiac tissue is composed of long cells called myofibers. At each beat, these myofibers are subject to to an electrical activation which is due to some ionic exchanges across the membrane, and roughly manifests as a planar wave running from the *apex* (i.e. the bottom of the heart) to the *base* (i.e. the top). This introduces the muscle contraction, defining the *systole*, a phase when the blood is rapidly ejected from the heart, as opposed to the *diastole*, a longer phase when the muscle relaxes and the blood fills the heart.

In order to simulate such a complex phenomenon, the heart model considered contains some fundamental ingredients, namely

- a constitutive law accounting for both the active and passive aspects in the behavior of the muscle fibres;
- a representation of the electrical activation, i.e. the input in the constitutive law, that can be obtained from modeling approaches of various types and complexities;
- a geometrical (or "anatomical") description of the myocardium incorporating the fibre directions;
- a simplified model of the blood circulation inside and outside of the heart cavities;
- and also a model describing the opening and closure of the valves that separate the cavities from each other and from the external circulation.

Excitation-contraction law for the myofibers

Each myofiber is modelled as a unidimensional structure, and is composed of $O(10^5)$ units of contraction called *sarcomeres*. It is associated with a constitutive law, linking the active stress σ_c to the corresponding strain e_c, that is a chemically controlled, and relies on a model of *actin-myosin bridges* occurring within the sarcomeres.

Huxley [25] proposed some evolution PDEs with respect to time and strain modelling the density of these bridges. The resulting proposed formula for the stress σ_c also integrates the *Starling effect*, which is the chemical mechanism that explains why the stretch of the fibers at the end of the diastole actually helps to increase the muscle contraction during the systole that follows it. Using a moment-scaling method proposed by Zahalak [54], and provided an electric input $u(t)$ (see [5] for a proposed form), the constitutive law hence appears as an integro-differential relation

$$\sigma_c(t) = \mathcal{F}(\epsilon_c, \dot{\epsilon}_c, |\dot{\epsilon}_c|, |u|)(t).$$

Using some convenient substitutions, it can be shown that this relation possesses the properties of thermomechanical and kinetic compatibility [48].

Mechanical model of the cardiac tissue

We describe the full mechanical model associated with the myofiber using the classical tensor notations, namely

- the displacement field $\underline{y}$;
- the deformation gradient $\underline{\underline{F}} = \underline{\underline{1}} + \underline{\underline{\nabla}}\,\underline{y}$;
- the right Cauchy-Green deformation tensor $\underline{\underline{C}} = \underline{\underline{F}}^T \cdot \underline{\underline{F}}$;
- the Green-Lagrange strain tensor $\underline{\underline{E}} = \frac{1}{2}\left(\underline{\underline{C}} - \underline{\underline{1}}\right)$;
- and the second Piola-Kirchhoff stress tensor $\underline{\underline{\Sigma}}$ i.e. the stress tensor which is energy-conjugate to $\underline{\underline{E}}$.

Based on the above modeling ingredients, the second Piola-Kirchhoff stress tensor $\underline{\underline{\Sigma}}$ contains the active cardiac fibre law, a viscous stress component and a hyperelastic potential accounting for passive effects, these components being combined by means of a rheological model of Hill-Maxwell type [11, 21, 45].

Using a total Lagrangian formulation and denoting by Ω_H the reference domain corresponding to cardiac tissue, while the part of the boundary corresponding to ventricular endocardium is denoted by Γ, the principle of virtual work then gives

$$\int_{\Omega_H} \rho\, \ddot{\underline{y}} \cdot \underline{v}\, \mathrm{d}\Omega + \int_{\Omega_H} \underline{\underline{\Sigma}} : \mathrm{d}_{\underline{y}}\, \underline{\underline{e}} \cdot \underline{v}\, \mathrm{d}\Omega + \int_{\Gamma} P_0 \underline{\nu} \cdot \underline{\underline{F}}^{-1} \cdot \underline{v} J\, \mathrm{d}\Gamma = 0 \quad \forall \underline{v} \in V,$$

where V denotes a suitable space of displacement test functions, ρ the mass per unit volume, $\mathrm{d}_{\underline{y}}\, \underline{\underline{e}}$ the differential of the Green-Lagrange strain tensor with respect to the displacement, while P_0 is a prescribed intraventricular pressure.

3.5.2 Numerical validation of the reduced-order heart model

For the simulations presented hereafter an idealized left ventricle embedded with active fibers has been considered. The discretization is performed with $\mathbf{P}_1$-Lagrange finite elements in space (with about 1000 degrees of freedom), and a Newmark scheme in time [4]. We show some snapshots of the solution for the full finite element model in Fig. 3.9.

Although we do not have a theoretical estimate for the reduction error in this complex non-linear case, the three error terms appearing in the linear estimation chain still feature excellent decreasing rate and correlation, see Figs. 3.10 and 3.11. Analyzing our indicators reveals that their magnitude may slightly differ from the linear one-dimensional case, but again shows the effectiveness of the POD reduction, namely,

- $l_{\max} = 36$;
- ρ_l and σ_l are almost identical, and numerically bounded;

– the sharpness indicator established for the linear case is still bounded, and more precisely

$$\mathcal{S}_{\text{Gal}} \in [0.6,\, 1.0].$$

In Figure 3.10, we also display the evolution in time of the relative residual e^l defined as

$$e^l(t) = \frac{\|\underline{y}(t) - \underline{y}^l(t)\|_{L^\infty(\Omega)}}{\|y\|_{C([0,T];L^\infty(\Omega))}}. \tag{3.35}$$

We observe an excellent behavior of e^l, which roughly decreases by an order of magnitude for each addition of 10 modes in the POD basis.

3.6 Conclusion

We have proposed Galerkin estimates for the Proper Orthogonal Decomposition reduction of some classical PDEs. The numerical implementation of the reduction and some verifications were presented in this article. We have also demonstrated reduced simulations of a complex three-dimensional electromechanical model of the heart, where the validity of similar Galerkin estimates is numerically verified, even though no formal proof can be given in this case.

A special emphasis was placed on the derivation of POD-reduction error estimates in convenient norms and which can be controlled in the construction of the POD basis.

The present study can be extended in many directions. Firstly, as far as POD reduction of PDEs is concerned, one of the major difficulties lies in achieving stability of the POD basis with respect to e.g. parameter variations, initial and boundary conditions, and so on. This subject needs be further investigated. Secondly, filtering and estimation techniques for inverse modeling are extremely costly from a computational standpoint. This justifies – or even often requires – the use of POD-based reduced models and/or reduced filters and hence, the derivation errors estimates for such problems is crucial.

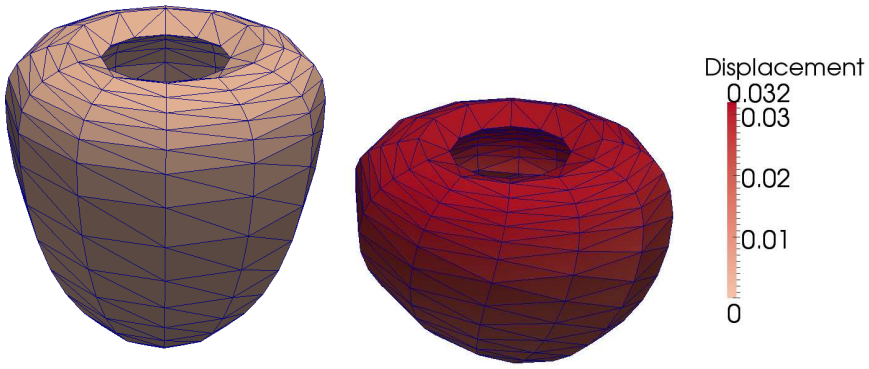

Figure 3.9: Model of a ventricle: snapshots of the displacement field at the beginning (left) and 40% (right) of the first cardiac cycle for the full model.

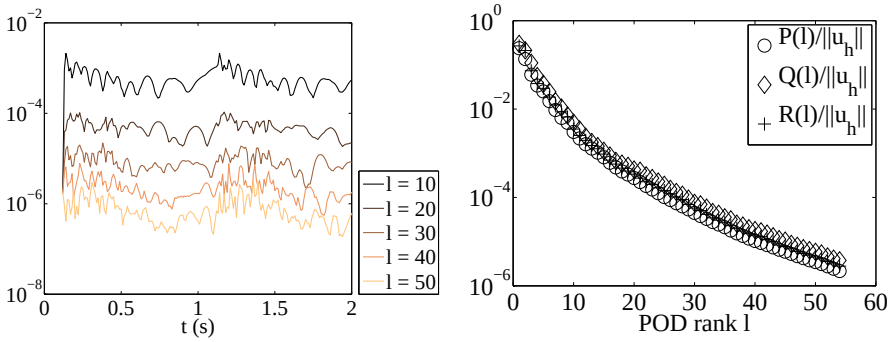

Figure 3.10: Model of a ventricle. Left: evolution in time of the residual e^l (see (3.35)), for several POD ranks. Right: relative errors of H_0^1-projection, L^2-projection and reduction.

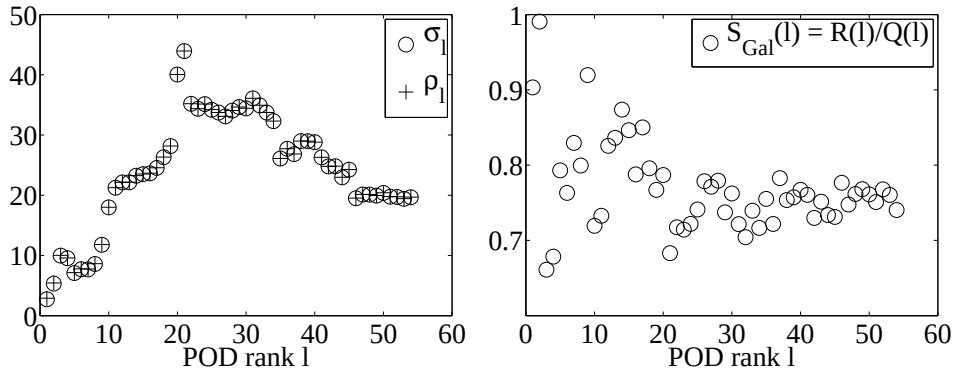

Figure 3.11: Model of a ventricle. Left: POD sequences (ρ_l) and (σ_l). Right: sharpness indicator $(\mathcal{S}_{\mathrm{Gal}}(l))$.

Chapter 4

Strategy of reduced-order modelling for parameter-dependent problems

In the above theoretical developments, the model reduction by Proper Orthogonal Decomposition was implicitly based on one solution of some partial differential equation with fixed (and hence not mentioned) parameter D. From now on, we tackle a particular extension to parameter-dependent equations. Indeed, the multiple-query situation, where we wish to compute quick and reliable approximations of the solutions for many points of the parameter space, would take advantage of the relevance and the efficiency of this reduced-order modelling.

Let some well-posed partial differential equation problem including a parameter vector $D \in \Theta \subset \mathbb{R}^p$, where Θ represents an admissible parametric set. Assume we wish to reduce it, and denote $u(t,x;D)$ its solution. The drawback of using a POD projector $\pi^l(D)$ associated with the solution $u(t,\cdot;D)$ over $t \in [0,T]$ does not only lie in the heaviness of a POD computation, which we shall limit to a minimum number of calls. Also, the projector evolves in a non-Euclidean space called a Grassmann manifold. This makes the question of its sensitivity with respect to D, namely the formal derivative $\frac{\partial \pi^l}{\partial D}$, difficult to describe. Hence, we shall not aim at defining and demarcating a certain domain of validity of the POD projector $\pi^l(D^\star)$ in a neighbourhood of $D^\star$ in Θ.

By contrast, we instead take the problem in the opposite direction. Given a simple, typically rectangular subdomain $\mathcal{D} \subset \Theta$, we aim at uniformly reducing the solutions $u(D)$, $D \in \mathcal{D}$, with a unique orthogonal projector $\overline{\pi}^l$. To that end, we investigate some cases of parametric dependence through the diffusion operator of some parabolic systems. We explain why a projector solution of an extended, *multi-POD* problem, appears in this way as an excellent candidate to control the maximum reduction error over

$\mathcal{D}$. We first provide a mathematical analysis with an extension of the previous estimates for linear parabolic equations. Secondly, we assess the power of this multi-POD methodology with a highly parameter-sensitive system known as the FitzHugh–Nagumo equations.

4.1 Galerkin error estimates for variations through a diffusion operator parameter

In order to fix the ideas and to corroborate a simple and general strategy based on POD, we introduce an abstract, linear, parametric parabolic equation, that is similar to Eqs. (3.1)-(3.2). Assume that only the diffusion operator is a regular function of a certain parameter $D \in \Theta \subset \mathbb{R}^p$, where Θ represents an admissible parametric domain, i.e.

$$\frac{\partial}{\partial t}(u(t;D),v) + a(u(t;D),v;D) = (f(t),v), \quad \forall v \in V, \tag{4.1}$$

$$u(0;D) = u_0, \tag{4.2}$$

and for any $D \in \Theta$, $a(D) \equiv a(\cdot,\cdot,D)$ is a symmetric bilinear form on V, continuous, and coercive, with constants now depending on D, i.e.

$$C_a(D) = \sup_{v,w\neq 0} \frac{a(v,w;D)}{\|v\| \cdot \|w\|}, \quad c_a(D) = \inf_{v\neq 0} \frac{a(v,v;D)}{\|v\|^2}.$$

Note that even when the operator $a(D)$ depends linearly on D, this simply exhibits a nonlinear dependence of the solution $u(D) \equiv u(\cdot;D)$ with respect to its parameter.

As in Section 3.3.1, considering a finite-dimensional subspace V^l of V, we define the spatial Galerkin approximation $u^l(D)$ of $u(D)$

$$u^l(t;D) \in V^l, \tag{4.3}$$

$$\frac{\partial}{\partial t}(u^l(t;D),v^l) + a(u^l(t;D),v^l;D) = (f(t),v^l), \quad \forall v^l \in V^l, \tag{4.4}$$

$$u^l(0;D) = u_0^l. \tag{4.5}$$

Provided that $f \in L^2(0,T;V)$, $u_0 \in H$ and $u_0^l \in V^l$, then Props. 4 and 5 guarantee that, for all $D \in \mathcal{D}$, there exists a unique solution u of Eqs. (4.1)-(4.2) such that

$$u(D) \in L^2(0,T;V) \cap C([0,T];H), \quad \frac{\partial u}{\partial t}(D) \in L^2(0,T;V'),$$

and a unique solution u^l of Eqs. (4.4)-(4.5) such that

$$u^l(D) \in C([0,T];V^l), \quad \frac{\partial u^l}{\partial t}(D) \in L^2(0,T;V^l).$$

Here, π^l denotes the V-orthogonal projector of V onto the reduction space V^l (we drop the former index V).

We start by adapting the main result of reduction, contained in Prop. 6 and Cor. 1. This consists in specifying the constants in the proof of Prop. 6 and passing to the supremum.

Proposition 17. *Let $\mathcal{D}$ a compact subset of Θ such that the infimum of the coercivity constants verifies*

$$c_a(\mathcal{D}) = \inf_{D\in\mathcal{D}} c_a(D) > 0.$$

Let also the supremum of the condition numbers be defined by

$$\kappa_a(\mathcal{D}) = \sup_{D\in\mathcal{D}} \frac{C_a(D)}{c_a(D)}.$$

Then, for all $T > 0$,

$$\begin{aligned}\|u - u^l\|_{C^0(\mathcal{D};L^2(0,T;V))} \leq \frac{1}{\sqrt{c_a(\mathcal{D})}}|\pi_H^l u_0 - u_0^l| \\ + \left(1 + \sqrt{\kappa_a(\mathcal{D})}\right)(1+\sigma_l)\|u - \pi^l u\|_{C^0(\mathcal{D};L^2(0,T;V))}.\end{aligned}$$

Proof. As in the proof of Prop. 6, we split $u - u^l$ into two parts

$$u - u^l = p^l + q^l,$$

where $p^l = u - \pi_H^l u$ and $q^l = \pi_H^l u - u^l$. With the same reasoning,

$$\int_0^T \|q^l(t)\|_{a(D)}^2 \,\mathrm{d}t \leq |q^l(0)|^2 + \int_0^T \|p^l(t)\|_{a(D)}^2 \,\mathrm{d}t.$$

Using the continuity and coercivity of $a(D)$,

$$\|q^l\|_{L^2(0,T;V)} \leq \frac{1}{\sqrt{c_a(D)}}|q^l(0)| + \sqrt{\frac{C_a(D)}{c_a(D)}}\|p^l\|_{L^2(0,T;V)},$$

which becomes, by a triangular inequality,

$$\|u - u^l\|_{L^2(0,T;V)} \leq \frac{1}{\sqrt{c_a(D)}}|q^l(0)| + \left(1 + \sqrt{\frac{C_a(D)}{c_a(D)}}\right)\|p^l\|_{L^2(0,T;V)}.$$

We end the proof by passing to the supremum in D over $\mathcal{D}$. □

Clearly, we need to push further this adaptation, because solving the saddle-point problem

$$\min_{\tilde{\pi}^l} \|u - \tilde{\pi}^l u\|_{C^0(\mathcal{D};L^2(0,T;V))}, \quad \text{i.e.} \quad \min_{\tilde{\pi}^l} \max_{D\in\Theta} \|u(D) - \tilde{\pi}^l u(D)\|_{L^2(0,T;V)}, \tag{4.6}$$

seems numerically very complex. Based on arguments from interpolation theory, we will instead consider a simpler, quadratic and parametrically discrete problem of the form

$$\min_{\tilde{\pi}^l} \sum_{m=1}^{M} \|u(D_m) - \tilde{\pi}^l u(D_m)\|^2_{L^2(0,T;V)}, \tag{4.7}$$

where the points D_m form a grid that is used for Lagrange interpolation, and for which can be computed. Indeed, using the interpolation error estimates for sufficiently regular functions on $\mathcal{D}$, we can control the minimum value of the saddle-point problem (4.6) with the minimum value of the quadratic problem (4.7) and the diameter of $\mathcal{D}$.

4.2 Proper orthogonal decomposition on parametric grids for interpolation

From now on, we assume that $\mathcal{D}$ is a rectangular subdomain of Θ, i.e.

$$\mathcal{D} = [a_1, b_1] \times \cdots \times [a_p, b_p], \quad a_i < b_i. \tag{4.8}$$

We begin by introducing a useful interpolation operator L.

4.2.1 Rectangular Lagrange interpolation operator

We build a Lagrange interpolation operator L of order $s \geq 1$, on a regular grid, onto the space $\mathcal{Q}_s$ of polynomials of p variables and of degree at most s in each variable, i.e.

$$\mathcal{Q}_s = \Big\{ \sum_{\substack{\alpha\in\mathbb{N}^p \\ 0\leq\alpha_i\leq s}} a_\alpha X^\alpha \ ; \ a_\alpha \in \mathbb{R} \Big\}, \quad \text{with } X^\alpha = X_1^{\alpha_1} \dots X_p^{\alpha_p},$$

of dimension s^p.

First, introducing the set of indices

$$I = (i_1, \dots, i_p) \in \{0, 1, \dots, s\}^p = \mathcal{I},$$

this regular subgrid of the rectangle $\mathcal{D}$ is defined by the $(s+1)^p$ points

$$D_I = \Big(a_1 + \frac{i_1}{s}(b_1 - a_1), \ \dots, \ a_p + \frac{i_p}{s}(b_p - a_p)\Big), \quad I \in \mathcal{I}.$$

In order to express an easy interpolation formula, let $\Sigma = \{\psi_I^\star\}_{I\in\mathcal{I}}$ be the finite set of linear forms on $\mathcal{Q}_s$ that are canonically defined by the subgrid $\{D_I\}_{I\in\mathcal{I}}$, i.e.

$$\psi_I^\star : \begin{array}{ccc} \mathcal{Q}_s & \to & \mathbb{R} \\ \psi & \mapsto & \langle \psi_I^\star, \psi\rangle = \psi(D_I). \end{array}$$

We verify below that Σ is a basis of $\mathcal{Q}_s^\star$, dual space of $\mathcal{Q}_s$, or in other words, since $\operatorname{Card}\Sigma = \dim\mathcal{Q}_s$, that *the parametric finite element* $(\Sigma,\Theta,\mathcal{Q}_s)$ *is unisolvent*.

Indeed, consider the reference, unidimensional Lagrange interpolation polynomials on $\left\{0,\frac{1}{s},\dots,\frac{s-1}{s},1\right\}$, i.e.

$$\hat{\psi}_m(x) = \prod_{\substack{n=0\\ n\neq m}}^{s} \frac{sx-n}{m-n}\,,$$

and their generalization to p variables on $\left\{0,\frac{1}{s},\dots,1\right\}^p = \frac{1}{s}\mathcal{I}$, indexed by $\mathcal{I}$ as

$$\hat{\psi}_I(X_1,\dots,X_p) = \prod_{j=1}^{p} \hat{\psi}_{i_j}(X_j), \quad I = (i_1,\dots,i_p)\in\mathcal{I}.$$

Then it is easy to check that the Lagrange polynomials corresponding to the current element $\mathcal{D}$, defined by substitution as

$$\psi_I(X_1,\dots,X_p) = \hat{\psi}_I\Big(\frac{X_1-a_1}{b_1-a_1},\dots,\frac{X_p-a_p}{b_p-a_p}\Big),$$

satisfy

$$\langle \psi_I^\star, \psi_J\rangle = \begin{cases} 1 & \text{if } I = J, \\ 0 & \text{otherwise,} \end{cases} \qquad \forall I,J\in\mathcal{I},$$

so that Σ is a dual basis of $\mathcal{Q}_s^\star$, with predual basis $\{\psi_I\}_{I\in\mathcal{I}}$ of $\mathcal{Q}_s$.

Hence, this naturally defines the linear interpolation operator L on the space $C^0(\mathcal{D})$ of continuous $\mathcal{D}\to\mathbb{R}$ functions by

$$Lv(D) = \sum_{i\in\mathcal{I}} \psi_I(D)v(D_I), \quad \forall v\in C^0(\mathcal{D}), \tag{4.9}$$

and it is by construction a projector onto $\mathcal{Q}_s$, i.e.

$$L\psi = \psi, \quad \forall\psi\in\mathcal{Q}_s.$$

With a view to deriving a classical interpolation error estimate in C^0 norm and directly on the projection error $u-\pi^I u$, the sequel mainly reconsiders the proof of [CR72, Th. 1]. This theorem was established for simplicial domains equipped with their corresponding polynomial spaces,

although it is actually valid for rectangular domains equipped with the space $\mathcal{Q}_s$ too. We also revisit this proof to underline that it principally relies on a strictly polynomial, moment-like property of the predual basis $\{\psi_I\}_{I\in\mathcal{I}}$, which can be stated as follows.

Lemma 4. *For all $\alpha \in \mathbb{N}^p$ such that $|\alpha| \le s$,*

$$\sum_{I\in\mathcal{I}} \psi_I(D)(D_I - D)^\alpha = 0.$$

Proof. Let $\psi \in \mathcal{Q}_s$. By the projection equation, for all $D \in \mathcal{D}$,

$$\psi(D) = \sum_{I\in\mathcal{I}} \psi_I(D)\psi(D_I). \tag{4.10}$$

On the one hand, testing (4.10) on $\psi = 1$ gives $\sum_{I\in\mathcal{I}} \psi_I = 1$. On the other hand, since ψ is a polynomial, then the expansion in finite Taylor series, for all $I \in \mathcal{I}$ and all $D \in \mathcal{D}$,

$$\psi(D_I) = \psi(D) + \sum_{s'=1}^{\deg\psi} \sum_{\substack{\beta\in\mathbb{N}^p \\ |\beta|=s'}} \frac{1}{\beta!}\partial_\beta\psi(D)(D_I - D)^\beta$$

holds. Hence, by substituting $\psi(D_I)$ by this expansion in (4.10), then

$$\sum_{I\in\mathcal{I}} \psi_I(D) \sum_{s'=1}^{\deg\psi} \sum_{\substack{\beta\in\mathbb{N}^p \\ |\beta|=s'}} \frac{1}{\beta!}\partial_\beta\psi(D)(D_I - D)^\beta = 0. \tag{4.11}$$

Now, consider, for any $\alpha \in \mathbb{N}^p$ such that $1 \le |\alpha| \le s$, the elementary polynomial $q_\alpha(D) = \frac{1}{\alpha!}D^\alpha$, that belongs to $\mathcal{Q}_s$. Remarking

$$\partial_{\alpha'} q_\alpha(D) = \begin{cases} 0 & \text{if } \alpha' \neq \alpha \text{ and } |\alpha'| \ge |\alpha|, \\ 1 & \text{if } \alpha' = \alpha, \end{cases}$$

it is easy to prove the statement

$$\mathcal{R}_{s'}: \quad \forall |\alpha| = s', \quad \sum_{I\in\mathcal{I}} \psi_I(D)(D_I - D)^\alpha = 0$$

by induction on $1 \le s' \le s$, by testing (4.11) with $\psi = q_\alpha$, $|\alpha| = s'$. □

We easily define in a similar way the operator L for vector-valued functions, and then, with a view to applying it to Proposition 17, as an endomorphism on $C^0(\mathcal{D}; L^2(0,T;V))$ by

$$Lw(t;D) = \sum_{I\in\mathcal{I}} \psi_I(D)w(t;D_I), \quad \forall w \in C^0(\mathcal{D}; L^2(0,T;V)). \tag{4.12}$$

4.2.2 Uniform projection error estimate by multi-POD criterion

We then obtain the following result. Note that it is valid for any sufficiently regular function $w : \mathcal{D} \to L^2(0,T;V)$.

Proposition 18. *Let $1 \le r \le s$ and $w \in C^{r+1}(\mathcal{D}; L^2(0,T;V))$, i.e.*

$$|w|_{C^{r+1}(\mathcal{D};L^2(0,T;V))} = \sup_{\substack{|\alpha|=r+1 \\ D\in\mathcal{D}}} \|\partial_\alpha w(D)\|_{L^2(0,T;V)} < \infty.$$

Then, for all $T > 0$,

$$\|w - \pi^l w\|_{C^0(\mathcal{D};L^2(0,T;V))} \le C_1(p,s)\,\delta_{r+1}(\mathcal{D})^{r+1}\,|w|_{C^{r+1}(\mathcal{D};L^2(0,T;V))} + C_2(p,s)\left\{\sum_{I\in\mathcal{I}} \|w(D_I) - \pi^l w(D_I)\|^2_{L^2(0,T;V)}\right\}^{1/2}.$$

with the constants

$$C_1(p,s) = \Big\|\sum_{I\in\mathcal{I}} |\hat\psi_I|\Big\|_{C^0([0,1]^p)}, \quad C_2(p,s) = \Big\|\Big\{\sum_{I\in\mathcal{I}} |\hat\psi_I|^2\Big\}^{1/2}\Big\|_{C^0([0,1]^p)},$$

and the measure of $\mathcal{D}$

$$\delta_{r+1}(\mathcal{D}) = \left\{\sum_{|\alpha|=r+1} \frac{(b_1-a_1)^{\alpha_1}}{\alpha_1!}\cdots\frac{(b_p-a_p)^{\alpha_p}}{\alpha_p!}\right\}^{1/(r+1)}.$$

Proof. Let $p^l = w - \pi^l w$. Clearly, $\pi^l w$ and then p^l have regularity C^{r+1} in the parameter. We use the triangular inequality

$$\|p^l\|_{C^0(\mathcal{D};L^2(0,T;V))} \le \|p^l - Lp^l\|_{C^0(\mathcal{D};L^2(0,T;V))} + \|Lp^l\|_{C^0(\mathcal{D};L^2(0,T;V))}.$$

We first estimate the interpolation error term. For all $I \in \mathcal{I}$ and all $D \in \mathcal{D}$, p^l admits the Taylor expansion

$$p^l(D_I) = p^l(D) + \sum_{s'=1}^{r}\sum_{|\alpha|=s'} \frac{1}{\alpha!}(D_I - D)^\alpha \partial_\alpha p^l(D) + \sum_{|\alpha|=r+1} \frac{1}{\alpha!}(D_I - D)^\alpha \partial_\alpha p^l(\eta_I(D,D_I)),$$

where $\eta_I(D,D_I) \in [D,D_I]$. We multiply this expression by $\psi_I(D)$ and take the sum for $I \in \mathcal{I}$. Then, by Lemma 4, this simply becomes

$$Lp^l(D) = p^l(D) + \sum_{|\alpha|=r+1} \frac{1}{\alpha!}\sum_{I\in\mathcal{I}} \psi_I(D)(D_I - D)^\alpha \partial_\alpha p^l(\eta_I(D,D_I)).$$

Taking the $L^2(0,T;V)$ norm and passing to the supremum in $D \in \mathcal{D}$ leads to

$$\|p^l - Lp^l\|_{C^0(\mathcal{D};L^2(0,T;V))} \leq C_1(p,s)\,\delta_{r+1}(\mathcal{D})^{r+1}\,|p^l|_{C^{r+1}(\mathcal{D};L^2(0,T;V))},$$

and we remark finally that $|p^l|_{C^{r+1}(\mathcal{D};L^2(0,T;V))} \leq |w|_{C^{r+1}(\mathcal{D};L^2(0,T;V))}$.

Then, taking the $L^2(0,T;V)$ norm on the interpolation formula (4.12), we estimate the second term by a function of the grid evaluations only. Finally, by a Cauchy–Schwarz inequality,

$$\|Lp^l\|_{C^0(\mathcal{D};L^2(0,T;V))} \leq C_2(p,s)\left\{\sum_{I\in\mathcal{I}} \|p^l(D_I)\|^2_{L^2(0,T;V)}\right\}^{1/2}.$$

□

In the next section, we solve the multi-POD problem (4.7) that appears in the right-hand side of Prop. 18, which itself helps as an estimate of the right-hand side of Prop. 17.

4.2.3 Multi-POD problem and final estimate

The solution of the multi-POD problem is a simple consequence of that of the standard POD problem solved in Chapter 2.

Let $M \geq 1$, and some functions $z_1, \ldots, z_M \in L^2(0,T;V)$. The multi-POD problem consists in finding the V-orthogonal projector $\pi^l \in \mathcal{L}(V)$ of rank l solution of

$$\min_{\tilde{\pi}^l} \sum_{m=1}^{M} \|z_m - \pi^l z_m\|^2_{L^2(0,T;V)}. \tag{4.13}$$

We introduce $\widehat{\mathrm{Cov}} : V \to V$ the *multi-covariance operator* defined by

$$\widehat{\mathrm{Cov}}\varphi = \sum_{m=1}^{M} \int_0^T ((z_m(t),\varphi))z_m(t)\,\mathrm{d}t.$$

We sum up the main result in the following proposition, which is easily proven.

Proposition 19. *There exists a unique sequence $(\lambda_i)_{i\in I}$, I either finite or infinite, of numbers λ_i such that*

$$\begin{gathered} \lambda_i > 0, \\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \quad \textit{if finite} \quad (I = \{1,\,2,\,,\ldots,\,N\}), \\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \ldots, \quad \lambda_i \underset{i\to\infty}{\longrightarrow} 0 \quad \textit{if infinite} \quad (I = \mathbb{N}\setminus\{0\}), \end{gathered}$$

and an orthonormal sequence $(\varphi_i)_{i\in I}$ of V of corresponding eigenvectors of $\widehat{\mathrm{Cov}}$, in finite number for each non-zero eigenvalue,

$$\widehat{\mathrm{Cov}}\varphi_i = \lambda_i\varphi_i, \quad \forall i \in I,$$

such that $(\varphi_i)_{i\in I}$ *is total in the orthogonal complement of the kernel of* $\widehat{\mathrm{Cov}}$ *i.e.*

$$V = \mathrm{Ker}\,\widehat{\mathrm{Cov}} \overset{\perp}{\oplus} \overline{\mathrm{Span}\{\varphi_i\}_{i\in I}}.$$

Then, for all $1 \le l \le \mathrm{Card}\, I$, *a solution* π^l *of Problem* (4.13) *is determined by*

$$\mathrm{Im}\,\pi^l = \mathrm{Span}(\varphi_1, \dots, \varphi_l).$$

Moreover, $(\lambda_i)_{i\in I}$ *is the only sequence such that the minimum value verifies*

$$\sum_{m=1}^{M} \|z_m - \pi^l z_m\|^2_{L^2(0,T;V)} = \min_{\tilde\pi^l} \sum_{m=1}^{M} \|z_m - \tilde\pi^l z_m\|^2_{L^2(0,T;V)} = \sum_{i>l} \lambda_i.$$

Proof. Let the $L^2(0,MT;V)$ function $\hat z$ be defined by

$$\hat z(t) = z_m(t-(m-1)T), \quad \forall 1 \le m \le M, \quad \forall (m-1)T \le t \le mT,$$

so that for all V-orthogonal projectors $\tilde\pi^l$ of rank l,

$$\|\hat z - \tilde\pi^l \hat z\|^2_{L^2(0,MT;V)} = \sum_{m=1}^{M} \|z_m - \tilde\pi^l z_m\|^2_{L^2(0,MT;V)}$$

Now apply Props. 1 and 3, by remarking also that the standard covariance matrix for $\hat z$ over $[0,MT]$ is exactly $\widehat{\mathrm{Cov}}$, i.e. the sum of the standard covariance matrices for z_m, $1 \le m \le M$. □

Finally, we can combine the previous inequalities to form, for the linear equations (4.1)-(4.2), an estimate of the maximum reduction error over $\mathcal{D}$ as follows.

Proposition 20. *Assume that* V^l *is equal to the range of the multi-POD projector* π^l *of order* s, *i.e.* π^l *is the minimizer of*

$$\sum_{I\in\mathcal{I}} \|u(D_I) - \pi^l u(D_I)\|^2_{L^2(0,T;V)}.$$

Then, for all $T > 0$ *and all* $1 \le r \le s$,

$$\begin{aligned}\|u - u^l\|_{C^0(\mathcal{D};L^2(0,T;V))} \le{}& \frac{1}{\sqrt{c_a(\mathcal{D})}} |\pi^l_H u_0 - u^l_0| \\ &+ \Big(1 + \sqrt{\kappa_a(\mathcal{D})}\Big)(1+\sigma_l)\Big(C_1(p,s)\,\delta_{r+1}(\mathcal{D})^{r+1}\,|u|_{C^{r+1}(\mathcal{D};L^2(0,T;V))} \\ &+ C_2(p,s)\Big\{\sum_{i>l}\lambda_i\Big\}^{1/2}\Big).\end{aligned}$$

4.2.4 Detailed general strategy for practical reduced-order modelling

We can now propose a general strategy for practical, i.e. dependent on say p parameters, reduced-order modelling based on nonlinear PDEs with a multi-proper orthogonal decomposition. Following all similar strategies in the literature, it decomposes into two parts : a costly *offline* one and a fast *online* one.

On the one hand, the online part simply stands for the execution of the prepared reduced-order models in a multiple-query context, with the objective of real-time performance in mind.

On the other hand, given a reasonable rectangular parametric subdomain $\mathcal{D}$, the offline part constructs all the fundamental data for a turn-key, precise and fast reduced-order model. We detail it in three steps and comment their advantages and drawbacks.

1.	Compute the full solutions $u(D_I)$, $I \in \mathcal{I}$ on the Lagrange subgrid of $\mathcal{D}$ of order $s \geq 1$.	$\oplus$ Provides a remarkable gain in precision. $\ominus$ Needs $(s+1)^p$ full solutions. High-order grids are profitable for very smooth solutions.
2.	Compute and store the corresponding multi-POD basis $(\varphi_i)_{i=1}^{l_{\max}}$ basis.	$\oplus$ Maximum rank $l_{\max}$ remains generally small for diffusive systems. $\pm$ For solutions discretized in $N_{\Delta t}$ timesteps and N_h degrees of freedom, requires the assembling and diagonalization of covariance matrix of $m \times m$ with $m = \min((s+1)^p N_{\Delta t}, N_h)$.

From here the spatial Galerkin approximation $u^l(D)$ are well defined. The last offline step treats the advantageous case of linear dependence of the operators with respect to both the parameter and the solution, e.g. of the form

$$a(w, v; D) = \sum_{q=1}^{p} D^{(q)} a_q(w, v) \qquad (\text{with } D = (D^{(1)}, \cdots, D^{(p)})).$$

It absolutely does not hold for nonlinear, even Lipschitz-continuous, operators.

3.	For linear operators N w.r.t. D and u, described by linear combinations of some constant elementary matrices $N_1, \ldots, N_q$, $q \leq p$, compute and store the reduced elementary matrices $(\varphi_l^\top N_{q'} \varphi_l)_{\substack{1 \leq l \leq l_{\max} \\ 1 \leq q' \leq q}}$.	$\oplus$ Minor gain but assembled once and for all.

4.3 Numerical validation with the electrophysiology FitzHugh–Nagumo system

We provide some results of reduced-order modelling by multi-POD on a range of solutions of a system inspired by the *FitzHugh–Nagumo equations*, which form a particularly parameter-sensitive system.

The scalar and spatially unidimensional FitzHugh–Nagumo equations

$$\frac{\partial u}{\partial t} - D\frac{\partial^2 u}{\partial x^2} = f(u) - \gamma w, \tag{4.14}$$

$$\frac{\partial w}{\partial t} = \alpha u - \beta w, \tag{4.15}$$

with $f(u)$ a third-degree polynomial in u, originally model the propagation of an *action potential* u in an axon. In this electrophysiology phenomenon, a particular threshold effect occurs: when an electric stimulus excites an end of the axon, if the amplitude of the signal is sufficient high, then a significant electric wave appears and propagates along the axon; otherwise, it quickly vanishes [20]. We can indeed tune the parameters D, α, β and γ and the initial conditions of Eqs. (4.14)–(4.15) to reproduce such propagating solutions, called *travelling waves* [27]. These parameters depend on the chemical and mechanical properties of the membrane of the axon.

In order to illustrate the previous developments and to investigate a multi-dimensional case of variation through a diffusion operator, we put the FitzHugh–Nagumo equations in variational form, and slightly modify them. In the sequel, $H = L^2(0,1)$ and $V = H_0^1(0,1)$. We keep the notation $(\cdot,\cdot)$ for the scalar product of H. We split the unit space interval into p parts by

$$0 = X_0 < X_1 < \cdots < X_{p-1} < X_p = 1, \tag{4.16}$$

$$\Omega_q = (X_{q-1}, X_q), \quad 1 \leq q \leq p, \tag{4.17}$$

so that $[0,1] = \bigcup_{i=1}^{p} \overline{\Omega}_q$. We associate independent piecewise-constant diffusion coefficients with each subinterval Ω_q, forming a p-dimensional parametric subspace $\mathcal{D}$ of the type (4.8), with $0 < a_i < b_i$. We then define the

resulting diffusion operator as $a(D)$ by

$$a(w,v;D) = \sum_{q=1}^{p} D^{(q)} \int_{\Omega_q} \frac{\partial w}{\partial x}\frac{\partial v}{\partial x} \tag{4.18}$$

Also, for consistency purposes with the previous chapter, we introduce in the following discretization descriptions a more general second term $f(t,u)$. Finally, Eq. (4.14) becomes

$$\frac{\partial}{\partial t}(u(t),v) + a(u(t),v;D) = (f(t,u(t)) - \gamma w(t), v), \quad \forall v \in V.$$

There is no particular reason to change the ODE (4.15).

4.3.1 Semi-discrete solutions and their POD reduced forms

In this section, we show as a first step how we discretize the FitzHugh–Nagumo equations for a direct simulation of the full and the reduced models.

As a system of coupled first-order, nonlinear equations that display a propagative phenomenon, several choices of discretization are available. In our case, we use a finite element approach for the spatial variable with a space step that is small compared to the active extent of the signal.

Let u_h and w_h be the $\mathbf{P}_1$ approximations of u and w on the regular mesh $(x_i)_{i=1}^{N_h}$

$$x_i = ih, \quad 1 \le i \le N_h, \quad h = \frac{1}{N_h+1},$$

associated with the basis of shape functions $(e_i)_{i=1}^{N_h}$. The discrete solution (u_h, w_h) is defined by

$$\frac{\partial}{\partial t}(u_h(t),e_i) + a(u_h(t),e_i;D) = (f(t,u_h(t)) - \gamma w_h(t), e_i), \quad 1 \le i \le N_h, \tag{4.19}$$

$$\frac{\partial w_h}{\partial t}(t) = \alpha u_h(t) - \beta w_h(t). \tag{4.20}$$

with the initial conditions

$$u_h(0) = U_{0,h}, \quad w_h(0) = 0.$$

Hence, with the zero initial condition for w_h, the unknown of the system (4.19)–(4.20) becomes simply u_h, and w_h is an auxiliary function. We note that differentiating Eq. (4.19) in time leads to a second-order differential equation where $w_h(t)$ disappears, but which is numerically less convenient to solve. This is why we preserve the system (4.19)–(4.20) as it is.

Let the non-reduced mass matrix M and stiffness matrix K be defined by

$$M = [(e_k, e_i)]_{1\le i,k\le N_h},$$
$$K(D) = [a(e_k, e_i; D)]_{1\le i,k\le N_h} = \sum_{j=1}^{p} D^{(j)} K_j, \tag{4.21}$$

with

$$(K_j)_{ik} = \int_{\Omega_j} \frac{\partial e_k}{\partial x}\frac{\partial e_i}{\partial x}, \tag{4.22}$$

and the reaction term application $F : \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ with coefficients defined by

$$(F(t,\beta))_i = \int_0^1 f\Big(t, \sum_{k=1}^{N_h} \beta_k e_k(x)\Big) e_i(x)\,\mathrm{d}x$$

Then the vectors $U_h(t)$ and $W_h(t) \in \mathbb{R}^{N_h}$ concatenating the coordinates of $u_h(x,t)$ and $w_h(x,t)$ respectively in $(e_i(x))_{i=1}^{N_h}$ satisfy

$$M\dot{U}_h(t) + K(D)U_h(t) = F(t, U_h(t)) - \gamma M W_h(t), \tag{4.23}$$
$$\dot{W}_h(t) = \alpha U_h(t) - \beta W_h(t), \tag{4.24}$$
$$U_h(0) = U_{h,0}, \quad W_h(0) = 0.$$

Let a V-orthonormal family $(\varphi_1, \dots, \varphi_l)$ of l vectors. We keep in mind that it plays the role of a certain POD basis that, because of the parametric dependence, can be defined in various ways and is specified below. Let $V^l = \mathrm{Span}(\varphi_1, \dots, \varphi_l)$. The corresponding reduced form u_h^l of u_h on V^l satisfies

$$\frac{\partial}{\partial t}(u_h^l(t), \varphi_i) + a(u_h^l(t), \varphi_i; D) = (f(t, u_h^l(t)) - \gamma w_h^l(t), \varphi_i), \quad 1 \le i \le l,$$
$$\frac{\partial w_h^l}{\partial t}(t) = \alpha u_h^l(t) - \beta w_h^l(t),$$
$$u_h^l(0) = u_{h,0}^l, \quad w_h^l(0) = 0,$$

where, following the estimates that naturally appear for linear parabolic equations, we choose $u_{h,0}^l$ as the H-projection of $u_{h,0}$ onto V^l.

By substituting the POD basis $(\varphi_i)_{i=1}^l$ for the finite element basis $(e_i)_{i=1}^{N_h}$ in the definitions of M, K and F, we obtain the reduced mass and stiffness matrices M^l and K^l, and the reduced reaction term F^l. We emphasize that although these reduced matrices are of limited size, they are full. This gives

$$M^l \dot{U}_h^l(t) + K^l(D)U_h^l(t) = F^l(t, U_h^l(t)) - \gamma M^l W_h^l(t), \tag{4.25}$$
$$\dot{W}_h^l(t) = \alpha U_h^l(t) - \beta W_h^l(t), \tag{4.26}$$
$$U_h^l(0) = U_{h,0}^l, \quad W_h^l(0) = W_{h,0}^l.$$

We call Φ^l the matrix

$$\Phi^l = [\varphi_1, \ldots, \varphi_l] \quad \in \mathbb{R}^{N_h \times l},$$

where vectors φ_i are expressed as column vectors of coordinates in $(e_i)_{i=1}^{N_h}$. Then we obtain the following relations between reduced and non-reduced operators

$$\begin{aligned} M^l &= (\Phi^l)^\top M \Phi^l, \\ K^l(D) &= (\Phi^l)^\top K(D) \Phi^l, \\ F^l(t, \beta^l) &= (\Phi^l)^\top F(t, \Phi^l \beta^l), \end{aligned}$$

and also the relation between differentials, useful for the Newton algorithm below,

$$\mathrm{d}_{\beta^l} F^l(t, \beta^l) = (\Phi^l)^\top \mathrm{d}_\beta F(t, \Phi^l \beta^l) \Phi^l.$$

4.3.2 Full discretization

We apply a semi-implicit time scheme by the θ-method. Again, since we work on a complex nonlinear system, we discretize finely enough in time to be able to neglect the matters of time-scheme influence.

For the non-reduced solution $(U_h(t), W_h(t))$, this gives

$$\begin{aligned} &M \frac{U_h^{n+1} - U_h^n}{\Delta t} + K(D)\Big(\theta U_h^{n+1} + (1-\theta) U_h^n\Big) \\ &\quad = \theta\Big(F(t^{n+1}, U_h^{n+1}) - \gamma W_h^{n+1}\Big) + (1-\theta)\Big(F(t^n, U_h^n) - \gamma W_h^n\Big), \end{aligned} \tag{4.27}$$

$$\frac{W_h^{n+1} - W_h^n}{\Delta t} = \theta(\alpha U_h^{n+1} - \beta W_h^{n+1}) + (1-\theta)(\alpha U_h^n - \beta W_h^n). \tag{4.28}$$

Substituting the expression for W_h^{n+1} from (4.28) into (4.27), we rewrite the system as

$$A(t^{n+1}, U_h^{n+1}) = B(t^n, U_h^n) - c_1 M W_h^n, \tag{4.29}$$

$$W_h^{n+1} = c_2 W_h^n + c_3(\theta U_h^{n+1} + (1-\theta) U_h^n), \tag{4.30}$$

with the constants

$$c_1 = \frac{\gamma \Delta t}{1 + \beta\theta\Delta t}, \quad c_2 = \frac{1 - \beta(1-\theta)\Delta t}{1 + \beta\theta\Delta t}, \quad c_3 = \frac{\alpha \Delta t}{1 + \beta\theta\Delta t}. \tag{4.31}$$

and, introducing the matrices appearing for the homogeneous case

$$A_0(D) = \left(1 + \frac{\alpha\gamma(\theta\Delta t)^2}{1 + \beta\theta\Delta t}\right) M + \theta\Delta t K(D), \tag{4.32}$$

$$B_0(D) = \left(1 - \frac{\alpha\gamma\theta(1-\theta)\Delta t^2}{1 + \beta\theta\Delta t}\right) M - (1-\theta)\Delta t K(D), \tag{4.33}$$

the nonlinear applications

$$A(t,\beta;D) = A_0(D)\beta - \theta\Delta t F(t,\beta),$$
$$B(t,\beta;D) = B_0(D)\beta + (1-\theta)\Delta t F(t,\beta).$$

Once (U_h^n, W_h^n) is known, we solve Eq. (4.29) for U_h^{n+1} using the Newton algorithm, and then Eq. (4.30) determines W_h^{n+1} immediately.

For the reduced solution $(U_h^l(t), W_h^l(t))$, the definition of the fully discrete, reduced solution $(U_h^{l,n}, W_h^{l,n})$ is very similar, i.e.

$$A^l(t^{n+1}, U_h^{l,n+1}) = B^l(t^n, U_h^{l,n}) - c_1 M^l W_h^{l,n},$$
$$W_h^{l,n+1} = c_2 W_h^{l,n} + c_3(\theta U_h^{l,n+1} + (1-\theta)U_h^{l,n}),$$

with, introducing the reduced matrices appearing for the homogeneous case

$$A_0^l(D) = (\Phi^l)^\top A_0(D)\Phi^l,$$
$$B_0^l(D) = (\Phi^l)^\top B_0(D)\Phi^l,$$

the nonlinear applications

$$A^l(t,\beta^l;D) = A_0^l(D)\beta^l - \theta\Delta t F^l(t,\beta^l),$$
$$B^l(t,\beta^l;D) = B_0^l(D)\beta^l + (1-\theta)\Delta t F^l(t,\beta^l).$$

4.3.3 Simulation of the action potential phenomenon and choice of two solution ranges of study

Before comparing the properties of the POD-reduced FitzHugh–Nagumo solutions with the original ones, we briefly begin by illustrating the efficiency of the latter to quantitatively reproduce the action potential phenomenon described above. We use the set of constants listed in Tab. 4.1, where P_{FHN}, used as the nonlinear source term, is the typical third-degree polynomial in u

$$P_{\mathrm{FHN}}(u;C,a) = -Cu(u-1)(u-a),$$

with $a \in (0,1)$; and η, used as the initial condition, is the following *pulse*,

$$\eta(x;A,m,\sigma) = \begin{cases} \frac{A}{2}\exp\left(1 - \frac{1}{1-(\frac{x-m}{\sigma})^2}\right) & \text{if } |x-m| < \sigma, \\ 0 & \text{otherwise,} \end{cases}$$

i.e. a localized regular cutoff function with:

– A, of nominal value 1 and not fixed here, representing its amplitude (maximum value);

Discretization	
N_h	199
$N_{\Delta t}$	200
Δt	$2.5 \cdot 10^{-2}$
θ	2/3

Fixed functions and coefficients	
$f(t,u)$	$P_{\mathrm{FHN}}(u;\, 20,\, 0.1)$
$u_0(x)$	$\eta(x;\, A,\, \frac{1}{2},\, \frac{1}{4})$
α	$5 \cdot 10^{-2}$
β	$1 \cdot 10^{-2}$
γ	20

Parameter subdomain	
p	2
(X_i)	$i/p,\ 0 \leq i \leq p$
$\mathcal{D}$	$[1 \cdot 10^{-3},\, 3 \cdot 10^{-3}]^2$
s_{POD}	1
s_{red}	5

Table 4.1: Set of constants for FitzHugh–Nagumo travelling pulse solutions

– m, its median point;
– and σ, the length of its support.

Hence, under these circumstances, slight variations of the amplitude around a certain critical amplitude $A_{\mathrm{c}} \approx 0.26$ brutally modify the qualitative behaviour of the FitzHugh–Nagumo solution. For some significant values of A around A_{c}, we display the corresponding FitzHugh–Nagumo solutions next to their associated self-reduction results in Fig. 4.1. Also, the associated standard PODs totally differ. In particular, the POD remainder decreases three times slower for upper values $A > A_{\mathrm{c}}$ than for lower values $A < A_{\mathrm{c}}$. We sum up these properties in Tab. 4.2.

In conjunction with Tab. 4.1, this forms two initial condition cases for a range of FitzHugh–Nagumo solutions, described by a particular range of parametric variation:

– the *nominal amplitude* case, corresponding to $A = 1$. We observe in this case that $\mathcal{D}$ describes a set of travelling pulse solutions. Figure 4.2 presents the solutions at the vertices of $\mathcal{D}$, namely

$$\begin{aligned} D_{(0,0)} &= (1 \cdot 10^{-3},\, 1 \cdot 10^{-3}), \quad D_{(0,1)} = (1 \cdot 10^{-3},\, 3 \cdot 10^{-3}), \\ D_{(1,0)} &= (3 \cdot 10^{-3},\, 1 \cdot 10^{-3}), \quad D_{(1,1)} = (3 \cdot 10^{-3},\, 3 \cdot 10^{-3}), \end{aligned} \tag{4.34}$$

  for a more detailed analysis below;
– the *critical amplitude* case, corresponding to $A = A_{\mathrm{c}}$. We observe in this case that a frontier representing the action potential threshold appears inside $\mathcal{D}$. In particular, taking the approximation $A_{\mathrm{c}} = 0.267$, we assert that the vertices $D_{(0,0)}$, $D_{(0,1)}$ and $D_{(1,0)}$ correspond to travelling pulse solutions, while the most diffusive parameter value $D_{(1,1)}$

	$A < A_c$	$A > A_c$
Type	Diffusive solution	Travelling pulse solution
Spatial distribution	Becomes global	Local
Quantitative behaviour	Diffuses the initial condition; converges in time towards a steady state	Amplificates the initial condition signal to a signal of amplitude $O(1)$; propagates it along the two directions
POD remainder decrease	4 orders of magnitude in 10 modes	4 orders of magnitude in 35 modes
Reduction error decrease	Same rate as standard POD	Same rate as standard POD

Table 4.2: Summary of the action potential phenomenon and its consequence on the self-reduced solutions

corresponds to a diffusive solution.

We use these two cases, that present opposite behaviours from many points of view according to Tab. 4.2, as benchmarks for the proposed multi-POD methodology throughout this memoir. The next section analyses the maximum reduction error over this range of solutions, while the numerical section of the next chapter focuses on the impact of the reduction on some inverse, parameter estimation problem.

4.3.4 Numerical comparison of efficiency between the standard POD and the multi-POD

We present a numerical comparison of the multi-POD methodology and a more simple approach using the classical POD method *(standard POD)* for spatial Galerkin approximation:

1. the first is the standard POD basis coming from the solution corresponding to the central point of $\mathcal{D}$, i.e.
$$D_c = (2 \cdot 10^{-3},\, 2 \cdot 10^{-3});$$
2. the second is the multi-POD basis of order $s_{\mathrm{POD}} = 1$ on $\mathcal{D}$, i.e. built with the solutions corresponding to the values on the vertices of $\mathcal{D}$, see (4.34).

We denote the corresponding V-orthogonal projectors by

$$\pi_1^{l_1}, \quad 1 \le l_1 \le l_{1,\max}\,, \qquad \pi_2^{l_2}, \quad 1 \le l_2 \le l_{2,\max}\,, \tag{4.35}$$

and the corresponding reduced solutions by

$$u_{1,h}^{l_1}, \quad 1 \le l_1 \le l_{1,\max}, \qquad u_{2,h}^{l_2}, \quad 1 \le l_2 \le l_{2,\max},$$

respectively.

Following the notations in Section 3.4.2, we analyse some error terms:

– the reduction errors $R_1(l_1)$ and $R_2(l_2)$

$$R_1(l_1) = \|u_h - u_{1,h}^{l_1}\|_{C^0(\mathcal{D};L^2(0,T;V))},$$
$$R_2(l_2) = \|u_h - u_{2,h}^{l_2}\|_{C^0(\mathcal{D};L^2(0,T;V))};$$

– the V-projection errors $P_1(l_1)$ and $P_2(l_2)$

$$P_1(l_1) = \|u_h - \pi_1^{l_1} u_h\|_{C^0(\mathcal{D};L^2(0,T;V))},$$
$$P_2(l_2) = \|u_h - \pi_2^{l_2} u_h\|_{C^0(\mathcal{D};L^2(0,T;V))};$$

– and the POD remainders $\epsilon_1(l_1)$ and $\epsilon_2(l_2)$, that correspond by construction to different discrete V-projections errors

$$\epsilon_1(l_1) = \Big\{ \sum_{i>l_1} \lambda_{1,i} \Big\}^{1/2} = \|u_h(D_c) - \pi_1^{l_1} u_h(D_c)\|_{L^2(0,T;V)},$$
$$\epsilon_2(l_2) = \Big\{ \sum_{i>l_2} \lambda_{2,i} \Big\}^{1/2} = \Big\{ \sum_{D\in\Theta_{\mathrm{POD}}} \|u_h(D) - \pi_2^{l_2} u_h(D)\|^2_{L^2(0,T;V)} \Big\}^{1/2},$$

where Θ_{POD} describes the points (4.34).

Numerically, we assimilate the $C^0(\mathcal{D})$ norms with the maximum values over a finer Lagrange subgrid of order $s_{\mathrm{err}} = 5$.

Figures 4.3 and 4.4 provide some opposite reduction behaviours with respect to the POD rank between the two pulse amplitude cases, and also between the two POD methods, even though the first three standard POD and multi-POD modes look very similar.

The standard POD method comes with fastly decreasing remainders $\epsilon_1(l_1)$, i.e. 4 orders of magnitude in 30 modes in both $A = 1$ and $A = A_c$ cases. This means that the self-reduction of $u(D_c)$ works as well as for the nonlinear solutions seen in Chapter 3. Nevertheless, it completely fails at properly reducing the solutions generated by the parameter vector values around D_c. This reflects large variations of the local shape, and therefore of the V-projection error

$$P_1(D, l_1) = \|u(D) - \pi_1^{l_1} u(D)\|_{L^2(0,T;V)}$$

with respect to D through this parametric window.

By contrast, the multi-POD method comes with a remainders $\epsilon_2(l_2)$ that decrease twice as slow, and then according to the criterion (3.33) allows access to twice as many POD modes. This decrease rate remains acceptable, since a relative remainder of 0.1 % appears from 60 modes for $A = 1$, and 50 modes for $A = A_c$. Since the multi-POD is built from the grid $\{D_I\}_{I\in\mathcal{I}}$, we may more properly compare these decrease rates with those of the standard PODs (or, more explicitly, the self-PODs) associated with each D_I, $I \in \mathcal{I}$. To that end, taking the example of the nominal amplitude, we provide in Fig. 4.2, for each $I \in \mathcal{I}$, the shape of the solution $u_h(D_I)$ and the associated relative self-POD remainders. We observe that, due to the optimal sense of these self-POD bases, the (spatially) asymmetrical solutions, corresponding to $D_{(0,1)}$ and $D_{(1,0)}$, need richer decompositions in comparison with the symmetrical solutions, corresponding to $D_{(0,0)}$ and $D_{(1,1)}$ for equal relative projection errors. More precisely, they need a basis twice as large as the self-POD basis associated with the most diffusive solution, i.e. with $D_{(1,1)}$. This reflects the nature of a symmetrical solution, containing twice as less information as a general non-symmetrical solution.

For the numerical efficiency now and for the nominal amplitude case, the major difference for the multi-POD method in comparison with the standard POD method appears in the maximum reduction error $R_2(l_2)$. Indeed, it follows a decrease rate close to that of the remainder $\epsilon_2(l_2)$, and shows no saturation phenomenon. Hence, for $A = 1$, while not locally reaching the efficiency of the self-POD bases, the multi-POD basis manages to capture the variation of $u_h(D)$ with respect to $D \in \mathcal{D}$, even with a coarse approach (recall that the Lagrange grid is of order 1), and offers very satisfying reduction properties. More precisely, from 80 modes, the maximum reduction error is smaller than $1 \cdot 10^{-3}$.

Yet, for the critical amplitude case, the multi-POD shows its limits. Recalling that one vertex of $\mathcal{D}$ generates a diffusive solution and the three others generate travelling pulse solutions, this drawback emanates from the action potential threshold, which makes the solution brutally vary on $\mathcal{D}$. Hence, the Lagrange grid of order 1, on which these multi-POD bases are built, do not provide enough information to accurately capture some first-order information in parameter of the solution.

4.4 Conclusion

Starting from the introduction of the multi-proper orthogonal decomposition, which is a natural extension of the standard POD on parametric regular grids of solutions, we built a general strategy of reduction for PDE-based models. We developed a mathematical analysis of this method and provided a numerical validation.

The analysis pursues the previous reduction error estimates for para-

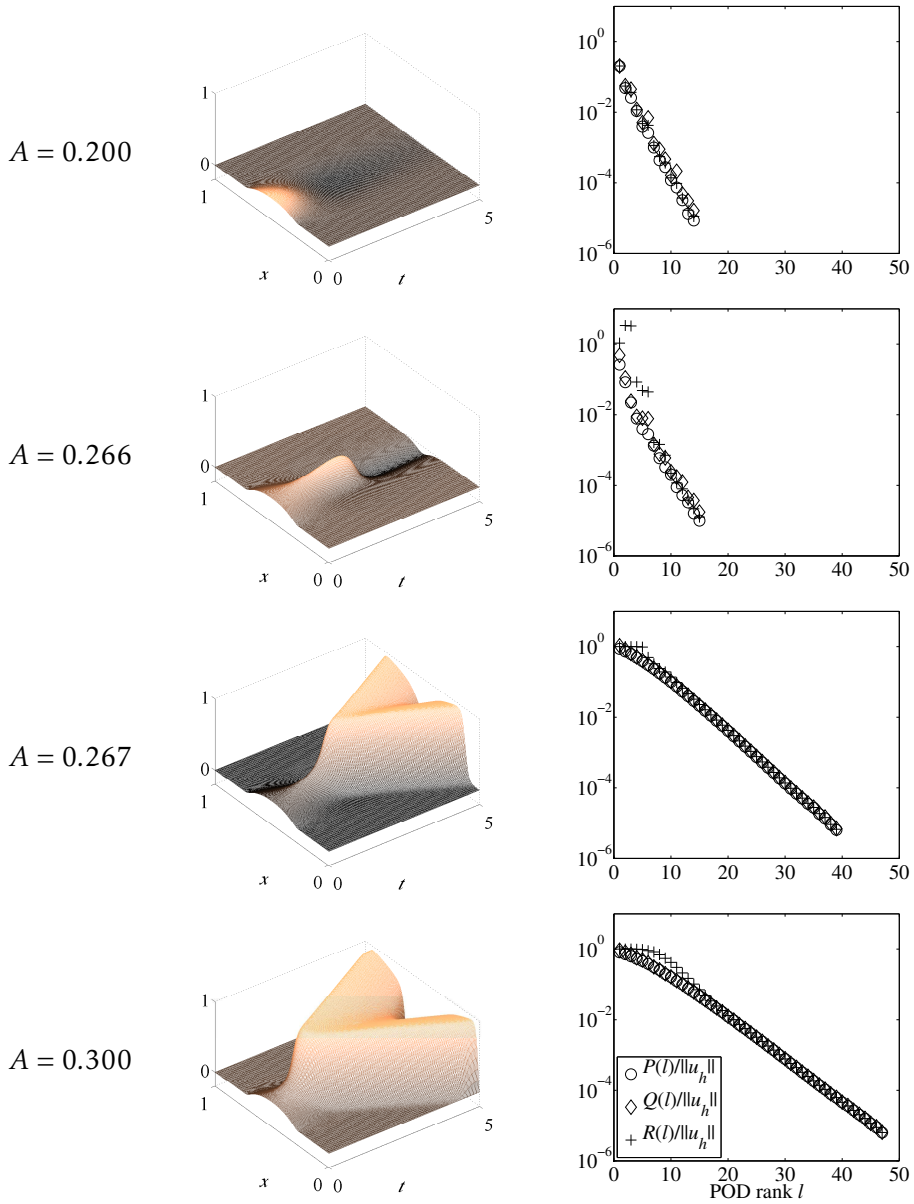

Figure 4.1: Action potential phenomenon: FitzHugh–Nagumo travelling pulse solutions and their self-reduction results, with parameter $D = D_c$ and various amplitudes

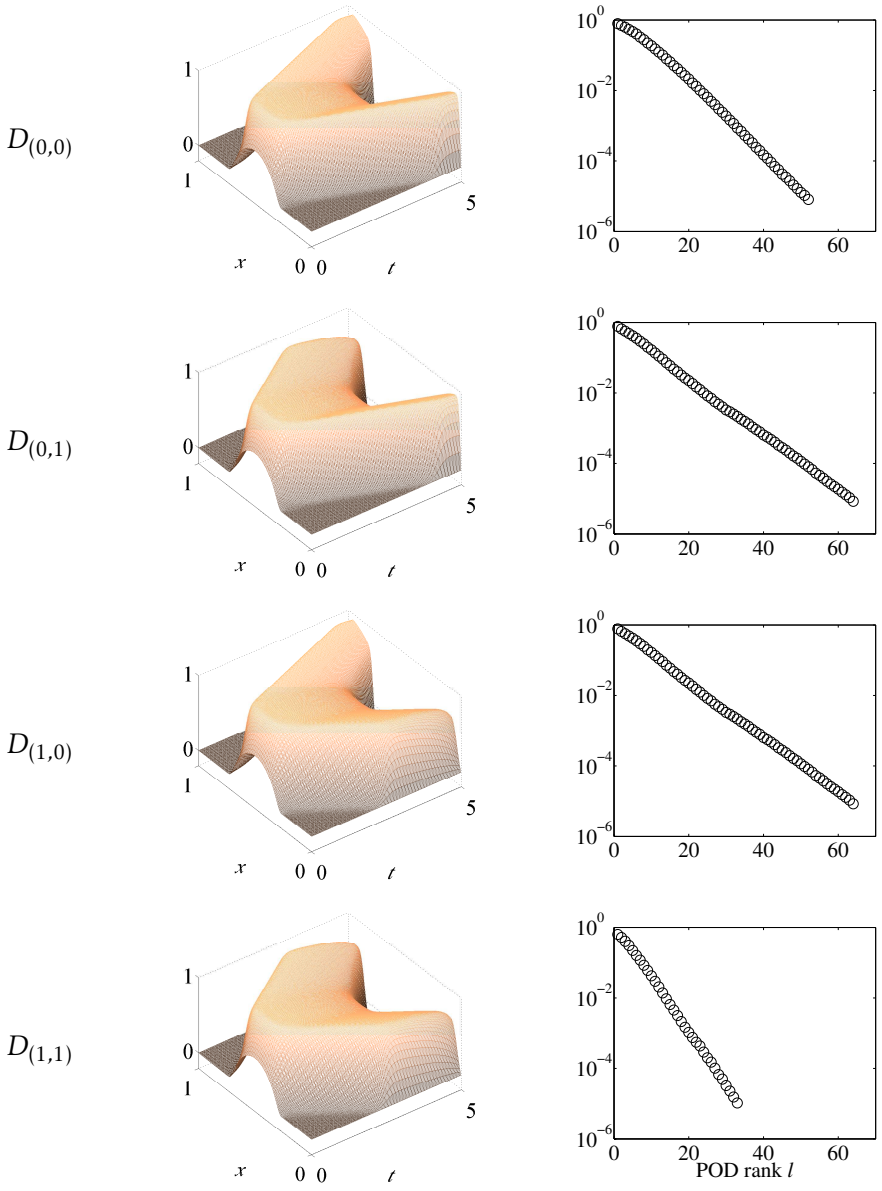

Figure 4.2: Multi-POD grid of degree 1: FitzHugh–Nagumo solutions and their relative self-POD remainders, with nominal amplitude $A = 1$ and parameter vector values (4.34)

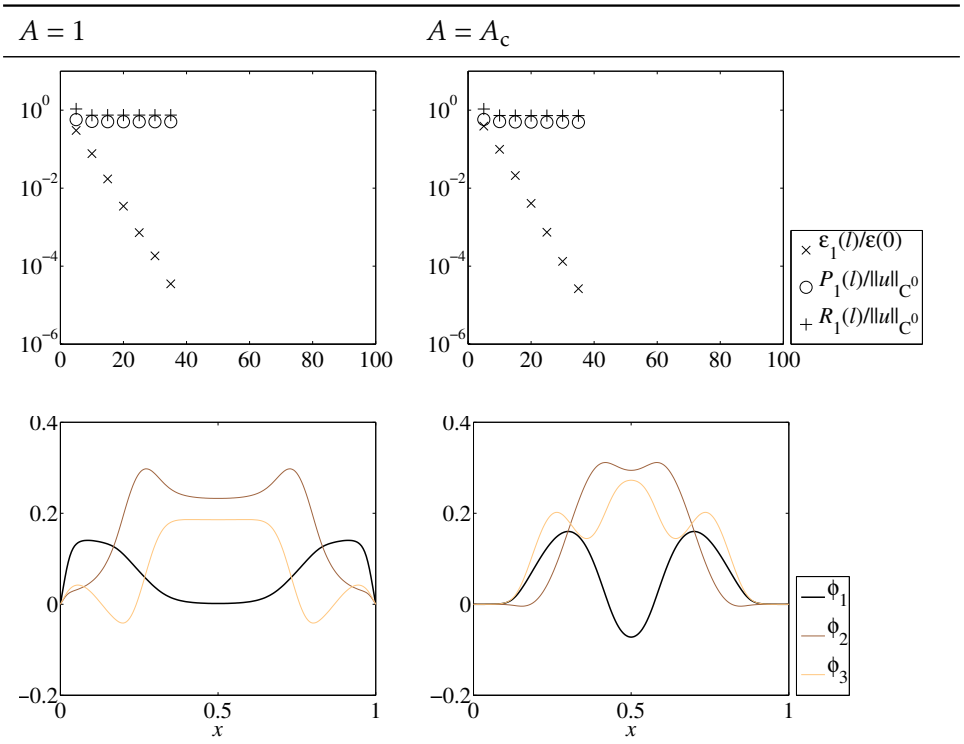

Figure 4.3: Standard POD method: maximum reduction error over $\mathcal{D}$ (top) and associated first 3 modes (bottom), for the two nominal amplitude and critical amplitude cases

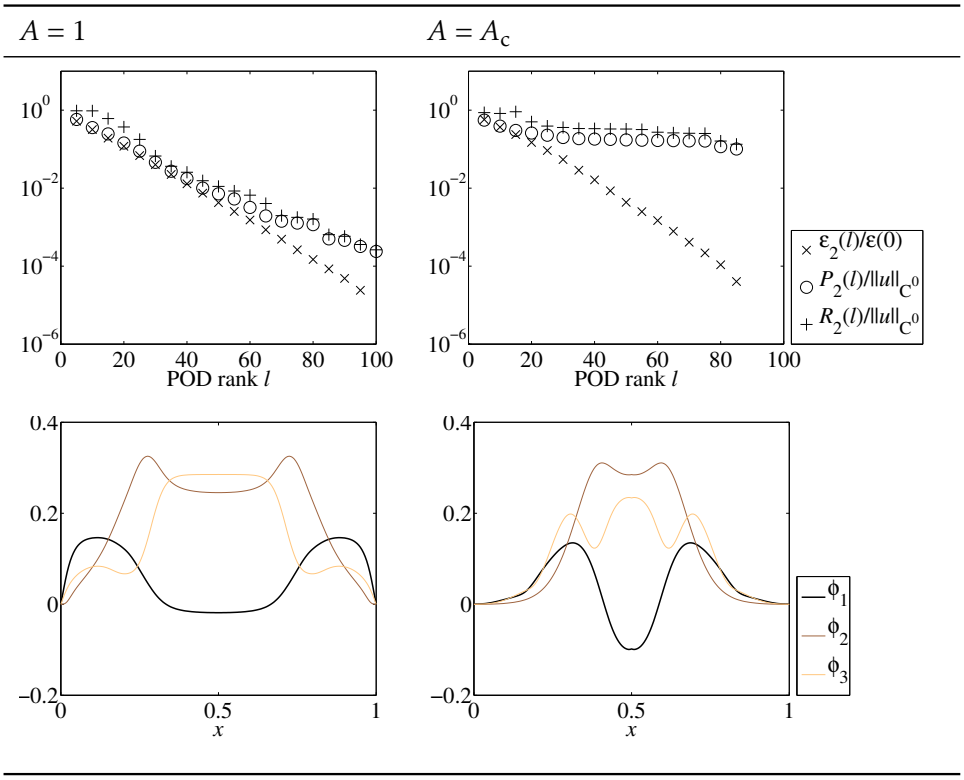

Figure 4.4: Multi-POD method: maximum reduction error over $\mathcal{D}$ (top) and associated first 3 modes (bottom), for the two nominal amplitude and critical amplitude cases

bolic linear equations with the standard POD. It inserts some parametric variation through the diffusion operator, and therefore non-trivial dependence of the solution with respect to its parameter. We bound a worst-case extension of these estimates with an adaptation and combination of error estimates from the interpolation theory literature. Eventually, we derived an estimate of the maximum reduction error over a given parameter domain, now controlled by a numerically accessible multi-POD remainder and a new term that depends on the size of this domain.

More numerical-oriented, the second part applies to a more complex situation arising in the electrophysiology models, with solutions of the FitzHugh–Nagumo system that feature the travelling wave phenomenon. The difficulty resides less in the nonlinear source term and the possibly steep front shape of the solutions, than in their propagative nature, known to be generally incompatible with POD-based model reduction. Nonetheless, still looking at parametric variation through the diffusion operator, the multi-POD methodology proves its superiority over a more intuitive technique that relies on the validity domain of a given standard POD. Indeed, in the first case, the maximum reduction error very satisfyingly decreases at the same rate as the multi-POD remainder, whereas in the second case, depleting all the accessible POD modes, the relative maximum reduction error hardly reaches the 50% threshold.

These results pave the way for the reduced solving of inverse problems of parameter estimation, using still the multi-proper orthogonal decomposition. These problems typically require heavier simulations, and may more remarkably benefit from model reduction. They are tackled in the next chapter.

Chapter 5

Reduced variational parameter estimation problem on an electrophysiology model

The parameters that rule Eqs. (4.14)–(4.15) are based on some constitutive laws that model the axon, and, from the viewpoint of the measurements, are generally out of reach. Given an electric event u, solution of these equations, we wish to retrieve the corresponding values of parameters that determine it from the only data of some measurements, that are intrinsically *imperfect* and *partial*. This establishes an inverse problem of *parameter estimation*. In this chapter, we apply our method of model reduction by multi-proper orthogonal decomposition to a *variational approach* of this inverse problem, i.e. to a formulation that involves a cost function on u to be minimized. This kind of approaches is generally known to lead to costly computations when manipulating finely discretized solutions.

Here, we only consider the problem for the semi-discrete and fully discrete solutions. We also limit the study to the estimation of initial conditions and the generalized diffusion coefficient. We start by discussing the related and common uncertainty models, from which we design a consistent cost function and also its immediate generic POD-reduced version. Then, with a view to minimizing these cost functions using first-order optimization methods, we provide efficient expressions of their respective gradients. The semi-discrete case features all the essential ingredients, while the fully discrete case is an adaptation containing some subtle differences. In both sections, we provide the result and the whole proof for the non-reduced cost function, and then the straightforward reduced version of the corresponding result.

Finally, we numerically verify whether the minimizers of the POD-reduced cost functions practically converge to the solution of the original problem. We again display a comparison between the standard POD-based and the multi-POD-based methods.

5.1 Semi-discrete parameter estimation problem and reduced form

We define the parameter-state estimation problem on the semi-discrete solutions of the system (4.23)-(4.24)

$$M\dot{U}_h(t) + K(D)U_h(t) = F(t, U_h(t)) - \gamma M W_h(t),$$
$$\dot{W}_h(t) = \alpha U_h(t) - \beta W_h(t),$$

with the initial conditions

$$U_h(0) = U_{h,0}, \quad W_h(0) = 0,$$

and detail how it is classically solved. Indeed, the semi-discrete problem already exposes, with no time-scheme, how a certain *adjoint state* plays a fundamental role. Also, still focusing on variation through the diffusion operator, we assume that, apart from D and the initial condition $U_{h,0}$, all the other parameters are fixed and known.

The principle sums up as follows. Provided some partial and imperfect *measures* $Z(t)$ of an *observed solution*

$$U_{h,\text{exact}}(t) = U_h(t;\, U_{h,0,\text{exact}}, D_{\text{exact}}),$$

the parameter-state estimation consists in finding an approximation of the *target* $(U_{h,0,\text{exact}}, D_{\text{exact}})$, i.e. the actual value of parameters from which the measurements derive. Here, as justified below, we shall actually approximate the state part to a fixed modal subspace. Straightaway, imagining that we possess too few measurements for a complex system, we understand that it is a priori an ill-posed problem.

First, we propose common models for the uncertainties arising in this problem, and then explain how they determine a regular cost function to be minimized.

5.1.1 Modelling of uncertainties

We begin by specifying a common linear interpretation of the measurement process. Inspired by the coarse resolution of some observation data such as MRI images, we model the partiality by a diminished number of degrees of freedom $N_{\text{obs}} \ll N_h$. More precisely, we imagine that we place N_{obs} sensors, with the convenient assumption

$$N_h + 1 = L(N_{\text{obs}} + 1), \quad L \in \mathbb{N},$$

on regularly spaced nodes

$$x_L,\, x_{2L},\, \cdots,\, x_{N_{\text{obs}}L}.$$

For the moment, we consider ideal sensors in the sense that they can capture a time-continuous signal. Also, we assume that the imperfection of these captures are of Gaussian white noise type. Hence we model the observation by

$$Z(t) = HU_{h,\text{exact}}(t) + \chi_{\text{exact}}(t),$$

where the *observation operator* $H \in \mathbb{R}^{N_{\text{obs}} \times N_h}$ is defined by

$$H = \begin{bmatrix} h_1 & \cdots & h_{N_h} \end{bmatrix},$$

$$h_i = \begin{cases} k\text{-th elementary vector of } \mathbb{R}^{N_{\text{obs}}} & \text{if } i = kL, \\ 0 & \text{otherwise,} \end{cases}$$

and $\chi_{\text{exact}}(t)$ is one realization of random vector process $\chi(t)$ concatenating N_{obs} independent random processes of identical Gaussian white noise distributions, i.e. for all $1 \le i \le N_{\text{modes}}$, at least formally,

$$\chi^{(i)}(t) \sim \mathcal{N}(0, \sigma_\chi^2), \quad 1 \le i \le N_{\text{obs}}, \quad \forall t, \tag{5.1}$$

$$\mathbb{E}[\chi(s)\chi(t)^\top] = \sigma_\chi^2 \delta(t-s)\, \mathrm{Id}_{N_{\text{obs}}}, \quad \forall t, s. \tag{5.2}$$

Moreover, it seems relevant and realistic at this stage to introduce some uncertainty in the initial condition $U_{h,0}$ for two reasons. First, the set of sensors $\{x_{kL}\}_{k=1}^{N_{\text{obs}}}$ is of course too weak to determine a unique spatial function shape. Also, it is highly unlikely for such an electrophysiology model that the initial electrical activation is fully known up to the precision of the mesh. In this case, with a view to passing easily to the limit in N_h, we do not attach independent random variables to the finite element nodes for this uncertain initial condition modelling. Instead, considering a spectral decomposition size $N_{\text{modes}} \le N_h$, we describe a random approximation $\widetilde{U}_{h,0}$ of $U_{h,0}$ by

$$\widetilde{U}_{h,0} = \sum_{i=1}^{N_{\text{modes}}} \xi^{(i)} \psi_i = \Psi \xi, \quad \Psi = \begin{bmatrix} \psi_1 & \cdots & \psi_{N_{\text{modes}}} \end{bmatrix},$$

where the ψ_i's are eigenvectors of the discrete Laplacian, e.g. those associated with the smallest eigenvalues, i.e.

$$K\psi_i = \omega_i^2 M \psi_i, \quad \psi_j^\top \psi_i = \delta_{ij}, \quad 1 \le i, j \le N_{\text{modes}},$$
$$0 < \omega_1 < \cdots < \omega_{N_{\text{modes}}}.$$

Let also $\xi_0 \in \mathbb{R}^{N_{\text{modes}}}$ be an *a priori* on the mean shape of the state, i.e. a vector intuitively placed near ξ_{exact}, that eventually helps to regularize the criterion that we build in the next section. Then, we attach to the coefficients $\xi^{(i)}$ some independent Gaussian distributions centered around the a priori, i.e.

$$\xi^{(i)} \sim \mathcal{N}(\xi_0^{(i)}, (\sigma_\xi^{(i)})^2), \quad 1 \le i \le N_{\text{modes}}.$$

This forms a *state* $\xi \in \mathbb{R}^{N_{\text{modes}}}$ to be estimated. However, the exact initial condition $U_{h,0,\text{exact}}$ may remain general, i.e. $U_{h,0,\text{exact}} \notin \operatorname{Im}\Psi$, which means that we transfer the problem of its estimation to, formally and in a sense precised below, finding its best limited modal approximation.

5.1.2 The cost function and its reduced form

We now formalize the problem in variational form. Consider

$$\begin{aligned} Q_\xi &= \operatorname{diag}((\sigma_\xi^{(1)})^{-2}, \dots, (\sigma_\xi^{(N_{\text{modes}})})^{-2}), \\ Q_\chi &= \sigma_\chi^{-2} \operatorname{Id}_{N_{\text{obs}}} \end{aligned} \tag{5.3}$$

the inverse covariance matrices of the random vectors associated with the state uncertainty ξ (or equivalently with $\xi - \xi_0$) and the noise $\chi(t)$. These matrices form scalar products that enable an appropriate comparison between the relative deviation of the state and that of the measurements. Indeed, we get the following general proposition.

Proposition 21. *Let Ω be a probability space, and $X : \Omega \to \mathbb{R}^N$ a random vector. Assume $\mathbb{E}[X] = 0$ and that the covariance matrix of X*

$$\operatorname{Cov}_X = \mathbb{E}[XX^\top]$$

is non-singular. Then

$$\mathbb{E}\Big[|X|^2_{\operatorname{Cov}_X^{-1}}\Big] = N.$$

Proof. By properties of the trace,

$$\begin{aligned} \mathbb{E}[|X|^2_{\operatorname{Cov}_X^{-1}}] &= \mathbb{E}\Big[\operatorname{tr}(\operatorname{Cov}_X^{-1} XX^\top)\Big] \\ &= \operatorname{tr}\Big(\operatorname{Cov}_X^{-1} \mathbb{E}[XX^\top]\Big) \\ &= \operatorname{tr}(\operatorname{Id}_N) = N. \end{aligned}$$

□

Hence, we state the quadratic expectation norms of the random process $\chi(t)$ and the random vector $\xi - \xi_0$

$$\begin{aligned} \mathbb{E}[|\chi(t)|^2_{Q_\chi}]^{1/2} &= \sqrt{N_{\text{obs}}}, \quad \forall t, \\ \mathbb{E}[|\xi - \xi_0|^2_{Q_\xi}]^{1/2} &= \sqrt{N_{\text{modes}}}. \end{aligned}$$

Let us again focus on the rectangular parametric range $\mathcal{D}$ defined by (4.8). Thus, in the same idea, consider $D_0 \in \mathbb{R}^p$, typically the center of the rectangle $\mathcal{D}$, an a priori for the parametric part, and, since we shall formally try to trap D in the rectangle $\mathcal{D}$, consider the matrix

$$Q_D = \operatorname{diag}\left(\left(\frac{b_1 - a_1}{2}\right)^{-2}, \dots, \left(\frac{b_p - a_p}{2}\right)^{-2}\right).$$

As a last step, approaching $U_{h,\text{exact}}(t)$ with the simulated $U_h(t;\xi,D)$, we define the artificial noise

$$\chi(t;\,\xi,D) = Z(t) - HU_h(t;\,\xi,D),$$

which is actually completely deterministic and accessible. Moreover, in the theoretical and unlikely case when the exact initial condition is described by

$$U_{h,0,\text{exact}} = \Psi\xi_{\text{exact}},$$

then the artificial noise verifies

$$\chi(t;\xi_{\text{exact}},D_{\text{exact}}) = \chi_{\text{exact}}(t).$$

Then, the main idea is that a good approximation produces an artificial realization $\chi(t;\xi,D)$ that is as conform as possible to the actual distribution defined by (5.1)–(5.2).

Finally, combining all the previous elements, we estimate the target $(U_{h,0\text{exact}},D_{\text{exact}})$ by minimizing the cost function

$$C(\xi,D) = \frac{\rho_\xi}{2}|\xi-\xi_0|^2_{Q_\chi} + \frac{\rho_D}{2}|D-D_0|^2_{Q_D} + \frac{\rho_\chi}{2}\int_0^T |\chi(t;\,\xi,D)|^2_{Q_\chi}\,\mathrm{d}t, \qquad (5.4)$$

where ρ_ξ, ρ_D and ρ_χ are some weights that represent the confidence put in each corresponding term, subject to (recall the form (4.18) of the operator $a(D)$) the constraint

$$D \in]0,\infty[^p, \qquad (5.5)$$

and no constraint for ξ. We understand that the main role of ξ_0 and D_0 is to help to "convexify" C. Assuming that a unique global minimum exists, we name it $(\xi^\star,D^\star)$.

The problem transforms into an unconstrained one by reparametrization, so that the optimization step becomes a lot easier [6]. Let $g:\mathbb{R}^p \to]0,\infty[^p$ the diffeomorphism

$$g(x) = g(x^{(1)},\ldots,x^{(p)}) = (e^{x^{(1)}},\ldots,e^{x^{(p)}}). \qquad (5.6)$$

Then we equivalently minimize

$$C_{\text{unc}}(\xi,D_{\text{unc}}) = C(\xi,g(D_{\text{unc}}))$$

with no more constraint on $(\xi,D_{\text{unc}}) \in \mathbb{R}^{N_{\text{modes}}}\times\mathbb{R}^p$, and the minimizer verifies

$$D^\star = (\ln D^{\star,(1)}_{\text{unc}},\ldots,\ln D^{\star,(p)}_{\text{unc}}).$$

In order to work with a black-box optimization method that uses first-order information, we need to express $\nabla C_{\text{unc}}(\xi,D_{\text{unc}})$ (where ∇ operates on

the concatenated variable (ξ, D)). Actually, expressing $\nabla C(\xi, D)$ suffices, since we verify easily the relation

$$\nabla_{(\xi, D_{\rm unc})} C_{\rm unc}(\xi, D_{\rm unc}) = \begin{bmatrix} \nabla_\xi C(\xi, g(D_{\rm unc})) \\ [\frac{{\rm d}g}{{\rm d}D_{\rm unc}}]^\top \nabla_D C(\xi, g(D_{\rm unc})) \end{bmatrix},$$

where the Jacobian matrix of g becomes, for our particular choice,

$$\frac{{\rm d}g}{{\rm d}D_{\rm unc}}(D_{\rm unc}) = {\rm diag}(e^{D_{\rm unc}^{(1)}}, \ldots, e^{D_{\rm unc}^{(p)}}).$$

Provided a POD basis $(\varphi_1, \cdots, \varphi_l)$, that plays the role of a standard POD or multi-POD basis, consider the reduced version of U_h^l of the solution U_h, defined by (4.25)–(4.26), except for the new parameterized initial conditions

$$U_h^l(0) = \Psi^l \xi, \quad W_h^l(0) = 0,$$

where $U_h^l(0)$ is formed of the coordinates of a H_0^1-projection of $u_h(0)$ onto $V^l = {\rm Span}(\varphi_1, \cdots, \varphi_l)$, i.e. with

$$\Psi^l = (\Phi^l)^\top K_{H_0^1} \Psi.$$

Indeed, these initial conditions are based on a fixed modal decomposition, which is asymptotically with respect to N_h independent of the space discretization. Hence, unlike in the case of a standard state directly based on a finite element decomposition, we shall not consider a parameter ξ of reduced size here.

Then, defining the reduction of the cost function (5.4) is rather immediate, since we naturally wish to replace U_h by its reduced equivalent $\Phi^l U_h^l$. Defining once and for all the reduced observation operators

$$H^l = H\Phi^l, \tag{5.7}$$

the reduced cost function to be minimized takes the form

$$C^l(\xi, D) = \frac{\rho_\xi}{2} |\xi - \xi_0|^2_{Q_\chi} + \frac{\rho_D}{2} |D - D_0|^2_{Q_D} + \frac{\rho_\chi}{2} \int_0^T |\chi^l(t; \xi, D)|^2_{Q_\chi} \, {\rm d}t.$$

with the reduced (artificial) noise

$$\chi^l(t; \xi, D) = Z(t) - H^l U_h^l(t; \xi, D).$$

Of course, the mapping g still makes this problem unconstrained, with a similar relation between the gradient of C^l and that of $C^l_{\rm unc}$.

5.1.3 Gradient of semi-discrete non-reduced cost function

We now prove a useful expression of the gradient $\nabla C(\xi, D)$. The key point is to introduce an *adjoint system*. It replaces the computation of the $p + N_{\text{modes}}$ solutions represented by $\nabla U_h(t;\xi,D)$, by the pre-computed matrices represented by $\nabla_D K(D)$ and the single solution of this adjoint system, which is a coupled system of one linear PDE and one linear ODE.

In this case, recalling (4.21) and (4.22), $\nabla_D K(D)$ has the simple components

$$\frac{\partial K}{\partial D^{(j)}} = K_j.$$

Let the right-hand side term

$$G(t;\xi,D) = H^\top Q_\chi \, \chi(t;\xi,D),$$

and the corrected stiffness matrix

$$K_{\text{adj}}(t;\xi,D) = K(D) - \mathrm{d}_U F(t, U_h(t;\xi,D))^\top.$$

We define the adjoint solution

$$(P_{h,T}(t;\xi,D),\ S_{h,T}(t;\xi,D)) \in \mathbb{R}^{N_h} \times \mathbb{R}^{N_h}$$

over $[0,T]$ by the backward-in-time system

$$-M\dot{P}_{h,T} + K_{\text{adj}} P_{h,T} - \alpha M S_{h,T} = G, \tag{5.8}$$

$$\dot{S}_{h,T} = \gamma P_{h,T} + \beta S_{h,T}, \tag{5.9}$$

with the zero terminal conditions

$$P_{h,T}(T;\xi,D) = 0, \quad S_{h,T}(T;\xi,D) = 0.$$

We then obtain the following expression.

Proposition 22.

$$\nabla C(\xi,D) = \begin{bmatrix} \rho_\xi Q_\xi(\xi - \xi_0) - \rho_\chi \Psi^\top M P_{h,T}(0) \\ \rho_D Q_D(D - D_0) + \rho_\chi \mathcal{I}_T(\xi,D) \end{bmatrix},$$

with the vector $\mathcal{I}_T(\xi,D)$ of coordinates

$$\mathcal{I}_T^{(j)}(\xi,D) = \int_0^T P_{h,T}(t;\xi,D)^\top K_j U_h(t;\xi,D)\,\mathrm{d}t, \quad 1 \le j \le p.$$

Proof. For the purpose of readability, we drop the subscripts h and T, and also the variable and parameters $(t;\xi,D)$ for the functions of time in this proof.

Let $(\xi,D)\in\mathbb{R}^{N_{\text{modes}}}\times\mathcal{D}$ be a parametric point, and $(\delta\xi,\,\delta D)\in\mathbb{R}^{N_{\text{modes}}}\times\mathbb{R}^p$ be any direction. Let then δC, $(\delta U,\,\delta W)$ and δK be the corresponding tangent values of the cost function C, the solution (U,W) and the stiffness matrix K. By differentiation, on the one hand,

$$\delta C=\rho_\xi(\xi-\xi_0)Q_\xi\,\delta\xi+\rho_D(D-D_0)Q_D\,\delta D-\rho_\chi I$$

with the integral

$$I=\int_0^T(Z-HU)^\top Q_\chi H\,\delta U\,\mathrm{d}t=\int_0^T G^\top\,\delta U\,\mathrm{d}t,$$

and on the other hand, $(\delta U,\,\delta W)$ is solution of

$$M\frac{\partial\,\delta U}{\partial t}+(K-\mathrm{d}_U F(t,U))\,\delta U=-[\delta K]U-\gamma M\,\delta W,$$
$$\frac{\partial\,\delta W}{\partial t}=\alpha\,\delta U-\beta\,\delta W,$$

with the initial conditions

$$\delta U(0)=\Psi\,\delta\xi,\quad \delta W(0)=0.$$

Introducing the augmented matrices of size $2N_h\times 2N_h$

$$\mathbf{M}=\begin{bmatrix}M&0\\0&M\end{bmatrix},\quad \mathbf{K}=\begin{bmatrix}K-\mathrm{d}_U F(t,U)&\gamma M\\-\alpha M&\beta M\end{bmatrix},\quad \mathbf{R}_\delta=\begin{bmatrix}-[\delta K]U\\0\end{bmatrix},$$

and the augmented vectors of size $2N_h$

$$\boldsymbol{\delta\mathbf{U}}=\begin{bmatrix}\delta U\\\delta W\end{bmatrix},\quad \mathbf{G}=\begin{bmatrix}G\\0\end{bmatrix},\quad \boldsymbol{\delta\mathbf{U}}_0=\begin{bmatrix}\Psi\,\delta\xi\\0\end{bmatrix},$$

we remark, on the one hand, that this system rewrites as

$$\mathbf{M}\frac{\partial\,\boldsymbol{\delta\mathbf{U}}}{\partial t}+\mathbf{K}\,\boldsymbol{\delta\mathbf{U}}=\mathbf{R}_\delta,$$
$$\boldsymbol{\delta\mathbf{U}}(0)=\boldsymbol{\delta\mathbf{U}}_0,$$

and on the other hand, that the adjoint system rewrites as

$$-\mathbf{M}\dot{\mathbf{P}}+\mathbf{K}^\top\mathbf{P}=\mathbf{G},$$
$$\mathbf{P}(T)=0.$$

Hence, the integral in the expression of δC becomes, by integration by parts,

$$\begin{aligned} I &= \int_0^T \mathbf{G}^\top \, \boldsymbol{\delta}\mathbf{U} \, \mathrm{d}t \\ &= \int_0^T (-\mathbf{M}\dot{\mathbf{P}} + \mathbf{K}\mathbf{P})^\top \, \boldsymbol{\delta}\mathbf{U} \, \mathrm{d}t \\ &= \mathbf{P}(0)\mathbf{M}\,\boldsymbol{\delta}\mathbf{U}_0 + \int_0^T \mathbf{P}^\top \Big(\mathbf{M}\frac{\partial\,\boldsymbol{\delta}\mathbf{U}}{\partial t} + \mathbf{K}\,\boldsymbol{\delta}\mathbf{U}\Big)\,\mathrm{d}t \\ &= \mathbf{P}(0)\mathbf{M}\,\boldsymbol{\delta}\mathbf{U}_0 + \int_0^T \mathbf{P}^\top \mathbf{R}_\delta \,\mathrm{d}t. \end{aligned}$$

Coming back to the non-augmented notations, we get

$$\begin{aligned} \mathbf{P}^\top \mathbf{R}_\delta &= -P^\top [\delta K] U, \\ \mathbf{P}(0)^\top \mathbf{M}\,\boldsymbol{\delta}\mathbf{U}_0 &= P^\top M \Psi\, \delta\xi, \end{aligned}$$

so that the tangent value δC simply becomes

$$\begin{aligned} \delta C = \delta\xi^\top \Big\{ &\rho_\xi Q_\xi (\xi - \xi_0) - \rho_\chi \Psi^\top M P(0) \Big\} \\ &+ \delta D^\top \rho_D Q_D (D - D_0) + \rho_\chi \int_0^T P^\top [\delta K] U \,\mathrm{d}t, \end{aligned}$$

where δK is linked to δD by the relation $\delta K = \sum_{j=1}^{p} \delta D^{(j)} K_j$. This ends the proof. □

5.1.4 Gradient of semi-discrete reduced cost function

We simply write down the analog result for the reduced cost function. Let the reduced right-hand side term

$$G^l(t;\xi,D) = (H^l)^\top Q_\chi \chi^l(t;\xi,D),$$

and the corrected reduced stiffness matrix

$$K^l_{\mathrm{adj}}(t;\xi,D) = K^l(D) - \mathrm{d}_{U^l} F^l(t, U^l_h(t;\xi,D))^\top.$$

We define this reduced adjoint solution

$$(P^l_{h,T}(t;\xi,D),\, S^l_{h,T}(t;\xi,D)) \in \mathbb{R}^l \times \mathbb{R}^l$$

over $[0,T]$ by the backward-in-time system

$$\begin{aligned} -M^l \dot{P}^l_{h,T} + K^l_{\mathrm{adj}} P^l_{h,T} - \alpha M^l S^l_{h,T} &= G^l, \\ \dot{S}^l_{h,T} &= \gamma P^l_{h,T} + \beta S^l_{h,T}, \end{aligned}$$

with the zero terminal conditions

$$P^l_{h,T}(T;\xi,D) = 0, \quad S^l_{h,T}(T;\xi,D) = 0.$$

We then obtain the following expression.

Proposition 23.

$$\nabla C^l(\xi,D) = \begin{bmatrix} \rho_\xi Q_\xi(\xi-\xi_0) - \rho_\chi {\Psi^l}^\top M^l P^l_{h,T}(0) \\ \rho_D Q_D(D-D_0) + \rho_\chi \mathcal{I}^l_T(\xi,D) \end{bmatrix},$$

with the vector $\mathcal{I}^l_T(\xi,D)$ *of coordinates*

$$\mathcal{I}^{l,(j)}_T(\xi,D) = \int_0^T P^l_{h,T}(t;\xi,D)^\top K^l_j U^l_h(t;\xi,D)\,\mathrm{d}t, \quad 1 \le j \le p,$$

and $K^l_j = (\Phi^l)^\top K_j \Phi^l$.

5.2 Fully discrete parameter estimation problem and reduced form

We easily adapt the results of the previous section for the fully discrete solution (U^n_h). We define a similar discretized cost function

$$C_{\Delta t}(D) = \frac{\rho_\xi}{2}|\xi-\xi_0|^2_{Q_\xi} + \frac{\rho_D}{2}|D-D_0|^2_{Q_D} + \frac{\rho_\chi}{2}\sum_{n=0}^{N-1}|\chi^n(\xi,D)|^2_{Q_\chi}\,\Delta t, \qquad (5.10)$$

where

$$\chi^n(\xi,D) = Z^n - HU^n_h(\xi,D),$$

subject to (5.5). The same remark about the reparametrization (5.6) holds, i.e. we can minimize

$$C_{\Delta t,\mathrm{unc}}(\xi,D_{\mathrm{unc}}) = C_{\Delta t}(\xi,g(D_{\mathrm{unc}})),$$

so that the problem becomes unconstrained. Moreover

$$\nabla C_{\Delta t,\mathrm{unc}}(\xi,D_{\mathrm{unc}}) = \begin{bmatrix} \nabla_\xi C_{\Delta t}(\xi,g(D_{\mathrm{unc}})) \\ [\frac{\mathrm{d}g}{\mathrm{d}D_{\mathrm{unc}}}]^\top \nabla_D C_{\Delta t}(\xi,g(D_{\mathrm{unc}})) \end{bmatrix}.$$

Again here, the reduced version of the discrete cost function naturally takes the form

$$C^l_{\Delta t}(D) = \frac{\rho_\xi}{2}|\xi-\xi_0|^2_{Q_\xi} + \frac{\rho_D}{2}|D-D_0|^2_{Q_D} + \frac{\rho_\chi}{2}\sum_{n=0}^{N-1}|\chi^{l,n}(\xi,D)|^2_{Q_\chi}\,\Delta t, \qquad (5.11)$$

with the discrete reduced (artificial) noise

$$\chi^{l,n}(\xi, D) = Z^n - H^l U^{l,n}(\xi, D), \tag{5.12}$$

and also admits an unconstrained form. A similar relation between the gradients of $C^l_{\Delta t}$ and $C^l_{\Delta t,\,\mathrm{unc}}$ still holds.

However, we need two specific modifications due to the discretization. We start by explaining how taking more realistic sensors change the inverse covariance matrix Q_χ. Then, we adapt Props. 22 and 23, and show that the corresponding discrete adjoint system differs from the direct discretization of the former time-continuous adjoint system, and that an additional term appears in the expression of the gradient.

5.2.1 Modelling of the noise for the fully discrete measures

By contrast with the ideal sensors described for the semi-discrete problem, we consider here sensors which can only treat the Gaussian white noise signal in a time-discrete way, and therefore better approximate the real measurement systems. We assume that, with a timestep Δt identical to the one that generates the discrete solution (U^n_h), it captures mean values of the signal over each corresponding time subinterval. Then, the random sequence (χ^n) that models the discrete noise naturally results from the same operation on the random process $\chi(t)$. In particular, the realization χ^n_{exact} of this sequence during the observation verifies

$$\chi^n_{\mathrm{exact}} = \frac{1}{\Delta t} \int_{(n-1)\Delta t}^{n\Delta t} \chi_{\mathrm{exact}}(t)\, \mathrm{d}t.$$

We verify easily the classical mean value and covariance results

$$\mathbb{E}[\chi^n] = 0,$$
$$\mathbb{E}[\chi^n (\chi^m)^\top] = \frac{\sigma_\chi}{\Delta t} \delta_{mn}\, \mathrm{Id}_{N_{\mathrm{obs}}} = \widetilde{\sigma}_\chi \delta_{mn}\, \mathrm{Id}_{N_{\mathrm{obs}}},$$

with standard deviation $\widetilde{\sigma}_\chi = \frac{\sigma_\chi}{\sqrt{\Delta t}}$ now depending on the timestep.

Hence, on the one hand, when modelling the measure of a given reference (i.e. exact) solution, we shall incorporate the noise using random sequences of identical probability laws $\mathcal{N}(0, \widetilde{\sigma}_\chi)$ on the sensors. On the other hand, the corresponding inverse covariance matrix $\widetilde{Q}_\chi$ becomes, according to (5.3),

$$\widetilde{Q}_\chi = \widetilde{\sigma}_\chi^{-2}\, \mathrm{Id}_{N_{\mathrm{obs}}} = \Delta t\, Q_\chi.$$

5.2.2 Gradient of fully discrete non-reduced cost function

Following the same path as in Prop. 22 for the time-continuous problem, we provide an analogous definition of the adjoint system, and the resulting expression of the gradient of the time-discrete cost function.

Let the right-hand side term

$$G^n(\xi, D) = H^\top Q_\chi \chi^n(\xi, D),$$

and the corrected stiffness matrix

$$K^n_{\text{adj}}(\xi, D) = K(D) - \mathrm{d}_U F(t^n, U^n_h(\xi, D))^\top.$$

We define the discrete adjoint state

$$(P^n_{h,N}(\xi, D),\, S^n_{h,N}(\xi, D)) \in \mathbb{R}^{N_h} \times \mathbb{R}^{N_h}$$

over $[0, T]$ by the scheme, with $0 \le n \le N-1$,

$$\begin{aligned} -M\frac{P^{n+1}_{h,N} - P^n_{h,N}}{\Delta t} + K^n_{\text{adj}}\Big(\theta P^n_{h,N} + (1-\theta)P^{n+1}_{h,N}\Big) \\ - \alpha M\Big(\theta S^n_{h,N} + (1-\theta)S^{n+1}_{h,N}\Big) = G^n, \end{aligned} \tag{5.13}$$

$$\frac{S^{n+1}_{h,N} - S^n_{h,N}}{\Delta t} = \gamma\Big(\theta P^n_{h,N} + (1-\theta)P^{n+1}_{h,N}\Big) + \beta\Big(\theta S^n_{h,N} + (1-\theta)S^{n+1}_{h,N}\Big), \tag{5.14}$$

with the zero terminal conditions

$$P^N_{h,N}(\xi, D) = 0, \quad S^N_{h,N}(\xi, D) = 0.$$

Of course, in (5.13), we can substitute $S^n_{h,N}$ by its expression in (5.14). Then, for computational purpose, we can rewrite the system as

$$\begin{aligned} A^n_{\text{adj}} P^n_{h,N} &= B^n_{\text{adj}} P^{n+1}_{h,N} + G^n_h + c_3 M S^{n+1}_{h,N}, \\ S^n_{h,N} &= c_2 S^{n+1}_{h,N} - c_1(\theta P^n_{h,N} + (1-\theta)P^{n+1}_{h,N}), \end{aligned}$$

with the constants defined in (4.31) and the matrices

$$\begin{aligned} A^n_{\text{adj}}(D) &= A_0(D) - \theta\Delta t\, \mathrm{d}_U F(t^n, U^n_h(D))^\top, \\ B^n_{\text{adj}}(D) &= B_0(D) + (1-\theta)\Delta t\, \mathrm{d}_U F(t^n, U^n_h(D))^\top, \end{aligned}$$

and A_0 and B_0 defined in (4.32)–(4.33).

Note also that, due to the form of the right-hand side in (5.13), the discrete adjoint system does not exactly derive from an application of the θ-method to the time-continuous one.

We then obtain the following expression, where an additional term, which vanishes when $\Delta t \to 0$, now appears on the state part of the gradient.

Proposition 24.

$$\nabla C_{\Delta t}(\xi, D) = \begin{bmatrix} \rho_\xi Q_\xi(\xi - \xi_0) - \rho_\chi \Psi^\top \mathcal{G}(\xi, D) \\ \rho_D Q_D(D - D_0) + \rho_\chi \mathcal{S}(\xi, D) \end{bmatrix},$$

with the vector

$$\mathcal{G}(\xi, D) = M\Big(P^0_{h,N} - \alpha\theta\Delta t S^0_{h,N}\Big) + \theta\Delta t K^0_{adj} P^0_{h,N},$$

and the vector $\mathcal{S}(\xi, D)$ *of coordinates*

$$\mathcal{S}^{(j)}(\xi, D) = \sum_{n=1}^{N-1} (P^n_{h,N})^\top K_j\Big(\theta U^n_h + (1-\theta)U^{n-1}_h\Big)\Delta t, \quad 1 \le j \le p.$$

Proof. For the purpose of readability, we drop the subscript h, and also the parameters (ξ, D) for the functions of time in this proof.

Let $(\xi, D) \in \mathbb{R}^{N_{\text{modes}}} \times \mathcal{D}$ be a parametric point, and $(\delta\xi, \delta D) \in \mathbb{R}^{N_{\text{modes}}} \times \mathbb{R}^p$ be any direction. Let then $\delta C_{\Delta t}$, $(\delta U^n, \delta W^n)$ and δK be the corresponding tangent values of the cost function $C_{\Delta t}$, the solution (U^n, W^n) and the stiffness matrix K. By differentiation, on the one hand,

$$\delta C_{\Delta t} = \rho_\xi (\xi - \xi_0)^\top Q_\xi \, \delta\xi + \rho_D (D - D_0)^\top Q_D \, \delta D - \rho_\chi S,$$

with the sum

$$S = \sum_{n=0}^{N-1} (Z^n - HU^n)^\top Q_\chi H \, \delta U^n \, \Delta t = \sum_{n=0}^{N-1} (G^n)^\top \delta U^n \, \Delta t,$$

and on the other hand, $(\delta U^n, \delta W^n)$ is solution of

$$\begin{aligned} M\frac{\delta U^{n+1} - \delta U^n}{\Delta t} &+ \theta (K^{n+1}_{\text{adj}})^\top \delta U^{n+1} + (1-\theta)(K^n_{\text{adj}})^\top \delta U^n \\ &= -[\delta K]\Big(\theta U^{n+1} + (1-\theta)U^n\Big) - \gamma M\Big(\theta\, \delta W^{n+1} + (1-\theta)\, \delta W^n\Big), \\ \frac{\delta W^{n+1} - \delta W^n}{\Delta t} &= \alpha(\theta\, \delta U^{n+1} + (1-\theta)\, \delta U^n) - \beta(\theta\, \delta W^{n+1} + (1-\theta)\, \delta W^n), \end{aligned}$$

with the initial conditions

$$\delta U^0 = \Psi \, \delta\xi, \quad \delta W^0 = 0.$$

Introducing the augmented matrices of size $2N_h \times 2N_h$

$$\mathbf{M} = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix}, \quad \mathbf{K}^n = \begin{bmatrix} K - \mathrm{d}_U F(t^n, U^n) & \gamma M \\ -\alpha M & \beta M \end{bmatrix}, \quad \mathbf{R}^n = \begin{bmatrix} -[\delta K]U^n \\ 0 \end{bmatrix},$$

and the augmented vectors of size $2N_h$

$$\boldsymbol{\delta}\mathbf{U}^n = \begin{bmatrix} \delta U^n \\ \delta W^n \end{bmatrix}, \quad \mathbf{G}^n = \begin{bmatrix} G^n \\ 0 \end{bmatrix}, \quad \boldsymbol{\delta}\mathbf{U}^0 = \begin{bmatrix} \Psi\, \delta\xi \\ 0 \end{bmatrix},$$

we remark, on the one hand, that this system rewrites as

$$\mathbf{M}\frac{\delta\mathbf{U}^{n+1}-\delta\mathbf{U}^{n}}{\Delta t}+\theta\mathbf{K}^{n+1}\,\delta\mathbf{U}^{n+1}+(1-\theta)\mathbf{K}^{n}\,\delta\mathbf{U}^{n}=\theta\mathbf{R}^{n+1}+(1-\theta)\mathbf{R}^{n},$$

and on the other hand, that the adjoint system rewrites as

$$-\mathbf{M}\frac{\mathbf{P}^{n+1}-\mathbf{P}^{n}}{\Delta t}+(\mathbf{K}^{n})^{\top}\Big(\theta\mathbf{P}^{n}+(1-\theta)\mathbf{P}^{n+1}\Big)=\mathbf{G}^{n}.$$

Hence, the sum in the expression of $\delta C_{\Delta t}$ becomes

$$\begin{aligned}
S &= \sum_{n=0}^{N-1}(\mathbf{G}^{n})^{\top}\,\delta\mathbf{U}^{n}\,\Delta t\\
&= \sum_{n=0}^{N-1}\Big\{-M(\mathbf{P}^{n+1}-\mathbf{P}^{n})+\Delta t\,(\mathbf{K}^{n})^{\top}\Big(\theta\mathbf{P}^{n}+(1-\theta)\mathbf{P}^{n+1}\Big)\Big\}^{\top}\,\delta\mathbf{U}^{n}\\
&= (\mathbf{P}^{0})^{\top}(\mathbf{M}+\theta\Delta t\mathbf{K}^{0})\,\delta\mathbf{U}^{0}\\
&\quad+\sum_{n=1}^{N-1}(\mathbf{P}^{n})^{\top}\Big\{\mathbf{M}(\delta\mathbf{U}^{n}-\delta\mathbf{U}^{n-1})+\Delta t\Big(\theta\mathbf{K}^{n}\,\delta\mathbf{U}^{n}+(1-\theta)\mathbf{K}^{n-1}\,\delta\mathbf{U}^{n-1}\Big)\Big\}\\
&= (\mathbf{P}^{0})^{\top}(\mathbf{M}+\theta\Delta t\mathbf{K}^{0})\,\delta\mathbf{U}^{0}+\Delta t\sum_{n=1}^{N-1}(\mathbf{P}^{n})^{\top}\Big(\theta\mathbf{R}^{n}+(1-\theta)\mathbf{R}^{n-1}\Big).
\end{aligned}$$

Coming back to the non-augmented notations, we get

$$(\mathbf{P}^{n})^{\top}\Big(\theta\mathbf{R}^{n}+(1-\theta)\mathbf{R}^{n-1}\Big)=-(P^{n})^{\top}[\delta K]\Big(\theta U^{n}+(1-\theta)U^{n-1}\Big),$$

$$(\mathbf{P}^{0})^{\top}(\mathbf{M}+\theta\Delta t\mathbf{K}^{0})\,\delta\mathbf{U}^{0}=\Big\{(P^{0}-\alpha\theta\Delta tS^{0})^{\top}M+\theta\Delta t(P^{0})^{\top}(K^{0}_{\mathrm{adj}})^{\top}\Big\}^{\top}\Psi\,\delta\xi,$$

so that the tangent value $\delta C_{\Delta t}$ simply becomes

$$\begin{aligned}
\delta C_{\Delta t} &= \delta\xi^{\top}\Big\{\rho_{\xi}Q_{\xi}(\xi-\xi_{0})-\rho_{\chi}\Psi^{\top}\mathcal{G}\Big\}\\
&\quad+\delta D^{\top}\rho_{D}Q_{D}(D-D_{0})+\rho_{\chi}\sum_{n=0}^{N-1}(P^{n})^{\top}[\delta K]\Big(\theta U^{n}+(1-\theta)U^{n-1}\Big)\Delta t.
\end{aligned}$$

This ends the proof. □

5.2.3 Gradient of fully discrete reduced cost function

Again, we simply write down the analog result of Prop. 24 for the fully discrete reduced cost function. Let the reduced right-hand side term

$$G_{h}^{l,n}(\xi,D)=(H^{l})^{\top}Q_{\chi}\chi^{l,n}(\xi,D),$$

with $\chi^{l,n}(\xi, D) = Z^n - H^l U^{l,n}$, and the corrected reduced stiffness matrix

$$K^{l,n}_{\text{adj}}(\xi, D) = K^l(D) - \mathrm{d}_U F^l(t^n, U^{l,n})^\top.$$

We define the discrete reduced adjoint solution

$$(P^{l,n}_{h,N}(\xi, D),\, S^{l,n}_{h,N}(\xi, D)) \in \mathbb{R}^l \times \mathbb{R}^l$$

over $[0, T]$ by the scheme, with $0 \le n \le N-1$,

$$\begin{aligned} -M^l \frac{P^{l,n+1}_{h,N} - P^{l,n}_{h,N}}{\Delta t} + K^{l,n}_{\text{adj}}\Big(\theta P^{l,n}_{h,N} + (1-\theta) P^{l,n+1}_{h,N}\Big) \\ - \alpha M^l \Big(\theta S^{l,n}_{h,N} + (1-\theta) S^{l,n+1}_{h,N}\Big) = G^{l,n}_h, \\ \frac{S^{l,n+1}_{h,N} - S^{l,n}_{h,N}}{\Delta t} = \gamma\Big(\theta P^{l,n}_{h,N} + (1-\theta) P^{l,n+1}_{h,N}\Big) + \beta\Big(\theta S^{l,n}_{h,N} + (1-\theta) S^{l,n+1}_{h,N}\Big), \end{aligned}$$

with the zero terminal conditions

$$P^{l,N}_{h,N}(\xi, D) = 0, \quad S^{l,N}_{h,N}(\xi, D) = 0.$$

We then obtain the following expression

Proposition 25.

$$\nabla C^l_{\Delta t}(\xi, D) = \begin{bmatrix} \rho_\xi Q_\xi(\xi - \xi_0) - \rho_\chi (\Psi^l)^\top \mathcal{G}^l(\xi, D) \\ \rho_D Q_D (D - D_0) + \rho_\chi \mathcal{S}^l(\xi, D) \end{bmatrix},$$

with the vector

$$\mathcal{G}^l(\xi, D) = M\Big(P^{l,0}_{h,N} - \alpha\theta\Delta t S^{l,0}_{h,N}\Big) + \theta\Delta t K^{l,0}_{adj} P^{l,0}_{h,N},$$

and the vector $\mathcal{S}^l(\xi, D)$ of coordinates

$$\mathcal{S}^{l,(j)}(\xi, D) = \sum_{n=1}^{N-1} (P^{l,n}_{h,N})^\top K^l_j \Big(\theta U^{l,n}_h + (1-\theta) U^{l,n-1}_h\Big) \Delta t, \quad 1 \le j \le p.$$

5.3 Numerical experiments of reduced-order parameter estimation

We have written above some practical expressions of the fully discrete cost function and its gradient, in both their non-reduced and reduced versions. Based on them, we show here the analysis of some numerical experiments performed on variational parameter-state estimation in POD reduced form. We aim at assessing the convergence, with respect to the POD

rank, of the minimizers of the reduced cost functions towards the minimizer $(\xi_\star, D_\star)$ of the complete cost function.

Here, we particularly focus on the parameter part of the estimated vector. Also, the notion of multi-POD is here only applied to the parameter part, the initial condition being fixed to $U_{h,0,\text{exact}}$ for this basis construction. In the tests of this section and in order to check the stability with respect to the initial condition, we therefore include an a priori state vector ξ_0 which corresponds to an initial condition $\widetilde{U}_{h,0}$ that is close enough to $U_{h,0,\text{exact}}$ to be compatible with limited variations of ξ, enforced by choosing a small variance σ_ξ during the minimization. In other words, we choose $\xi_0 = \xi_{\text{proj}}$ where

$$\xi_{\text{proj}} = \Psi^\top K_{H_0^1} U_{h,0,\text{exact}} \tag{5.15}$$

concatenates the coefficients of the H_0^1-projection of $U_{h,0,\text{exact}}$ onto $\operatorname{Im}\Psi$. In accordance with this assumption, the choice of a small variance σ_ξ reflects the great confidence that we shall put in the state. The next chapter, where we apply sequential methods of estimation on the mechanical model of a heart, introduces a more general framework.

As a continuation of the comparison exposed in Section 4.3.4 for direct simulations, we come back to the range of solutions described by the parameter values of Tab. 4.1, where the amplitude A is undetermined. Like in Chapter 4, this enables us to consider the nominal amplitude case corresponding to $A = 1$, and the critical amplitude case corresponding to $A = A_c$. For each case, the corresponding standard POD and multi-POD are hence identical to those of Chapter 4. We sum up the new constants of interest in Tab. 5.1, including those introduced by the observation process.

With regard to the minimization, we define and justify the chosen iterative process, that apply to all the involved minimizer computations in the sequel, for both the non-reduced and the reduced cost functions. Namely,

- we always initialize the minimization algorithm with the a priori vector value (ξ_0, D_0), since it represents the virtually only available knowledge of the state-parameter vector;
- we use a black-box algorithm as advised by [6]. In this case, since we can compute here the mathematically exact gradient of the discrete cost functions, we take a first-order, *trust-region* method based on the interior-reflective Newton method [14, 15];
- and we use the following stopping criterion: when the relative variation of (ξ, D) between consecutive iterations is smaller than $1 \cdot 10^{-6}$ in norm, the algorithm stops and *has numerically converged*. Indeed, this value corresponds the smallest relative projection error implied by the criterion (3.33), and is consistent with the assessed reduction errors.

We underline that all the involved minimization processes in the sequel succeed in converging.

Parameter	
p	2
ρ_D	1
D_{exact}	$[1.1 \cdot 10^{-3}, 1.4 \cdot 10^{-3}]$
D_0	$[2.0 \cdot 10^{-3}, 2.0 \cdot 10^{-3}]$

State uncertainty	
N_{modes}	10
ρ_ξ	1
σ_ξ	$5 \cdot 10^{-3}$
$u_{h,0,\text{exact}}$	$\eta(x; A, \frac{1}{2}, \frac{1}{4})$
ξ_0	see (5.15)

Observation noise	
N_{obs}	9
ρ_χ	1
$\widetilde{\sigma}_\chi$	$2 \cdot 10^{-3}/\sqrt{\Delta t}$

Table 5.1: Observation constants, used with the constants of Table 4.1

From a wider point of view, given the complexity of the PDEs involved in the cost functions, we make two other comments to expose the possible shortcomings of this iterative method itself, and justify its use here.

First, we verify that, with the configuration of constants defined by Tabs. 4.1 and 5.1, and given several tests corresponding to different generations of random noises (χ^n_{exact}), the algorithm never seems to be trapped in any local minimum. We observed than the stopping on local minima can actually occur when choosing incompatible observation constants, e.g. taking an a priori ξ_0 that is too far from ξ_{exact} in comparison with the distance defined by σ_ξ. In particular, in such a case, we found that for the same set of constants, the algorithm may stop on several unrelated local minimizers when performing the estimation with different generated noises.

Also, a convergence towards a seemingly unique minimum may still cause a wrong estimation if, e.g. for an ill-posedness reason due to the lack of measurements, the exact vector $(\xi_{\text{exact}}, D_{\text{exact}})$ is not *observable*. Indeed, the existence of this minimum might be only due to the "convexifying" terms of the cost functions. We verify that, in our case, our set of sensors is sufficient to perform an accurate and *unique* estimation.

5.3.1 Results for the non-reduced estimation problem

We begin by describing the results for the non-reduced estimation problem. Name $u_{h,\star}$ the solution corresponding to the optimal state-parameter value $(\xi_\star, D_\star)$. We define the relative estimation error in state, parameter

	$A = 1$	$A = A_c$
Iterations	10	34
τ_ξ (%)	0.02	0.12
τ_D (%)	0.17	0.27
τ_{u_h} (%)	0.50	0.48

Table 5.2: Number of iterations and error values for the non-reduced estimation problems corresponding to the two cases of nominal amplitude and critical amplitude

and solution by

$$\tau_\xi = \frac{|\xi_\star - \xi_{\text{proj}}|}{|\xi_{\text{proj}}|},$$
$$\tau_D = \frac{|D_\star - D_{\text{exact}}|}{|D_{\text{exact}}|},$$
$$\tau_{u_h} = \frac{\|u_{h,\star} - u_{h,\,\text{exact}}\|_{L^2(0,T;V)}}{\|u_{h,\,\text{exact}}\|_{L^2(0,T;V)}},$$

respectively, and group the corresponding values for the two cases considered $A = 1$ and $A = A_c$ in Tab. 5.2.

A global comparison between the two cases shows that here, while the case of critical amplitude seems to attain slightly more accurate results, the associated estimation problems appears more delicate to solve. Indeed, it requires four times more iterations for the minimization process to converge than for the case of nominal amplitude. This reflects severe oscillations of the cost function value around the states corresponding to action potential thresholds.

For the numerical values now, we first verify that, since we give the projection of the exact initial condition as the a priori state, the errors in state remain small. This shows that for the choice of the standard deviation $\widetilde{\sigma}_\chi$ displayed in Tab. 5.1, the estimation problem is stable with respect to the state.

Then, the optimal parameter value $D_\star$ departs from the exact vector value D_{exact} with a greater relative error. This property is due to the presence of the "convexifying" term $|D - D_0|^2$ in the functional $C_{\Delta t}$ and occurs even when taking a null measure noise. The observed order of magnitude $O(10^{-2})$ sets the magnitude of the largest tolerable reduction error in parameter for the next section.

Finally, we observe that the error in solution is of the same order.

5.3.2 Comparison of efficiency between the standard POD and the multi-POD

We now display the reduction results for the reduced versions of this estimation problem. Here, we recall that we minimize the reduced discrete cost function $C^l_{\Delta t}$, defined by (5.11) and (5.12). We study the dependence of the result with respect to l and for two different POD bases.

We now fix the notations for the reduced-order optimal quantities, and for the purpose of simplicity, without distinction between the $A = 1$ and $A = A_c$ cases. (Of course, each case will take its own set of values). Let $j \in \{1, 2\}$ determine whether we choose the standard POD at D_c ($j = 1$) or the multi-POD of degree 1 ($j = 2$), and $l_j \in \{1, \ldots, l_{j,\max}\}$ a corresponding POD rank. We name

- $C^{l_j}_{\Delta t,j}$ the reduced cost function $C^l_{\Delta t}$ where we have substituted the projector $\pi^{l_j}_j$ for the dummy projector π^l;
- $(\xi^{l_j}_{j,\star}, D^{l_j}_{j,\star})$ the minimizer of $C^{l_j}_{\Delta t,j}$, that we shall name the *reduced-order minimizer;*
- and $u^{l_j}_{h,j,\star}$ the reduced solution $u^{l_j}_h$ (of reduction space $\operatorname{Im}\pi^{l_j}_j$) corresponding to the reduced-order minimizer $(\xi^{l_j}_{j,\star}, D^{l_j}_{j,\star})$.

Since we wish to isolate the effect of reduction in itself, we here specifically evaluate and analyse the *convergence (with respect to POD rank) error* between

$$(\xi^{l_j}_{j,\star}, D^{l_j}_{j,\star}) \quad \text{and} \quad (\xi_\star, D_\star),$$

i.e. between the reduced minimizers and the original one, instead of the effective *reduced-order estimation error* between

$$(\xi^{l_j}_{j,\star}, D^{l_j}_{j,\star}) \quad \text{and} \quad (\xi_{\text{proj}}, D_{\text{exact}}),$$

i.e. between the reduced minimizers and the closest state-parameter vector value to the exact observed solution. We define, analogously to the non-reduced problem, the relative convergence errors in state, parameter and solution by

$$\eta^{l_j}_{j,D} = \frac{|D_\star - D^{l_j}_{j,\star}|}{|D_\star|},$$

$$\eta^{l_j}_{j,\xi} = \frac{|\xi_\star - \xi^{l_j}_{j,\star}|}{|\xi_\star|},$$

$$\eta^{l_j}_{j,u_h} = \frac{\|u_{h,\star} - u^{l_j}_{h,j,\star}\|_{L^2(0,T;V)}}{\|u_{h,\star}\|_{L^2(0,T;V)}}.$$

Figures 5.1 and 5.2 display the corresponding results for the nominal amplitude $A = 1$ and the critical amplitude $A = A_c$ cases, respectively. Both

figures place the results for the standard POD at the top, and those for the multi-POD at the bottom. Also, the charts for the number of iterations before convergence in each reduced minimization process, on the right, accompany those for the relative convergence errors, on the left.

To begin, note that both methods feature equivalent results for the state part, i.e. an almost constant behaviour with respect to the POD rank, of a similarly small order $O(10^{-3})$ to that of the complete state estimation error τ_ξ. We interpret this phenomenon with the same reason invoked above about the the a priori state, equated to the a priori initialization point.

For the parametric part now, here again, a similar conclusion to that of Chapter 4 follows for the comparison of efficiency between the standard POD and the multi-POD, in this inverse problem situation, see Tab. 5.3 for the attained minimum values of relative convergence error in parameter and in solution. First, the reduced-order minimizers with the standard POD fail in all cases at converging towards the minimizer of the complete functional. On the contrary, the reduced-order minimizers with the multi-POD show better reduction results, although a significant improvement only appears in the nominal amplitude case, where the whole parameter subdomain $\mathcal{D}$ describes travelling pulse solutions. We explain this low reduction quality for the critical amplitude case with the same arguments developed in Chapter 4.

With regard to the nominal amplitude case treated with the multi-POD method now, the decrease rates of the convergence errors and the POD remainder are satisfyingly close, reaching relative errors of magnitude as small as $O(10^{-4})$. More precisely:

- from 85 modes, $\eta_{2,D}^{l_2} < 1 \cdot 10^{-3}$;
- and from 91 modes, $\eta_{2,u_h}^{l_2} < 1 \cdot 10^{-3}$.

5.3.3 Limitation of the multi-POD basis with a reduction-to-estimation tolerance

In this application, we shall not actually aim at a reduction error that is as low as possible, but one that instead meets the order of magnitude of the non-reduced estimation problem. In other words, an $O(10^{-4})$ relative reduction error is so small compared to an $O(10^{-2})$ relative non-reduced estimation error that we may significantly weaken the basis and remain with relevantly small reduction errors.

To that end, we may consider a reduction-to-estimation tolerance α, say 50% or 10%, which defines a smallest multi-POD rank $\bar{l}_{2,D}$ (resp. $\bar{l}_{2,u_h}$) such that for all $l_2 \geq \bar{l}_{2,D}$ (resp. $l_2 \geq \bar{l}_{2,u_h}$),

$$\frac{|D_\star - D_{2,\star}^{l_2}|}{|D_\star - D_{\text{exact}}|} \leq \alpha \quad \Big(\text{resp. } \frac{\|u_{h,\star} - u_{h,2,\star}^{l_2}\|_{L^2(0,T;V)}}{\|u_{h,\star} - u_{h,\text{exact}}\|_{L^2(0,T;V)}} \leq \alpha\Big).$$

		$A = 1$	$A = A_c$
Standard POD method			
$l_{1,\max}$		39	39
$\min_{l_1} \eta_{1,D}^{l_1}$	(%)	11.85	11.98
$\min_{l_1} \eta_{1,u_h}^{l_1}$	(%)	30.90	33.02
Multi-POD method			
$l_{2,\max}$		100	87
$\min_{l_2} \eta_{2,D}^{l_2}$	(%)	< 0.01	1.30
$\min_{l_2} \eta_{2,u_h}^{l_2}$	(%)	0.02	10.78

Table 5.3: Convergence results with the standard POD and the multi-POD methods

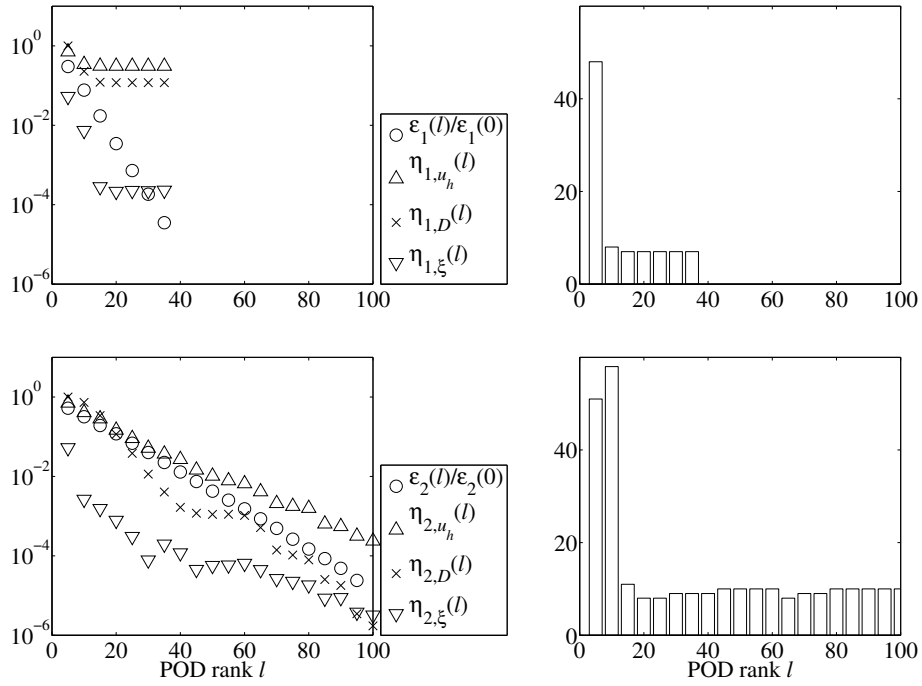

Figure 5.1: Case of nominal amplitude $A = 1$ for the initial pulse. Reduction errors of parameter-state estimation (left) with corresponding number of optimization iterations (right). Top: using a standard POD at central parameter value. Bottom: using a multi-POD of degree 1

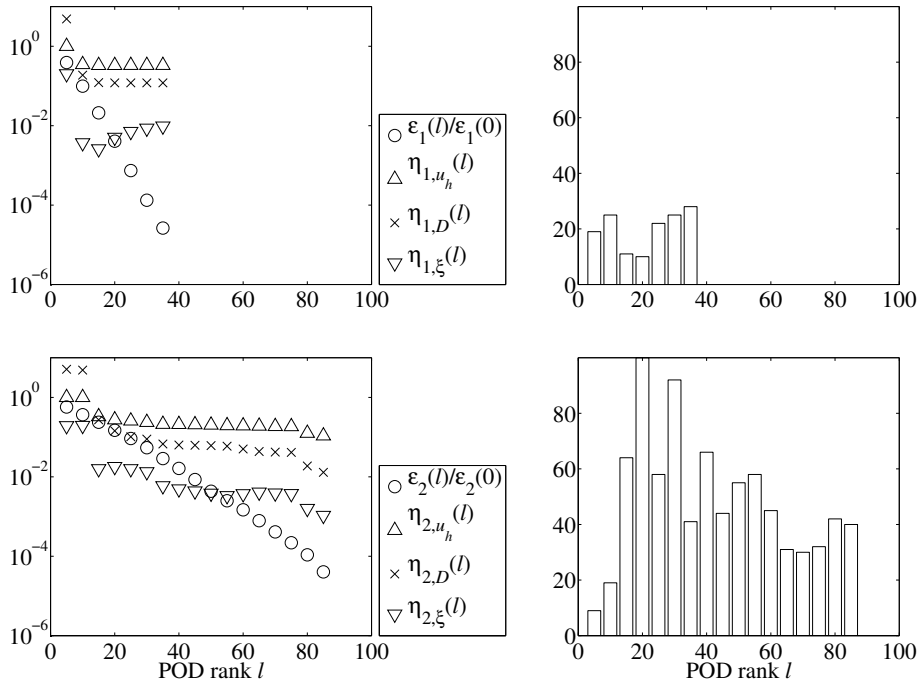

Figure 5.2: Case of critical amplitude $A = A_c$ for the initial pulse. Reduction errors of parameter-state estimation (left) with corresponding number of optimization iterations (right). Top: using a standard POD at central parameter value. Bottom: using a multi-POD of degree 1

α (%)	50	10
$\bar{l}_{2,D}$	65	70
$\bar{l}_{2,u_h}$	70	95

Table 5.4: Case of nominal amplitude $A = 1$: multi-POD basis limitation with a reduction-to-estimation tolerance

Table 5.4 displays the corresponding results for the nominal amplitude case reduced with a multi-POD. We deduce that, given a sensible tolerance of $\alpha = 50\,\%$, a basis of only 70 modes brings a sufficient reduction error for this problem. In the present mesh configuration, this result means that we can actually decrease the number of degrees of freedom by a 65 % factor.

5.4 Conclusion

We proposed a variational approach to a specific parameter-state estimation problem on the highly sensitive FitzHugh–Nagumo equations, followed by its immediate reduced-order version. We detailed the models of uncertainty that justify the different choices of norm attached to each term of the standardly quadratic cost functions, in complete and reduced form. With a view to numerically applying a minimization method that makes use of gradient information, we proved convenient formulae for the gradient of the resulting a priori non-convex cost functions, in both semi-discrete and fully discrete case, and both complete and reduced cases.

Then, we extended the analysis on the same sets of constants used for the direct reduction error analysis of Chapter 4, with the additional difficulty implied by the nature of the observations. We chose to focus more on the variation through the generalized diffusion coefficient D. With artificial, partial and imperfect measurements of a solution at the parameter-state vector value $(D_{\text{exact}}, \xi_{\text{exact}})$, given a clue of an a priori point near D_{exact}, we tested the ability of the previous multi-POD and standard POD methodologies, to be employed in the associated reduced-order methods to retrieve that parameter value with a sufficient accuracy.

On the one hand, the standard POD still suffers from the lack of parameter variability information in its basis, and fails at approximating the original problem. On the other hand, the multi-POD confirms its reliability for building reduced-order, here inverse, models. Indeed, more precisely, we aimed at making the minimizers of the reduced cost functions approach the minimizer of the original cost function, within an error comparable to the parameter estimation error associated with the latter. The multi-POD achieves this goal using about 70 multi-POD modes. Hence, its compatibility with first-order information in parameter, represented by

the differentiation of the reduced cost functions during the optimization, is once again attested.

Now that the multi-POD methodology proves its efficiency for correctly reducing complex PDE-based optimization problems, the natural next step would be its application on optimal control problems. In this case, the next difficulty appears in the very variable of the cost function, that is now infinite-dimensional, since it takes the form of a time-dependent function (the command). Hence, the question of integrating all the necessary parametric variability in its basis takes a new twist. We leave here this problematic as a possible perspective.

Chapter 6

Reduced sequential parameter estimation problem on a mechanical model for the heart

The previous chapter has presented a successful application of our multi-POD reduction method, proposed in Chapter 4, to variationally posed problems of state-parameter estimation. However, while the POD reduction then constitutes an improvement for the computational cost of large systems, one may still question the efficiency of this very type of formulation. Indeed, while the minimization of cost functions is analytically handy, the resulting simulations raise several problems. Technically, they require two full solutions over $[0, T]$ at each minimization step, namely the direct and adjoint ones. Then, from a methodological point of view, first, the stability analysis, with respect to the parameters to be estimated, is hard to derive, and even more for a nonconvex cost function, that shall face local minima. Secondly, the solutions associated with optimal parameters do not present an "extensivity" in time. This means that no easy mathematical link exists between such a solution corresponding to a time interval $[0, T]$ and another corresponding to $[0, T + \epsilon]$, whereas one do not wish to start the simulations from scratch because of some needed addition of a small interval $[T, T + \epsilon]$. We may also add, in a viewpoint that exceeds the application scope of this report though, that the variational methods also exclude the idea of real-time estimation.

On the contrary, the sequential approach of *filtering* induces these computational benefits. The context is still made of some measurements of a system and a model based on some parameter-dependent PDEs. Filtering consists in adding in the equations some feedback corrective term, which is essentially controlled by the *covariance* associated with the solution uncertainties. The resulting new solution is called an *observer*, that follows its specific dynamics, so that the former optimization step then becomes a simple, easier to handle, *timestep*. Then, we can sum up as follows the main idea: the more the observer progresses, the more it chronologically

collects some measurements, and hence, the more it approaches the exact solution. Therefore, the *storage problem* is naturally solved, since not all the measurements corresponding to a fixed interval $[0, T]$ are needed at each progression. Also, the former *continuation problem* here disappears by construction. However, the overall computational cost of the observer usually remains excessive even for reasonably large systems, because of the actual coupling with a dynamics of covariance matrix with full profile that has the order of the solution vector. This is why we study in this chapter the intervention of the above mentioned multi-POD reduction method, that has displayed promising results in the previous chapters.

We begin by presenting and building for linear systems the standard *Kalman filter* [33], which is a cornerstone in the estimation literature. Because of the linear assumption, it is of little use in itself, but forms a basis from which many other types of filters derive. We then rapidly move on to its application by linearization of nonlinear systems, called the *extended Kalman filter* [26]. Despite its practical appeal among the most studied estimation methods, it hardly guarantees a robustness properties. Nevertheless, we propose a systematic way of applying the reduced-order modelling on this filter in order to make the computational gain explicit.

In order to avoid any ambiguity, we underline that the concept of reduction used throughout this chapter only refers to that of *reduced-order modelling*. By contrast, a notion that we may call *reduction of uncertainties*, and that appears e.g. through the method of *singular evolution extended Kalman filter* [42] (see also Moireau et al. [40]), is very different and relies on an assumption of singularity of the covariance matrix observer. However, while it is actually used in the simulations of the heart model that we study below, *we do not specifically develop this reduction of uncertainties here.*

To finish, we develop and particularly focus on a concrete medical application of joined filtering and reduced-order modelling methods. Namely, through the numerical heart model presented in Section 3.5, we aim at estimating a contractility field on a pig heart using real clinical data, and especially, at automatically retrieving an infarcted zone. Indeed, as a reference, such an estimation with a non-reduced approach and computationally lighter filtering method has already been conducted and provided satisfying results, see Chabiniok et al. [10].

To that end, we present another existing derived filtering method, the *unscented Kalman filtering* [28], which is more statistics-oriented and avoids some linearization problems posed by the extended Kalman filter. We then apply its natural reduced-order version to the reduced-order heart model, and assess the effect of reduction on the accuracy of the estimated contractility field, and also on the dynamics of the observers themselves. We underline that Ph. Moireau (also [41, 39]) substantially contributed to this part.

6.1 POD reduction of a Kalman observer for a semi-discrete linear system

Consider a general finite-dimensional, first-order ODE in $\mathbb{R}^{N_h}$

$$\dot{X}(t;\xi) = \mathcal{A}(t, X(t;\xi)) + R(t), \tag{6.1}$$

$$X(0;\xi) = X_0 + \xi, \tag{6.2}$$

with a nonlinear mapping

$$\mathcal{A} : \begin{array}{ccc} [0,\infty[\times \mathbb{R}^{N_h} & \to & \mathbb{R}^{N_h} \\ (t,\beta) & \mapsto & \mathcal{A}(t,\beta), \end{array}$$

a source term $R(t) \in \mathbb{R}^{N_h}$, an a priori initial condition $X_0 \in \mathbb{R}^{N_h}$, and where $\xi \in \mathbb{R}^{N_h}$ represents an unknown of the system, modelling some uncertainty. The solution $X(t)$ represents the spatial discretization of some partial differential system by a finite element method, that involves N_h degrees of freedom. For the purpose of simplicity, we assume that no other parameter is unknown, and that the system collects some time-continuous measurements $Z(t) \in \mathbb{R}^{N_{\text{obs}}}$. However, we now consider a general and nonlinear associated model of observation. More precisely, with and observed solution $X_{\text{exact}}(t)$ and a measurement noise $\chi_{\text{exact}}(t)$, that is a realization of some random vector process $\chi(t)$, we consider

$$Z(t) = \mathcal{H}(X_{\text{exact}}(t)) + \chi_{\text{exact}}(t),$$

with $\mathcal{H}$ a nonlinear, non-invertible $\mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ mapping.

We momentarily simplify the problem to a case of linear dynamic operator and linear observation operator, in order to introduce the *Kalman filter* (KF), widely used, studied and adapted in the estimation theory, and originated in [33]. It is a certain system, the solution $\widehat{X}$ of which we name the *KF observer*, that tries to retrieve *in real-time* the exact trajectory X_{exact} according to the given observations $Z(t)$.

Although the KF is classically built with a Bayesian methodology (e.g. [26]), we shall not investigate this approach here. Yet, in a view to continuity with the previous chapters, we define it through a variational estimation with continuously varying time interval, so that that a particular equivalence between the variational and sequential strategies appears.

We then naturally extend the resulting equations and definitions to the initial case of nonlinear operators $\mathcal{A}$ and $\mathcal{H}$, leading to the *extended Kalman filter* (EKF). Also, and still in continuity with the previous study, we introduce an uncertainty in the dynamic operator, namely becoming of the form $\mathcal{A}(t,\beta;\theta)$. We use a canonical and simple way to estimate it *over time* with little modifications of the preceding framework. As is, the computational cost of the resulting EKF observer still is prohibitive. We therefore naturally introduce its reduced-order form, that becomes completely numerically realizable.

6.1.1 Variational construction of the Kalman filter for a linear system and a linear observation operator

In this section, we simplify both the mapping $\mathcal{A}$ and the observation operator $\mathcal{H}$ into linear ones, i.e.

$$\mathcal{H}(\beta) = H\beta,$$
$$\mathcal{A}(t,\beta) = A(t)\beta,$$

where $H \in \mathbb{R}^{N_{\text{obs}} \times N_h}$ and $A(t) \in \mathbb{R}^{N_h \times N_h}$ represent the hence β-independent Jacobian matrices of $\mathcal{H}$ and $\mathcal{A}(t,\cdot)$, respectively.

At each time t, we may solve, with the knowledge of the measurements on the whole time interval $[0,t]$, the problem of finding an *optimal approximation* of the initial condition uncertainty ξ in a variational sense. Naming, like in Chapter 5, Q_ξ and Q_χ the inverse covariance matrices associated with ξ and the observation noise χ, we minimize the regular cost function

$$C_t(\xi) = \frac{\rho_\xi}{2}|\xi|^2_{Q_\xi} + \frac{\rho_\chi}{2}\int_0^t |\chi(s;\xi)|^2_{Q_\chi}\,\mathrm{d}s,$$

where

$$\chi(s;\xi) = Z(s) - HX(s;\xi)$$

is a function that C_t aims at keeping as conform as possible to the actual noise distribution. Let $\xi_t^\star$ be a minimizer, that we assume to be unique, then building a certain function of time.

We define an adjoint solution $P_t(t;\xi) \in \mathbb{R}^{N_h}$ over $[0,t]$ by the backward-in-time system

$$-\dot{P}_t(s;\xi) = A(s)^\top P_t(s;\xi) + G(s;\xi), \quad 0 \le s \le t,$$
$$P_t(t) = 0,$$

with the right-hand side term $G(s;\xi) = H^\top Q_\chi \chi(s;\xi)$. We then get the following result.

Lemma 5. *For all t,*

$$\xi_t^\star = \frac{\rho_\chi}{\rho_\xi} Q_\xi^{-1} P_t(0;\xi_t^\star).$$

Proof. Let $\xi \in \mathbb{R}^{N_h}$ be a parametric point, and $\delta\xi$ be any direction. Let then δC_t and δX be the corresponding tangent values of the cost function C_t and the solution X. By differentiation, on the one hand,

$$\delta C_t = \rho_\xi \xi^\top Q_\xi\, \delta\xi - \rho_\chi I,$$

with the integral

$$I = \int_0^T (Z - HX)^\top Q_\chi H\, \delta X\,\mathrm{d}t = \int_0^T G^\top\, \delta X\,\mathrm{d}t,$$

and on the other hand, δX is the solution of

$$\frac{\partial \delta X}{\partial t} = A\,\delta X,$$
$$\delta X(0) = \delta\xi.$$

By integration by parts, the integral I becomes

$$\begin{aligned} I &= \int_0^t (-\dot{P} + A^\top P)^\top \delta X \,\mathrm{d}t \\ &= P(0)^\top \delta X(0) + \int_0^t P^\top \Big(\frac{\partial \delta X}{\partial t} + A\,\delta X\Big)\,\mathrm{d}t \\ &= P(0)^\top \delta\xi, \end{aligned}$$

so that the tangent value verifies

$$\delta C = \delta\xi^\top(\rho_\xi Q_\xi \xi - \rho_\chi P(0)).$$

The first-order optimality condition, given by $\delta C = 0$ on any direction $\delta\xi$, is then

$$\rho_\xi Q_\xi \xi^\star - \rho_\chi P^\star(0;\xi^\star) = 0.$$

This ends the proof. □

Substituting this result for $\xi_t^\star$ in the initial condition, we remark that, for a fixed t, the augmented variable $(X_t^\star, P_t^\star) = (X, P_t)(\xi_t^\star)$ verifies the particular boundary problem on $[0,t]$

$$\dot{X}_t^\star = AX_t^\star + R,$$
$$-\dot{P}_t^\star = A^\top P_t^\star + G(\xi_t^\star),$$

with the initial and terminal conditions

$$X_t^\star(0) = X_0 + \frac{\rho_\chi}{\rho_\xi} Q_\xi^{-1} P_t^\star(0),$$
$$P_t^\star(t) = 0.$$

Moreover, the covariance matrix $Q_\xi^{-1} = \mathrm{Cov}_\xi$ appears in the initial condition.

We now introduce two solutions of systems that are independent of the uncertainty ξ, and that play a fundamental role in the description of the optimal solution $X_t^\star$. Note also that whereas the definition of $X_t^\star$ over $[0,t]$ depends on t, (it is in general not a continuation of the solutions $X_s^\star$ for $s \le t$), the following two definitions are not interval-dependent.

Let $\Sigma(t)$, $t \ge 0$, play the role of a dynamic covariance matrix, and be the solution of the Riccati equation on $N_h \times N_h$ matrices

$$\dot{\Sigma}(t) - A(t)\Sigma(t) - \Sigma(t)A(t)^\top + \Sigma(t)H^\top Q_\chi H\Sigma(t) = 0,$$

starting with

$$\Sigma(0) = \mathrm{Cov}_\xi .$$

We now build $\widehat{X}(t)$, $t \geq 0$, the *observer*, solution of the following system, where a new term appears in the right-hand side in comparison with the original system, and where the initial condition is now fixed, i.e.

$$\dot{\widehat{X}}(t) = A(t)\widehat{X}(t) + R(t) + K(t)\hat{\chi}(t),$$
$$\widehat{X}(0) = X_0,$$

with the corrective term

$$\hat{\chi}(t) = Z(t) - H\widehat{X}(t),$$

named the *innovation*, and the matrix

$$K(t) = \Sigma(t)H^\top Q_\chi,$$

named the *Kalman gain.* We get the following proposition.

Proposition 26. *For all* $0 \leq s \leq t$,

$$X_t^\star(s) = \widehat{X}(s) + \Sigma(s)P_t^\star(s).$$

Proof. Consider the two functions

$$\epsilon_1(s) = X_t^\star(s) - \widehat{X}(s),$$
$$\epsilon_2(s) = \Sigma(s)P_t^\star(s).$$

On the one hand, by subtracting the equation for $\widehat{X}$ to the equation for $X_t^\star$, ϵ^1 solves on $[0,t]$

$$\dot{\epsilon}_1 = (A - KH)\epsilon_1 - K\chi(\xi_t^\star).$$

On the other hand, differentiating ϵ_2 and using the definitions of $P_t^\star$ and Σ, ϵ_2 also solves

$$\begin{aligned}\dot{\epsilon}_2 &= \dot{\Sigma}P_t^\star + \Sigma\dot{P}_t^\star \\ &= [A\Sigma + \Sigma A^\top - \Sigma H^\top Q_\chi H\Sigma]P_t^\star - \Sigma[A^\top P_t^\star + H^\top Q_\chi \chi(\xi_t^\star)] \\ &= (A - KH)\epsilon_2 - K\chi(\xi_t^\star).\end{aligned}$$

Then, remarking also that $\epsilon_1(0) = \epsilon_2(0)$, we identify the two Cauchy problems, and therefore $\epsilon_1 = \epsilon_2$. □

Given the zero terminal condition on $P_t^\star$, we immediately derive the following property of equivalence between the optimal solution $X_t^\star$ and the independently defined solution $\widehat{X}$.

Corollary 2. *For all* $t \geq 0$,

$$X_t^\star(t) = \widehat{X}(t).$$

Hence, even if $X_t^\star$ is optimal under a $L^2(0,t)$-like norm, its final pointwise value also has a specific sense. Indeed, in a sequential point of view now, the solution $\widehat{X}(t)$ is optimal on $[0,t]$ in the sense that the system has learnt all the available information on this interval.

6.1.2 Extended Kalman observer and effect of POD reduction

We now come back to the nonlinear system (6.1)–(6.2). The above definition of the Kalman observer naturally extends to the nonlinear case, and still consists in the solution of a similar correction of the original system. The main difference lies in the dependence of the involved Jacobian matrices $\frac{\partial \mathcal{H}}{\partial X}(\cdot)$ and $\frac{\partial \mathcal{A}}{\partial X}(t,\cdot)$ (formerly H and $A(t)$) on the spatial variable. This implies that the resulting Kalman gain $K(t)$ and solution $\Sigma(t)$ of the Riccati equation become coupled with $\widehat{X}(t)$. Also, the equivalence with the variational approach is actually lost.

This process is named the *extended Kalman filtering* (EKF), and its associated observer $\widehat{X}(t) \in \mathbb{R}^{N_h}$ is solution of

$$\begin{aligned} \dot{\widehat{X}}(t) &= \mathcal{A}(t,\widehat{X}(t)) + R(t) + K(t)\hat{\chi}(t), \\ \widehat{X}(0) &= X_0, \end{aligned}$$

with the (nonlinear) innovation $\hat{\chi}(t) = Z(t) - \mathcal{H}(\widehat{X}(t))$, and the extended Kalman gain

$$K(t) = \Sigma(t)\,\frac{\partial \mathcal{H}}{\partial X}(\widehat{X}(t))^{\top} Q_{\chi},$$

determined by the coupled Riccati equation in $\mathbb{R}^{N_h \times N_h}$

$$\begin{aligned} \dot{\Sigma}(t) - \frac{\partial \mathcal{A}}{\partial X}(t,\widehat{X}(t))\,\Sigma(t) - \Sigma(t)\,\frac{\partial \mathcal{A}}{\partial X}(t,\widehat{X}(t))^{\top} \\ + \Sigma(t)\,\frac{\partial \mathcal{H}}{\partial X}(\widehat{X}(t))^{\top} Q_{\chi}\,\frac{\partial \mathcal{H}}{\partial X}(\widehat{X}(t))\,\Sigma(t) = 0, \\ \Sigma(0) = \mathrm{Cov}_{\xi}\,. \end{aligned}$$

Consider now a linearly independent family $(\varphi_1, \cdots, \varphi_l)$ of $\mathbb{R}^{N_h}$, the orthogonality of which we do not specify at the moment, that plays the role of a certain POD basis, and the associated matrix

$$\Phi^l = \begin{bmatrix} \varphi_1 & \cdots & \varphi_l \end{bmatrix} \in \mathbb{R}^{N_h \times l}.$$

This defines the reduced operator $\mathcal{A}^l(t,\cdot)$ and the reduced source term $R^l(t)$ as

$$\begin{aligned} \mathcal{A}^l(t,\beta^l) &= (\Phi^l)^{\top}\,\mathcal{A}(t,\Phi^l\beta^l), \quad \forall \beta^l \in \mathbb{R}^l, \\ R^l(t) &= (\Phi^l)^{\top} R(t). \end{aligned}$$

We also introduce the reduced a priori initial condition $X_0^l \in \mathbb{R}^l$, the definition of which depending on X_0 and the chosen orthogonality for the family $(\varphi_1, \cdots, \varphi_l)$. Finally, for the purpose of consistency [1], the unknown of the

1. Note the different context from Chapter 5, where we considered a state ξ belonging to a fixed modal subspace, and hence where no directly reduced form of the state appeared.

initial condition also becomes in reduced form $\xi^l \in \mathbb{R}^l$. The reduced-order model associated with (6.1)–(6.2) then writes as

$$\widetilde{X}^l(t;\xi^l) = \sum_{i=1}^{l} X^{l,(i)}(t;\xi^l)\varphi_i \in \mathbb{R}^{N_h},$$

with $X^l(\xi^l) = (X^{l,(1)}(\xi^l), \ldots, X^{l,(l)}(\xi^l))^\top \in \mathbb{R}^l$ solution of

$$\dot{X}^l(t;\xi^l) = \mathcal{A}^l(t, X^l(t;\xi^l)) + R^l(t),$$
$$X^l(0) = X_0^l + \xi^l.$$

Therefore, considering also the reduced observation operator

$$\mathcal{H}^l(\beta^l) = \mathcal{H}(\Phi^l\beta^l), \quad \forall \beta^l \in \mathbb{R}^l,$$

the reduced EKF observer $\widehat{X}^l(t)$ of $X^l(t)$ is solution of

$$\dot{\widehat{X}}^l(t) = \mathcal{A}^l(t, \widehat{X}^l(t)) + R^l(t) + K^l(t)\hat{\chi}^l(t),$$
$$\widehat{X}^l(0) = X_0^l,$$

with the reduced innovation $\hat{\chi}^l(t) = Z(t) - \mathcal{H}^l(\widehat{X}^l(t))$, and the reduced EKF gain

$$K^l(t) = \Sigma^l(t)\,\frac{\partial \mathcal{H}^l}{\partial X^l}(\widehat{X}^l(t))^\top Q_\chi,$$

determined by the coupled reduced Riccati equation in $\mathbb{R}^{l\times l}$

$$\dot{\Sigma}^l(t) - \frac{\partial \mathcal{A}^l}{\partial X^l}(t, \widehat{X}^l(t))\,\Sigma^l(t) - \Sigma^l(t)\,\frac{\partial \mathcal{A}^l}{\partial X^l}(t, \widehat{X}^l(t))^\top$$
$$+ \Sigma^l(t)\,\frac{\partial \mathcal{H}^l}{\partial X^l}(\widehat{X}^l(t))^\top Q_\chi \frac{\partial \mathcal{H}^l}{\partial X^l}(\widehat{X}^l(t))\,\Sigma^l(t) = 0,$$
$$\Sigma^l(0) = (\Phi^l)^\top \Sigma(0)\Phi^l.$$

Note the relations between Jacobian matrices, for all $\beta^l \in \mathbb{R}^l$,

$$\frac{\partial \mathcal{A}^l}{\partial X^l}(t,\beta^l) = (\Phi^l)^\top \frac{\partial \mathcal{A}}{\partial X}(t, \Phi^l\beta^l)\Phi^l,$$
$$\frac{\partial \mathcal{H}^l}{\partial X^l}(t,\beta^l) = (\Phi^l)^\top \frac{\partial \mathcal{H}}{\partial X}(t, \Phi^l\beta^l)\Phi^l.$$

6.2 POD reduction of an unscented Kalman filter observer

The *unscented Kalman filter* (UKF) is a method that derives from the KF with a more statistical approach. Several examples have shown its robustness with respect to the parameters it tries to estimate, and also in some

cases a better accuracy than the EKF [28]. It addresses some issues raised by the EKF, while remaining of the same order of computational cost.

Namely, first, the EKF needs the computation of the tangent operators (Jacobian matrices) $\frac{\partial \mathcal{A}}{\partial X}$ and $\frac{\partial \mathcal{H}}{\partial X}$ associated with the dynamic and observation operators, respectively. This inherently restricts its application to regular systems that effectively admit such Jacobian matrices, while some systems of interest might contain some irregularities, or even discrete variables. Moreover, whenever these matrices exist, the very computation of these tangent operators induces a major difficulty, since it requires the differentiation with respect to the state of complex implemented processes. Also, the operators $\mathcal{A}$ and $\mathcal{H}$ need to present some linearizations

$$\mathcal{A}(t,X) \approx \mathcal{A}(t,X_0) + \frac{\partial \mathcal{A}}{\partial X}(t,X_0)(X-X_0),$$
$$\mathcal{H}(X) \approx \mathcal{H}(X_0) + \frac{\partial \mathcal{H}}{\partial X}(X_0)(X-X_0),$$

that numerically hold, which is generally hard to analyze (see [30] and the corrections [29]). By contrast, the UKF avoids this linearization and applies to the Kalman filtering a method coming from statistics, fully detailed in [28], named *the unscented transform*, that offers a computationally light and high-order approximation of a nonlinearly transformed random vector.

We begin by rapidly describing the unscented transform. Then, we move on to the derived UKF, written for a fully discrete system. We justify the corresponding scheme and particularly its high-order consistence. Finally, we adapt the UKF scheme to Galerkin approximation reduction.

6.2.1 Statistical method of unscented transform

In order to fix the ideas, let θ be a random vector with values in the parameter (here *state-parameter*[2]) domain $\Theta \subset \mathbb{R}^q$, subject to a nonlinear operator $f : \Theta \to F$. Instead of approximating $f(\theta)$ using a series expansion of f, and then operating on the codomain of f, the unscented transform, operating on the domain of definition of f, introduces a few *sigma-points* $\theta_1, \cdots, \theta_{N_\sigma}$, which are deterministic, strategically placed points of Θ, built below.

Let $(\omega_i)_{i=1}^{N_\sigma} > 0$ some weights such that $\sum_{i=1}^{N_\sigma} \omega_i = 1$. We define the associated *empirical mean* by

$$\overline{(\nu_i)} = \sum_{i=1}^{N_\sigma} \omega_i \nu_i, \tag{6.3}$$

and, assuming that (ν_i) is the realization of a random variable, the associ-

2. Hence, the notation Θ differs from that of Chapter 4.

ated *empirical covariance matrix* by

$$\Sigma(\nu_i) = \sum_{i=1}^{N_\sigma} \omega_i \Big(\nu_i - \overline{(\nu_i)}\Big)\Big(\nu_i - \overline{(\nu_i)}\Big)^\top. \tag{6.4}$$

Let also some *particle directions* $(e_i)_{i=1}^{N_\sigma}$ be some fixed family of Θ verifying

$$\overline{(e_i)} = 0 \quad \text{and} \quad \Sigma(e_i) = \mathrm{Id}_q.$$

The sigma-points are then defined by

$$\theta_i = \mathbb{E}[\theta] + \sqrt{\mathrm{Cov}_\theta}\, e_i,$$

where $\sqrt{\mathrm{Cov}_\theta}$ can be any matrix square root of Cov_θ. Then we get the following proposition [31, III, Th. 1 & Appx. I].

Proposition 27.

$$\overline{(\theta_i)} = \mathbb{E}[\theta] \quad \textit{and} \quad \Sigma(\theta_i) = \mathrm{Cov}_\theta,$$

and also, provided sufficient regularity for f,

$$\overline{f(\theta_i)} \approx \mathbb{E}[f(\theta)] \quad \textit{and} \quad \Sigma(f(\theta_i)) \approx \mathrm{Cov}_{f(\theta)}$$

constitute second-order approximations with respect to $\Big|\theta - \overline{(\theta_i)}\Big|$.

We shall also interpret this transform from the viewpoint of the interpolation theory. Note the analogy with the interpolation operator L (4.9) defined in Chapter 4. Indeed, that operator also performed a high-order approximation of continuous functions by means of pointwise evaluations on special Lagrange grids, without any derivative computation on f in its definition.

6.2.2 Unscented Kalman filter prediction-correction scheme

Consider the time discretization of the system (6.1)–(6.2), according to a certain time scheme. We of course assume that there exists an implicit nonlinear function $\mathcal{F}$ (that shall not depend on ξ) such that, for each timestep,

$$X^{n+1}(\xi) = \mathcal{F}(t^n, X^n(\xi)),$$

with the initial condition

$$X^0 = X_0 + \xi.$$

We build the UKF observer $(\widehat{X}^n)$ by the following *prediction-correction* iterative process. Like the EKF being associated with the Riccati solution $\Sigma(t)$, the UKF also evolves with a dynamics (Σ^n), measuring the covariance of the estimation error $X^{n+1} - \widehat{X}^{n+1}$. When we apply the unscented

transform, we assume that the independently chosen weights $(\omega_i)_{i=1}^{N_\sigma}$ and particle directions $(e_i)_{i=1}^{N_\sigma}$ are fixed.

We start with the initial conditions

$$\widehat{X}^0 = X_0 \quad \text{and} \quad \Sigma^0 = \mathrm{Cov}(X^0 - \widehat{X}^0) = Q_\xi.$$

At the timestep t^n, $n \geq 0$, the current state observer $\widehat{X}^n$ and associated covariance Σ^n are known, and the measurement Z^n has already been treated. Given the next piece of information Z^{n+1} available at t^{n+1}, we define $(\widehat{X}^{n+1}, \Sigma^{n+1})$ using the notion of *best linear unbiased estimator* (BLUE). It linearly defines an observer $\widehat{X}^{n+1}$ such that

$$\begin{aligned} \mathbb{E}[\widehat{X}^{n+1}] &= \mathbb{E}[X^{n+1}], \\ \widehat{X}^{n+1} - \mathbb{E}[\widehat{X}^{n+1}] &= \Lambda(Z^{n+1} - \mathbb{E}[Z^{n+1}]), \end{aligned}$$

where the matrix $\Lambda \in \mathbb{R}^{N_h \times N_{\mathrm{obs}}}$ is determined by the minimizer of

$$\min_{\widetilde{\Lambda}} \mathbb{E}\Big[\big|X^{n+1} - \mathbb{E}[X^{n+1}] - \widetilde{\Lambda}(Z^{n+1} - \mathbb{E}[Z^{n+1}])\big|^2\Big].$$

If $\mathrm{Cov}\, Z$ is invertible, which we assume, the solution is unique and given by

$$\Lambda = \mathrm{Cov}(X^{n+1}, Z^{n+1})(\mathrm{Cov}\, Z^{n+1})^{-1},$$

so that

$$\widehat{X}^{n+1} = \mathbb{E}[X^{n+1}] + \mathrm{Cov}(X^{n+1}, Z^{n+1})(\mathrm{Cov}\, Z^{n+1})^{-1}(Z^{n+1} - \mathbb{E}[Z^{n+1}]). \tag{6.5}$$

Furthermore, assuming independence between the uncertainty ξ^{n+1} of the observation noise χ^{n+1}, we can verify that, in this case,

$$\begin{aligned} \Sigma^{n+1} = \;& \mathrm{Cov}(X^{n+1} - \widehat{X}^{n+1}) \\ =\; & \mathrm{Cov}\, X^{n+1} \\ & - \mathrm{Cov}(X^{n+1}, Z^{n+1})(\mathrm{Cov}\, Z^{n+1})^{-1}\, \mathrm{Cov}(X^{n+1}, Z^{n+1})^\top. \end{aligned} \tag{6.6}$$

Hence, the method naturally falls into two parts. The first one *predicts* $\widehat{X}^{n+1}$ and Σ^{n+1} by simply propagating the probabilistic information associated with $(\widehat{X}^n, \Sigma^n)$ through the dynamics. The second one *corrects* these values with the additive terms appearing in (6.5) and (6.6).

Prediction

In order to approach the corresponding expectation and covariance matrix, we apply the unscented transform on

$$X^{n+1} = \mathcal{F}(t^n, X^n).$$

Since $\widehat{X}^n$ and Σ^n are assumed to provide good approximations of the mean and covariance of X^n (given all the observations up to Z^n), introducing the independently computed particles

$$X_i^{n+1|n} = \mathcal{F}(t^{n+1}, \widehat{X}^n + \sqrt{\Sigma^n} e_i), \quad 1 \le i \le N_\sigma,$$

and using Prop. 27, we define the *predictors*

$$\begin{aligned} X^{n+1|n} &= \overline{(X_i^{n+1|n})} \quad &\approx \mathbb{E}[X^{n+1}], \\ \Sigma^{n+1|n} &= \Sigma(X_i^{n+1|n}) \quad &\approx \mathrm{Cov}(X^{n+1}), \end{aligned}$$

where the approximations are of second-order.

Correction

We now need to approximate the expectation of measurement

$$\mathbb{E}[Z^{n+1}] = \mathbb{E}[\mathcal{H} \circ \mathcal{F}(t^n, X^n)],$$

and also the crossed state-measurement covariance and measurement covariance matrices. Using the same assumption of independence,

$$\begin{aligned} \mathrm{Cov}(X^{n+1}, \chi^{n+1}) &= 0, \\ \mathrm{Cov}\, Z^{n+1} &= \mathrm{Cov}\, \mathcal{H}(X^{n+1}) + Q_\chi, \end{aligned}$$

so that

$$\begin{aligned} \mathrm{Cov}(X^{n+1}, Z^{n+1}) &= \mathbb{E}\Big[(\mathcal{F}(X^n) - \mathbb{E}[\mathcal{F}(X^n)]) \cdot \Big(\mathcal{H} \circ \mathcal{F}(X^n) - E[(\mathcal{H} \circ \mathcal{F}(X^n))\Big)^\top\Big], \\ \mathrm{Cov}\, Z^{n+1} &= \mathrm{Cov}\, \mathcal{H} \circ \mathcal{F}(X^n) + Q_\chi, \end{aligned}$$

Hence, introducing the *zero-noise measurement* particles

$$Z_{0,i}^{n+1} = \mathcal{H}(X_i^{n+1|n}) = \mathcal{H} \circ \mathcal{F}(t^n, X^n + \sqrt{\Sigma^n} e_i), \quad 1 \le i \le N_\sigma, \tag{6.7}$$

we define the intermediate second-order approximations

$$\begin{aligned} Z_0^{n+1} &= \overline{(Z_{0,i}^{n+1})} \quad &\approx \mathbb{E}[Z^{n+1}], \\ \Sigma_{XZ}^{n+1} &= \sum_{i=1}^{N_\sigma} \omega_i (X_i^{n+1|n} - X^{n+1|n})(Z_{0,i}^{n+1} - Z_0^{n+1})^\top \quad &\approx \mathrm{Cov}(X^{n+1}, Z^{n+1}), \\ \Sigma_Z^{n+1} &= \Sigma(Z_{0,i}^{n+1}) + Q_\chi \quad &\approx \mathrm{Cov}\, Z^{n+1}, \end{aligned}$$

and, finally, the new state observer $\widehat{X}^{n+1}$ and associated covariance Σ^{n+1} by

$$\begin{aligned} \widehat{X}^{n+1} &= X^{n+1|n} - \Sigma_{XZ}^{n+1} (\Sigma_Z^{n+1})^{-1} (Z^{n+1} - Z_0^{n+1}), \\ \Sigma^{n+1} &= \Sigma^{n+1|n} - \Sigma_{XZ}^{n+1} (\Sigma_Z^{n+1})^{-1} \Sigma_{XZ}^{n+1}. \end{aligned}$$

6.2.3 POD reduced version

Following exactly the same path, we may define the reduced-order version of this unscented Kalman filter for the reduced discrete system

$$X^{l,n+1}(\xi^l) = \mathcal{F}^l(t^n, X^{l,n}(\xi^l)),$$
$$X^{l,0} = X_0^l + \xi^l,$$

where $\mathcal{F}^l$ is defined with $\mathcal{F}$ and the chosen POD basis $(\varphi_1, \ldots, \varphi_l)$. We display below the corresponding algorithm, the justifications being analogous to those for (X^n).

Initial condition

$$\widehat{X}^{l,0} = X_0^l \quad \text{and} \quad \Sigma^{l,0} = (\Phi^l)^\top \operatorname{Cov}_\xi \Phi^l.$$

Prediction

Independent particles

$$X_i^{l,n+1} = \mathcal{F}^l(t^n, \widehat{X}^{l,n} + \sqrt{\Sigma^{l,n}}\, e_i), \quad 1 \le i \le N_\sigma.$$

Predictors

$$X^{l,n+1|n} = \overline{(X_i^{l,n+1|n})},$$
$$\Sigma^{l,n+1|n} = \Sigma(X_i^{l,n+1|n}).$$

Correction

Independent particles

$$Z_{0,i}^{l,n+1} = \mathcal{H}^l(X_i^{l,n+1}), \quad 1 \le i \le N_\sigma.$$

Intermediate expectation and covariance approximations

$$Z_0^{l,n+1} = \overline{(Z_{0,i}^{l,n+1})},$$
$$\Sigma_Z^{l,n+1} = \Sigma(Z_{0,i}^{l,n+1}) + Q_\chi,$$
$$\Sigma_{XZ}^{l,n+1} = \sum_{i=1}^{N_\sigma} \omega_i (X_i^{l,n+1|n} - X^{l,n+1|n})(Z_{0,i}^{l,n+1} - Z_0^{l,n+1})^\top.$$

Update for the state observer and covariance

$$\widehat{X}^{l,n+1} = X^{l,n+1|n} + \Sigma_{XZ}^{l,n+1}(\Sigma_Z^{l,n+1})^{-1}(Z^{n+1} - Z_0^{l,n+1}),$$
$$\Sigma^{l,n+1} = \Sigma^{l,n+1|n} - \Sigma_{XZ}^{l,n+1}(\Sigma_Z^{l,n+1})^{-1}(\Sigma_{XZ}^{l,n+1})^\top.$$

6.3 POD reduced parameter estimation on an electromechanical heart model for assessing an infarct

In this section, we come back to the electromechanical heart model introduced in Section 3.5. Regarding both the microscopic, biochemistry-based constitutive law and the macroscopic, large displacement elastodynamic model, we presented the essential ideas, and referred to the corresponding papers for further details. Using the reduced UKF algorithms developed and justified in the previous section, we display and discuss the results of a specific *contractility parameter* estimation problem, performed by confronting the heart model to real clinical data collected on a pig heart. We briefly summarize below the needed model-data interaction involved to prepare the estimation, and we redirect to [10] for a more extensive technical overview.

We recall that, while we showed that the UKF algorithm features a certain mathematical consistence, it relies on several approximations. Hence, we shall keep in mind that the consistency of these filters essentially becomes in this context of heuristic nature.

Nevertheless, a successful estimation *with non-reduced UKF for parameters* and *a Luenberger filter for the state* [36] has been performed for a seven-dimensional parameter and a large number of degrees of freedom, as presented in [10], see also [41]. Indeed, due to the computational cost of manipulating full covariance matrices, the use of a full UKF for the joint parameter-state vector is intractable. This is why the lighter Luenberger filter is employed for the state. Of course, the assessment of a "successful" estimation has no meaning as a comparison between the exact parameter values and the estimated ones, because in this case, the exact ones are concretely unknown. Instead, it means that, compared to a manual *a priori* calibration of the model, the *a posteriori* retrieved parameter improved the fitting of the direct simulations to the observed data, also the estimator is more consistent than the initial direct model with various other source of information (data not used in estimation, medical knowledge). Therefore, this experimental investigation constitutes, as a first validation step, an illustration of a *patient-specific* modelling goal.

By contrast here, the application of a UKF on the vector formed by the parameter and the reduced-order state becomes feasible. Based on the effective quality of the estimation explained above, even though it involves a different filtering strategy, we again try to focus on the effect of POD reduction in the accuracy of the resulting reduced-order observer, for this intricate real-life inverse problem.

We begin by presenting the main steps of the experimental protocol and its integration in the model, raising the inherent difficulties of the communication between the two processes. Then, we show some reduction results

corresponding to a two-dimensional parameter UKF estimation. In particular, we focus on the reduction error in displacement, which is directly analogous to the exhibited reduction errors of the previous chapters, and that related to the evolution in time of the parameter observer $\widehat{\sigma}$.

6.3.1 Contractility field and model adaptation with the experimental process

Following the mathematical scope and modest medical content of this report, we shall simply limit the description of the clinical case to the necessary general key points for our study.

With a division of the heart into a finite number of regions, we aim at estimating the contractility values, assumed to be uniform per region and constant with respect to time, associated with the mechanical active stress field σ_c, that derives from the myofiber constitutive law. With regard to the human heart, this contractility field, due to differences in age, gender, corpulence, physical activity and personal medical history, significantly varies from one patient to another. Indeed, it reflects the ability of the heart to eject blood at each beat, and then determines the general cardiac function. In particular, a localized low contractility may characterize an infarcted zone on the cardiac tissue. For these reasons, whereas one cannot directly access the value of this mechanical property, it turns out to be a major factor with regard to patient-specific models. Hence, the accurate estimation of the contractility field appears as an essential contribution to that end.

With this patient-specific aim in mind, and as a piece of validation for the developed heart model, an ethically approved clinical trial has been conducted on a farm pig. Initially with a healthy heart, the pig had an infarct created by an artificial temporary localized coronary occlusion during two hours, hence damaging the properties of the corresponding zone. Then, after 38 days of evolution, a modification of the mechanical property of the tissue and a change of its geometry, due to its remodelling, was observed. The experimentalists then took some measurements at this stage. Then, without the use of a reduced-order model strategy, a simulation of the observers was run with these data in order to verify that they were able to detect the infarcted zone [10].

Independently of the chosen method, this simply formulated idea of estimation actually requires several technical calibration steps to ensure, before the considerations of accuracy and robustness for the estimation method, the very feasibility of the computations.

First, the estimation process should be initiated with a geometry, represented by three-dimensional mesh, that matches as well as possible the observed one. This constitutes a *compatibility* problem. We detail the necessary preparation from the two standpoints:

– with regard to the observations, the measurements needs to be geometrically compatible. They are performed by MRI snapshots around the heart region that produce, at each timestep, $5 \cdot 10^5$ voxels. The resulting three-dimensional matrix is then subject to a segmentation to detect the surface of the heart, from which a mesh can be built using common reconstruction tools;
– with regard to the model, that mesh needs to be physically acceptable with the mechanical model. Since the solution to the mechanical equations tends in time to a certain periodic attractor, we run a first stabilization cardiac cycle with the reconstructed mesh, to cut out the transient phenomena due to small aberrant geometric approximations, before using it for estimation.

Also, as an analogy with forced harmonic oscillators, the electrical activation, which is modelled separately and forms the input of the constitutive law, needs to be finely calibrated so that the cardiac cycles run synchronously with the observations.

6.3.2 Numerical assessment of the reduction error on the UKF state-parameter observers

We present the results for the reduction of the UKF observer that aims at detecting the contractility field corresponding to the infarcted pig heart after 38 days. We group the major constants determining the simulations in Tab. 6.1, and detail their justification below.

Numerical configuration and multi-POD construction

We refer to [10] for considerations corresponding to the spatial and time discretization of the model, but also for the nonlinear observation model represented by $\mathcal{H}$ in the previous analytic developments. Yet, we need to specify that

– first, since the elastodynamic model is of second-order in time, the reduced solution X^l rewritten in the generic form (6.1)–(6.2) actually decomposes as

$$X^l = \begin{bmatrix} X^l_{\text{state}} \\ \sigma \end{bmatrix},$$

with X^l_{state} concatenating:

1. the reduced displacement Y^l;
2. the reduced velocity $\dot{Y}^l$;
3. some additional variables, in particular the pressure values in the various cavities. The *reduction of uncertainties* indicated in

the introduction and referred to as *"reduced filtering"* in the literature, is performed for these variables, see [40] and also [39] with corrections [38];

and with σ following a null dynamic (an artifical point of view that canonically defines the observer $\widehat{\sigma}$). Nevertheless, we only collect displacement measurements through the MRI images, so that only the Y part is concerned by the observation model;

– secondly, the innovation term is slightly changed and replaced by a similar *discrepancy operator*, that differently assesses the distance between the measurements Z and the observer $\widehat{X}$ in the correction steps [41].

We now discuss the other constant values. First, the set of measurements (Z^n) was collected with a frequency limited by the MRI acquisition equipment, corresponding to an observation timestep of $\Delta t_{\text{obs}} = 26$ ms. We need by contrast to set the simulation timestep for the UKF observer to a much lower value, and in our case to $\Delta t_{\text{sim}} = 1$ ms. The simulations are run on a physical interval $[0, T]$ with $T = 600$ ms, corresponding to a full cardiac cycle.

With regard to the parameter subdomain and as a preliminary trial (recall that our method of multi-POD reduction depends on the dimension of the parameter value), we limited the division of the heart into a number of two regions of homogeneous contractility, based on the segmentation of the infarcted region as seen in the *late enhancement IRM*. Namely, it consists in a *healthy* region Ω_1, and the complement, *infarcted* one Ω_2, forming a contractility vector $\sigma = (\sigma_1, \sigma_2) \in \Theta = \mathbb{R}^p$ with a dimension $p = 2$. Then, we explain the choice for the parametric subdomain $\mathcal{D}$, that defines the likely variability region of the contractility vector σ, by some nominal values of contractility, namely, in normalized nondimensional unit

– $\sigma_{1,\text{ref}} = 1$ for a healthy region;
– and $\sigma_{2,\text{ref}} = 0.25$ for an infarcted one.

Also, we need to define the initial condition Σ^0 for the dynamics (Σ^n) of the covariance matrix, that, assuming the independence of uncertainties in initial state and that in parameter, we decompose Σ^0 by block by

$$\Sigma^0 = \begin{bmatrix} \Sigma^0_{\text{state}} & 0 \\ 0 & \Sigma^0_\sigma \end{bmatrix},$$

with square matrices Σ^0_{state} and Σ^0_σ of orders corresponding to the sizes of X^l_{state} and σ, respectively. On the one hand, the low value put in the state (that one should interpret as a distance by taking the square root) reflects the *strong confidence* that we put in X^l_0, which is legitimate because of the pre-calibration, namely the stabilization of the mesh and the synchronization of the activation. Indeed, the direct full cardiac cycle run before use of the mesh causes indeed makes the error in state less detectable, and of

Discretization	
N_h	$2.4 \cdot 10^4$
$N_{\Delta t_{\text{sim}}}$	600
Δt_{sim}	$1.0 \cdot 10^{-3}$
Time scheme	cen. Newmark

Measurements	
N_{obs}	$1.1 \cdot 10^3$
$N_{\Delta t_{\text{obs}}}$	24
Δt_{obs}	$2.6 \cdot 10^{-2}$

Parameter subdomain (for σ)	
p	2
$\mathcal{D}$	$[0.75, 1.25] \times [0, 1]$
s_{POD}	1

Initial dynamic covariance	
Σ^0_{state}	$1 \cdot 10^{-7}\ \text{Id}_{N_h}$
Σ^0_{σ}	$2\ \text{Id}_p$

Table 6.1: Set of constants for the reduced UKF estimation problem using clinical measurement data

reduced influence. On the other hand, the relatively large value used for the σ part induces large variability tolerance for the observer $\widehat{\sigma}$.

Finally, for the multi-POD construction, we use the scalar product defined by the stiffness matrix K_{stiff} attached to the initial reference mesh configuration. Also, we restrict the grid to a degree $s_{\text{POD}} = 1$, since it implied convincing reduction errors values for the variational estimations on Chapter 5. However, by contrast with those previous one-dimensional numerical experiments, the storage of the snapshots matrices here requires some further limitations. Then, we only stored 12 snapshots, namely forming the set of snapshot timesteps

$$J_{\text{snap}} = \{50j \ ; \ 0 \leq j \leq 11\},$$

so that the (alternate) covariance matrix to be diagonalized, while costly to assemble, is only of order 48. According to the criterion (3.33), the obtained maximum rank is 44. We also display the corresponding multi-POD remainder decrease in Fig. 6.1.

POD reduction results for the UKF estimation with low covariance in the state

According to Fig. 6.1, given the very low $O(10^{-6})$ value of relative POD remainder for $l_{\text{ref}} = 44$ modes, we assume that the corresponding reduced UKF state-parameter observer reaches a satisfactory *numerical convergence* with respect to the POD rank. This is why we assume its role of *reference reduced solution*, with displacement observer and parameter observer

$$(\widehat{Y}^n_{\text{ref}}, \widehat{\sigma}^n_{\text{ref}}) \equiv (\Phi^l \widehat{Y}^{l_{\text{ref}},n}, \widehat{\sigma}^{l_{\text{ref}},n}) \quad \in \mathbb{R}^{N_h} \times p, \quad \forall 0 \leq n \leq N_{\Delta t_{\text{sim}}} - 1,$$

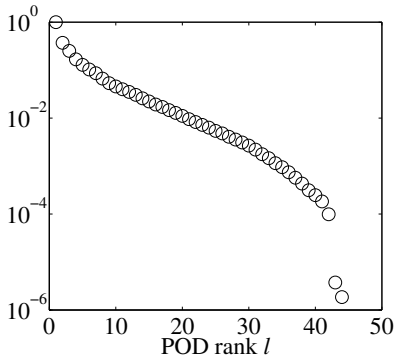

Figure 6.1: Relative multi-POD remainder decrease

to which we shall compare the other observers reduced with lower POD ranks.

For a first study of the effect of reduction, we conducted the reduced estimations with the following POD ranks:

- $l_{\text{ref}} = 44$ for the reference reduced one;
- $l = 6$, 12, 20 and 32 for the other ones.

We shall qualify the corresponding results as *pseudo-errors* ones, because of the lack of non-reduced matching reference. We group the reduction results in displacement in Fig. 6.2, where:

- $\epsilon(l)$ is the multi-POD remainder;
- $\|\widehat{y_{\text{ref}}}\|$ is an approximated $L^2(0,T;V)$ value

$$\|\widehat{y_{\text{ref}}}\| = \Big\{ \sum_{j\in J_{\text{snap}}} |Y^j_{\text{ref}}|^2_K \Big\}^{1/2} \Delta t;$$

- $P(l)$ and $R(l)$ are pseudo-projection and pseudo-reduction errors, i.e.

$$P(l) = \Big\{ \sum_{j\in J_{\text{snap}}} |Y^j_{\text{ref}} - \Pi^l Y^j_{\text{ref}}|^2_K \Big\}^{1/2} \Delta t,$$

$$R(l) = \Big\{ \sum_{j\in J_{\text{snap}}} |Y^j_{\text{ref}} - \Phi^l Y^{l,j}|^2_K \Big\}^{1/2} \Delta t.$$

We observe that the reduced observer with 32 modes reaches a satisfactory performance of 4 % relative pseudo-reduction error. We also note that the projection and reduction errors are very close albeit substantially larger than $\epsilon(l)$, which may be attributed to the coarseness of the parameter domain.

In order to visualize the evolution of *the distance between the observed IRM data and the POD reduced models*, we compare them for a representative cross-section and for regular set of timesteps in Fig. 6.4. The curves

l		6	12	20	32
$\eta_{\widehat{\sigma_1}}(T)$	(%)	9.9	5.7	5.4	2.6
$\eta_{\widehat{\sigma_1}}(T)$	(%)	30.7	0.6	4.4	3.6

Table 6.2: Pseudo-reduction errors for cardiac cycle end observers

correspond to the contour of the heart model deformed with the estimated displacements $\widehat{y}^l$, for $l = 12$, 32 and 44. First of all, we underline that all the observers $\widehat{y}^l$ run synchronously with the cardiac cycle, and follow quite closely the ventricle contours. Better accuracy is achieved though for the left ventricle, corresponding to the main cavity of roughly circular cross-section. This actually occurs because *the data reconstruction had actually been performed for this ventricle only.* In particular, this means that the estimation is performed without any actual data on the right ventricle, corresponding to the small half-ellipse part. This then shows the robustness of the heart model and predictive character of the estimation, since the estimated displacement field for the right ventricle, relying on the model only, remains very acceptable.

With regard to the convergence with respect to the POD rank now, indeed, one can hardly distinguish the POD reduced displacement observers for $l = 32$ and $l = 44$ on Fig. 6.4, which is confirmed by the pseudo-reduction error given above. We provide some zooms of the frame for $t = 400$ ms, near the end of systole (contraction), in Fig. 6.5, as an evidence of their most distant position from each other.

Also, as this study is motivated by the estimation of the contractility values for the two regions Ω_1 and Ω_2, we display the relative pseudo-reduction errors in observers $\widehat{\sigma_1}$ and $\widehat{\sigma_2}$ in Fig. 6.3, where one should read $j \in \{1, 2\}$,

$$\eta^l_{\widehat{\sigma_j}}(t) = \frac{|\widehat{\sigma}_{\mathrm{ref},j}(t) - \widehat{\sigma}^l_j(t)|}{\|\widehat{\sigma}_{\mathrm{ref},j}\|_{C^0(0,T)}}.$$

Note that these trajectories can be easily stored at Δt_{sim} resolution. We notice that on the one hand, the relative pseudo-reduction error of the healthy region observer significantly decreases by one order of magnitude between $l = 6$ and $l = 32$, which is also same order observed in Fig. 6.2 for the pseudo-reduction error in displacement. On the other hand, the analogous pseudo-error for the infarcted region does not have a clear behaviour, since the timeline of $\widehat{\sigma}_2^{12}$ seems to match much more accurately that of $\widehat{\sigma}_2^{44}$ (i.e. $\widehat{\sigma}_{\mathrm{ref},2}$) than that of the intermediate solutions $\widehat{\sigma}_2^{20}$ and $\widehat{\sigma}_2^{32}$. However, the same pseudo-errors in final cardiac cycle, i.e. at time T, are globally as satisfying for the two regions, see Tab. 6.2, if we assume that estimating the contractility with two significant digits is sufficient.

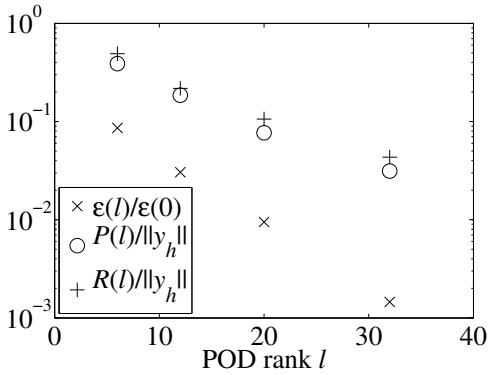

Figure 6.2: Pseudo-relative reduction error in displacement

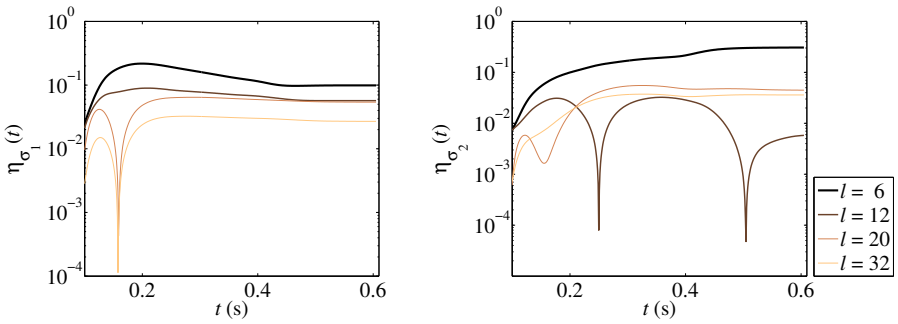

Figure 6.3: Timelines of relative pseudo-reduction errors associated with the two contractility values

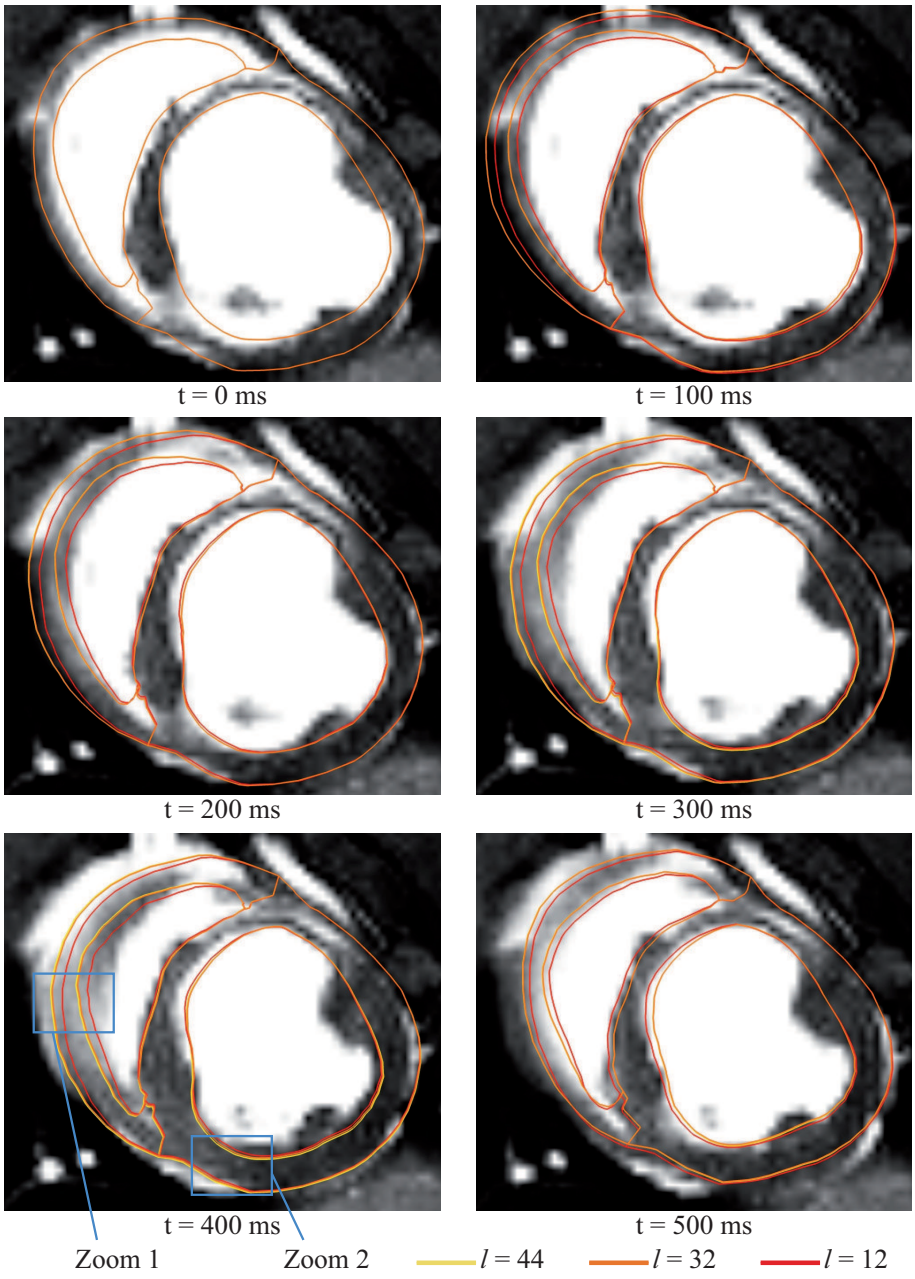

Figure 6.4: Horizontal slice of IRM images for the heart, periodically collected during one beat

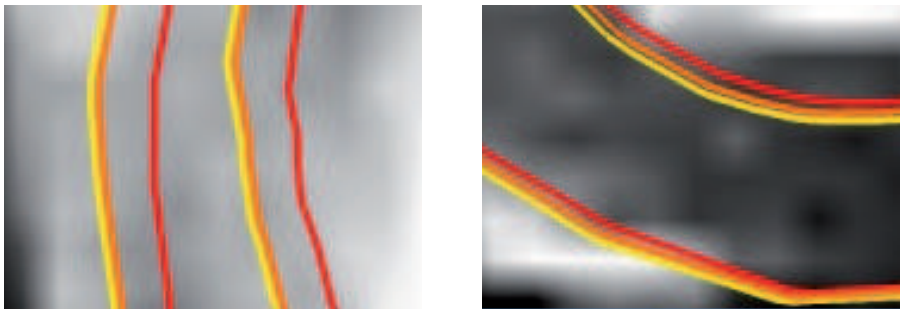

Figure 6.5: Zooms of the frame for $t = 400$ ms in Fig. 6.4

6.4 Conclusion

This chapter was driven by the objective of testing a generic multi-POD reduction method developed in Chapter 4 on a concrete estimation procedure using some clinical data, and in particular, detect some anatomically distinguishable infarcted zone in a heart. It encompassed some filtering methods that were regularly studied in the estimation community. We progressively introduced some Kalman filtering-based observers and adapted them to this reduction method.

First, we developed a standalone construction for a general linear system of the Kalman filter, core of a significant range of filtering methods. For the sake of simplicity in this applied chapter, we avoided the yet fundamental stochastic foundations of this filter. In particular, we showed the classical equivalence link in this linear case with the variational estimation methods, that were the main matter of Chapter 5. Then, this enabled us to derive the immediate extension to the nonlinear case with the computationally heavy, regarding both storage and manipulation, extended Kalman filter. Hence, this context was suitable for the application of our multi-POD reduction, and we proposed a generic way, still based on the Galerkin approximation, to formulate the reduced EKF observer.

However, the EKF has been debated for forty years and unadvised for many applications, since the system linearization on which it relies may lead to uncontrolled, poor quality observers [30]. We hence detailed and justified the alternative unscented Kalman filter, that transfers the main assumption on the probability distribution, with respect to the system uncertainties, of the solution, and leaves, like an interpolation operator, the possibly nonlinear dynamics and observation operators untouched. Yet, in practice the UKF hardly addresses the computation cost issues of the EKF and remains as is intractable for large systems. Then, again, we adapted the UKF observer, written specifically for a time-discretized system and solving at each step a particular prediction-correction scheme, for a reduced

version.

Finally, favoring the reduced UKF algorithms and leaving the reduced EKF ones for a possible future comparison, we tackled the concrete estimation of a not directly observable contractility field in a living heart. With a view to detecting cardiac pathologies in a systematic way, some MRI observations were performed on this heart, that was purposely placed in a local infarcted state. The ability of a non-reduced Luenberger filtering model on the state, with no clue on the concerned region, to localize the infarct has already been shown in [10]. For this report, by reducing the heart model with a multi-POD on a likely variability domain of the contractility field of dimension 2, we performed a full UKF on the joint state-parameter vector. By comparison with a *numerically converged* (with respect to the POD rank) reduced parameter observer, the reduction by a multi-POD basis of size 32, in comparison with the large $O(10^5)$ number of degrees of freedom, has shown its efficiency, since the final resulting estimated values of contractility were numerically converged up to a relative $O(10^{-2})$ error.

This encouraging reduction result may make us move towards an extended and more targeted reduction method, with e.g. the new aim of a geometrical division of the heart advised by the AHA into 17 regions [8]. However, the number of direct non-reduced simulations needed for the offline multi-POD construction exponentially grows with respect to the dimension of the parameter of interest, and hence constitutes a first limitation. Also, by contrast to the very satisfying results of Chapter 5 for a variational parameter estimation applied to the complex FitzHugh–Nagumo equations, no obvious numerical control of the maximum projection error, and *a fortiori* of the maximum reduction error, by the multi-POD remainder appears here (see Fig. 6.2). This brings crucial questions of robustness concerning the UKF, and also the multi-POD, e.g. for a finer description of the parameter space, that might still need to be investigated.

Conclusion

Ce rapport a principalement poursuivi deux objectifs, complémentaires du point de vue de l'analyse numérique, autour de la méthode de réduction par *proper orthogonal decomposition*. Premièrement, on a montré des estimations d'erreur nouvelles et performantes pour des problèmes non-linéaires. Deuxièmement, on a formalisé et analysé une méthode de réduction qui étend le principe de réduction par POD, en vue de son application pratique et efficace. On a ensuite numériquement validé l'application de cette méthode étendue sur une série de modèles biomathématiques, notamment pour la résolution de problèmes inverses.

Par rapport aux estimations d'erreur existant dans la littérature, celles qu'on a proposées, pour des problèmes paraboliques sous-linéaires et pour une équation des ondes abstraite, en modifient deux aspects. D'une part, elles suppriment dans le membre de droite le terme d'erreur de projection POD des dérivées en temps les plus hautes des équations. D'autre part, elles multiplient le reste POD par une suite de normes particulières d'opérateurs de projection. Puisqu'on a constaté que cette suite est numériquement bornée, et puisqu'on apporte des résultats abstraits qui confortent cette hypothèse, on justifie qu'il n'est pas nécessaire d'intégrer des informations de dérivée en temps dans la construction des espaces POD.

Ensuite, on a écrit une extension de la réduction par POD à la situation de dépendance paramétrique. On rappelle que, pour des raisons de coût numérique, la nécessité d'évaluer une solution pour plusieurs valeurs de paramètres peut constituer un facteur limitant. Cette méthode, de réduction par *multi-POD*, calcule un espace POD uniforme pour tout un sous-domaine paramétrique $\mathcal{D}$, en faisant appel pour cela à un nombre restreint de résolutions complètes. Dans un premier temps, on a adapté les estimations d'erreur précédentes pour celle-ci. Ainsi on contrôle l'erreur maximale de réduction sur $\mathcal{D}$ par un *reste multi-POD*, accessible par construction et à rapide décroissance.

On a validé l'efficacité de la réduction par multi-POD, sa performance significativement meilleure que la méthode de réduction par POD standard, ainsi que la rigueur des estimations d'erreur précédentes, sur plusieurs exemples numériques complexes.

Tout d'abord, on a choisi d'illustrer ces propriétés sur le système de

FitzHugh–Nagumo, qui modélise en électrophysiologie le phénomène de potentiel d'action, et qui montre par essence une grande sensibilité paramétrique. Pour les fenêtres les plus génériques de variation de paramètres, qui s'éloignent des effets de seuil dus au potentiel d'action, on observe rapidement de faibles erreurs avec des rangs POD raisonnables, tant pour le contrôle indiqué par les estimations d'erreur, qui ne couvrent pas ce cas surlinéaire, que pour les erreurs de réduction elles-mêmes. Ces résultats témoignent alors de la stabilité de la méthode de réduction par multi-POD. En effet, par des considérations d'interpolation sur l'espace des paramètres, on s'aperçoit que les projecteurs multi-POD prennent en compte de manière sous-jacente des informations sur les dérivées des solutions par rapport au paramètre.

Ensuite, pour ces mêmes équations de FitzHugh–Nagumo, et en introduisant cette fois une incertitude grande sur le coefficient de diffusion, ainsi qu'une incertitude faible sur la condition initiale, on a voulu confronter la méthode à la résolution d'un problème inverse. Muni d'observations partielles et bruitées d'une solution donnée, on cherche à retrouver le coefficient de diffusion correspondant. On a choisi pour cette partie une approche variationnelle, c'est-à-dire par minimisation de fonction coût. Si la formulation mathématique du problème se prête assez naturellement aux méthodes de réduction, l'obtention dans ce cadre d'estimations d'erreur semble en revanche beaucoup plus délicate. On a donc adopté une approche essentiellement empirique ici. Les constatations sont similaires à ceux du problème direct, dans le sens où l'erreur entre les paramètres estimés avec la réduction par multi-POD et ceux estimés sans réduction converge rapidement vers zéro. Pour appuyer encore la qualité de la méthode, on observe au contraire des résultats médiocres de convergence en rang POD avec une technique POD standard.

Pour une plus large possibilité de recouvrement de l'espace des paramètres, ces résultats ouvrent la voie à des techniques de maillage de l'espace paramétrique, pour lequel on munirait chaque maille d'un projecteur multi-POD calculé sur celui-ci. Cela pourrait permettre à un algorithme de minimisation tel qu'on l'a employé dans cette partie d'utiliser, pour chaque maille rencontrée au cours des itérations, un projecteur mieux adapté.

Enfin, on a confronté la réduction par multi-POD à un modèle mécanique de cœur, complexe et très non-linéaire, sur un problème particulier d'estimation de paramètre par filtrage. On prend la situation où on veut retrouver un champ de contractilité musculaire depuis la donnée d'images IRM successives, par essence très bruitées, d'un cœur, afin de retrouver une zone infarcie. Le filtre utilisé, "*unscented Kalman filter*", ne nécessite pas de linéarisation du système et repose sur une méthode statistique de propagation de moyennes d'état et de covariances d'incertitude. Également ici, les résultats de convergence en rang POD des champs de contractilité estimés et réduits par multi-POD, ainsi que ceux des dynamiques d'observateurs de

déplacement, sont très satisfaisants. En effet, bien que le reste multi-POD contrôle moins bien l'erreur de réduction ici que dans les tests précédents, une trentaine de modes POD suffit à discriminer sur un cœur une région saine d'une région infarcie. L'erreur relative par rapport à une solution de référence, numériquement convergée, est de l'ordre de quelques %. Étant donnée la complexité des considérations multi-échelles du modèle, de sa géométrie et du nombre important de variables internes dynamiques comprises, la multi-POD permet au final de réaliser une prouesse numérique de réduction.

Outre les questions de description médicale plus fine et propres au modèle de cœur, la même proposition de maillage paramétrique formulée plus haut constitue une perspective d'application dans ce cadre aussi. De plus, comme les techniques de filtrage définissent des systèmes dynamiques d'observateurs de même nature que les systèmes originaux, on peut envisager et essayer d'étendre les estimations d'erreurs précédentes pour les cas simples de filtrage réduit par POD.

Appendix A

Existence and uniqueness of solutions of variational equations with a Lipschitz continuous reaction term

Although some results pertaining to this type of problem exist in the literature, for the sake of completeness we provide the sketch of a self-contained proof for the specific result that we need in our case, namely, Proposition 7.

Since we assume the embedding $V \hookrightarrow H$ to be compact, we can use the Hilbertian bases (w_i) and $(\tilde{w}_i)$ of H and V – respectively – made up by the eigenvectors, as already introduced in Section 3.3.5. Let W_k, $k \geq 1$, denote the subspace

$$W_k = \mathrm{Span}(w_1, \ldots, w_k) = \mathrm{Span}(\tilde{w}_1, \ldots, \tilde{w}_k),$$

and P_k the orthogonal projector from H onto W_k, i.e.

$$P_k h = \sum_{i=1}^{k} (h, w_i) w_i, \quad \forall h \in H.$$

It coincides with the orthogonal projector from (V, a) onto W_k defined by

$$P_k^a v = \sum_{i=1}^{k} a(v, \tilde{w}_i) \tilde{w}_i = P_k v, \quad \forall v \in V.$$

We will show the existence result by a Galerkin approach using the sequence of eigenspaces.

Proposition 28. *For all $k \geq 1$, there exists a unique global solution u_k in the space $C^1(\mathbb{R}^+; W_k)$ such that*

$$\frac{\mathrm{d}}{\mathrm{d}t}(u_k(t), v_k) + a(u_k(t), v_k) = (f(t, u_k(t)), v_k), \quad \forall v_k \in W_k, \tag{A.1}$$
$$u_k(0) = P_k u_0.$$

Moreover, for all $T > 0$, (u_k) is bounded in $C([0,T];H)$.

Proof. Let us first prove uniqueness. Consider two solutions u_k^1, u_k^2, then

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_k^1 - u_k^2|^2(t) \le 2L|u_k^1 - u_k^2|^2(t) \le 0,$$

and since $u_k^1(0) = u_k^2(0)$, we infer $u_k^1 = u_k^2$.

We now tackle the global existence. By the Peano existence theorem (see e.g. [19, 2.4.4]), we have local existence. By uniqueness, the maximum time of existence $T_k^\star \in (0,\infty]$ is well-defined. We test the equation with $v_k = u_k(t)$, i.e.

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u_k(t)|^2 + \|u_k(t)\|_a^2 = (f(t,u_k(t)) - f(t,0), u_k(t)) + (f(t,0),u_k(t)).$$

By Young's inequality,

$$\frac{\mathrm{d}}{\mathrm{d}t}|u_k(s)|^2 \le 4L|u_k(t)|^2 + \frac{1}{2L}|f(t,0)|^2.$$

Then, by Gronwall's lemma, for all $T > 0$,

$$\lim_{t\to T^-}|u_k(t)| \le C(T) < \infty,$$

with $C(T) = |u_0| + \frac{1}{2L}\int_0^T e^{4L(T-s)}|f(t,0)|^2\,\mathrm{d}t$. Finally, on the one hand we deduce global existence, i.e. $T_k^\star = \infty$ (e.g. [19, 2.4.3, 2.4.4]), and on the other hand we get the boundedness in $C([0,T];H)$. □

In order to show that (u_k) is a Cauchy sequence in the Banach spaces $C([0,T];H)$ and $L^2(0,T;V)$, let us consider the decomposition

$$u_{k+p} - u_k = P_k(u_{k+p} - u_k) + (\mathrm{Id} - P_k)u_{k+p}. \tag{A.2}$$

For the first term in the right-hand side, we get the following estimates.

Lemma 6. *For all $0 \le t_0 \le t_1$ and all $k,p \ge 1$,*

$$\max\Big(\|P_k(u_{k+p}-u_k)\|_{C([t_0,t_1];H)},\ \sqrt{2c_a}\|P_k(u_{k+p}-u_k)\|_{L^2(t_0,t_1;V)}\Big) \le |(u_{k+p}-u_k)(t_0)| + \sqrt{2L(t_1-t_0)}\|u_{k+p}-u_k\|_{C([t_0,t_1];H)}.$$

Proof. For all $v_k \in W_k$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(u_{k+p}(t) - u_k(t), v_k) + a(u_{k+p}(t) - u_k(t), v_k) = (f(t,u_{k+p}(t)) - f(t,u_k(t)), v_k). \tag{A.3}$$

Testing this equation with $v_k = P_k(u_{k+p} - u_k)$, using orthogonality properties of P_k, and finally integrating on $[t_0, t]$, $t \in [t_0, t_1]$, we have

$$\frac{1}{2}|P_k(u_{k+p} - u_k)(t)|^2 + c_a\|P_k(u_{k+p} - u_k)(t)\|^2_{L^2(t_0,t;V)} \leq \frac{1}{2}|P_k(u_{k+p} - u_k)(t_0)|^2 + L\|u_{k+p} - u_k\|^2_{L^2(t_0,t;H)}.$$

□

Using the diagonalisation of a, we obtain the following estimate for the second term in (A.2) in the $C([t_0, t_1]; H)$-norm.

Lemma 7. *For all $0 \leq t_0 \leq t_1$ and all $k, p \geq 1$*

$$\|(\mathrm{Id} - P_k)u_{k+p}\|_{C([t_0,t_1];H)} \leq |(\mathrm{Id} - P_k)u_{k+p}(t_0)| + \frac{C}{\omega_{k+1}}.$$

Proof. Applying now (A.1) for u_{k+p} with $v^l = (\mathrm{Id} - P_k)u_{k+p}(t)$ yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|(\mathrm{Id} - P_k)u_{k+p}|^2(t) + \|(\mathrm{Id} - P_k)u_{k+p}(t)\|_a^2 = \Big(f(t, u_{k+p}(t)), (\mathrm{Id} - P_k)u_{k+p}(t)\Big). \tag{A.4}$$

Note that

$$\|(\mathrm{Id} - P_k)u_{k+p}(t)\|_a^2 \geq \omega_{k+1}^2|(\mathrm{Id} - P_k)u_{k+p}(t)|^2,$$

so that by Young's inequality on the right-hand side of (A.4), we infer

$$\Big(f(t, u_{k+p}(t)), (\mathrm{Id} - P_k)u_{k+p}(t)\Big) \leq \frac{1}{4\omega_{k+1}^2}|f(t, u_{k+p}(t))|^2 + \|(\mathrm{Id} - P_k)u_{k+p}(t)\|_a^2$$

We conclude using the Lipschitz character of f and the boundedness of (u_k) in $C([0, T]; H)$. □

We are ready to show the first convergence result.

Proposition 29. *For all $T > 0$, (u_k) converges in $C([0, T]; H)$ to some limit u.*

Proof. Let $\tau = \frac{1}{8L}$ and $j \geq 1$. Lemmas 6 and 7 lead to

$$\frac{1}{2}\|u_{k+p} - u_k\|_{C([(j-1)\tau, j\tau];H)} \leq |(u_{k+p} - u_k)((j-1)\tau)| + |(\mathrm{Id} - P_k)u_{k+p}((j-1)\tau)| + \frac{C}{\omega_{k+1}}. \tag{A.5}$$

We prove by induction that the statement

$$\mathcal{P}(j): \quad (u_k) \text{ is a Cauchy sequence in } C([0, j\tau]; H)$$

holds for all $j \geq 1$.

We easily show $\mathcal{P}(1)$. Assume now that $\mathcal{P}(j-1)$ holds for some $j \geq 2$. Let u be the limit of (u_k) in $C([0,(j-1)\tau];H)$. Then we decompose in (A.5)

$$|(\mathrm{Id}-P_k)u_{k+p}((j-1)\tau)| \leq |(u_{k+p}-u)((j-1)\tau)| + |(\mathrm{Id}-P_k)u((j-1)\tau)|,$$

which proves that (u_k) is a Cauchy sequence in $C([(j-1)\tau, j\tau];H)$, and hence that $\mathcal{P}(j)$ holds. □

Remark that we directly obtain

$$u(0) = u_0. \tag{A.6}$$

Next, we get an estimate for the second term in (A.2) in the $L^2(0,T;V)$ norm.

Lemma 8. *For all $T > 0$ and all $k, p \geq 1$,*

$$\|(\mathrm{Id}-P_k)u_{k+p}\|_{L^2(0,T;V)} \leq C(|(\mathrm{Id}-P_k)u_0| + g_{k,p}), \tag{A.7}$$

where the sequence $g_{k,p}$ is defined as

$$g_{k,p} = \|(\mathrm{Id}-P_k)f(\cdot, u_{k+p})\|_{L^2(0,T;H)},$$

and verifies

$$\forall \varepsilon > 0,\ \exists k_0,\ \forall k \geq k_0,\ \forall p \geq 0, \quad g_{k,p} \leq \varepsilon. \tag{A.8}$$

Proof. We consider again (A.4) that we rewrite as

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|(\mathrm{Id}-P_k)u_{k+p}|^2(t) + \|(\mathrm{Id}-P_k)u_{k+p}(t)\|_a^2 = \Big((\mathrm{Id}-P_k)f(t,u_{k+p}(t)), (\mathrm{Id}-P_k)u_{k+p}(t)\Big).$$

Then by integration,

$$c_a\|(\mathrm{Id}-P_k)u_{k+p}\|^2_{L^2(0,T;V)} \leq \frac{1}{2}|(\mathrm{Id}-P_k)u_0|^2 + g_{k,p}\|(\mathrm{Id}-P_k)u_{k+p}\|_{L^2(0,T;V)},$$

and by Young's inequality, we obtain the estimate (A.7).

Now, by strong convergence in $C([0,T];H)$ and continuity of f,

$$|(\mathrm{Id}-P_k)f(s,u_n(s))| \underset{n\to\infty}{\longrightarrow} |(\mathrm{Id}-P_k)f(s,u(s))|.$$

Also,

$$|(\mathrm{Id}-P_k)f(s,u_n(s))| \leq L\overline{C} + |f(s,0)| \leq C,$$

so that by the dominated convergence theorem,

$$\bar{g}_{k,n} = \|(\mathrm{Id} - P_k) f(\cdot, u_n)\|_{L^2(0,T;H)} \underset{n\to\infty}{\longrightarrow} \bar{g}_k = \|(\mathrm{Id} - P_k) f(\cdot, u)\|_{L^2(0,T;H)}.$$

Moreover, by the Parseval theorem, $\bar{g}_{k,n}^2$ and $\bar{g}_k^2$ are the remainders of some positive converging series, so that in particular $\bar{g}_k \underset{k\to\infty}{\longrightarrow} 0$, and $(\bar{g}_{k,n})_k$ is a decreasing sequence for each n. Finally for $\varepsilon > 0$, there exists K such that $\bar{g}_K \leq \frac{\varepsilon}{2}$, and n_0 such that for all $n \geq n_0$, $|\bar{g}_{K,n} - \bar{g}_K| \leq \frac{\varepsilon}{2}$. We conclude by taking $k_0 = \max(K, n_0)$. □

This entails the second convergence result.

Proposition 30. *For all $T > 0$, (u_k) converges in $L^2(0,T;V)$ and its limit is u.*

Proof. By the decomposition (A.2) and Lemma 6,

$$\|u_{k+p} - u_k\|_{L^2(0,T;V)} \leq C\Big(|(\mathrm{Id} - P_k)u_0| + \|u_{k+p} - u_k\|_{C([0,T];H)} + g_{k,p}\Big).$$

Using (A.8), (u_k) is also a Cauchy sequence in $L^2(0,T;V)$. Let $\tilde{u}$ be its limit. Since $L^2(0,T;V)$ and $C([0,T];H)$ are both continuously embedded in $L^2(0,T;H)$, then $\tilde{u} = u$. □

We can finally conclude.

Proof of Proposition 7. Using the previous convergence results we can reinterpret the limit u as satisfying Equation (3.9) in the distribution sense. Given the regularity of u, we have that $-a(u(t), v) + (f(t, u(t)), v)$ is in $L^2(0,T)$ for any $v \in V$, hence it directly follows that $\frac{\mathrm{d}u}{\mathrm{d}t} \in L^2(0,T;V')$. We finally prove the uniqueness as in Proposition 28. □

Références

[1] D. Amsallem and C. Farhat. Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA Journal*, 46(7), 2008.

[2] A. Astolfi. Model reduction by moment matching for linear and nonlinear systems. *Automatic Control, IEEE Transactions on*, 55(10):2321 –2336, 2010.

[3] A.V. Balakrishnan. *Applied Functional Analysis*. Applications of Mathematics. Springer Verlag, 2nd edition, 1981.

[4] K.J. Bathe. *Finite Element Procedures*. Prentice Hall, 1996.

[5] J. Bestel, F. Clément, and M. Sorine. A biomechanical model of muscle contraction. In *Lectures Notes in Computer Science*, volume 2208. Eds W.J. Niessen, M.A. Viergever, Springer, 2001.

[6] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical optimization: Theoretical and practical aspects – Second Edition*. Springer, 2006.

[7] H. Brezis. *Analyse Fonctionnelle : Théorie et Applications*. Mathématiques appliquées pour le master. Dunod, 1999.

[8] M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan, and M. S. Verani. Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professionals from the cardiac imaging committee of the council on clinical cardiology of the american heart association. *American Heart Association Writing Group on Myocardial Segmentation and Registration for Cardiac Imaging*, 2002. AHA Scientific Statement.

[9] R. Chabiniok, D. Chapelle, P.-F. Lesault, A. Rahmouni, and J.-F. Deux. Validation of a biomechanical heart model using animal data with acute myocardial infarction. In *MICCAI Workshop on Cardiovascular Interventional Imaging and Biophysical Modelling (CI2BM09)*, 2009.

[10] R. Chabiniok, Ph. Moireau, P.-F. Lesault, A. Rahmouni, J.-F. Deux, and D. Chapelle. Estimation of tissue contractility from cardiac Cine-

MRI using a biomechanical heart model. *Biomechanics and Modeling in Mechanobiology*, 2011.

[11] D. Chapelle, F. Clément, F. Génot, P. Le Tallec, M. Sorine, and J. Urquiza. A physiologically-based model for the active cardiac muscle. In *Lectures Notes in Computer Science*, volume 2230. Eds T. Katila, I.E. Magnin, P. Clarysse, J. Montagnat, J. Nenonen, Springer-Verlag, 2001.

[12] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1987.

[13] P. Clément. Approximation by finite element functions using local regularization. *R.A.I.R.O., Anal. Numér.*, 8:77–84, 1975.

[14] T.F. Coleman and Y. Li. On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 67(2):189–224, 1994.

[15] T.F. Coleman and Y. Li. An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optimization*, 6(2):418–445, 1996.

[16] L. Daniel, C.S. Ong, L.S. Chay, H.L. Kwok, and J. White. A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 23(5):678 – 693, May 2004.

[17] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology*, volume 5. Springer Verlag, 1992.

[18] B.F. Feeny and R. Kappagantu. On the physical interpretation of proper orthogonal modes in vibrations. *Journal of Sound and Vibration*, 211(4):607–616, 1998.

[19] T.M. Flett. *Differential Analysis*. Cambridge University Press, 1980.

[20] S. P. Hastings. Some mathematical problems from neurobiology. *The American Mathematical Monthly*, 82(9):881–895, 1975.

[21] A.V. Hill. The heat of shortening and the dynamic constants in muscle. *Proc. Roy. Soc. London (B)*, 126:136–195, 1938.

[22] M. Hinze and S. Volkwein. Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control. In T.J. Barth, M. Griebel, D.E. Keyes, R.M. Nieminen, D. Roose, T. Schlick, P. Benner, D.C. Sorensen, and V. Mehrmann, editors, *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lecture Notes in Computational Science and Engineering*, pages 261–306. Springer, 2005.

[23] M. Hinze and S. Volkwein. Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition. *Comput. Optim. Appl.*, 39(3):319–345, 2008.

[24] P. Holmes, J. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, Cambridge, 1996.

[25] A.F. Huxley. Muscle structure and theories of contraction. In *Progress in Biophysics and Biological Chemistry*, volume 7, pages 255–318. Pergamon press, 1957.

[26] A. H. Jazwinski. *Stochastic processes and filtering theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, 1970.

[27] Ch. K. R. T. Jones. Stability of the travelling wave solution of the FitzHugh–Nagumo system. *Transactions of the American Mathematical Society*, 286(2):431–469, 1984.

[28] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *AeroSense: the 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls – Multi Sensor Fusion, Tracking and Resource Management II*. SPIE, 1997.

[29] S.J. Julier and J.K. Uhlmann. Corrections to [30]. In *Proceedings of the IEEE*, volume 92, page 1958, 2004.

[30] S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. In *Proceedings of the IEEE*, volume 92, pages 401–422, 2004.

[31] S.J. Julier, J.K. Uhlmann, and H.F. Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3), 2000.

[32] M. Kahlbacher and S. Volkwein. Galerkin proper orthogonal decomposition methods for parameter dependent elliptic systems. *Discuss. Math. Differ. Incl. Control Optim.*, 27(1):95–117, 2007.

[33] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[34] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic problems. *Numer. Math.*, 90(1):117–148, 2001.

[35] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. *M2AN Math. Model. Numer. Anal.*, 42(1):1–23, 2008.

[36] D. G. Luenberger. Observing the state of a linear system. *IEEE Transactions on Military Electronics*, pages 74–80, 1964.

[37] Y. Maday, A.T. Patera, and G. Turinici. A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. *J. Sci. Comput.*, 17:437–446, December 2002.

[38] Ph. Moireau and D. Chapelle. Erratum of article [39]. *COCV*, 17:406–409, 2011.

[39] Ph. Moireau and D. Chapelle. Reduced-order Unscented Kalman Filtering with application to parameter identification in large-dimensional systems. *COCV*, 17:380–405, 2011.

[40] Ph. Moireau, D. Chapelle, and P. le Tallec. Joint state and parameter estimation for distributed mechanical systems. *Comput. Methods Appl. Mech. Engrg.*, 197:659–677, 2008.

[41] Ph. Moireau, D. Chapelle, and P. Le Tallec. Filtering for distributed mechanical systems using position measurements: Perspectives in medical imaging. *Inverse Problems*, 25(3):035010 (25pp), 2009.

[42] D. T. Pham, J. Verron, and M. C. Roubeaud. A singular evolutive interpolated Kalman filter data assimilation in oceanography. *J. Mar. Syst.*, 16:323–341, 1997.

[43] C. Prud'homme, D.V. Rovas, K. Veroy, and A.T. Patera. A mathematical and computational framework for reliable real-time solution of parametrized partial differential equations. *M2AN Math. Model. Numer. Anal.*, 36(5):747–771, 2002. Programming.

[44] P.-A. Raviart and J.-M. Thomas. *Introduction à l'Analyse Numérique des Equations aux Dérivées Partielles*. Collection Mathématiques Appliquées pour la Maîtrise (*in French*). Masson, 1983.

[45] E. Rohan and R. Cimrman. Sensitivity analysis and material identification for activated smooth muscle. *Computer Assisted Mechanics and Engineering Science*, 9:519–541, 2002.

[46] D.V. Rovas, L. Machiels, and Y. Maday. Reduced-basis output bound methods for parabolic problems. *IMA J. Numer. Anal.*, 26(3):423–445, 2006.

[47] 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.

[48] J. Sainte-Marie. *Models and numerical schemes for free surface flows - Beyond the Saint-Venant system*. Habilitation à diriger des recherches, Université Pierre et Marie Curie, 2010.

[49] J. Sainte-Marie, D. Chapelle, R. Cimrman, and M. Sorine. Modeling and estimation of the cardiac electromechanical activity. *Computers & Structures*, 84:1743–1759, 2006.

[50] G. Serkan and A.C. Athanasios. A survey of model reduction by balanced truncation and some new results. *Internat. J. Control*, 77(8):748–766, 2004.

[51] T. Stykel. Balanced truncation model reduction for semidiscretized Stokes equation. *Linear Algebra and its Applications*, 415(2-3):262–289, 2006. Special Issue on Order Reduction of Large-Scale Systems.

[52] K. Veroy, C. Prud'homme, and A.T. Patera. Reduced-basis approximation of the viscous Burgers equation: rigorous a posteriori error bounds. *C. R. Math. Acad. Sci. Paris*, 337(9):619–624, 2003.

[53] K. Willcox and J. Peraire. Balanced model reduction via the proper orthogonal decomposition. *AIAA Journal*, pages 2323–2330, 2002.

[54] G.I. Zahalak. A distribution moment approximation for kinetic theories of muscular contraction. *Mathematical Biosciences*, 55:89–114, 1981.