Towards well-posed and versatile numerical solutions of scalar-tensor theories of gravity with screening mechanisms : applications at sub-Solar system scales

Hugo Lévy

THÈSE DE DOCTORAT

NNT : 2024UPASP119

université PARIS-SACLAY

Towards well-posed and versatile numerical solutions of scalar-tensor theories of gravity with screening mechanisms: applications at sub-Solar system scales

Vers des solutions numériques bien posées et polyvalentes pour les théories tenseur-scalaires de la gravité avec écrantage : applications aux échelles sub-système Solaire

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 564, physique en Île-de-France (PIF)
Spécialité de doctorat : Physique
Graduate School : Physique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **DPHY Physique, Instrumentation, Environnement Espace** (Université Paris-Saclay, ONERA) et à l'**Institut d'Astrophysique de Paris** (Sorbonne Université, CNRS), sous la direction de **Joël BERGÉ**, Chargé de recherche, et la co-direction de **Jean-Philippe UZAN**, Directeur de recherche

Thèse soutenue à Paris, le 29 octobre 2024, par

Hugo LÉVY

Composition du Jury

Membres du jury avec voix délibérative

Philippe BRAX Directeur de recherche, IPhT, CEA	Président
Clare BURRAGE Professeure, University of Nottingham	Rapporteure & Examinatrice
Meike LIST Professeure, German Aerospace Center	Rapporteure & Examinatrice
Patrick JOLY Directeur de recherche, ENSTA Paris	Examinateur
Gilles MÉTRIS Astronome, Observatoire de la Côte d'Azur	Examinateur

université PARIS-SACLAY | ÉCOLE DOCTORALE Physique en Île-de-France (PIF)

Titre : Vers des solutions numériques bien posées et polyvalentes pour les théories tenseur-scalaires de la gravité avec écrantage : applications aux échelles sub-système Solaire

Mots clés : gravitation, gravité tenseur-scalaire, méthode des éléments finis, mécanismes d'écrantage, géodésie spatiale, relativité générale

Résumé : Les théories tenseur-scalaires de la gravité font partie des alternatives à la Relativité Générale les plus convaincantes, résilientes, et riches en termes de phénoménologie. Les modèles encore viables aujourd'hui reposent sur des mécanismes d'écrantage afin d'être compatibles avec les tests locaux de la gravité, tout en conservant une certaine pertinence physique. La recherche de ces champs scalaires hypothétiques dépend alors de notre capacité à concevoir des expériences adaptées à leur phénoménologie. Hélas, cette tâche est grandement entravée par la difficulté de modéliser suffisamment précisément les effets de cinquième force dans des configurations réalistes. En effet, cela nécessite de résoudre des équations aux dérivées partielles semi-linéaires en présence de distributions de masse non-triviales, ce pour quoi les méthodes purement analytiques ne sont que d'un usage limité.

Dans cette perspective, le présent travail de thèse traite ce problème via le développement d'un outil numérique polyvalent visant à obtenir des solutions bien posées aux équations de Klein–Gordon non-linéaires qui apparaissent dans de tels modèles de gravité modifiée. L'outil en question, nommé *femtoscope*, s'appuie sur la méthode des éléments finis. Celle-ci permet de représenter des géométries arbitrairement complexes et des problèmes multi-échelles par le biais de raffinement locaux du maillage. Les non-linéarités sont quant à elles traitées par la méthode de Newton.

La nouveauté majeure apportée par *femtoscope* est sa gestion des conditions aux limites asymptotiques — i.e. lorsque le comportement du champ n'est connu qu'infiniment loin des sources — dont la prise en compte de manière appropriée est souvent essentielle en vue d'obtenir des solutions numériques pourvues de sens physique. Pour ce faire, nous utilisons la méthode des éléments finis inversés.

Nous nous appuyons ensuite sur *femtoscope* pour étudier la gravité tenseur-scalaire aux échelles sub-système Solaire. En utilisant un modèle réaliste de la Terre, nous traitons la question relative à la détectabilité d'une cinquième force de type caméléon, au moyen de missions de géodésie spatiale telles que GRACE-FO. L'influence de l'atmosphère terrestre ainsi que la rétroaction d'un satellite sur le champ scalaire sont toutes deux prises en compte. Nous constatons que la cinquième force a un effet supposément mesurable sur la dynamique orbitale d'un point matériel, mais que la connaissance imparfaite de la distribution de masse à l'intérieur de la Terre donne lieu à des dégénérescences qui réduisent considérablement le pouvoir contraignant de ce type de mission. Ces dégénérescences peuvent en principe être levées en réalisant l'expérience à deux altitudes différentes.

Enfin, nous ouvrons de nouvelles perspectives en explorant la possibilité de tester les théories tenseur-scalaires avec écrantage en se servant d'horloges atomiques. L'idée des expériences que nous décrivons est d'exploiter la contribution du champ scalaire sur le décalage vers le rouge gravitationnel, cette dernière étant absente en Relativité Générale. On souligne le fait que de telles expériences sont de nature profondément différente des recherches de cinquième force.

université PARIS-SACLAY | ÉCOLE DOCTORALE Physique en Île-de-France (PIF)

Title: Towards well-posed and versatile numerical solutions of scalar-tensor theories of gravity with screening mechanisms: applications at sub-Solar system scales

Keywords: gravitation, scalar-tensor gravity, finite element method, screening mechanisms, space geodesy, general relativity

Abstract: Scalar-tensor theories of gravity are among the most compelling, resilient and phenomenologically-rich alternatives to General Relativity. Viable models make use of screening mechanisms in order to be consistent with local tests of gravity whilst still retaining physical relevance. The hunt for such hypothetical scalar fields therefore hinges on the design of sophisticated model-dependent experiments. Alas, this task is greatly hampered by the difficulty of accurately modeling fifth force effects in realistic setups. Indeed, the latter requires solving semi-linear partial differential equations in the presence of complex matter distributions, for which analytical approaches are clearly insufficient.

In this perspective, the present PhD work tackles this issue by developing a versatile numerical tool devoted to obtaining well-posed solutions to the nonlinear Klein–Gordon equations arising in such modified gravity models. The tool, called *femtoscope*, builds on the finite element method which allows one to deal with arbitrarily complex geometries and multi-scale problems through local mesh refinement. Nonlinearities, on the other hand, are handled via Newton's method.

The novelty and most important feature of *femtoscope* is its careful treatment of asymptotic boundary conditions — i.e. when the field's behavior is only known infinitely far away from the sources — which is often essential to obtain physically meaningful numerical solutions. This is achieved by employing the inverted finite element method.

We then make use of *femtoscope* to investigate screened scalar-tensor gravity at sub-Solar system scales. Using a realistic model of the Earth, we address the question of the detectability of a putative chameleon fifth force in orbit by means of GRACE-FO-like space geodesy missions. The influence of the atmosphere as well as the backreaction of spacecraft on the scalar field are both considered. We find that, although the fifth force has a supposedly measurable effect on the dynamics of a point-like spacecraft, the imperfect knowledge of the mass distribution inside the Earth gives rise to degeneracies, which in turn severely limit the constraining power of such space missions. These degeneracies can in principle be lifted by performing the experiment at two different altitudes.

Finally, we open up new perspectives by exploring the possibility of testing screened scalar-tensor theories with atomic clocks, exploiting the distinctive imprint of the scalar field on the gravitational redshift with respect to General Relativity. It is emphasized that such experiments are profoundly different in nature from fifth force searches.

Remerciements / Acknowledgement

Français

J'ai souvent entendu dire, avant de me lancer moi-même dans cette aventure, que la thèse représente un travail de longue haleine, mené en *solitaire*. Il y a bien sûr du vrai, mais je ne peux m'en tenir ici à cette description, car ce serait occulter la contribution de *tant de personnes* au succès de cette entreprise ! Dans les quelques lignes qui suivent, j'aimerais donc prendre le temps de les remercier dûment.[1]

En premier lieu, je tiens à remercier chaleureusement mes directeurs de thèse, Joël Bergé et Jean-Philippe Uzan, pour leur précieux encadrement tout au long de ces trois années. Au-delà de leur constante bienveillance, j'ai profondément apprécié la liberté et l'autonomie dont j'ai pu disposer dans ma recherche, tout en sachant que leurs portes m'étaient toujours ouvertes. Ce subtil équilibre n'aurait pas pu se construire sans une forme de confiance mutuelle, et je leur suis ainsi reconnaissant de m'avoir accordé la leur.

L'évaluation par les pairs est un pilier fondamental de la démarche scientifique, aussi voudrais-je exprimer ma gratitude à l'égard des membres de mon jury de thèse, qui ont accepté d'examiner mon travail. Je remercie tout particulièrement Meike List et Clare Burrage d'avoir endossé le rôle plus exigeant et minutieux de rapportrices. Merci également à Patrick Joly, mathématicien dans une assemblée majoritairement physicienne, d'avoir apporté son regard critique sur les aspects relatifs aux mathématiques appliquées et à l'analyse numérique présents dans ma thèse.

En plus des personnes déjà mentionnées, je tiens à exprimer ma profonde gratitude envers toutes celles et ceux qui m'ont apporté leur aide et leurs conseils sur le plan scientifique. En particulier, merci à Manuel Rodrigues de m'avoir prêté son expertise sans faille sur la mission MICROSCOPE, à Gilles Esposito-Farèse pour les discussions relatives au redshift en théories tenseur-scalaires et ses relectures attentives de la partie du manuscrit correspondante, à Tahar Boulmezaoud — à l'origine de la méthode des éléments finis inversés — de m'avoir donné de son temps, et à Clare Burrage de m'avoir aimablement accueilli à l'université de Nottingham le temps de quelques jours. Merci également à Antoine Ait-Mehdi pour ses conseils avisés en programmation dans le développement de *femtoscope* (et sans qui je ne me serais jamais lancé dans un si important *code refactoring*), à Phuong-Anh Huynh pour son aide au portage de mon code sur les supercalculateurs de l'ONERA, et à Matthieu Dellavalle d'en avoir été le tout premier bêta-testeur.

Si ma thèse a pu se dérouler avec autant de fluidité, c'est aussi grâce au concours de nombreuses personnes, parmi lesquels Nassim Zahzam et Jérôme Perez qui ont suivi — c'est le cas de le dire — mon parcours de jeune thésard. En particulier, j'aimerais remercier Jérôme qui, de manière tout à fait spontanée, a trouvé des interlocuteurs à l'unité de mathématiques appliquées de l'ENSTA Paris pour répondre à mes questions au moment où j'en ai eu besoin. Mes pensées vont aussi au secrétariat du DPHY, qui œuvre au quotidien pour offrir aux doctorants les meilleures conditions de travail.

La thèse a aussi été pour moi l'occasion de m'essayer à l'enseignement. À cet égard, merci notamment à Stéphanie Lizy-Destrez de m'avoir offert l'opportunité de revenir dans mon ancienne école en tant que PC-man de méca spa'. Toutes ces excursions hors de ma zone de confort, en plus de rythmer mon quotidien de thèse — où les jours se suivent mais ne ressemblent pas — ont été très enrichissantes, sinon gratifiantes !

Sur une note plus personnelle à présent, je voudrais dire un grand merci à mes collègues et amis d'IEA, qui ont tous contribué, à leur manière, à ce que je me sente bien sur mon lieu de travail. Vu de l'extérieur, il est vrai que le centre de l'ONERA Châtillon peut sembler quelque peu austère et triste, mais cela n'est qu'apparence ! La réalité est bien différente une fois poussée la porte du bâtiment F, où la bonne humeur que j'ai tenté d'apporter chaque jour m'a été pleinement rendue. Ayant été le seul doctorant de l'unité pendant l'essentiel de ma thèse, je tiens à adresser une mention spéciale aux stagiaires — maintenant devenus à leur tour doctorants pour certains — avec qui j'ai cohabité, car ils ont apporté de la vie dans un bureau bien trop vide autrement. Je n'oublierai certainement pas les montées d'adrénaline, à 15 heures, autour du babyfoot !

À l'IAP, un grand merci à l'ensemble des doctorantes et doctorants que j'ai eu la chance de côtoyer de créer cette ambiance si singulière, solidaire, et appréciable au quotidien. Malgré ma présence quelque peu sporadique au 98bis boulevard Arago, je me suis senti, et ce dès le départ, pleinement intégré au groupe. Un merci tout

[1] La langue française a beau être d'une grande richesse, peut-être manque-t-elle de synonymes pour le mot 'merci'. Décliné sous toutes ses formes dans ce qui suit, c'est avec une égale sincérité que je le répète.

particulier à Mathieu qui a accepté, le jour de ma soutenance, la lourde responsabilité de 'garant technique', me libérant ainsi un volume significatif de RAM-mentale.

Je voudrais faire savoir à mes amis combien leur soutien indéfectible au cours de ces trois dernières années me touche. Entre les matchs de volley-ball du lundi soir, les footings du mercredi entre midi et deux, les soirées du samedi soir[2] et les répets' de musique du dimanche avec mes amis les plus métalleux (\m/) , j'ai vraiment été comblé ! Je ne peux qu'espérer que ces amitiés, essentielles à mon équilibre de vie, perdurent encore longtemps.

Enfin, et pour conclure cette liste déjà longue, j'adresse mes remerciements les plus sincères à toute ma famille[3] pour son soutien inconditionnel sur tous les plans, pendant ma thèse bien sûr, mais surtout pendant les vingt-quatre années qui l'ont précédée. Un merci tout particulier à mes parents qui — excusez-moi pour la déformation professionnelle — m'ont donné les justes conditions initiales pour que j'évolue sur une trajectoire épanouie et heureuse dans la vie. Mes derniers mots — et ils sont bien trop peu pour lui exprimer toute ma reconnaissance — iront à Carole, devenue malgré elle experte en gravité caméléon ! Merci d'avoir été à mes côtés dans les moments de stress, de frustration, ou de doute que j'ai pu ressentir. Merci bien plus encore d'avoir partagé avec moi les moments de joie, d'excitation, voire d'euphorie liés à cette aventure un peu folle, mais certainement pas *solitaire*, qu'a été cette thèse.

English

I have often heard, before embarking on this journey myself, that a PhD is a long and *solitary* haul. There is, of course, some truth in this statement; but it actually overlooks the contribution of *so many people* to the success of the endeavor! In the few paragraphs that follow, I would like to take the time to thank them all properly.

First and foremost, I warmly thank my thesis supervisors, Joël Bergé and Jean-Philippe Uzan, for their invaluable guidance throughout these three years. Beyond their constant benevolence, I greatly appreciated the freedom and autonomy they granted me to conduct my research work, knowing that their doors were always open. This subtle balance could not have been achieved without a form of mutual trust, and I am grateful to them for placing theirs in me.

Peer review is a cornerstone of the scientific process, and I want to express my gratitude to the members of my thesis jury who agreed to review my work. I particularly thank Meike List and Clare Burrage for taking on the more demanding role of reviewers. I also extend my thanks to Patrick Joly, a mathematician in an otherwise predominantly physicist assembly, for bringing his critical perspective to the applied mathematics and numerical analysis aspects of my thesis.

In addition to those mentioned, I wish to express my sincere gratitude to everyone who provided me with scientific advice and assistance. In particular, my thanks go to Manuel Rodrigues for his invaluable expertise on the MICROSCOPE mission, to Gilles Esposito-Farèse for discussions on redshift in scalar-tensor theories and his careful review of the corresponding section of the manuscript, to Tahar Boulmezaoud — author of the inverted finite element method — for sharing his time with me, as well as to Clare Burrage for kindly welcoming me at the University of Nottingham for a few days. My thanks also go to Antoine Ait-Mehdi for his enlightening programming advice during the development of *femtoscope* (and without whom I may never have embarked on such an ambitious code refactoring during my final year), to Phuong-Anh Huynh for assisting me with the migration of my code to ONERA's supercomputers, and to Matthieu Dellavalle for being its very first beta-tester.

If my PhD journey has been so smooth, it is also thanks to the support of numerous people, including Nassim Zahzam and Jérôme Perez, who followed my early research career as members of my so-called "comité de suivi de thèse". I particularly thank Jérôme, who, entirely on his own initiative, found contacts in the ENSTA Paris applied mathematics department to answer my questions at a crucial time. My thoughts also go out to the DPHY secretariat, whose daily efforts ensure that doctoral students have the best possible working conditions.

My PhD also gave me the opportunity to try my hand at teaching. In this regard, my thanks go to Stéphanie Lizy-Destrez for the chance to return to my *alma mater* as a space mechanics PC-man. All these excursions outside my comfort zone added rhythm to my PhD experience — where no two days where alike — and proved both enriching and fulfilling!

On a more personal note, I want to extend my heartfelt thanks to my colleagues and friends at IEA, each of whom, in their own way, contributed to making me feel at home in my workplace. Admittedly, from the outside, ONERA's Châtillon center may seem somewhat austere and dull, but this is only an illusion! The reality is quite different once you step inside Building F, where the good cheer I tried to bring each day was fully reciprocated. Being the only PhD student in the unit for most of my program, I want to give a special mention to the interns — some of whom have since become doctoral students themselves — with whom I shared my workspace, as they brought life to an otherwise far too quiet office. I will most certainly remember the adrenaline-fueled foosball matches at 3 p.m.!

[2]Lorsque celles-ci ne sont pas mises à profit pour terminer l'écriture du présent manuscrit toutefois...

[3]Sans vouloir être exhaustif : mes parents, ma sœur, Maxime, mes grands-parents, mon oncle, mes tantes, mes cousin·es, sans oublier les chats et les toutous !

At IAP, my thanks go to all the PhD students I had the pleasure of meeting, and who contributed to creating such a unique, supportive, and enjoyable atmosphere. Despite my somewhat sporadic presence at 98bis Boulevard Arago, I felt fully welcomed by the group from the outset. A special thanks to Mathieu for taking on the crucial role of 'technical guarantor' on the day of my defense, thus freeing up a significant portion of my mental RAM.

I want my friends to know how much their unwavering support over the past three years has meant to me. Between Monday night volleyball games, Wednesday lunchtime runs, Saturday night gatherings[4] and Sunday band practices with my most metalhead friends (\m/), I've truly been fulfilled! I can only hope these friendships, essential to my work-life balance, will continue for years to come.

Lastly, and to conclude this already lengthy list, I extend my deepest thanks to my family for their unconditional support in all respects, not only during my PhD but also over the twenty-four years leading up to it. A special thanks to my parents, who — pardon my professional bias — gave me the ideal initial conditions to pursue a fulfilling and happy path in life. My final thanks — and they are far too little to convey the full extent of my gratitude — go to Carole, who, in spite of herself, has become an expert on chameleon gravity! Thank you for standing by me through moments of stress, frustration, or doubt. Thank you even more for sharing with me the joy, excitement, and even euphoria of this somewhat crazy but certainly not *solitary* adventure that has been my PhD.

[4] Except for those nights spent finishing this very manuscript...

Résumé substantiel en langue française

Contexte général de la thèse

La Relativité Générale (RG) est la théorie géométrique de la gravitation publiée par Albert Einstein en 1915 et constitue la description actuelle de la gravitation en physique moderne. La RG est l'une des théories physiques les plus éprouvées, puisqu'aucun des nombreux tests expérimentaux effectués depuis le début du XX$^{\text{ème}}$ siècle n'a pu la mettre en défaut. Ceux-ci comprennent d'abord les tests dits 'classiques' — la précession du périhélie de Mercure, la déflexion de la lumière par le Soleil, le décalage vers le rouge gravitationnel — et plus généralement tous les tests post-newtoniens de la gravité. Plus récemment, les détections directes par les interféromètres *LIGO* et *Virgo* d'ondes gravitationnelles produites par la coalescence de systèmes binaires compacts, ainsi que l'imagerie par l'*Event Horizon Telescope* de trous noirs, ont contribué à asseoir plus encore la théorie. Aujourd'hui, la RG fait partie intégrante de la physique moderne. En particulier, elle est à la base même du *modèle standard de la cosmologie* qui, à l'heure actuelle, décrit de la façon la plus satisfaisante l'histoire de l'univers dans lequel nous vivons. Dans ce modèle, la majeure partie du contenu masse/énergie de l'espace-temps est constituée de matière noire ($\sim 27\,\%$) et d'énergie sombre ($\sim 68\,\%$), dont le choix des noms renvoie directement à leur nature élusive. Quand bien même on accepterait que $95\,\%$ de l'univers nous demeure de nature inconnue, le nombre croissant de relevés astronomiques et cosmologiques ces dernières décennies a fait émerger plusieurs *tensions*. Ces tensions font référence à des observations contradictoires dans le cadre du modèle standard — lui-même reposant sur la RG — la plus célèbre étant la *tension de Hubble*. Au-delà de ces anomalies de plus en plus dérangeantes, l'apparente incompatibilité avec la *mécanique quantique* est un indice supplémentaire quant au fait que la RG n'est certainement pas le fin mot sur la gravitation.

Face à ces difficultés, il semble légitime et même nécessaire de s'autoriser à étudier des modèles alternatifs à la RG, regroupés sous la dénomination générique de 'gravité modifiée'. Pour être jugé pertinent, tout modèle alternatif doit non seulement rendre compte des observations qui motivent son existence (e.g. expliquer de manière cohérente l'accélération de l'expansion de l'univers), mais aussi rester en accord avec la physique connue. Cette entreprise est d'autant plus difficile pour les physiciens théoriciens que la RG est aujourd'hui fortement contrainte par une myriade de tests.

Dans cette thèse, je m'intéresse aux théories *tenseur-scalaires* de la gravitation. Celles-ci constituent l'une des extensions les plus naturelles et résilientes de la RG, où la gravité est décrite mathématiquement par la combinaison d'un tenseur métrique $g_{\mu\nu}$ et d'un champ scalaire ϕ. Cet ajout, par rapport à la RG, d'un degré de liberté scalaire dans le secteur gravitationnel offre un assez large spectre de phénoménologies possibles, le champ scalaire pouvant jouer différents rôles suivant les motivations physiques sous-jacentes : de l'accélération de l'expansion de l'univers aux candidats pour la matière noire, en passant par le paradigme d'inflation. De plus, ces particules de spin nul et de faible masse apparaissent naturellement dans le cadre de théories plus fondamentales, comme par exemple en théorie des cordes, dans la limite de faibles énergies.

On ne peut pas parler de théories tenseur-scalaires sans introduire la notion de *cinquième force*, un terme inventé dans les années 1980 par E. Fischbach pour désigner une nouvelle force — aux côtés des quatre interactions fondamentales connues — dont le champ scalaire est le médiateur. Du point de vue du lagrangien de la théorie, une telle cinquième force apparaît lorsque (*i*) le champ scalaire est couplé de manière conforme aux champs de matière mais demeure minimalement couplé à la métrique ; ou, de manière équivalente, lorsque (*ii*) le champ scalaire est couplé à la courbure scalaire tout en restant minimalement couplé à la matière. De telles cinquièmes forces constituent des déviations à la RG, qui sont par conséquent fortement contraintes, notamment par nos observations à l'échelle du système Solaire et expériences de laboratoires.

Pour demeurer simultanément viables vis-à-vis des contraintes sus-mentionnées et intéressantes sur le plan physique, les théories tenseur-scalaires doivent se doter de *mécanismes d'écrantage* : des non-linéarités astucieusement introduites au niveau du lagrangien dans le but de 'cacher' les effets du champ scalaire aux échelles sub-système Solaire, tout en autorisant des déviations à la RG aux échelles astrophysiques et cosmologiques, où la gravité est bien moins contrainte. Concrètement, ces non-linéarités ont pour effet de rendre dynamique une des propriétés du champ — comme sa masse (e.g. modèle *caméléon*) ou son couplage à la matière (e.g. modèle *symmetron*) — de manière à réduire largement les effets de cinquième force dans les milieux suffisamment denses. Toutefois, même dans le cadre de ces modèles avec écrantage, la cinquième force n'est jamais strictement nulle.

L'empreinte du champ scalaire sur la gravité, suivant son ampleur, reste donc susceptible d'être mesurée.

Comme toujours en physique, c'est la comparaison entre les prédictions d'un modèle donné d'une part, et des données expérimentales acquises d'autre part, qui permet de tirer des conclusions quant à la viabilité du modèle en question. Pour mener à bien cette démarche, il est nécessaire de comprendre — qualitativement et quantitativement — comment se comportent le champ scalaire et la cinquième force qui lui est associée. Dans le cadre des théories tenseur-scalaires avec écrantage, cette tâche est complexe à bien des égards. D'abord, l'importance d'une modélisation réaliste des géométries et de la distribution de matière a été soulignée à plusieurs reprises dans des travaux antérieurs, notamment dans la thèse de Martin Pernot-Borràs. Une autre difficulté est celle des non-linéarités : le calcul de la cinquième force passe par la résolution d'une équation aux dérivées partielles (EDP) de Klein–Gordon semi-linéaire, ce qui limite d'autant plus l'utilisation de techniques analytiques et d'ansätze. Ces deux considérations m'ont naturellement poussé à me tourner vers les méthodes numériques, plus particulièrement la *méthode des éléments finis* (FEM). En effet, celle-ci présente l'avantage qu'elle repose sur un maillage dont la résolution peut être adaptée localement (h-adaptivité), ce qui permet de représenter des géométries arbitrairement complexes et de concentrer les ressources computationnelles là où elles sont le plus nécessaires. De plus, l'utilisation complémentaire de méthodes itératives — comme la méthode de Newton — permet d'étendre le cadre de la FEM aux problèmes non-linéaires qui nous intéressent.

La dernière difficulté majeure est relative aux *conditions aux limites*, dont la spécification est nécessaire pour obtenir un problème dit *bien posé*. Dans de nombreux contextes physiques, on utilise des conditions de *Dirichlet*, où la valeur de l'inconnue est fixée sur le bord du domaine numérique. Or dans notre cas, le comportement du champ scalaire n'est en général connu qu'infiniment loin des sources. En d'autres termes, cela signifie que le domaine spatial sur lequel le problème est posé n'est pas borné : on parle alors de *conditions aux limites asymptotiques*. Naturellement, la mémoire d'un ordinateur étant finie, il est impossible de mailler une région non bornée de $\mathbb{R}^n$, $n \in \{1, 2, 3\}$. Une solution naïve consisterait à tronquer le domaine à une distance finie et à utiliser la valeur asymptotique du champ comme condition de Dirichlet. Cela n'est malheureusement pas satisfaisant car (i) le domaine résultant de la troncature doit alors être suffisamment grand ce qui donne lieu à de gros systèmes linéaires à résoudre, et (ii) imposer une condition de nature asymptotique à une distance finie peut engendrer une erreur non-négligeable — et surtout non-quantifiable aisément — sur la solution numérique.

Le premier objectif de ma thèse a été de développer un outil numérique polyvalent, basé sur la méthode des éléments finis, pour pouvoir résoudre les EDPs qui apparaissent dans l'étude des théories tenseur-scalaires de la gravitation avec écrantage. Cet outil est nommé *femtoscope*. Un soin particulier est apporté à la gestion des conditions aux limites asymptotiques, et plus généralement au caractère bien-posé des solutions numériques obtenues. Ces problématiques font l'objet des chapitres 2, 3 et 4. Ensuite, *femtoscope* est mis à profit pour explorer des scénarios en gravité modifiée qui étaient inaccessibles jusqu'alors. Spécifiquement, on s'intéresse dans les chapitres 4 et 5 au mouvement d'un satellite en orbite autour de la Terre dans le contexte de la gravité caméléon comme prototype de modèle tenseur-scalaire avec écrantage. Dans le chapitre 6, on étudie la possibilité de tester le modèle caméléon via des mesures du décalage vers le rouge gravitationnel — une idée qui n'avait pas été envisagée dans la littérature jusqu'alors.

Dans ce qui suit, on propose un résumé de la thèse chapitre par chapitre, mettant en avant les résultats principaux obtenus.

Chapitre 1 : théories tenseur-scalaires de la gravité

Ce premier chapitre est une introduction aux théories tenseur-scalaires de la gravité comme extension de la RG. On se concentre sur la classe de modèles dite 'traditionnelle', dont l'action S peut s'écrire génériquement comme

$$S = S_{\rm EH} + S_\phi + S_{\rm mat}[\tilde{g}_{\mu\nu}] \,.$$

Dans cette expression, $S_{\rm EH}$ est l'action de Einstein–Hilbert usuelle, tandis que les actions du champ scalaire et des champs de matière, en unités naturelles pour lesquelles $c = \hbar = 1$, sont données respectivement par

$$S_\phi = -\int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi + V(\phi)\right], \qquad S_{\rm mat}[\tilde{g}_{\mu\nu}] = \int \mathrm{d}^4x\sqrt{-\tilde{g}}\mathcal{L}_{\rm mat}\big(\tilde{g}_{\mu\nu},\,\psi_{\rm mat}\big)\,.$$

La métrique d'Einstein $g_{\mu\nu}$ est liée à la métrique de Jordan $\tilde{g}_{\mu\nu}$ via une transformation de Weyl $\tilde{g}_{\mu\nu} = \Omega^2(\phi)g_{\mu\nu}$, où Ω et V sont deux fonctions du champ scalaire. En particulier dans la section 1.1, on ré-établit à partir de cette action toutes les expressions qui jouent un rôle important dans ce travail, à savoir :

- les équations de champ pour la métrique et pour le champ scalaire, ainsi que leurs limites newtoniennes ;
- les équations de la cosmologie [Eqs. (1.81–1.92)] ;
- l'équation des géodésiques, sa limite newtonienne et l'expression de la cinquième force qui en découle.

Pour assurer une certaine cohérence et dans un objectif d'autosuffisance, la plupart de ces équations sont données dans les représentations d'Einstein et de Jordan. Voir la Table 6.1 pour une compilation de ces équations.

La section 1.2 présente les différents mécanismes d'écrantage connus (voir la Table 1.3 pour une classification), avec un focus sur le mécanisme *caméléon*, où la masse du champ scalaire varie dynamiquement selon la densité du milieu ambiant : dans les milieux de forte densité, le champ acquiert une importante masse, limitant ainsi la portée de la cinquième force qu'il occasionne ; tandis qu'il devient léger dans les milieux de densité plus faible.

On revient dans la section 1.3 sur la mission spatiale MICROSCOPE qui, au-delà d'être le test du principe d'équivalence faible le plus précis jamais réalisé à ce jour, a permis de contraindre toute une panoplie de modèles alternatifs à la RG. Nous revisitons les résultats de thèse de Martin Pernot-Borràs sur la recherche de cinquième force de type caméléon dans les données de MICROSCOPE. À la lumière de ces résultats, il apparaît que la testabilité des modèles avec écrantage dépend de façon cruciale du développement de nouveaux outils numériques pour une modélisation réaliste de leurs caractéristiques. On établit alors dans la section 1.4 une liste des spécifications que devra posséder un tel outil, en mettant en évidence le fait qu'aucun des codes numériques existants ne remplit le cahier des charges. Ces principales spécifications sont récapitulées dans la Table 1.4.

Chapitre 2 : méthode des éléments finis

Les EDPs que l'on souhaite pouvoir résoudre numériquement sont (*i*) l'équation de Poisson qui régit le potentiel newtonien, et (*ii*) les équations de Klein–Gordon semi-linéaires auxquelles obéissent e.g. les champs caméléon et symmetron. On choisit d'utiliser la méthode des éléments finis pour résoudre ces EDPs, dont les bases sont exposées dans ce deuxième chapitre introductif.

La section 2.1 commence avec une vue d'ensemble de la FEM, illustrée sur une EDP linéaire elliptique posée sur un domaine $\Omega \subset \mathbb{R}^n$ borné, avec conditions de Dirichlet imposées sur le bord $\Gamma \coloneqq \partial\Omega$. On y introduit en particulier les concepts de formulation variationnelle, de problème bien-posé au sens de Hadamard ainsi que tous les outils mathématiques associés : inégalité de Poincaré, inégalité de trace, espaces de Sobolev, théorème de Lax–Milgram, etc. Une fois le cadre fonctionnel établi, on procède à la discrétisation du problème par éléments finis qui permet sa résolution sur une machine à mémoire finie. Bien que cette thèse ne s'intéresse qu'à des problèmes stationnaires, on inclut une discussion sur la résolution d'EDPs dépendantes en temps qui pourra éventuellement s'avérer utile pour de futurs travaux.

Dans la section 2.2, on montre comment la FEM peut être étendue pour traiter des EDPs non-linéaires via des méthodes itératives. Celles-ci reposent sur l'algorithme suivant :

1. linéariser l'EDP autour d'une première estimation de la solution ;
2. résoudre l'EDP linéarisée par la FEM ;
3. actualiser l'estimée à partir de la solution précédente et reprendre à l'étape 2 jusqu'à convergence.

Les méthodes de Picard et de Newton sont illustrées sur l'équation de Klein–Gordon régissant le champ caméléon.

Enfin, il est courant en physique de rencontrer des symétries continues globales (e.g. symétrie sphérique, invariance par translation, etc.), auquel cas il est possible de réduire la dimension effective de l'EDP en jeu. D'un point de vue numérique, la réduction de la dimension du problème est un atout non-négligeable en termes de complexité temporelle et spatiale, c'est pourquoi j'explique dans la section 2.3 comment tirer parti de ces symétries, lorsqu'elles existent, dans le cadre de la FEM. En particulier, je prouve le caractère bien-posé des formulations faibles qui résultent de la réduction dimensionnelle pour les symétries sphériques et cylindriques.

Chapitre 3 : problèmes posés sur des domaines non-bornés

Pour pleinement satisfaire le cahier des charges établi au chapitre 1, on doit s'affranchir de l'hypothèse Ω borné. Le chapitre 3 est ainsi dédié au traitement des conditions aux limites asymptotiques et constitue une synthèse des diverses approches explorées au cours de cette thèse.

On commence dans la section 3.2 par spécifier un cadre fonctionnel adéquat aux problèmes posés sur $\mathbb{R}^n$. Suivant la forme spécifique de l'EDP en jeu, il peut être nécessaire d'avoir recours à des espaces de Sobolev à poids pour donner un sens à la formulation faible sous-jacente, où le choix d'un poids adéquat permet d'imposer le comportement asymptotique désiré sur la solution. On vérifie dans les deux cas d'intérêt — les équations de Poisson et de Klein–Gordon — que le théorème de Lax–Milgram s'applique. En particulier, l'inégalité de Poincaré, qui servait à démontrer la coercivité de la forme bilinéaire associée à la formulation faible en domaine borné, est remplacée par une inégalité de Hardy généralisée.

Une fois le décor fonctionnel planté, on revient dans la section 3.3 à des considérations numériques davantage pratiques. Un moyen de conserver le caractère non-borné du domaine en machine consiste à utiliser des *compactifications* : des changements de coordonnées par lesquels l'image de $\mathbb{R}^n$ est un borné. La méthode retenue est la suivante. On décompose d'abord le domaine Ω en un domaine intérieur borné $\Omega_{\text{int}} = \mathcal{B}(R_c)$ et un domaine extérieur non-borné $\Omega_{\text{ext}} = \Omega \setminus \bar{\Omega}_{\text{int}}$, où $R_c > 0$ et $\mathcal{B}(R_c)$ est la boule ouverte de rayon R_c et centrée à l'origine. Le

domaine extérieur est ensuite *inversé* via la *transformation de Kelvin* $\mathcal{K}\colon \mathbb{R}^n \setminus \{\mathbf{0}\} \ni \mathbf{x} \mapsto (R_c/\|\mathbf{x}\|^2)\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, et l'on note $\tilde{\Omega}_{\text{ext}} = \mathcal{K}(\Omega_{\text{ext}})$. On se ramène ainsi à un problème posé sur $\bar{\Omega}_{\text{int}} \cup \tilde{\Omega}_{\text{ext}}$ borné. La section 3.4 est consacrée aux détails et subtilités associées à la discrétisation éléments finis de ce problème. Cette première méthode est assimilable à la *méthode des éléments finis inversés* (*ifem*) introduite par T. Boulmezaoud.

Dans la section 3.5, je propose une nouvelle méthode — la *méthode des éléments finis inversés alternée* (*a-ifem*) — s'inspirant en grande partie de la méthode *ifem* et d'une méthode de décomposition de domaine proposée par Marini et Quarteroni. Enfin, on réalise dans la section 3.6 une série d'expériences numériques visant à comparer les performances des méthodes *ifem* et *a-ifem* sur des exemples pour lesquels la solution analytique est connue. Influence des méta-paramètres, vitesse de convergence des itérations de la méthode *a-ifem*, complexité temporelle et courbes de convergence font partie, entre autres, des sujets abordés en détail à cette occasion.

Chapitre 4 : modélisation de la gravité tenseur-scalaire avec *femtoscope*

Les chapitres 2 et 3 ont permis de couvrir l'ensemble des points critiques du cahier des charges dressé au chapitre 1. Le chapitre 4 constitue le point culminant de cet effort avec l'implémentation de *femtoscope*, dont un aperçu global est donné dans la section 4.1. Ce code Python, logiciel libre sous licence MIT et disponible sur la plate-forme GitHub,[5] repose sur l'ensemble des techniques numériques abordées aux chapitres précédents pour résoudre des EDPs non-linéaires elliptiques posées sur des domaines bornés ou non de $\mathbb{R}^n$, $n \in \{1, 2, 3\}$. Dans sa version actuelle (09/2024), *femtoscope* pré-implémente :

- l'équation de Poisson (4.1) associée au potentiel newtonien ;
- l'équation de Klein–Gordon linéaire (4.2) relative au potentiel de Yukawa ;
- les équations de Klein–Gordon non-linéaires (4.4, 4.6) associées au caméléon et au symmetron respectivement.

L'architecture du programme, qui suit le paradigme de la programmation orientée objet, est détaillée à travers les figures 4.3 à 4.5. En vue des applications en orbite terrestre qui font l'objet du chapitre 5, *femtoscope* implémente un modèle de densité de la Terre (PREM) et de son atmosphère (US76).

La section 4.2 est dédiée à la validation du code. Plusieurs leviers sont utilisés à cette fin :

- Implémentation de tests unitaires et tests d'intégration écrits en `pytest`.
- Pour l'équation de Poisson, analyse de l'erreur sur le cas de l'ellipsoïde de révolution oblate pour lequel il existe une expression analytique du potentiel newtonien [Eqs. (4.8–4.10)] — voir Fig. 4.8.
- Pour l'équation de Klein–Gordon associée au caméléon, comparaisons de la solution éléments finis aux approximations analytiques proposées dans la littérature et au code SELCIE.

Enfin, la section 4.3 met *femtoscope* en valeur sur deux cas d'usage. La première application concerne l'étude de la gravité caméléon en orbite terrestre, où l'on utilise les modèles de densité réalistes PREM et US76 mentionnés ci-dessus. En guise de seconde application, on s'intéresse au problème à deux corps, toujours en gravité caméléon. En imaginant une expérience de laboratoire, je montre que l'ajout de la cinquième force a pour effet de déplacer le point de Lagrange $L1$ du système par rapport au cas purement newtonien à un niveau non-négligeable devant la taille caractéristique du système, lorsqu'au moins une des deux sphères est écrantée. Cela constitue, au moins sur le plan théorique, un moyen de discriminer l'attraction newtonienne des effets de cinquième force.

Soulignons que l'intérêt de disposer d'un outil comme *femtoscope* est double : il peut être utilisé pour traduire les données d'une expérience en contraintes sur un certain modèle tenseur-scalaire, mais peut aussi revêtir un rôle plus prospectif. En effet, on peut s'en servir pour évaluer la pertinence d'un dispositif expérimental donné, ou encore pour optimiser ce dernier.

Chapitre 5 : effets de cinquième force en orbite terrestre

Le chapitre 5 fait suite à la première application de *femtoscope* à la gravité caméléon en orbite terrestre présentée au chapitre précédent. L'objectif est d'évaluer quantitativement la détectabilité d'une cinquième force de type caméléon en orbite terrestre par le biais de techniques de géodésie spatiale. Comme précédemment, on utilise les modèles PREM et US76 pour assigner une densité à la Terre et son atmosphère respectivement. On introduit de plus une déviation à la symétrie sphérique incarnée par une montagne, qu'on choisit axisymétrique de manière à pouvoir réaliser nos calculs éléments finis en dimension deux et non trois. Ensuite, on utilise *femtoscope* pour calculer numériquement le potentiel newtonien et le champ scalaire, qui permettent d'accéder à la gravité

[5] https://github.com/onera/femtoscope

standard et modifiée. Dans les régions encore viables de l'espace des paramètres du modèle caméléon, la Terre est écrantée, ce qui signifie que seules ses couches les plus externes contribuent à la cinquième force, par opposition à la gravité newtonienne de portée infinie. Ainsi, le champ caméléon laisse une signature distinctive sur le champ gravitationnel terrestre, qu'on retrouve bien dans sa décomposition en harmoniques sphériques. Dès lors, la question est de savoir si cette signature est mesurable en pratique.

Dans cette perspective, on commence par aborder la question de l'influence de l'atmosphère sur la cinquième force. À paramètres du modèle caméléon (Λ, n) fixés, quatre régimes distincts apparaissent lorsqu'on augmente progressivement le paramètre de couplage β : (i) pour les petites valeurs de β, l'atmosphère est totalement *transparente* pour le champ scalaire, (ii) au-delà d'un certain seuil, elle agit comme un atténuateur de la cinquième force, (iii) pour des couplages encore plus forts, toute dépendance non-radiale du champ scalaire est supprimée, de sorte que la montagne est tout à fait invisible, et (iv) l'atmosphère elle-même finit par être écrantée.

Nous nous penchons aussi sur la question de la rétroaction d'un satellite sur le champ caméléon d'arrière-plan. Pour la première fois, on dépasse les approximations faites dans la littérature en calculant le champ créé par le système complet {Terre + satellite}. On montre en particulier que la transition entre les régimes écranté et non-écranté pour le satellite s'opère sur une région très étroite de l'espace des paramètres du modèle caméléon. Lorsque ce dernier est écranté, la cinquième force résultante est supprimée de manière extrêmement efficace.

Dans un troisième temps, on retient le scénario le plus favorable (i.e. qui maximise la cinquième force, sans atmosphère) et on simule la dynamique d'une paire de satellites placés sur une même orbite, à l'image de la mission de géodésie spatiale GRACE-FO. On fait l'hypothèse que les satellites ne sont pas écrantés, ce qui permet de les traiter comme des points matériels dans le système {Terre sphérique + montagne}. Conformément au principe de la mission GRACE-FO, on s'intéresse à la variation de la distance entre les deux satellites au cours du temps, en gravité newtonienne d'une part et en gravité modifiée d'autre part. Contre toute attente, l'anomalie causée par la cinquième force est non-négligeable, plusieurs ordres de grandeur supérieure à la limite de sensibilité offerte par la technologie spatiale actuelle. Malheureusement, l'existence d'incertitudes dans le modèle — notamment le fait que la distribution de masse à l'intérieur de la Terre demeure mal connue — réduit considérablement le pouvoir contraignant de ce type de mission. Toutefois, ces dégénérescences peuvent en principe être levées en réalisant l'expérience à plusieurs altitudes différentes.

Chapitre 6 : tester la gravité tenseur-scalaire avec des horloges atomiques

Dans ce sixième et dernier chapitre, j'explore la possibilité de tester les théories tenseur-scalaires au moyen d'expériences basées sur le phénomène de décalage vers le rouge, ou *redshift*, gravitationnel. Contrairement aux effets de cinquième force, qui dépendent principalement du gradient du champ scalaire, la contribution scalaire au redshift total est essentiellement proportionnelle à la différence entre les valeurs prises par le champ en différents points de l'espace. Ceci donne ainsi lieu à des concepts de tests reposant sur l'utilisation d'horloges atomiques, aux antipodes des recherches de cinquième force.

Dans la section 6.1, j'obtiens l'expression théorique de la contribution scalaire au redshift total. Comme toute théorie métrique de la gravité, les théories tenseur-scalaires satisfont à l'*invariance de position locale*. Pour autant, on montre qu'il est possible de distinguer ces dernières de la RG dans des expériences basées sur le redshift. C'est l'objet de la section 6.2 où l'on esquisse une première expérience de pensée reposant sur l'emploi d'horloges plongées dans des milieux de densités distinctes et dont il serait possible de comparer les fréquences. Nous passons de plus en revue l'état de l'art en matière d'horloges atomiques de manière à pouvoir calculer des premiers ordres de grandeur sur les contraintes auxquelles on pourrait s'attendre sur le modèle caméléon via ce type d'expériences. Cela fournit des contraintes 'optimales' sur le modèle caméléon qui s'avère être compétitives avec l'état de l'art.

Ces résultats préliminaires nous encouragent à pousser plus loin le degré de réalisme d'une telle expérience dans la section 6.3. Dans le laboratoire, des horloges atomiques suffisamment modulaires permettraient de tester le modèle caméléon pour des couplages à la matière très grands ($\beta \gg 10^5$). Toutefois, la faisabilité d'un tel montage expérimental est discutable et l'écrantage des noyaux atomiques eux-mêmes, aux très grands couplages, pourrait s'avérer limitant.

Le recours aux missions spatiales avec horloge atomique embarquée est envisagé pour tester le caméléon à de plus faibles couplages, typiquement $\beta \leq 10^3$. En effet, la très faible densité qui règne aux hautes orbites est avantageux à deux égards. D'abord, cela favorise le non-écrantage du satellite, condition *sine qua non* pour que l'horloge embarquée 'voit' la valeur du champ scalaire dans ce presque-vide. D'autre part, cela permet de maximiser la contribution scalaire au redshift. La principale limitation identifiée dans ce concept de mission spatiale est la précision des meilleures horloges embarquées, encore quelques ordres de grandeur trop basse pour espérer obtenir des contraintes compétitives.

Annexes

Le présent manuscrit comprend cinq annexes. L'annexe A est relative à l'utilisation d'unités naturelles et aux conversions avec le système international d'unités. L'annexe B établit les connexions qu'il existe entre les théories tenseur-scalaires d'une part, et les modèles $f(R)$ ou la théorie de Kaluza–Klein en dimension cinq d'autre part. Dans l'annexe C, je m'intéresse aux questions d'existence et d'unicité des solutions aux EDPs semi-linéaires qui régissent la dynamique du champ caméléon et symmetron en régime stationnaire. Je m'appuie pour cela sur des résultats classiques de la littérature en analyse des EDPs. Dans l'annexe D, j'examine le comportement asymptotique de la solution radiale d'une équation de Klein–Gordon semi-linéaire où la valeur de la solution à l'infini est imposée. En particulier, je montre que sous certaines hypothèses, la dérivée radiale de la solution tend nécessairement vers zéro. Enfin, l'annexe E présente une *technique de projection* pour la résolution d'équations différentielles ordinaires avec conservation de l'énergie. Cette technique est utilisée dans l'étude du chapitre 5, lorsqu'on calcule la trajectoire d'un satellite en orbite en imposant la conservation de l'énergie mécanique.

Contents

General introduction

Gravitation is by far the weakest of all four known fundamental interactions, yet it is the one whose effects are perhaps the most appreciable in our experience of life on Earth. It is the common denominator to phenomena as diverse as the sensation of weight, tides, or the motion of celestial bodies in the Solar system, up to the way matter is scattered across the largest scales of the universe. A major milestone in the history of our understanding of gravity was reached in the XVII$^{\text{th}}$ century when Isaac Newton, upon realizing that all these seemingly unrelated phenomena could be explained by a same force, formulated his famous *law of universal gravitation*. This so-called 'inverse-square law' provided the first comprehensive mathematical description of gravity as an attractive force between massive bodies, acting instantaneously in vacuum across distances. The advent of modern physics in the early XX$^{\text{th}}$ century, most notably the Michelson–Morley experiment and the discovery of the 'anomalous' perihelion precession of Mercury, led Albert Einstein to rethink the very concepts of space and time, culminating in the formulation of his *theory of General Relativity* (GR) in 1915. GR describes gravity not as a force but as the curvature of spacetime, itself caused by the mere presence of mass and energy, and in this sense constitutes a genuine paradigm shift with respect to Newton's law.

This truly groundbreaking theory was initially met with skepticism, but its acceptance grew among the scientific community in the subsequent decades following its publication, as numerous observations and experiments confirmed its predictions, one by one. Still today, GR remains our best understanding of gravity, consistently passing all experimental tests thrown at it with flying colors. The latter include the so-called 'classical tests' — namely the perihelion precession of Mercury's orbit, the deflection of light by the Sun and the gravitational redshift of light — and more generally all the post-Newtonian tests of gravity. More recently, the direct detections of gravitational waves and imaging of black holes further consolidated GR. Today, it is fully integrated into the mainstream of physics. Especially, it underlies the contemporary *standard model of cosmology*, which currently constitutes the most satisfactory description of the history of the universe we live in.

Despite being one of the most tried and tested theories in all physics, GR is most certainly not the final word on gravity. New questions always arise as science progresses along the endless road towards a deeper understanding of the laws of nature. As far as GR is concerned, most unresolved conundra revolve around the difficulty of marrying it with *quantum mechanics* and the universe's *dark sector*. The latter refers to *dark matter* and *dark energy* namely, two ingredients of unknown nature which nevertheless must be included in the universe's total mass-energy budget, for otherwise astronomical and cosmological observations would not be consistent with GR. Even under the hypothesis that our universe is 95 % dark,[6] the standard model of cosmology is plagued with several *tensions* — inconsistencies between the values of some cosmological parameters obtained with different datasets. The so-called *Hubble tension* and S_8 *tension* are perhaps the two most famous examples of such hot topics of modern cosmology.

In face of these challenges, it seems legitimate to examine alternative models to GR — also known as *modified gravity* models — which do not suffer from the aforementioned issues. Actually, even beyond the pragmatic purpose of finding a model that will better fit our observations, exploring alternatives to GR is interesting in itself. Not only does that help tighten the noose on possible alternative theories, it is also a good way to pin down which physical effects are specific to GR and which are not. If two models perform equally well, Occam's razor can ultimately be invoked to pick the simplest one.

Now, if the study of modified gravity models is well motivated, what does it mean exactly to 'modify' gravity? There is not a single answer to this question — in fact, the landscape of alternative theories to GR is so broad that it would be both futile and unenlightening to try to be exhaustive in listing them here. Because GR works so well, most alternative models do not start from scratch but rather build on top of it, in various manners. A convenient framework for classifying those is to relax one or several hypotheses of *Lovelock's theorem* — a mathematical result providing a set of conditions under which Einstein's field equations are the only viable gravitational field equations. This includes, but is not limited to, the addition of new field contents involved in mediating the gravitational force, the addition of extra spatial dimensions or allowing the presence of greater-than-second-derivative terms in equations of motion.

Given physical and mathematical tools we have at our disposal (field theory, action principle, differential

[6] Dark energy and dark matter must account for 68 % and 27 % of this total mass-energy budget, respectively.

geometry, etc.), we dare to say that it is relatively easy to construct a mathematically consistent gravitational theory, on paper. It is only when one starts doing *physics* that the real difficulties come in. Indeed, to be deemed relevant, any alternative model must not only fix one (or several) of the identified shortcomings of GR, but also remain consistent with known physics at the same time. This endeavor is all the more challenging for theoretical physicists as GR is tightly constrained by a myriad of tests by now.

One of the most natural and resilient extensions of GR are *scalar-tensor* theories of gravity. In this class of models, gravity is mediated by both a rank-2 tensor field (the metric) and a scalar field. The addition of this extra scalar degree of freedom in the gravitational sector, compared to GR, allows for a wide range of different phenomenologies. The scalar field can indeed be made to play various roles depending on the underlying physical motivations — from driving the universe's accelerated expansion to dark matter candidates. Light spin-0 particles also naturally arise in more fundamental theoretical contexts, e.g. in the low energy limit of string theories.

One cannot introduce scalar-tensor theories of gravity without mentioning *fifth forces*, a term coined in the 1980s by E. Fischbach to designate an additional putative force that would extend beyond the four fundamental interactions known in physics. As a matter of fact, when the scalar field is coupled to the matter sector (at the level of the action),[7] it automatically gives rise to a gravity-like fifth force, resulting in deviations from GR in the predicted outcome of gravitational phenomena. This leads to a genuine theoretical deadlock: natural couplings to matter[8] would inevitably result in violations of the known bounds on the existence of fifth forces in the Solar system or in the laboratory. In that respect, many scalar-tensor models went extinct in the second half of the XX$^{\text{th}}$ century as the accuracy and precision of our tests improved, in conjunction with the development of the powerful *parameterized post-Newtonian formalism* in the early 1970s.

Scalar-tensor theories were given a new lease of life with the subsequent development of *screening mechanisms* — theoretical constructs cleverly engineered to hide (or 'screen') the effects of the scalar-mediated fifth force in Earth-based and Solar system experiments, while allowing for larger deviations from GR at astrophysical and cosmological scales (where gravity is much less constrained). In models featuring screening mechanisms, the scalar field has to dynamically adapt its properties from one place to another — like its mass, which relates to the fifth force range (e.g. chameleon model), or its coupling to matter (e.g. symmetron model) — in order to evade Solar system bounds. Yet, however efficient the mechanisms, fifth forces are never totally suppressed. The scalar field still leaves an imprint on gravity which, depending on its magnitude, may be detected.

Ultimately, it is the confrontation between the theoretical predictions of a given model on the one hand, and experiment on the other hand, which allows conclusions to be drawn. Specifically, if the model turns out to be inconsistent with the data, it is henceforth ruled-out. We stress that performing this kind of analysis is crucially dependent on our ability to determine how the scalar field, together with its associated fifth force, behave in a given experimental setup. In models featuring screening mechanisms, accessing the fifth force — which directly derives from the scalar field itself — is partly impeded by a number of difficulties. At the level of the scalar field equation, screening mechanisms all have in common the fact that they are enabled through *nonlinearities*. Nonlinear partial differential equations (PDEs) are generally more challenging to work with than linear ones. For one thing, analytical approaches are more restricted in scope, and solutions (provided they exist!) behave in a somewhat less predictable way. While reasonable approximations might be obtained for simple configurations, spherical symmetry and its cousins are a far-off dream in real experimental setups. Yet, it has been shown in previous studies that the testability of such theories was highly dependent on the environment in which the experiment takes place. Constraining scalar-tensor models with screening mechanisms therefore requires accurate solutions to the PDEs at stake — simple approximations being, more often than not, unsatisfactory in this regard.

The necessity to solve nonlinear PDEs where mass distribution — which acts as a source term — can be very complex, calls for the use of numerical techniques. The finite element method (FEM) appears to be particularly well suited in this regard. On the one hand, it hinges on meshes (tessellations composed of simple cells) that can fit virtually any given geometry. On the other hand, while it is generally first taught on linear problems, its framework can readily be extended to the realm of nonlinear problems by means of iterative techniques. To form a well-posed problem, any given PDE must be supplemented with a set of suitable *boundary conditions* (as well as initial conditions when dealing with time-dependent problems). Usually, this is done by fixing the value of the unknown at the boundary of the simulation box. From a numerical perspective, this seemingly harmless fact can become a major obstacle in obtaining physically meaningful solutions to the model equations. The problem is most easily seen when trying to solve the Poisson equation governing the Newtonian potential. The value taken by the potential in the immediate vicinity of the sources is a priori not known. Rather, we demand that physical solutions should steadily decay to zero infinitely far away from the sources. By the same token, the behavior of scalar fields is usually known *asymptotically* at best. The issue with this mere observation is that such asymptotic boundary conditions appear to be at odds with FEM, where meshes cannot extend up to

[7] This is expected to be the case even if no such coupling term appears in the Lagrangian. Indeed, from a quantum mechanical perspective, the introduction of a scalar field in the gravity sector *always* generates interactions between this scalar and matter fields.

[8] Among other things, couplings that make the scalar-tensor model at stake physically relevant in addressing GR's shortcomings.

infinity.

The first goal of my PhD thesis is to develop a versatile numerical tool based on FEM for solving the PDE problems that arise in the study of scalar-tensor theories of gravity — no such tool being publicly available at its inception. As underlined above, sometimes asymptotic conditions are the only physical property that can be stated about the unknown fields. In this respect, the tool has to ensure these asymptotic conditions are correctly taken into account, for otherwise the numerical solutions could turn out to be, slightly off at best, or *nonphysical* at worst. Obtaining trustworthy solutions is an ongoing source of concern throughout this work.[9] We shall therefore lay emphasis on the well-posedness character of the PDE problems whe aim at tackling numerically. In a second phase we use this numerical tool to tackle problems which, prior to this PhD work, could not be wholly addressed. Specifically, we conduct a thorough study of the motion of spacecraft in orbit around the Earth in the framework of the chameleon as the prototypical screening mechanism for scalar-tensor models. This allows us to assess quantitatively the testability of such a model with space geodesy. A third objective is to analyze the possibility of testing scalar-tensor theories with screening mechanisms via redshift measurements, which is fundamentally different from what is done in fifth force searches.

The structure of this manuscript is as follows. Chapter 1 serves as an introduction to the physics of the problem. Namely, we present the class of scalar-tensor theories of gravity and shed light on the various ways in which they extend GR — both mathematically and phenomenologically. We also review the legacy of the MICROSCOPE space mission in order to comment on the relevance of space-based experiments for testing scalar-tensor theories, which provides further insight for establishing a precise list of specifications for the envisioned numerical tool.

Chapter 2 lays the foundations of the finite element method for solving elliptic PDE problems posed on bounded domains, including the nonlinear case. We do not lose sight of the physical problem of interest as FEM techniques are illustrated either with the Poisson equation governing the Newtonian potential, or with the nonlinear Klein–Gordon equation driving the dynamics of the chameleon scalar field. The way time-dependent problems might be eventually handled is discussed here.

Chapter 3 is a further excursion into applied mathematics. It builds on top of the previous chapter in order to adapt the FEM framework to the case of unbounded domains, which is a necessary step for being able to solve the physical problem of interest. Guided by both theoretical and numerical considerations, we delve into techniques based on compactification transforms. Building on top of the so-called inverted finite element method (*ifem*) and a specific domain decomposition scheme, we propose a novel method for solving elliptic PDE problems on the whole space — the alternate inverted finite element method (*a-ifem*). Beyond establishing a firm mathematical ground for both *ifem* and *a-ifem*, we conduct several numerical experiments, notably to control their respective implementations and to study their rate of convergence.

The culmination of the first three chapters is the actual implementation of the numerical code, called *femtoscope*, which is presented in Chapter 4. This PYTHON code solves semi-linear elliptic PDEs on bounded and unbounded domains of $\mathbb{R}^3$, which encompasses the physical problems of interest. Specifically, we provide a clear overview of *femtoscope*: its architecture, its main features and limitations. Emphasis is laid on the points that make it stand out from other existing numerical tools. It is then showcased on the two main examples of interest — the Newtonian potential and the chameleon scalar field. In particular, we solve both the one-body and two-body problems in the modified gravity setting, which sparks several physical discussions.

Chapter 5 takes the one-body problem a step further compared to what is presented in the previous chapter. Its ultimate goal is to quantitatively assess the testability of fifth force effects in Earth orbit. To this end, we use *femtoscope* in order to model chameleon gravity, solving for both the Newtonian potential and the scalar field. In particular, numerical simulations allow us to go beyond the simplifying assumptions and modeling traditionally found in the literature. Building on these FEM computations, we study the dynamics of satellites in orbit around the Earth with and without the putative chameleonic force, which roughly amounts to comparing geodesics of the Einstein-frame metric *vs* those of the Jordan-frame metric, respectively. Given the level of precision achieved by recent space geodesy missions, we look whether it is possible to discriminate between the two in the presence of model uncertainties. This whole chapter was published in Physical Review D.

Finally, Chapter 6 puts forward a novel idea for testing screened scalar-tensor models exploiting the gravitational redshift (or equivalently, gravitational time-dilation). Building on the theoretical aspects laid out in Chapter 1, we derive the redshift expression in the framework of scalar-tensor models and single out the scalar contribution in the Newtonian limit. As in Chapter 5, we focus our discussion on the chameleon model. Unlike fifth force effects, which are mainly dependent on the magnitude of the gradient of the scalar field, it appears that the scalar contribution to the total redshift depends, for the most part, on the field's value itself. We then endeavor to show that precise redshift measurements could reveal the presence of the scalar field. For

[9] To us, this is all the more important as we are about to go to yet unexplored physical situations. For such numerical simulations where one has only limited insights into the 'expected' solution, it is crucial to be able to rely on some theoretical results from the fields of applied mathematics and PDE analysis (well-posedness, a priori error estimates, etc.).

this purpose, we imagine a thought experiment which guides us towards more realistic experimental setups, in the laboratory and in space. We conclude with some perspectives afterwards.

Chapter 1

Scalar-tensor theories of gravity

Outline of the current chapter

To quote from Will's book [1], *"Scalar-tensor theories have proven to be the most interesting, compelling and resilient of alternatives to general relativity."*. This first chapter is dedicated to their presentation, shedding light on their phenomenology and the various ways in which they extend General Relativity. We also review the recent results from the MICROSCOPE space mission and use this example for discussing the relevance of space-based experiments for testing scalar-tensor theories of gravity. In this perspective, it appears that numerical tools are part of the answer, which motivates their development.

1.1 A phenomenologically rich framework to go beyond General Relativity

The theory of General Relativity (GR) was proposed by Albert Einstein in 1915 [2] and is still our best understanding of gravity to this day. Unlike Newton's earlier inverse-square law, which depicted gravity as a force acting instantaneously in vacuum across distances, GR describes gravity as the manifestation of the curvature of spacetime caused by the presence of energy. So far, it has passed all the experimental tests thrown at it with flying colors [3] and underlies the contemporary standard model of cosmology [4]. Does that mean that *"there is nothing new to be discovered in physics now"*?[1] Most certainly not, not even in gravitational physics as there remains many unsolved conundra, most notably associated with *(i)* the accelerated expansion of the universe, *(ii)* the apparent presence of dark matter accounting for 85% of the total matter, or *(iii)* the challenge of reconciling GR with quantum mechanics. Whether they take the form of tensions in observations or originate from more theoretical grounds, these hints all point toward the same conclusion: GR is not the final word on gravity.

Consequently, it seems reasonable to explore alternative theories of gravity. Even beyond the pragmatic considerations laid out above, it is important to compare GR's predictions against those of alternative models. Not only does that help shed light on the features specific to GR, but it also contributes to the effort of narrowing down the landscape of viable models. The history of alternative theories dates back as early as the time when GR was being established; in that respect let us mention Nordström theories [5–7] (1912, 1913) which describe gravity by a scalar field in flat spacetime, Kaluza–Klein theory [8–10] (1920s) which is a generalization of GR in a 5-dimensional manifold that aims to encompass electromagnetism, and Weyl vector-metric theory [11] (1919). Over the past century, theoretical physicists have continued to propose new models, with various *raisons d'être*, which makes the field of alternative theories of gravity rich and complex. Because GR works so well, most alternative models do not start from scratch but are rather extensions of GR, as we shall see in this section. Irrespective of the path taken to extend Einstein's theory of gravity, such models have a destiny with few possible outcomes: they can end up being ruled-out by some experiments or observations, they can turn out to be indistinguishable from GR itself or even fail to produce verifiable predictions, they can also fail to spark interest among the scientific community and just fall into oblivion...

This PhD work is focused on scalar-tensor theories of gravity, although the word 'focus' is perhaps ill-chosen insofar they constitute a very wide area of active research on their own. Indeed, scalar-tensor theories have surely received more attention than other models owing to the fact that they are one of the most natural[2] extensions of GR, where gravity is mediated by both a rank-2 tensor field and a scalar field. Adding a new scalar degree of freedom in the gravitational sector is a phenomenologically rich idea, as the scalar field can be made to play different roles depending on the underlying physical motivations.

This section serves several purposes, the first one being to succinctly introduce the theory of general relativity together with its mathematical framework. Shedding light on the modern physics challenges GR faces motivates the introduction of scalar-tensor theories of gravity. After showing how such theories are constructed, we restrict our description to a particular subclass of models whose phenomenology is discussed.

1.1.1 General Relativity theory and challenges

Conventions

Before diving straight into the actual matter, let us set some conventions to be used consistently throughout this chapter. We work in natural units, for which $c = 1$ and $\hbar = 1$, where c is the speed of light and $\hbar$ is the reduced Planck constant. With such a choice, mass and energy have the same units, while length and time acquire the units of reciprocal energy. Consequently, any kinematical variable (i.e. not involving any other dimensions than M, L, T) can be expressed in powers of an arbitrarily chosen unit of energy. In this regard, we will use the electron-volt [eV]. Furthermore, it will be convenient to use a fixed energy scale in the subsequent computations. In this perspective, we define the reduced Planck mass

$$M_{\rm Pl} = \sqrt{\frac{\hbar c}{8\pi G}} \simeq 4.34 \times 10^{-9}\,\text{kg} \simeq 2.44 \times 10^{27}\,\text{eV}\,. \tag{1.1}$$

This specific choice allows for a smoother connection between GR and physical phenomena other than gravity. Last but not least, one may legitimately be surprised to see $\hbar$ appearing in a purely classical context. In fact, despite $\hbar$ being embedded in the definition of the reduced Planck mass (and thus, in its numerical value), it *is not really there* in the sense that it necessarily cancels out in all the equations to be written in this chapter.

[1] For some reason, this popular statement has been widely misattributed to Lord Kelvin since the 1980s.
[2] Here, it would certainly be overly reductive to translate 'natural' into 'simple'.

Appendix A provides further insights into the use of natural units, including practical considerations for switching back and forth between SI units and natural units.

We employ the Einstein summation convention. Greek indices (μ, ν, ρ, σ, etc.) run from 0 to 3 while Latin indices (i, j, k, etc.) run from 1 to 3. We further adopt the $(-,+,+,+)$ metric signature.

The theory of general relativity

Einstein's theory of general relativity is way more than a mere 'update' of Newton's law of universal gravitation. GR is a theory whereby gravitation is interpreted in terms of an elegant mathematical structure: the differential geometry of curved spacetime. To a certain extent, it is profoundly different from our modern description of other forces of nature, which are represented by quantum fields defined *on* spacetime. As such, spacetime is the stage on which physics plays out, while gravity is *inherent in* spacetime itself. The mathematical structure of spacetime is that of a 4-dimensional pseudo-Riemannian manifold equipped with a symmetric metric tensor denoted $g_{\mu\nu}$. This object is at the heart of the mathematical description of GR. Among its multiple roles, it notably allows for the computation of path lengths via the line element (or equivalently, via the proper time)

$$ds^2 = -\mathrm{d}\tau^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu \,. \tag{1.2}$$

In the above, the quantities $\mathrm{d}x^\mu$ are being regarded as the components of an infinitesimal coordinate displacement 4-vector, expressed in local coordinates x^μ, while τ is the proper time. In terms of units, $\mathrm{d}x^\mu$ has the dimension of a length and is thus expressed in eV^{-1}. Components of the metric tensor being dimensionless, the line element and proper time [Eq. (1.2)] have units of eV^{-2}.

From a physical point of view, GR should provide answers to, at the very least, two questions. On the one hand, spacetime is not empty but filled with matter and energy (galaxies, dark matter halos, electromagnetic field, etc.). GR must somehow describe how such fields evolve through spacetime. In the language of classical mechanics, we need an equivalent of

$$\mathbf{a} = -\boldsymbol{\nabla}\Phi \tag{1.3}$$

for the 3-acceleration experienced by a massive test body in a gravitational potential Φ. Here $\boldsymbol{\nabla}$ denotes the gradient operator in flat space, i.e. $\boldsymbol{\nabla} = (\partial_i)_{1\leq i\leq 3}$ in Cartesian coordinates. Conversely, the mere presence of matter and energy in spacetime is what causes it to be curved, which also has to be expressed mathematically. Following the classical mechanics analogy, we need a general relativistic equivalent for the Poisson equation

$$\Delta\Phi = 4\pi G\rho \,, \tag{1.4}$$

whereby the gravitational potential Φ is sourced by the matter density ρ. Here Δ denotes the Laplace operator in flat space, i.e. $\Delta = \delta^{ij}\partial_i\partial_j$ in Cartesian coordinates. As it turns out, these two questions are two sides of the same coin.

Following J. A. Wheeler famous quote,[3] let us first examine how "spacetime tells matter how to move", and be a bit more inclusive by considering massless particles as well. The answer is given by the geodesic equation

$$\frac{\mathrm{d}^2x^\mu}{\mathrm{d}\lambda^2} + \Gamma^\mu_{\rho\sigma}\frac{\mathrm{d}x^\rho}{\mathrm{d}\lambda}\frac{\mathrm{d}x^\sigma}{\mathrm{d}\lambda} = 0 \,, \tag{1.5}$$

where λ parameterizes the geodesic curve $x^\mu(\lambda)$ and $\Gamma^\mu_{\rho\sigma}$ are the Christoffel symbols of the second kind, given in terms of the metric by

$$\Gamma^\alpha_{\mu\nu} = \frac{1}{2}g^{\alpha\sigma}(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\sigma\mu} - \partial_\sigma g_{\mu\nu}) \,. \tag{1.6}$$

Note that having the rhs of Eq. (1.5) equal to zero actually constrains λ to be an *affine parameter* (see e.g. Carroll's book [12] Chapt. 3.4 or Wald's book [13] Chapt. 3.3). There are several ways to obtain Eq. (1.5), which are all insightful in their own respect. From a rather geometrical point of view, a geodesic can be interpreted either as a curve along which the tangent vector is parallel-transported, or as the timelike path that maximizes the proper time between two timelike-separated events (see e.g. Ref. [12] Chapt. 3.3). It can also be derived in a more physical way by invoking the equivalence principle (see e.g. Ref. [14]). Test particles — massive or massless — travel along such curves in spacetime provided they are *freely falling*, i.e. not subjected to any interaction but gravity. For massive test particles, it is possible to make the proper time τ play the role of the affine parameter,[4] and the geodesic equation can be rewritten in terms of the 4-velocity $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau$ as

$$\frac{\mathrm{d}^2u^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{\rho\sigma}u^\rho u^\sigma = u^\alpha\nabla_\alpha u^\mu = 0 \,, \qquad \text{with} \quad g_{\alpha\beta}u^\alpha u^\beta = -1 \,. \tag{1.7}$$

[3] *"Spacetime tells matter how to move; matter tells spacetime how to curve."* — John Archibald Wheeler.

[4] For timelike geodesics, affine parameters are of the form $a\tau + b$, a, $b \in \mathbb{R}$.

We have made use of the covariant derivative operator ∇_α, which acts on u^μ (and more generally on any vector) as $\nabla_\alpha u^\mu = \partial_\alpha u^\mu + \Gamma^\mu_{\alpha\beta} u^\beta$. Eq. (1.7) conveys more clearly the idea that freely falling particles move in the direction in which their 4-velocity vector (or 4-momentum vector $p^\mu = m u^\mu$) is pointing. Were that test particle to be submitted to additional non-gravitational forces, the rhs of Eq. (1.7) would be non-zero. For instance, Eq. (1.7) for a particle of charge q would read

$$u^\alpha \nabla_\alpha u^\mu = \frac{q}{m} F^\mu{}_\nu \frac{\mathrm{d}x^\nu}{\mathrm{d}\tau}\,, \tag{1.8}$$

where $F_{\mu\nu}$ is the electromagnetic tensor. Massless particles on the other hand travel along null paths of spacetime for which $\mathrm{d}\tau = 0$ which means that the proper time cannot be used as an affine parameter. Although there is no real preferred choice for λ, it is sometimes handy to normalize it so that $\mathrm{d}x^\mu/\mathrm{d}\lambda$ is equal to the 4-wavevector k^μ, in which case Eq. (1.5) becomes

$$\frac{\mathrm{d}k^\mu}{\mathrm{d}\lambda} + \Gamma^\mu_{\rho\sigma} k^\rho k^\sigma = k^\alpha \nabla_\alpha k^\mu = 0\,, \qquad \text{with} \quad g_{\alpha\beta} k^\alpha k^\beta = 0\,. \tag{1.9}$$

At this stage, we have to admit that the link between Eq. (1.3) and Eq. (1.7) is still a bit obscure. It will be made clearer when we examine the so-called Newtonian limit of the former equation.

The other side of the coin, that is how "matter tells spacetime how to curve" is the hard part of GR and is known as the Einstein's field equations. There are different routes that lead to them. The way Einstein himself derived them arguably involved proceeding by trial-and-error at certain stages in the development of the theory, but always with powerful guiding principles in mind: the equivalence principle, the principle of general covariance (the laws of physics should appear the same to all observers), the conservation of energy. Here, we do not go down this historical yet rather long road and make the deliberate choice to go through the Lagrangian formulation instead. Just like for other classical field theories (in flat spacetime), we formulate the theory in terms of an action, and derive the field's equations by applying the principle of least action. The action of GR, which Hilbert was the first to figure out, is

$$S = S_{\mathrm{EH}} + S_{\mathrm{mat}} \tag{1.10}$$

where

$$S_{\mathrm{EH}} = \frac{M_{\mathrm{Pl}}^2}{2} \int \mathrm{d}^4x \sqrt{-g} R\,, \tag{1.11a}$$

$$S_{\mathrm{mat}} = \int \mathrm{d}^4x \sqrt{-g} \mathcal{L}_{\mathrm{mat}}\Big(g_{\mu\nu}, \psi^{(i)}_{\mathrm{mat}}\Big)\,. \tag{1.11b}$$

Eq. (1.11a) is the Einstein–Hilbert action (without cosmological constant), featuring the determinant of the metric tensor $g = \det(g_{\mu\nu})$, and the Ricci scalar R constructed from $g_{\mu\nu}$. On the other hand, Eq. (1.11b) defines the action for matter. In this expression, $\psi^{(i)}_{\mathrm{mat}}$ denotes the matter fields (labelled by i), $\sqrt{-g}\mathcal{L}_{\mathrm{mat}}$ is a Lagrange density and $\mathcal{L}_{\mathrm{mat}}$ alone is a scalar. Note that the latter is assumed to be independent of the derivatives of the metric. The field's equations are obtained by applying the stationary-action principle, i.e. by finding the stationary points of S with respect to the metric. A fairly standard way to proceed consists in varying the action S with respect to the inverse metric[5] $g^{\mu\nu} \to g^{\mu\nu} + \delta g^{\mu\nu}$, yielding δS, and demanding that δS vanishes for any $\delta g^{\mu\nu}$. Doing so with some amount of care results in the sought equation

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = \frac{1}{M_{\mathrm{Pl}}^2} T_{\mu\nu}\,. \tag{1.12}$$

In the above, $R_{\mu\nu}$ is the Ricci tensor and $T_{\mu\nu}$ denotes the energy-momentum tensor (also called the stress-energy tensor), which is defined by

$$T_{\mu\nu} = \frac{-2}{\sqrt{-g}} \frac{\delta S_{\mathrm{mat}}}{\delta g^{\mu\nu}} = \frac{-2}{\sqrt{-g}} \frac{\delta(\sqrt{-g}\mathcal{L}_{\mathrm{mat}})}{\delta g^{\mu\nu}} \implies T^{\mu\nu} = \frac{2}{\sqrt{-g}} \frac{\delta(\sqrt{-g}\mathcal{L}_{\mathrm{mat}})}{\delta g_{\mu\nu}}\,, \tag{1.13}$$

where $\delta S_{\mathrm{mat}}/\delta g^{\mu\nu}$ is the functional derivative of S_{mat} with respect to $g^{\mu\nu}$. This derivation can be found in e.g. Ref. [12] Chapt. 4.3, or Ref. [15] which comes with many insightful remarks (in French).

Now that the most important equations are written out, some comments are in order. At first sight, the Lagrangian formulation Eqs. (1.10–1.11) can seem to appear out of the blue. As elements of a response, let us stress that R is the only independent scalar which can be constructed from no higher than second order derivatives of the metric. As such, the action Eq. (1.10) is practically the unique way to end up with covariant, second order equations of motion that link spacetime curvature to the mass-energy content contained in it

[5] The stationary points of S with respect to variations in $g_{\mu\nu}$ are equivalent to those with respect to variations in $g^{\mu\nu}$. Nonetheless, computations are slightly more handy when using the inverse metric.

and respect the energy conservation for matter $\nabla_\mu T^{\mu\nu} = 0$. Once we acknowledge this, we have a powerful formalism at our disposal with several advantages: *(i)* the Lagrangian is a scalar (simpler object than, say, rank-2 tensors) on which the symmetries of the theory can be easily read / imposed, and *(ii)* it offers a very convenient framework for studying beyond-GR models. Aside these technical remarks, let us remind that Eq. (1.12) is the general relativistic generalization of Poisson equation (1.4). This will also be made clearer when deriving its Newtonian limit.

Newtonian limit

The Newtonian limit is the combination of three approximations under which GR's equations (1.5, 1.12) boil down to Newton's equations (1.3, 1.4) respectively. Namely, these approximations are:

1. the gravitational field is weak, in the sense that it can be considered as a small perturbation of flat spacetime;
2. it is unchanging with time;
3. objects are moving slowly compared to the speed of light.

Applying this set of approximations to Eqs. (1.5, 1.12) will provide valuable insights into the way GR encompasses Newtonian physics.

Geodesic equation Let us start with the geodesic equation (1.7) for massive particles. We use a spacetime coordinate system $\{x^\mu\}$ where an event is specified by one time coordinate t and three spatial coordinates $\mathbf{x} = \{x^1, x^2, x^3\}$ with $\mathrm{d}x^0 = \mathrm{d}t$. Hypothesis 3 is written mathematically as

$$v^i \equiv \frac{\mathrm{d}x^i}{\mathrm{d}t} \ll 1 \quad \text{so that} \quad \frac{\mathrm{d}x^i}{\mathrm{d}\tau} \ll \frac{\mathrm{d}t}{\mathrm{d}\tau} \iff u^i \ll u^0 \,. \tag{1.14}$$

As a consequence, the geodesic equation simplifies to

$$\frac{\mathrm{d}^2 x^\mu}{\mathrm{d}\tau^2} + \Gamma^\mu_{00} \left(\frac{\mathrm{d}t}{\mathrm{d}\tau}\right)^2 = 0 \,. \tag{1.15}$$

Hypothesis 2 means that partial derivatives of the metric with respect to the time coordinate are null, while hypothesis 1 allows the decomposition of the metric into the Minkowski metric $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$ plus a small perturbation as $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, with $|h_{\mu\nu}| \ll 1$. At first order in h, the connection component appearing in Eq. (1.15) can be approximated as

$$\Gamma^\mu_{00} = -\frac{1}{2}\eta^{\mu\sigma}\partial_\sigma h_{00} \,. \tag{1.16}$$

The zeroth component of Eq. (1.15) provides $\mathrm{d}t/\mathrm{d}\tau =$ constant. As for the three other components, we end up with

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = \frac{1}{2}\partial_i h_{00} \,. \tag{1.17}$$

We have finally arrived at a form that is quite reminiscent of Eq. (1.3). All that is left is to perform the identification $h_{00} = -2\Phi$.

Field equations We have shown that, provided the 00-component of the metric tensor can be written in the form $g_{00} = -(1 + 2\Phi)$ with $|\Phi| \ll 1$, the Newtonian limit of the geodesic equation for massive particles gives the expected form for the 3-acceleration (1.17). However, there remains to check that the field equations (1.12) together with the three hypotheses listed above actually lead to the form $g_{00} = -(1 + 2\Phi)$. Before going any further, we recast Eq. (1.12) in the so-called *trace-reversed* form

$$R_{\mu\nu} = \frac{1}{M_{\rm Pl}^2}\left(T_{\mu\nu} - \frac{1}{2}T g_{\mu\nu}\right) , \tag{1.18}$$

which comes from the fact that contracting both sides of Eq. (1.12) with $g^{\mu\nu}$ yields $M_{\rm Pl}^2 R = T$, where T is the trace of the stress-energy tensor. We consider a perfect-fluid with 4-velocity u^μ source of energy-momentum, for which

$$T_{\mu\nu} = (\rho + p)u_\mu u_\nu + p g_{\mu\nu} \,. \tag{1.19}$$

A perfect fluid is a fluid that can be entirely characterized by its rest frame energy density ρ, and isotropic pressure p (also called the momentum density). Nevertheless, by virtue of the assumption that the fluid's particles are moving slowly with respect to the speed of light, pressure is negligible and the stress-energy tensor

is well approximated by $T_{\mu\nu} = \rho\, u_\mu u_\nu$. We can do even better. If the 'fluid' we are considering represents some rigid body (the particles constituting the body do not move with respect to each other) — in plain language, a planet for instance — we can choose a coordinate system $\{x^\mu\}$ attached to the body such that the 4-velocity reads $u^\mu = u^0 \delta^\mu_0$. Now, the weak field approximation implies that ρ is already small (spacetime is almost flat). Consequently, it is legitimate to do the zeroth-order approximation $u^0 = -u_0 = 1$, so that the only non-zero component of the stress-energy tensor is $T_{00} = \rho$ and $T = g^{00} T_{00} = -\rho$. So far, Eq. (1.18) takes the simplified form $2M_{\rm Pl}^2 R_{00} = \rho$. Using hypotheses 1 and 2, the 00-component of the Ricci tensor simplifies to $2R_{00} = -\Delta h_{00}$, where $\Delta = \delta^{ij}\partial_i\partial_j$ is the usual Laplace operator in flat space. The Newtonian limit of the field equations is thus one single equation

$$M_{\rm Pl}^2 \Delta h_{00} = -\rho\,. \tag{1.20}$$

Again, this form is highly reminiscent of Poisson equation (1.4). All we have to do is perform the identification $h_{00} = -2\Phi$, which is consistent with what we found above in the derivation of the Newtonian limit of the geodesic equation. The spatial part of the metric boils down to $g_{jk} \sim \delta_{jk}$, $g_{0k} \sim 0$.

Finally, the Newtonian limit justifies the factor $M_{\rm Pl}^2/2$ in the Einstein–Hilbert action (1.11a) and thereby fixes the only free parameter of the theory. GR is therefore a theory with no free parameters.

FLRW cosmology

As stated in the introduction of the present section, GR underlies the standard model of cosmology. Despite being our current best mathematical model to retrace the history of the universe, some simulations and observations challenge its validity to a certain extent. As such, it is no wonder that many alternative models to GR attempt at addressing these challenges — which will be explained at more length at the end of this sub-section. In order to appreciate why a given model might be more attractive than another from a cosmologist's point of view, it is relevant to briefly review what GR has to tell us about cosmology.

By considering the universe's content as a homogeneous and isotropic perfect fluid [with energy-momentum tensor given by Eq. (1.19)] on the largest spatial scales,[6] only evolving in time, GR provides the adequate framework to compute the evolution of such a universe. Indeed, this assumption allows us to decompose spacetime as $\mathbf{R} \times \Sigma$ (referred to as a *spacetime foliation*), where $\mathbf{R}$ is the time direction and Σ is a maximally symmetric 3-dimensional manifold. It can be shown (see e.g. Ref. [12], Chapt. 8.2) that the metric on spacetime can be put in the form

$$ds^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-\kappa r^2} + r^2 \mathrm{d}\Omega^2\right], \qquad \kappa \in \{-1,\, 0,\, +1\}\,. \tag{1.21}$$

This metric is famously known as the Friedmann–Lemaître–Robertson–Walker metric, or FLRW metric for short, and is the generic metric that meets the conditions of spatial homogeneity and isotropy. Here, the time coordinate t is called the *cosmic time*, and corresponds to the proper time that would be measured by clocks at rest in the Hubble flow, i.e. at constant spatial coordinates x^i. The spatial part of the metric (between square brackets) is expressed in spherical coordinates where r is the radial coordinate while $\mathrm{d}\Omega^2$ is the usual metric on the two-sphere. It is weighted by the square of the so-called scale factor $a(t)$, which is itself a function of the cosmic time. Note that the scale factor is the only dynamical variable in the FLRW metric. The proper distance between two comoving observers evolves as $d(t) = a(t) d_0$, where d_0 refers to the proper distance at some reference time t_0. Finally, $\kappa \in \{-1,\, 0,\, +1\}$ is the only discrete free parameter and maps to three distinct topologies for the universe: negative, zero and positive curvature on Σ respectively.

Plugging the FLRW metric into the field equations (1.12) yields the Friedmann equations

$$H^2 \equiv \left(\frac{\dot a}{a}\right)^2 = \frac{\rho}{3M_{\rm Pl}^2} - \frac{\kappa}{a^2}\,, \tag{1.22a}$$

$$\frac{\ddot a}{a} = -\frac{\rho + 3p}{6M_{\rm Pl}^2}\,, \tag{1.22b}$$

where H is the Hubble parameter. The conservation of energy $\nabla_\mu T^{\mu\nu} = 0$ gives the continuity equation

$$\dot\rho + 3H(\rho + p) = 0\,. \tag{1.23}$$

Note that the latter does not constitute a new independent equation as it can be obtained from the two Friedmann equations (1.22). For a fluid which has an equation of state $p = w\rho$ (w being a constant), this continuity equation can be integrated and provide the insightful relation

$$\rho \propto a^{-3(1+w)}\,, \tag{1.24}$$

which describes the dilution of the various forms of energy in an expanding universe. Non-relativistic matter has essentially zero pressure so that $w_{\rm mat} = 0$. The equation of state for radiation can be obtained by looking at

[6] This is more or less the *cosmological principle*.

Parameter	Symbol	Value	Units	Evidence
Radiation	$\Omega_{\mathrm{r},0}$	$\sim 9\times10^{-5}$	—	CMB temperature
Baryonic matter	$\Omega_{\mathrm{b},0}$	~ 0.05	—	CMB measurements
Dark matter	$\Omega_{\mathrm{dm},0}$	~ 0.27	—	CMB measurements
Curvature	$\Omega_{K,0}$	~ 0	—	CMB anisotropy
Vacuum	$\Omega_{\mathrm{vac},0}$ or $\Omega_{\Lambda,0}$	~ 0.68	—	Cosmic acceleration
Hubble constant	H_0	~ 70	km/s/Mpc	
Cosmological constant	Λ	$\sim 1.1\times10^{-52}$	m^{-2}	Cosmic acceleration

Table 1.1: Values of the main parameters of the ΛCDM model [4, 17]. They are purposely given with few significant digits and without their associated uncertainties as these quantities depend on the cosmological survey being used.

the specific form of the stress-energy tensor for electromagnetism (which involves the field strength $F^{\mu\nu}$) and reads $w_{\mathrm{rad}} = 1/3$. Finally, vacuum energy — an energy density characteristic of empty space — also takes the form of a perfect fluid, with energy-momentum tensor $T^{(\mathrm{vac})}_{\mu\nu} = -\rho_{\mathrm{vac}} g_{\mu\nu}$ and equation of state $\rho_{\mathrm{vac}} = -p_{\mathrm{vac}}$. This specific form of the energy-momentum tensor for vacuum energy allows one to equivalently recast the field equations (1.12) with a *cosmological constant* Λ as

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{1}{M_{\mathrm{Pl}}^2}\hat{T}_{\mu\nu}\,. \tag{1.25}$$

For the equivalence to hold, we demand that $\hat{T}_{\mu\nu}$ account for all forms of energy but vacuum and set $\Lambda = \rho_{\mathrm{vac}}/M_{\mathrm{Pl}}^2$. At the action level, the Lagrange density of the Einstein–Hilbert action Eq. (1.11a) is simply replaced by $M_{\mathrm{Pl}}^2\sqrt{-g}(R-2\Lambda)/2$. When vacuum energy is accounted for in the lhs of the field equations as in Eq. (1.25), it is common to use the term 'cosmological constant'; whereas we employ the term 'vacuum energy' when it is implicitly included in the stress-energy tensor in the rhs. At last, it is convenient to introduce the density parameters

$$\Omega_i = \frac{\rho_i}{3H^2M_{\mathrm{Pl}}^2} \qquad \text{and} \qquad \Omega_K = -\frac{\kappa}{H^2a^2} \tag{1.26}$$

for the fluids (labeled by i) and for the spatial curvature respectively. By doing so, the first Friedmann equation (1.22a) can be rearranged as

$$H^2 = H_0^2\big(\Omega_{\mathrm{r},0}a^{-4} + \Omega_{\mathrm{mat},0}a^{-3} + \Omega_{K,0}a^{-2} + \Omega_{\mathrm{vac},0}\big) \qquad \text{with} \quad \Omega_{K,0} = 1 - \sum_i \Omega_{i,0}\,, \tag{1.27}$$

where the zero subscript indicates that the quantities are taken at $t = t_0$ which we set to 'now' by convention. Eq. (1.27) is a convenient (yet idealized) way to describe the expansion history of our universe using four independent constants — the density parameters at present time.

While a lot more could be said about FLRW cosmology, we do not take our discussion much further than this since we are now in a position to outline our current best cosmological model: the ΛCDM model. It postulates the existence of two additional ingredients that supplement baryonic (ordinary) matter and radiation, namely cold dark matter (abbreviated CDM) and a cosmological constant Λ which has already been introduced. From there, turning the "Ω-knobs" in Eq. (1.26) to the right values results in a history for the scale factor that is in fair agreement with our various observations of the sky. Most contemporary methods are consistent with the values provided in Table 1.1. In particular, let us note that, while the global topology of the universe is unknown, we have evidence that it is spatially flat,[7] i.e. $\Omega_0^c \sim 0$.

The reason why the ΛCDM model has been adopted as the standard model of cosmology is at least twofold. On the one hand, it has the ability to explain and predict a wide range of observed phenomena in the universe (CMB fluctuations, large-scale structure formation, cosmic acceleration, etc.) using a restricted number of parameters. This last point is worth emphasizing as one could easily construct a cosmological model with additional *ad hoc* parameters. While such a model could manifestly better fit observations, that would come at the cost of *(i)* further obscuring its underlying physical principles, *(ii)* loosing interpretability, and *(iii)* introducing degeneracies, making it difficult to constrain the values of individual parameters accurately. On the other hand, the ΛCDM model builds on top of GR which, at the risk of repeating ourselves, is one of the most thoroughly tested theory in all physics.

[7] This is suggested by WMAP, BOOMERanG, or Planck data. See e.g. Ref. [16] for a review.

A century-old theory still standing

Physicists widely regard GR as one of the most elegant theories in physics [18]. Yet, however 'beautiful' the mathematics and deep ideas underlying GR may be, elegance alone is not enough make it a 'good' physical theory. Among the necessary criteria to be deemed so, it must possess the ability to make testable predictions which in turn must be accurate. Here, we provide evidence that GR actually checks the latter criterion by briefly reviewing the most up-to-date tests. For the paragraphs to come not to be a long unordered list, GR's tests are grouped in four categories: axioms' tests, parameterized post-Newtonian bounds, gravitational waves and strong-field regime tests — with of course some unavoidable overlaps.

Axioms' tests By axioms' tests, we are referring to those experiments that closely examine the very foundations of GR. As any other metric theory, GR embodies the Einstein Equivalence Principle (EEP) which has three pillars:

1. All uncharged, freely falling test particles follow the same trajectories, once an initial position and velocity have been prescribed — this is the Weak Equivalence Principle (WEP), also known as the universality of free fall;

2. The outcome of any local nongravitational test experiment is independent of the velocity of the (freely falling) apparatus — this is Local Lorentz Invariance (LLI);

3. The outcome of any local nongravitational test experiment is independent of where and when in the universe it is performed — this is Local Position Invariance (LPI).

Put another way, the WEP states that the inertial mass $m_{\rm I}$ of a given test body is equal to its gravitational mass $m_{\rm G}$.[8] Considering two test bodies labeled by the index $i \in \{1, 2\}$, we generally parameterize violations of the WEP using the so-called Eötvös parameter

$$\eta_{1,2} \equiv 2\frac{|a_1 - a_2|}{|a_1 + a_2|} = 2\left|\frac{m_{\rm G}^1}{m_{\rm I}^1} - \frac{m_{\rm G}^2}{m_{\rm I}^2}\right| \left(\frac{m_{\rm G}^1}{m_{\rm I}^1} + \frac{m_{\rm G}^2}{m_{\rm I}^2}\right)^{-1} , \tag{1.28}$$

where a_i denotes the acceleration of the body i and the expression involving masses assumes an extended Newtonian framework. The WEP then holds if and only if $\eta_{i,j} \equiv 0$, for all pairs of bodies (i, j), regardless of their mass or composition. Recent bounds on the Eötvös parameter include torsion-balance tests led by the Eöt-Wash Group [19, 20] and Lunar laser ranging (LLR) measurements [21] which probe the free fall of the Earth and the Moon in the Sun's gravity field [22]. The best bound is currently held by the MICROSCOPE experiment with a precision of roughly one part in 10^{15} [23, 24] — this space mission is presented in more details in Sec. 1.3.

Similarly, modern experiments looking for LLI violations use the so-called 'c^2-formalism' (c is the speed of light here), introducing the dimensionless parameter

$$\delta_0 = \left|c^{-2} - 1\right| . \tag{1.29}$$

A slight violation of LLI would alter the speed of the electromagnetic interactions, leading to $\delta_0 \neq 0$ (recall that we work in natural units for which $c = 1$). However, most modern analyses now employ a different framework — the 'Standard Model Extension' (SME) — which extends the possibility of LLI violations to the entire standard model of particle physics.

Finally, tests of LPI split into two classes: gravitational redshift experiments and measurements of the constancy of the fundamental (non-gravitational) constants. The former also conveniently relies on parameterized violations of the form

$$z_{12} = (1 + \alpha)\Delta_{12}U \qquad \text{with} \quad \Delta_{12}U = U_2 - U_1 , \tag{1.30}$$

where U is the gravitational potential whose gradient is related to the acceleration of test bodies.[9] The most stringent bounds on α are set at the 10^{-5} level [25–27]. Such precise tests have been made possible in part by advances in atomic clocks and frequency standards over the past few decades. As a side note, we will come back to the parameterized form Eq. (1.30) in Chapt. 6 when we discuss redshift-based experiments in the framework of scalar-tensor theories. To put things into perspective, the history of WEP, LLI and LPI tests, parameterized by Eqs. (1.28–1.30), is depicted in Fig. 1.1.

The picture of axioms' tests would not be complete without mentioning tests of the strong equivalence principle (SEP). The SEP goes beyond the EEP in the sense that the universality of free fall is generalized to

[8] Actually, it would be more accurate to say that the ratio of the two masses $m_{\rm I}/m_{\rm G}$ is the same for all test bodies — physically measurable quantities are dimensionless ratios. But because the resulting proportionality constant can be absorbed in the the gravitational constant G, it is legitimate to equal the two masses.

[9] Note that we purposely do not use Φ, the Newtonian potential given by Eqs. (1.3–1.4), in the definition of the parameter α Eq. (1.30) as it is not a directly measurable quantity.

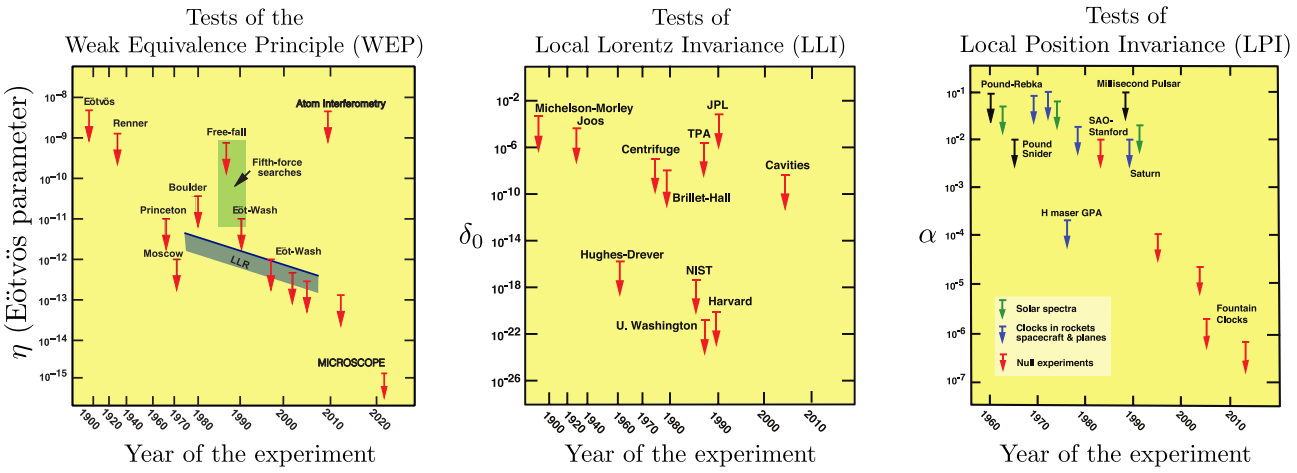


Figure 1.1: Evolution of upper bounds set on WEP, LLI, and LPI violations (from left to right). Figure adapted from Ref. [3].

extended bodies with gravitational self-energy (revision of statement 1 of the EEP above) and that it includes gravitational experiments (revision of statements 2 and 3). The SEP can be tested by *(i)* looking for the (non-)existence of the Nordtvedt effect [28–30], notably in LLR data [31], *(ii)* searching for time-variations of the gravitational constant G over the course of the universe's lifespan [21, 32, 33] or *(iii)* searching for variations and anisotropies in the locally-measured value of G owing to the existence of preferred frames and preferred locations. As a side remark, let us stress that GR, together with Nordström theory [6, 7], are the only known field theories verifying the SEP [34], for it is incompatible with extra fields supplementing the metric (and a fortiori fifth forces).

The parameterized post-Newtonian formalism Moving on to the tests of GR's predictions, it would be a serious omission not to begin by introducing the parameterized post-Newtonian (PPN) formalism — see e.g. Will's book [1]. This formalism applies not only to GR but nearly to any metric theory of gravity in the slow-motion, weak-field limit. For this class of gravitational theories, no matter what their Lagrangians, "matter responds only to the metric" [3] as highlighted by the geodesic equation (1.5), even if $g_{\mu\nu}$ is not the metric tensor from GR [i.e. the one given by the field equations (1.12)]. Under these assumptions, it is possible to expand the metric about the Minkowsky metric[10] $\eta_{\mu\nu}$ in terms of dimensionless gravitational potentials which are constructed from the matter variables (e.g. density, coordinate velocity, pressure, etc.). In this framework, any given metric theory is then characterized by the numerical value of the coefficients that weigh the metric potentials. In the canonical convention fixed by Ref. [35], there are 10 potentials and ten coefficients — the latter are called PPN parameters and are reported in Table 1.2. Once laid down, the PPN formalism provides an efficient way to test GR, by comparing the measured values of the PPN parameters to those predicted by GR. This encompasses all the 'classical' tests — deflection of light, Shapiro time delay, perihelion advance of Mercury — and other relativistic effects involving spinning bodies (e.g. Lense–Thirring precession), the de Sitter precession and tests of post-Newtonian conservation laws. Table 1.2 also provides current bounds on the PPN parameters. Note that, beyond enabling us to test GR, the PPN formalism has proved to be a powerful tool to 'kill' alternative theories owing to its agnostic approach with respect to gravity models. It is also worth noting that this formalism can pinpoint SEP violations through the PPN parameters α_1, α_2, α_3 and ξ (see Table 1.2).

Gravitational waves and strong-field regime Recently, the direct detection of gravitational waves (GW) from the inward spiral and merger of compact objects [46], namely neutron stars and black holes, has opened a new window for testing GR — see e.g. Refs. [47, 48] for an overview of such tests. In particular, the detection in 2017 of gravitational wave signal emitted by the merger of a binary neutron stars together with its electromagnetic counterpart [49] puts severe constraints on a whole class of alternatives to GR (see e.g. Refs. [50–53]).

This brings us to tests of GR in the strong-field regime, since these GW detections also allow for probing the strong-field and dynamical regimes of GR, by comparing the measured signal close to the merger phase against numerical relativity simulations. Other tests of the strong regime include observations of binary pulsar systems (which provided evidence for the existence of GW) [54], the direct observation of black holes with the Event Horizon Telescope [55, 56].

[10]This is more or less what we did when we studied the Newtonian limit of GR, which formally corresponds to a PN expansion at zeroth order. In contrast, the PPN-formalism is developed at the first PN order.

Parameter	Significance	Effects	Value in GR	Constraints
$\gamma - 1$	Spatial curvature produced by mass	Time delay, light deflection	0	2.3×10^{-5} [36] 2×10^{-4} [37, 38]
$\beta - 1$	Nonlinearity in superposition of gravity	perihelion precession Nordtvedt effect	0	8×10^{-5} [39] 2×10^{-5} [40]
ξ	Preferred-location effects	spin precession	0	4×10^{-9} [41]
α_1	Preferred-frame effects	orbital polarization	0	4×10^{-5} [42]
α_2		spin precession	0	2×10^{-9} [41]
α_3		self-acceleration	0	4×10^{-20} [43, 44]
ζ_1	Violation of conservation of total momentum	—	0	2×10^{-2}
ζ_2		binary-pulsar acceleration	0	4×10^{-5} [45]
ζ_3		Newton's 3rd law	0	1×10^{-8}
ζ_4		—	0	6×10^{-3}

Table 1.2: PPN parameters, their physical significance and experimental constraints.

So why modify gravity?

In the few preceding pages, we have endeavored to show that GR, despite being a century-old theory, works extremely well: so far, it has passed all the tests thrown at it and underpins the ΛCDM model which is in agreement with cosmological observations. So why modify gravity? Here, we expose the main physical motivations for studying alternative models which, for the most part, fall into either one the two following categories:

1. solution to cosmological conundra;

2. attempts to construct a quantum theory of gravity.

Cosmological motivations Let us begin with cosmology. As we have seen previously, the standard model of cosmology ΛCDM is based upon the assumptions that the cosmological principle holds and that GR is the correct description of gravity. While most observations are explained within this model [57], this comes at the price of having to set $\Omega_{\mathrm{dm},0} = 0.27$ and $\Omega_{\Lambda,0} = 0.68$ (see Table 1.1) — that is the composition of the universe features 27% of dark matter and 68% of dark energy. The picture is rather disturbing: the universe in which we live would contain only 5% of ordinary matter, the remaining being unknown yet necessary ingredients for the model to hold. Indeed, we have strong evidence that the standard model of particle physics has to be supplemented by this dark sector. On the one hand, the observed accelerated cosmic expansion hints towards the existence of a fluid with equation of state $w = -1$ filling the universe. This lack of dynamics implied by a constant energy density contrasts with our understanding of another period of acceleration in the early universe, namely *inflation* [58]. On the other hand, other evidences such as galaxy clustering [59], galaxy rotation curves [60], gravitational lensing and the CMB power spectrum point to the presence of a kind of matter that does not interact with baryonic matter and radiation except though gravity. All such evidences are nonetheless *indirect*, meaning that dark matter and dark energy are only known through their effects at astrophysical and cosmological scales, their exact nature remains elusive.

Beyond our ignorance of this dark sector, the ΛCDM model still faces challenges. In particular, observations of different kinds can lead to tensions when they are interpreted within that model. The most famous example is of course the Hubble tension (also known as 'the crisis in cosmology'), which refers to the discrepancy between the locally measured value of H_0 and its value inferred from the CMB, with the difference now reaching a statistically significant level of around 5σ. The model is plagued with other, albeit more modest, tensions — see e.g. Ref. [61] for a comprehensive review of those.

One way around this unpleasant truth is to boldly abandon one of the assumptions we made in the first place, i.e. to state that GR is not the correct description of gravity. A good example of this paradigm is the MOND model (modified Newtonian dynamics) whereby phenomena usually attributed to the presence of dark matter result from a modification of GR whose Newtonian limit differs from Newton's inverse-square law. Another example is given by the class of $f(R)$ models. In these modified gravity models, the Ricci scalar R in the Einstein–Hilbert action Eq. (1.11a) is replaced by $f(R)$ where f is an arbitrary real function. This freedom in the choice of f can be leveraged in attempts to explain the late time acceleration and structure formation of the universe.

Distinguishing between modified gravity and adding new fields Before going any further a legitimate question may arise: what is the difference between actually modifying gravity [e.g. presumably MOND or $f(R)$ theories] and simply adding new fields to the theory (e.g. dark matter and dark energy)? To illustrate our point, we saw earlier when introducing FLRW cosmology that constructing a model that features cosmic acceleration could be explained equivalently by

- adding a cosmological constant Λ in the Einstein–Hilbert action, which leads to a modification of the lhs of the Einstein's field equations, or
- considering 'new' fields in the matter sector (in this case, vacuum energy), which effectively modifies the content of the stress-energy tensor in the rhs.

In the light of this remark, the line between modified gravity models and theories involving new forms of energy densities may justifiably seem blurry. To draw a more rigorous, unambiguous distinction, we rely on the SEP (in the same vein as Ref. [62]): models that violate the SEP belong to the former class while models that comply with the SEP belong to the latter class.

Unification with quantum field theory Let us now turn to the second incentive driving the development of alternatives to GR, namely the construction of a quantum theory of gravity. From a historical perspective, quantum field theory has successfully managed to unify three of the four (known) fundamental interactions in nature — the strong, weak and electromagnetic interactions — under the same umbrella that we call the standard model of particle physics. Gravity is the only fundamental interaction that is left out of the picture, the "black sheep that does not want to unite with the others" [63]. We could be tempted to just leave things as they are, since both models work stunningly well in their own physical scope of applications. However, this is not very satisfying for several reasons:

- It is more than hinted that our world is fundamentally quantum-mechanical. So from a theoretical standpoint, the fact that GR is purely classical provides internal evidence that the theory is incomplete. Having quantum fields evolving on a spacetime whose dynamics is classically described by the field equations can lead to thorny questions. For instance, what is the gravitational field sourced by an object put in a superposition of two spatially-separated states?
- Beyond this rather conceptual argument, there are actual physical situations for which GR cannot teach us anything, specifically when solutions to the field equations yield singularities. The two well-known examples of such singularities are black hole centers and the Big Bang. More generally, the description of any situation involving conditions where both quantum effects and strong gravitational fields are present would require a theory that encompasses quantum mechanics and GR (black hole thermodynamics and Hawking radiation, quantum fluctuations in the very early universe, etc.).

Reconciling GR with quantum mechanics is thus as important as it is a challenging task. The difficulty comes partly from the fact that the two theories are written in mathematical languages foreign to each other — differential geometry on pseudo-Riemannian manifolds *vs* vectors in a Hilbert space. One important obstacle in bridging them is the fact that GR is a non-renormalizable theory.

Ultimately, beyond all these aforementioned physical motivations for going beyond GR, there are also reasons of a more conceptual nature to consider alternative models of gravity. Indeed, exploring other classes of models can help identify which physical effects are specific to GR and which are not. It also tightens the noose on possible alternatives to GR, effectively reducing the space of viable theories.

1.1.2 Addition of a scalar degree of freedom in the gravitational sector

Following this introduction to GR, we are in a position to take a closer look at the mathematical construction of the 'alternative theories' that we kept mentioning elusively. In particular, and because the space of such theories is way too vast to be covered here, we focus on scalar-tensor theories, where a new scalar field ϕ is introduced in addition to the metric tensor $g_{\mu\nu}$. Here, we provide the mathematical grounds of scalar-tensor models. Their discussion from a physical standpoint is postponed to the next sub-section.

Lovelock's theorem

An insightful way to build modified gravity theories is provided by Lovelock's theorem, reported in Box A (see Refs. [64, 65] for the original papers and Ref. [66] for the modern version of it reported here). This theorem provides a set of assumptions under which the Einstein's field equations are the only viable gravitational field equations.

Box A: Lovelock's Theorem — Uniqueness of GR

The only possible second-order, local gravitational field equations derivable from an action containing solely the 4D metric tensor are the Einstein's field equations with a cosmological constant.

This theorem is useful because it provides five ways in which GR can be modified, by relaxing its assumptions respectively. These options are:

1. Add new field contents involved in mediating the gravitational force, i.e. new fields coupled to the metric tensor in the Einstein–Hilbert action Eq. (1.11a) — e.g. scalar-tensor theories.

2. Consider more than four dimensions — e.g. string theory, Kaluza–Klein theory, or the inclusion of a Gauss–Bonnet term in the action which only becomes relevant in higher dimensions.

3. Build a higher-order theory whose field equations contain greater than second-order derivatives — see e.g. the class of DHOST theories [67, 68] which are free from ghost-like instabilities.

4. Give up locality — e.g. actions containing the inverse d'Alembertian operator $\Box^{-1}$.

5. Give up on the action principle.

Of course, this merely constitutes a convenient way of classifying modified gravity models into different categories. Yet, this is an idealized picture as *(i)* there is no reason why we could not relax two or more hypotheses simultaneously, and *(ii)* these paths are intertwined in the sense that a given modified gravity model might fall into different categories depending on how it is written mathematically (see e.g. Appendix B).

In this PhD work, we follow the first path that consists in extending GR with the addition of new fields. One can indeed supplement the Einstein–Hilbert action $S_{\rm EH}$ with a scalar field ϕ (scalar-tensor theories), or a vector field A^μ (vector-tensor theories), or another rank-2 tensor $\hat{g}_{\alpha\beta}$ (bimetric theories). Some more complex proposals even include all three ingredients together (e.g. Tensor-Vector-Scalar theories, or TeVeS for short)! From now on, we will specialize to the class of scalar-tensor models.

Horndeski theory

Scalar-tensor theories are perhaps the simplest extensions of GR, where the metric tensor $g_{\mu\nu}$ is supplemented with a new scalar degree of freedom ϕ. Indeed, scalars are arguably easier to work with than tensors of rank greater than one (as is the case in vector-tensor or bimetric theories mentioned above). A good place to start is to introduce the so-called class of Horndeski theory [69], which is the most general 4-dimensional scalar-tensor theory whose Lagrangian leads to second-order equations of motion, i.e. not involving higher than second derivatives of the metric and the scalar field.[11]. Its action can be put in the form

$$S = \int \mathrm{d}^4x\sqrt{-g}\sum_{i=2}^{5}\mathcal{L}_i + S_{\rm mat}[g_{\mu\nu}]\,, \tag{1.31}$$

where $S_{\rm mat}[g_{\mu\nu}]$ is given by Eq. (1.11b) and the four scalars $(\mathcal{L}_i)_{2\leq i\leq 5}$ correspond to combinations of four functions $(G_i)_{2\leq i\leq 5}$ of the Ricci scalar, the Einstein tensor $G_{\mu\nu}$, the scalar field ϕ and its kinetic energy $X = -g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi/2$. The scalars read

$$\begin{aligned}
\mathcal{L}_2 &= G_2(\phi, X)\,,\\
\mathcal{L}_3 &= G_3(\phi, X)\,\Box\phi\,,\\
\mathcal{L}_4 &= G_4(\phi, X)R + \partial_X G_4(\phi, X)\left[(\Box\phi)^2 - (\partial_\mu\partial_\nu\phi)(\partial^\mu\partial^\nu\phi)\right]\,,\\
\mathcal{L}_5 &= G_5(\phi, X)G_{\mu\nu}\partial^\mu\partial^\nu\phi - \frac{1}{6}\partial_X G_5(\phi, X)\left[(\Box\phi)^3 + 2(\partial^\mu\partial_\alpha\phi)(\partial^\alpha\partial_\beta\phi)(\partial^\beta\partial_\mu\phi) - 3(\partial_\mu\partial_\nu\phi)(\partial^\mu\partial^\nu\phi)\Box\phi\right].
\end{aligned} \tag{1.32}$$

The d'Alembertian is defined as $\Box = g^{\mu\nu}\nabla_\mu\nabla_\nu$, where ∇_μ refer to the covariant derivatives.

The gravitational wave event GW170817 [49], by constraining the speed of gravitational waves to be practically equal to the speed of light, has greatly reduced to set of viable Horndeski actions — the surviving models featuring only $\mathcal{L}_2$ and $\mathcal{L}_3$ [52, 53].

[11] In that respect, DHOST theories are even more general than Horndeski theories as they relax this second-order derivatives constraint, without generating ghosts. See Fig. 1.2.

The 'traditional' scalar-tensor theory subclass

In the subsequent parts, we restrict ourselves to a simpler subclass of Horndeski models. For lack of a better name, we refer to this subclass as the 'traditional' scalar-tensor models, because they constitute the simplest yet phenomenologically interesting models, and in this sense are thoroughly studied in the literature. Their action can be put in the form

$$S = S_{\rm EH} + S_\phi + S_{\rm mat}[\tilde{g}_{\mu\nu}]\,, \tag{1.33}$$

where $S_{\rm EH}$ is still given by Eq. (1.11a) while

$$S_\phi = -\int \mathrm{d}^4x\sqrt{-g}\left[\frac{1}{2}g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi + V(\phi)\right]\,, \tag{1.34a}$$

$$S_{\rm mat}[\tilde{g}_{\mu\nu}] = \int \mathrm{d}^4x\sqrt{-\tilde{g}}\,\mathcal{L}_{\rm mat}\big(\tilde{g}_{\mu\nu},\,\psi_{\rm mat}\big)\,. \tag{1.34b}$$

Here, $g_{\mu\nu}$ will be referred to as the 'Einstein-frame metric'. The reason behind this designation is that the kinetic term for the metric in Eq. (1.33) is of the Einstein–Hilbert form. Eq. (1.34a) is the action of the scalar field with a canonical kinetic term $X = g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi$ and potential $V(\phi)$. Eq. (1.34b) is the matter action which differs from Eq. (1.11b) since the matter fields $\psi_{\rm mat}$ are minimally coupled to a different metric $\tilde{g}_{\mu\nu}$ (instead of $g_{\mu\nu}$) that is called the 'Jordan-frame metric'. The latter is chosen to be related to the Einstein-frame metric through the Weyl transformation

$$\tilde{g}_{\mu\nu} = \Omega^2(\phi)g_{\mu\nu}\,, \tag{1.35}$$

where Ω is called the conformal factor function. Note that Eqs. (1.33, 1.34a) imply that the scalar field has units of eV.

At this point, it is not immediately clear that the scalar-tensor model defined by the action Eq. (1.33) actually constitute a subclass Horndeski's theories. That is because the former involves two metric tensors $g_{\mu\nu}$ and $\tilde{g}_{\mu\nu}$ instead of a single one. In order to make the connection with Horndeski's theories more transparent, we should rewrite the 'Einstein-frame action' Eq. (1.33) so that only the Jordan-frame metric $\tilde{g}_{\mu\nu}$ appears in the action. This is a fairly standard procedure that we outline here so that the end result does not completely come out of the blue.

Straightforwardly, Eq. (1.35) implies that $\sqrt{-g} = \Omega^{-4}(\phi)\sqrt{-\tilde{g}}$ and $g^{\mu\nu} = \Omega^2(\phi)\tilde{g}^{\mu\nu}$. Less trivially, the Ricci scalar transforms as (see e.g. Ref. [12] Appendix G)

$$R = \Omega^2\tilde{R} - 6\tilde{g}^{\alpha\beta}\Big[\tilde{\nabla}_\alpha\tilde{\nabla}_\beta\left(\Omega^{-1}\right)\Big]\Omega^3, \tag{1.36}$$

where $\tilde{R}$ and $\tilde{\nabla}_\mu$ denotes respectively the Ricci scalar and the covariant derivatives, both constructed from the Jordan-frame metric $\tilde{g}_{\mu\nu}$. Consequently, the Einstein–Hilbert action becomes

$$\begin{aligned} S_{\rm EH} = \frac{M_{\rm Pl}^2}{2}\int \mathrm{d}^4x\sqrt{-g}R &= \frac{M_{\rm Pl}^2}{2}\int \mathrm{d}^4x\sqrt{-\tilde{g}}\left\{\Omega^{-2}\tilde{R} - 6\,\tilde{g}^{\alpha\beta}\,\Omega^{-1}\left[\tilde{\nabla}_\alpha\tilde{\nabla}_\beta\left(\Omega^{-1}\right)\right]\right\} \\ &= \frac{M_{\rm Pl}^2}{2}\int \mathrm{d}^4x\sqrt{-\tilde{g}}\left\{\Omega^{-2}\tilde{R} + 6\,\tilde{g}^{\alpha\beta}\partial_\alpha\big(\Omega^{-1}\big)\partial_\beta\big(\Omega^{-1}\big)\right\}, \end{aligned} \tag{1.37}$$

where the last equality follows from an integration by parts and makes use of the fact that $\tilde{\nabla}_\mu(\Omega^{-1}) = \partial_\mu(\Omega^{-1})$. Similarly, we obtain for the scalar action

$$S_\phi = -\int \mathrm{d}^4x\sqrt{-\tilde{g}}\left[\frac{\Omega^{-2}}{2}\tilde{g}^{\alpha\beta}\partial_\alpha\phi\,\partial_\beta\phi + \Omega^{-4}V(\phi)\right]\,. \tag{1.38}$$

At this stage, we can perform a field redefinition $\phi \to \varphi$ by making use of three functions F, U and Z, such that

$$F(\varphi) = \Omega(\phi)^{-2}\,, \tag{1.39a}$$

$$U(\varphi) = \Omega(\phi)^{-4}\,V(\phi)\,, \tag{1.39b}$$

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\right)^2 = \frac{Z(\varphi)}{F(\varphi)} + \frac{3}{2}M_{\rm Pl}^2\left(\frac{\mathrm{d}\ln F}{\mathrm{d}\varphi}\right)^2\,. \tag{1.39c}$$

In particular, these definitions lead to

$$\partial_\alpha\big(\Omega^{-1}\big)\partial_\beta\big(\Omega^{-1}\big) = \frac{F(\varphi)}{4}\left(\frac{\mathrm{d}\ln F}{\mathrm{d}\varphi}\right)^2\partial_\alpha\varphi\,\partial_\beta\varphi \qquad \text{and} \qquad \partial_\mu\phi\,\partial_\nu\phi = \partial_\mu\varphi\,\partial_\nu\varphi\left(\frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\right)^2\,. \tag{1.40}$$

All in all, we end up with the action

$$S = \int \mathrm{d}^4x\sqrt{-\tilde{g}}\left[\frac{M_{\rm Pl}^2}{2}F(\varphi)\tilde{R} - \frac{1}{2}Z(\varphi)\tilde{g}^{\mu\nu}\partial_\mu\varphi\,\partial_\nu\varphi - U(\varphi)\right] + \int \mathrm{d}^4x\sqrt{-\tilde{g}}\mathcal{L}_{\rm mat}\big(\tilde{g}_{\mu\nu},\,\psi_{\rm mat}\big)\,. \tag{1.41}$$

With Eq. (1.41), we have managed to rewrite the action of the theory using the Jordan-frame metric $\tilde{g}_{\mu\nu}$ (together with its derived quantities, $\tilde{g}$, $\tilde{R}$) and scalar field φ. In contrast, its initial form Eqs. (1.33–1.34) was written in terms of the Einstein-frame metric $g_{\mu\nu}$ and scalar field ϕ. A few remarks are in order:

- The action S put in the Jordan frame as Eq. (1.41) maps to the Horndeski class where $\mathcal{L}_2$ is the only non-zero Lagrangian.
- From Eq. (1.39c), we see that the scalar field φ is dimensionless whereas ϕ has the dimension of an energy.
- In the Jordan frame, the special case $Z = U = 0$ turns off the pure scalar terms in the action (1.41). Nonetheless, the Einstein-frame action (1.33–1.34) actually includes the conventional kinetic term (with no potential), even if it was absent from the Jordan frame action. Therefore, the degrees of freedom of this theory include a propagating scalar as well as the metric.
- We said earlier that matter is minimally coupled to the Jordan-frame metric. This is to be understood in the sense that the Lagrange density $\sqrt{-\tilde{g}}\mathcal{L}_{\rm mat}(\tilde{g}_{\mu\nu},\,\psi_{\rm mat})$ is the simplest choice for adding matter fields to the theory that ensures diffeomorphism covariance. In return, φ is directly coupled to the curvature scalar in the Jordan frame.
- Following the preceding remark, matter fields "see" the Jordan-frame metric $\tilde{g}_{\mu\nu}$, not the Einstein-frame metric $g_{\mu\nu}$. In the absence of other forces, matter test particles thus follow geodesics of the Jordan-frame metric. As such, $\tilde{g}_{\mu\nu}$ is sometimes called the physical metric — it is the metric to which matter is universally coupled and thus defines the lengths and times measured by material rods and clocks [70].
- In the following, we will only talk about the Einstein frame and the Jordan frame. However, it is easy to see from the transformation that we performed [Eqs. (1.39–1.41)] that there is actually a continuum of frames in between the two. The reason for this dichotomy is a practical one: the Einstein frame and the Jordan frame are the special cases for which the scalar field is minimally coupled to curvature and matter respectively — granting them a higher status with respect to other frames which are consequently non-minimally coupled to both curvature and matter.
- Theories of the form (1.33–1.35) fall within the category of *metric theories* of gravity. As such, they automatically satisfy the EEP [1].

Now that we have clearly defined the action of the theory, let us study the field equations. Namely, we derive the equations of motion for $(g_{\mu\nu},\,\phi)$ in the Einstein frame, and for $(\tilde{g}_{\mu\nu},\,\varphi)$ in the Jordan frame. We then see how these field equations simplify in the Newtonian limit.

Field equations and Newtonian limit in the Einstein frame

Similarly to GR — see Sec. 1.1.1 — the field equations are obtained by varying the action in the Einstein frame [Eq. (1.33)] with respect to $g^{\mu\nu}$ and ϕ.

Field equations in the Einstein frame Varying the action with respect to the inverse metric yields a modified version of the Einstein's field equations

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \frac{1}{M_{\rm Pl}^2}\left(T_{\mu\nu} + T_{\mu\nu}^{(\phi)}\right)\,. \tag{1.42}$$

The stress-energy tensor of matter $T_{\mu\nu}$ is given by

$$T_{\mu\nu} = \frac{-2}{\sqrt{-g}}\frac{\delta S_{\rm mat}}{\delta g^{\mu\nu}}\,, \tag{1.43}$$

while $T_{\mu\nu}^{(\phi)}$ denotes the scalar field stress-energy tensor whose expression reads

$$T_{\mu\nu}^{(\phi)} = \frac{-2}{\sqrt{-g}}\frac{\delta S_\phi}{\delta g^{\mu\nu}} = \partial_\mu\phi\,\partial_\nu\phi - \frac{1}{2}g_{\mu\nu}g^{\rho\sigma}\partial_\rho\phi\,\partial_\sigma\phi - g_{\mu\nu}V(\phi)\,. \tag{1.44}$$

Likewise, varying the action with respect to the scalar field yields the following Klein–Gordon equation

$$\Box\phi \equiv g^{\mu\nu}\nabla_\mu\nabla_\nu\phi = \frac{\mathrm{d}V}{\mathrm{d}\phi} - \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}T\,, \tag{1.45}$$

where $T = g^{\mu\nu}T_{\mu\nu}$ denotes the trace of the stress-energy tensor of matter. Here let us recall that the electromagnetic stress-energy tensor[12] is traceless, and so a conformally coupled scalar field does not have a classical coupling to photons.

Newtonian limit in the Einstein frame In order to derive the Newtonian limit of the field equations (1.42), we proceed as we did in Sec. 1.1.1 for GR. We still assume that the Einstein-frame metric can be expanded about the Minkowski metric as $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, with $|h_{\mu\nu}| \ll 1$. Moreover, gauge freedom allows us to put the metric in the form (see e.g. Ref. [1] Chapt. 4.2.2)

$$ds^2 = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = (\eta_{\mu\nu} + h_{\mu\nu})\mathrm{d}x^\mu\mathrm{d}x^\nu = -(1+2\Phi)\mathrm{d}t^2 + g_{ij}\mathrm{d}x^i\mathrm{d}x^j\,. \tag{1.46}$$

Before starting the computations, be aware that from now on, the symbol "Δ" is *defined* as[13]

$$\Delta \equiv g^{ij}\partial_i\partial_j\,. \tag{1.47}$$

The trace-reversed form of Eq. (1.42) reads

$$M_{\rm Pl}^2 R_{\mu\nu} = T_{\mu\nu} + T^{(\phi)}_{\mu\nu} - \frac{1}{2}\left(T + T^{(\phi)}\right)g_{\mu\nu}\,, \tag{1.48}$$

where $T^{(\phi)}$ is the trace of the scalar field stress-energy tensor and is given by

$$T^{(\phi)} = g^{\mu\nu}T^{(\phi)}_{\mu\nu} = g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi - \frac{1}{2}g^{\mu\nu}g_{\mu\nu}g^{\rho\sigma}\partial_\rho\phi\,\partial_\sigma\phi - g^{\mu\nu}g_{\mu\nu}V(\phi) = -\partial^\alpha\phi\,\partial_\alpha\phi - 4V(\phi)\,. \tag{1.49}$$

We are interested in the 00-component of Eq. (1.48). The computation is the same as in GR, except we have the additional scalar field contribution to the rhs

$$\begin{aligned} T^{(\phi)}_{00} - \frac{1}{2}T^{(\phi)}g_{00} &= (\partial_t\phi)^2 - \frac{1}{2}g_{00}\partial^\alpha\phi\,\partial_\alpha\phi - g_{00}V(\phi) + \frac{1}{2}g_{00}\partial^\alpha\phi\,\partial_\alpha\phi + 2g_{00}V(\phi) \\ &\simeq 2g_{00}V(\phi)\,. \end{aligned} \tag{1.50}$$

In the above computation, the kinetic term X vanishes exactly whereas partial derivatives with respect to time are dropped because of the quasi-static approximation. We end up with the following modified Poisson equation

$$2M_{\rm Pl}^2\Delta\Phi = \rho\Big[2\Omega^{-2}(\phi) + g_{00}\Big] + 2g_{00}V(\phi)\,, \tag{1.51}$$

where the scalar field ϕ contributes to the rhs through its potential and conformal factor. The presence of the latter in Eq. (1.51) — and more importantly, our definition of the Einstein frame 'density' ρ — are explained in Box B. It should be noted that at no point during the above derivations did we assume that the conformal factor Ω was close to one. Once again, we refrain from making the approximation $g_{00} \simeq -1$ in Eq. (1.51) because that would mean losing track of powers of Ω if we ever want to go to the Jordan frame from this Einstein frame approximation.[13]

The Newtonian limit of the Klein–Gordon equation (1.45) is somewhat easier to derive. As a matter of fact, the d'Alembertian $\Box$ boils down to the Laplacian $\Delta = g^{ij}\partial_i\partial_j$ [see Eq. (1.47)] in the quasi-static limit, while the trace of the energy-momentum tensor of matter is approximated as $T \simeq -\rho$ (see Box B), leading to

$$\Delta\phi = \frac{\mathrm{d}V}{\mathrm{d}\phi} + \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\rho\,. \tag{1.52}$$

Interestingly, the Klein–Gordon equation governing the scalar field ϕ in the Newtonian limit [Eq. (1.52)] is fully decoupled from metric-related quantities. That means it is possible to solve it first to obtain the ϕ profile, so that the potential Φ obeys a Poisson's equation (1.51) with known rhs.

[12] This tensor is $T^{\mu\nu}_{\rm EM} = \frac{1}{\mu_0}\left(F^{\mu\alpha}F^{\nu}{}_{\alpha} - \frac{1}{4}\eta^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta}\right)$ is SI units.

[13] Of course, one may argue that, given the approximation we just stated, $g^{ij} \simeq \delta^{ij}$. While this is true, we keep the definition Eq. (1.47) as is. This will prove to be crucial when we try to match the Newtonian limits in the Einstein frame *vs* Jordan frame later on. Without this precaution, powers of the scale factor Ω can be lost along the way...

Field equations and Newtonian limit in the Jordan frame

Field equations in the Jordan frame Conversely, the Jordan frame field equations are obtained by varying the action Eq. (1.41) with respect to $\tilde{g}^{\mu\nu}$ and φ.

Varying the action with respect to $\tilde{g}^{\mu\nu}$ yields

$$F(\varphi)\left(\tilde{R}_{\mu\nu}-\frac{1}{2}\tilde{R}\tilde{g}_{\mu\nu}\right)=\frac{1}{M_{\rm Pl}^2}\left(\tilde{T}_{\mu\nu}+\tilde{T}^{(\varphi)}_{\mu\nu}\right). \tag{1.53}$$

In this frame, the stress-energy tensor of matter is given by

$$\tilde{T}_{\mu\nu}=\frac{-2}{\sqrt{-\tilde{g}}}\frac{\delta S_{\rm mat}}{\delta\tilde{g}^{\mu\nu}}, \tag{1.54}$$

and $\tilde{T}^{(\varphi)}_{\mu\nu}$ is a convenient notation for the scalar field contribution

$$\tilde{T}^{(\varphi)}_{\mu\nu}=Z(\varphi)\left[\partial_\mu\varphi\,\partial_\nu\varphi-\frac{1}{2}\tilde{g}_{\mu\nu}\tilde{g}^{\alpha\beta}\partial_\alpha\varphi\,\partial_\beta\varphi\right]-\tilde{g}_{\mu\nu}U(\varphi)+M_{\rm Pl}^2\big(\tilde{\nabla}_\mu\tilde{\nabla}_\nu F-\tilde{g}_{\mu\nu}\tilde{\Box}F\big). \tag{1.55}$$

The equation of motion for the scalar field is obtained by varying the action with respect to φ, yielding

$$Z(\varphi)\tilde{\Box}\varphi=\frac{{\rm d}U}{{\rm d}\varphi}-\frac{M_{\rm Pl}^2}{2}\frac{{\rm d}F}{{\rm d}\varphi}\tilde{R}-\frac{1}{2}\frac{{\rm d}Z}{{\rm d}\varphi}\tilde{g}^{\alpha\beta}\partial_\alpha\varphi\,\partial_\beta\varphi, \tag{1.56}$$

where $\tilde{\Box}=\tilde{g}^{\mu\nu}\tilde{\nabla}_\mu\tilde{\nabla}_\nu$ and $\tilde{\nabla}_\mu$ refers to the covariant derivatives constructed from $\tilde{g}_{\mu\nu}$. It is striking that both the metric equation (1.53) and the scalar field equation (1.56) appear significantly more complex than their Einstein frame counterparts, Eq. (1.42) and Eq. (1.45) respectively.[14] Note that the Jordan frame equations could also have been obtained by applying the conformal transformation $\tilde{g}_{\mu\nu}=\Omega^2 g_{\mu\nu}$ together with the field's redefinition (1.39) directly in the Einstein frame field equations, thus bypassing the variation of the action as we did. We lay emphasis on the fact that at the equation level, it is strictly equivalent to work with Einstein frame quantities $(g_{\mu\nu},\,\phi)$ governed by Eqs. (1.42, 1.45) or with Jordan frame quantities $(\tilde{g}_{\mu\nu},\,\varphi)$ governed by Eqs. (1.53, 1.56). Of course depending on the context, one frame might turn out to be more handy than the other to perform certain calculations, but the computation of observable quantities shall provide the same results.

Box B: Energy-momentum tensors of matter in both frames

The definitions of the stress-energy tensor of matter in the Einstein frame [Eq. (1.43)] and in the Jordan frame [Eq. (1.54)] imply the following relations

$$\tilde{T}_{\mu\nu}=\Omega^{-2}\,T_{\mu\nu}=F\,T_{\mu\nu}\,,\qquad \tilde{T}^{\mu\nu}=\Omega^{-6}\,T^{\mu\nu}=F^3\,T^{\mu\nu}\,,\qquad \tilde{T}=\Omega^{-4}\,T=F^2\,T\,. \tag{1.57}$$

For a perfect-fluid source of energy-momentum with 4-velocity $\tilde{u}^\mu$, $\tilde{\rho}$ and $\tilde{p}$ as rest-frame energy and momentum densities — in the Jordan frame —, we have

$$\tilde{T}_{\mu\nu}=(\tilde{\rho}+\tilde{p})\,\tilde{u}_\mu\tilde{u}_\nu+\tilde{p}\tilde{g}_{\mu\nu}=\Omega^2\big[(\tilde{\rho}+\tilde{p})\,u_\mu u_\nu+\tilde{p}g_{\mu\nu}\big]=\Omega^{-2}T_{\mu\nu}\,. \tag{1.58}$$

Therefore, it makes sense to define $(\rho,\,p)=\Omega^4(\tilde{\rho},\,\tilde{p})$ so that we recover the canonical form

$$T_{\mu\nu}=(\rho+p)u_\mu u_\nu+pg_{\mu\nu}\,. \tag{1.59}$$

When going to the Newtonian limit, approximating $\tilde{T}_{00}$ by $\tilde{\rho}$ on the one hand, and T_{00} by ρ on the other hand, would be incompatible with the above definitions (which in contrast do not rely on any approximation). Because the Jordan frame is the physical frame (i.e. $\tilde{\rho}$ corresponds to the *measured* density of the fluid), we give it preference over the Einstein frame for computing the Newtonian limit of the energy-momentum tensor of matter. As such Eq. (1.51) is consistent with setting $\tilde{T}_{00}\simeq\tilde{\rho}$ and $\tilde{T}\simeq-\tilde{\rho}$, which readily implies $T_{00}\simeq\Omega^{-2}\rho$ and $T\simeq-\rho$. Of course, in practice $\Omega(\phi)$ will always be very close to unity so that $\rho\simeq\tilde{\rho}$ is a legitimate approximation at 0 PN. Footnote 2 from Ref. [71] provides further insights into the commonly used 'densities' in the literature.

[14] This is in part due to the fact that the Einstein frame equations are *meant* to look like GR, for which we already have a condensed way of writing the objects appearing in the field equations.

Newtonian limit in the Jordan frame For the Newtonian limit, we proceed as before by putting the Jordan-frame metric in the Newtonian gauge, i.e.

$$d\tilde{s}^2 = \tilde{g}_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = (\eta_{\mu\nu} + \tilde{h}_{\mu\nu})\mathrm{d}x^\mu \mathrm{d}x^\nu = -(1+2\tilde{\Phi})\mathrm{d}t^2 + \tilde{g}_{ij}\mathrm{d}x^i \mathrm{d}x^j \,, \tag{1.60}$$

where $\{x^\mu\}$ denotes the same set of coordinates as in Eq. (1.46) and $\tilde{h}_{\mu\nu}$ is a small metric perturbation i.e. $|\tilde{h}_{\mu\nu}| \ll 1$. Likewise, we *define* the operator $\tilde{\Delta}$ by

$$\tilde{\Delta} \equiv \tilde{g}^{ij}\partial_i\partial_j \,. \tag{1.61}$$

As for the derivation of the Newtonian limit in the Einstein frame, we refrain from approximating $\tilde{g}_{00} \simeq -1$, $g_{ij} \simeq \delta_{ij}$ in places where one would usually do — again, the reason for this will appear clear when comparing the Newtonian limits in both frames. We also write down the trace-reversed version of the field equations (1.53)

$$M_{\rm Pl}^2 F(\varphi)\tilde{R}_{\mu\nu} = \tilde{T}_{\mu\nu} + \tilde{T}^{(\varphi)}_{\mu\nu} - \frac{1}{2}\left(\tilde{T} + \tilde{T}^{(\varphi)}\right)\tilde{g}_{\mu\nu} \,. \tag{1.62}$$

Again, we consider the 00-component of this equation. Straightforwardly, the matter terms simplify to $\tilde{T}_{00} - \tilde{T}\tilde{g}_{00}/2 \simeq \tilde{\rho}(1+\tilde{g}_{00}/2)$ (see Box B). For the scalar field contribution, we have

$$\begin{aligned}
\tilde{T}^{(\varphi)}_{00} - \frac{1}{2}\tilde{T}^{(\varphi)}\tilde{g}_{00} &= Z(\varphi)\left[(\partial_t\varphi)^2 - \frac{1}{2}\tilde{g}_{00}\partial^\alpha\varphi\,\partial_\alpha\varphi\right] - \tilde{g}_{00}U(\varphi) + M_{\rm Pl}^2(\tilde{\nabla}_t\partial_t F - \tilde{g}_{00}\tilde{\Box}F) \\
&\quad + \frac{1}{2}\left[Z(\varphi)\partial^\alpha\varphi\,\partial_\alpha\varphi + 4U(\varphi) + 3M_{\rm Pl}^2\tilde{\Box}F\right]\tilde{g}_{00} \\
&\simeq \left[\frac{1}{2}M_{\rm Pl}^2\tilde{\Delta}F + U(\varphi)\right]\tilde{g}_{00} \,,
\end{aligned} \tag{1.63}$$

where we discarded the time derivatives and approximated $\tilde{\Box} \sim \tilde{\Delta}$ [see Eq. (1.61)] in the last line (the kinetic terms cancel exactly). We thus end up with a modified Poisson equation for the potential $\tilde{\Phi}$

$$M_{\rm Pl}^2\left[F(\varphi)\tilde{\Delta}\tilde{\Phi} - \frac{1}{2}\tilde{g}_{00}\tilde{\Delta}F\right] = \tilde{\rho}\left(1 + \frac{1}{2}\tilde{g}_{00}\right) + \tilde{g}_{00}U(\varphi) \,. \tag{1.64}$$

On the other hand, the Klein–Gordon equation (1.56) involves the Jordan frame Ricci scalar, which is obtained by taking the trace of Eq. (1.53)

$$\tilde{R} = \frac{-1}{F(\varphi)M_{\rm Pl}^2}\left(\tilde{T} + \tilde{T}^{(\varphi)}\right) = \frac{1}{F(\varphi)M_{\rm Pl}^2}\left[\tilde{\rho} + Z(\varphi)\tilde{g}^{\alpha\beta}\partial_\alpha\varphi\,\partial_\beta\varphi + 4U(\varphi) + 3M_{\rm Pl}^2\tilde{\Box}F\right]. \tag{1.65}$$

As a result, the Newtonian limit of Eq. (1.56) reads

$$Z(\varphi)\tilde{\Delta}\varphi = \frac{\mathrm{d}U}{\mathrm{d}\varphi} - \frac{1}{2}\frac{\mathrm{d}\ln F}{\mathrm{d}\varphi}\left[\tilde{\rho} + 4U(\varphi) + 3M_{\rm Pl}^2\tilde{\Delta}F\right], \tag{1.66}$$

where we have discarded terms of the form $\partial_i\varphi\,\partial_j\varphi$ as they represent higher order terms. As for the scalar field ϕ equation (1.52), the equation of motion for φ in the Newtonian limit Eq. (1.66) is decoupled from the metric tensor $\tilde{g}_{\mu\nu}$.

Comparison of the Newtonian limits

To conclude this rather computationally-involved part, it is insightful to outline the link between the Newtonian limits in the two frames, and check whether they match as hoped. The issue with the expansions in the Newtonian gauge [Eqs. (1.46, 1.60)] we made is that they are not equivalent, in the sense that they correspond to distinct approximations. To see it, let us write

$$\tilde{g}_{\mu\nu} = \eta_{\mu\nu} + \tilde{h}_{\mu\nu} = \Omega^2 g_{\mu\nu} = \Omega^2(\eta_{\mu\nu} + h_{\mu\nu}) = \eta_{\mu\nu} + \left[(\Omega^2-1)\eta_{\mu\nu} + \Omega^2 h_{\mu\nu}\right]. \tag{1.67}$$

For the Einstein frame computations we have assumed $|h_{\mu\nu}| \ll 1$, whereas for the Jordan frame computations we have assumed $|\tilde{h}_{\mu\nu}| \ll 1$. These two assumptions are not equivalent, precisely because we refrained from making the approximation $\Omega \sim 1$ until now. The latter assumption is therefore necessary for the two expansions

to be mathematically equivalent at first order.[15] We write this down as

$$\Omega(\phi) = 1 + \omega(\phi)\,, \qquad \text{with} \quad |\omega(\phi)| \ll 1\,. \tag{1.68}$$

Moreover, we have to choose which metric is *truly* expanded around Minkowski, and which metric carries the conformal factor weight, as writing

$$\begin{cases} \tilde{g}_{\mu\nu} = \eta_{\mu\nu} + \tilde{h}_{\mu\nu} \\ g_{\mu\nu} = (1-2\omega)(\eta_{\mu\nu} + \tilde{h}_{\mu\nu}) \end{cases} \qquad \text{and} \qquad \begin{cases} g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu} \\ \tilde{g}_{\mu\nu} = (1+2\omega)(\eta_{\mu\nu} + h_{\mu\nu}) \end{cases} \tag{1.69}$$

simultaneously would inevitably lead to inconsistencies. In the light of this remark, we arbitrarily decide to pick the second version of Eq. (1.69) for which the conformal factor weight $(1+2\omega)$ is placed on the Jordan-frame metric and stick to this choice all the way through.[16] Recalling that $\tilde{h}_{00} = -2\tilde{\Phi}$, $h_{00} = -2\Phi$ [see Eqs. (1.46, 1.60)], we have

$$1 + 2\tilde{\Phi} = \Omega^2(1+2\Phi) \implies \tilde{\Delta}\tilde{\Phi} = \frac{1}{2}\tilde{\Delta}\big(\Omega^2\big) + \tilde{\Delta}\big(\Omega^2\Phi\big)\,. \tag{1.70}$$

Moreover, looking at the definitions of $\tilde{\Delta}$ [Eq. (1.61)] and Δ [Eq. (1.47)], we see that the two operators are related via $\tilde{\Delta} = \Omega^{-2}\Delta$. To show that the two Newtonian limits are equivalent, we go from the Jordan frame approximations [Eqs. (1.64, 1.66)] to the Einstein frame ones [Eqs. (1.51, 1.52)]. Using Eq. (1.70), the lhs of Eq. (1.64) becomes

$$\begin{aligned} M_{\rm Pl}^2\left[F(\varphi)\tilde{\Delta}\tilde{\Phi} - \frac{1}{2}\tilde{g}_{00}\tilde{\Delta}F\right] &= M_{\rm Pl}^2\left[\frac{1}{2}\Omega^{-4}\Delta\big(\Omega^2\big) + \Omega^{-4}\Delta\big(\Omega^2\Phi\big) + \frac{1}{2}\Delta\big(\Omega^{-2}\big)\right] \\ &\simeq M_{\rm Pl}^2\Big[(1-4\omega)\Delta\omega + \Omega^{-4}\Delta\big(\Omega^2\Phi\big) - \Delta\omega\Big] \\ &\simeq M_{\rm Pl}^2\big[\Omega^{-2}\Delta\Phi + \text{higher-order terms}\big]\,. \end{aligned} \tag{1.71}$$

Here, higher-order terms include $\{\partial_i\phi\,\partial_j\phi,\ \omega\Delta\phi,\ \Phi\Delta\phi,\ \partial_i\Phi\,\partial_j\phi,\ \dots\}$.[17] Likewise, the rhs boils down to

$$\tilde{\rho}\left(1 + \frac{1}{2}\tilde{g}_{00}\right) + \tilde{g}_{00}U(\varphi) \simeq \frac{1}{2}\tilde{\rho}\big(1-2\omega\big) - \Omega^2 U(\varphi) \simeq \frac{1}{2}\Omega^{-2}\big[\tilde{\rho} - 2\Omega^4 U(\varphi)\big] \tag{1.72}$$

Putting back the two sides [Eqs. (1.71, 1.72)] together, we recover Eq. (1.51) since

$$\rho\left[2\Omega^{-2} + g_{00}\right] \simeq \rho\big(1-4\omega\big) \simeq \tilde{\rho}\,. \tag{1.73}$$

We have just shown that the 00-component of Eqs. (1.48, 1.62) are equivalent in the Newtonian limit together with the approximation Eq. (1.68).

The same goes for the scalar field equations, although one has to be extra careful about the terms that are now negligible and the ones that must be kept. The partial derivatives of the scalar fields are considered small, so any term of the form $\{\partial_i\phi\,\partial_j\phi,\ \partial_i\varphi\,\partial_j\varphi\}$ can be safely discarded. In view of this remark, we can approximate

$$\tilde{\Delta}\varphi = \tilde{g}^{ij}\partial_i\partial_j\varphi = \tilde{g}^{ij}\partial_i\left(\partial_j\phi\frac{\mathrm{d}\varphi}{\mathrm{d}\phi}\right) = \tilde{g}^{ij}\partial_i\partial_j\phi\frac{\mathrm{d}\varphi}{\mathrm{d}\phi} + \tilde{g}^{ij}\partial_i\phi\,\partial_j\phi\frac{\mathrm{d}^2\varphi}{\mathrm{d}\phi^2} \simeq \frac{\mathrm{d}\varphi}{\mathrm{d}\phi}\tilde{\Delta}\phi\,, \tag{1.74}$$

$$\tilde{\Delta}\omega = \tilde{g}^{ij}\partial_i\partial_j\omega = \tilde{g}^{ij}\partial_i\big[\partial_j\phi\,\omega'(\phi)\big] = \tilde{g}^{ij}\omega'(\phi)\partial_i\partial_j\phi + \tilde{g}^{ij}\partial_i\phi\,\partial_j\phi\,\omega''(\phi) \simeq \omega'(\phi)\tilde{\Delta}\phi\,, \tag{1.75}$$

$$\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi} \simeq \omega'(\phi) - \omega(\phi)\omega'(\phi)\,. \tag{1.76}$$

Note that $\omega'(\phi)$ is a priori not small.[17] The function $Z(\varphi)$ may be approximated from Eq. (1.39c) as

$$\Omega^2(\phi)Z(\varphi) \simeq \left(\frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\right)^2\left[1 - 6M_{\rm Pl}^2\omega'(\phi)^2\right]\,. \tag{1.77}$$

[15] This important remark is also underlined in Ref. [72], Sec. IV.

[16] Of course, we could have equivalently chosen the first metric formulation of Eq. (1.69) and derive the same self-consistent conclusions.

[17] One may find it helpful to keep in mind the common form for the conformal factor $\Omega(\phi) = \exp(\alpha\phi)$ (where α is just a coupling constant), in which case $\omega(\phi) \sim \alpha\phi$.

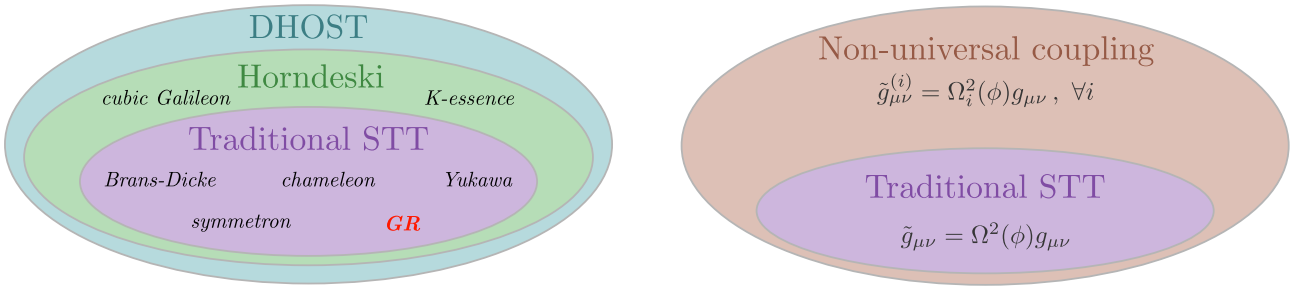


Figure 1.2: Inclusions in scalar-tensor theories. The left set satisfies the WEP at the Lagrangian level.

Using Eqs. (1.74, 1.76, 1.77), the lhs of Eq. (1.66) boils down to

$$Z(\varphi)\tilde{\Delta}\varphi \simeq \frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\left[1 - 6M_{\rm Pl}^2\omega'(\phi)^2\right]\Omega^{-4}\Delta\phi\,. \tag{1.78}$$

Similarly, the rhs simplifies to

$$\begin{aligned}\frac{\mathrm{d}U}{\mathrm{d}\varphi} - \frac{1}{2}\frac{\mathrm{d}\ln F}{\mathrm{d}\varphi}\left[\tilde{\rho} + 4U(\varphi) + 3M_{\rm Pl}^2\tilde{\Delta}F\right] &\simeq \frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\left\{\Omega^{-4}V'(\phi) + \omega'(\phi)\left[\Omega^{-4}\rho - 6M_{\rm Pl}^2\tilde{\Delta}\omega\right]\right\} \\ &\simeq \Omega^{-4}\frac{\mathrm{d}\phi}{\mathrm{d}\varphi}\left[\frac{\mathrm{d}V}{\mathrm{d}\phi} + \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\rho - 6M_{\rm Pl}^2\omega'(\phi)^2\Delta\phi\right],\end{aligned} \tag{1.79}$$

where we have only dropped terms of the form $\{\omega\Delta\phi\}$, consistently with the previous approximations (note that the $\Omega^{-4}V'$ term is recovered exactly). As hoped, putting Eqs. (1.78, 1.79) back together (multiplied by $\Omega^4\mathrm{d}\varphi/\mathrm{d}\phi$) yields the Klein–Gordon equation (1.52) that we obtained in the Einstein frame.

These computations, despite being quite heavy, help shed light on the bridge between the Einstein frame and the Jordan frame. Similar computations are undertaken in Refs. [71, 72], with notations different from ours and sometimes taking alternative routes towards the end results. They were conducted here for the sake of having all the relevant equations in a single place, in particular Eqs. (1.42, 1.45, 1.51, 1.52) for the Einstein frame, and Eqs. (1.53, 1.56, 1.64, 1.66). Note that we have linearized the metric perturbations, but not the scalar field perturbations, i.e. we have not made approximations of the form $\phi = \phi_0(1+\delta\phi)$ with $|\delta\phi| \ll 1$ or $\varphi = \varphi_0(1+\delta\varphi)$ with $|\varphi| \ll 1$. As a matter of fact, these approximation are far from being valid in some scalar-tensor models with screening mechanisms due to their crucial nonlinear effects, which will be discussed in Sec. 1.2.

The case of non-universal coupling

So far, we have considered that all matter fields $\psi_{\rm mat}$ couple in the same way to the scalar field ϕ, through the conformal factor $\Omega(\phi)$. This statement is merely the literal translation of Eq. (1.34b), the matter action of the model we set out to study. We could however imagine a scenario in which the scalar field does not couple universally to the matter sector. Mathematically, this is easily achieved by labeling the matter fields $\psi^{(i)}_{\rm mat}$ and associating each of them with a different conformal factor Ω_i. In doing so, we define as many new metrics via $\tilde{g}^{(i)}_{\mu\nu} = \Omega_i^2(\phi)g_{\mu\nu}$, and the matter action can be decomposed into

$$S_{\rm mat} = \sum_i \int \mathrm{d}^4x\sqrt{-\tilde{g}^{(i)}}\mathcal{L}^{(i)}_{\rm mat}\left(\tilde{g}^{(i)}_{\mu\nu},\,\psi^{(i)}_{\rm mat}\right). \tag{1.80}$$

As such, there is no longer one single 'Jordan frame'. Instead, there is one preferred frame per matter field (j), which is obtained by following the derivation in Eqs. (1.35–1.41) with $\Omega \to \Omega_i$ and $\tilde{g}_{\mu\nu} \to \tilde{g}^{(i)}_{\mu\nu}$. At this point, it is important to remark that the case of non-universal coupling straightforwardly leads to WEP violation, as different particle species follow different geodesics in spacetime. More generally, tweaks in couplings are severely limited because they lead to variations of the physical constants, see Ref. [73].

From a physical point, such non-universal couplings arise in a variety of theoretical models. In many string theory models, the dilaton is a spin-0 field that couples with different strengths to different types of matter, see e.g. Refs. [74–76]. In cosmology, one can introduce a scalar field that couples differently to ordinary matter *vs* dark matter [77, 78].

1.1.3 Observational consequences and tests

In the previous sub-section, we introduced the mathematical framework of scalar-tensor theories with the focus laid on the field equations. With these derived expressions at hand, we go back to the realm of physics with two

applications. First, we reinvest our discussion of FLRW cosmology above and take a look at the cosmological implications of having a scalar field alongside the metric. Second, we examine a question that has been postponed so far: what happens to geodesics in scalar-tensor theories? Specifically, the fact that test particles follow geodesics of the Jordan-frame metric means that the geodesic equation in the Einstein frame will no longer read Eq. (1.5).

But before delving into these two applications, let us briefly discuss the question of the equivalence between frames.

Equivalence between frames

In the above, we have insisted upon the fact that there was a one-to-one correspondence between the Einstein-frame fields $(g_{\mu\nu}, \phi)$ and the Jordan-frame fields $(\tilde{g}_{\mu\nu}, \varphi)$. As such, one is free to write down the action of the theory and the field equations that follow in either of the two frames — or both, as we did for the sake of completeness. That being said, it is possible to shift from one frame to the other at any stage of any given computation. As we have seen however, inconsistencies may arise when one starts making approximations so as to simplify expressions, see e.g. our derivation of the Newtonian limit of the field equations in both frames. In that case, one has to be extra careful about how the 'smallness' of a given quantity translates from one frame to the other.

As stated above, it is not uncommon to read the Jordan frame being referred to as the 'physical frame' in the literature. The reason for this name is that the matter part of the action is the standard one: it is covariantly conserved $\tilde{\nabla}_\mu \tilde{T}^{\mu\nu} = 0$ and all particle physics' properties (e.g. masses, cross sections, decay rates, etc.) can be computed 'as usual', without having to care about the spacetime dependence of the scalar field φ. In turn, the equations describing gravitational phenomena look much more complicated than in pure GR [see e.g. Eqs. (1.53–1.55)]. In this regard, the Einstein frame looks more appealing as the absence of coupling between the curvature R and the scalar field ϕ makes the field equations less intimidating. The price to be paid for this apparent simplicity is that all the aforementioned particle physics' properties become spacetime dependent, due to their ϕ-dependence.

Yet, laws of nature do not care about which frame we pick for making our calculations. In other words, the computation of observable quantities must somehow yield frame-independent results. This is ensured by the fact that physical observables are always dimensionless ratios between physical quantities and the appropriate units of measure [79] — which turns out to be a frame invariant quantity. See e.g. Refs. [80–83] for a more in-depth discussion regarding frame-invariant observables. This 'frame-freedom' is a double-edged sword. On the one side, extra care has to be taken when confronting the theory against experimental data. On the other side, this duality not only allows one to take whatever route is the most convenient computationally speaking, but also offers more room for physical interpretation, mixing views from the two frames.

Finally before moving on, it should be mentioned that, while this equivalence holds for classical phenomena, it may not hold for quantum phenomena, see e.g. Refs. [84–86].

Cosmological implications

In Sec. 1.1.1, we showed how GR is underlying the standard model of cosmology, by providing some basic elements of FLRW cosmology. Here, we follow the same path for a generic scalar-tensor model given by an action Eq. (1.33) [or Eq. (1.41) equivalently in the Jordan frame]. In particular, our aim is to show how scalar-tensor models can provide alternatives to the cosmological constant to account for the observed cosmic acceleration — which is perhaps the reason why these models have been so trendy in the literature since the late 1990s. Therefore, we limit ourselves to the background evolution of the universe in the scalar-tensor framework laid out through Eqs. (1.33–1.79). For the sake of simplicity and without loss of generality, we assume a spatially flat universe [note that this property is frame-independent through the Weyl rescaling Eq. (1.35)].

We continue to discuss the relation between the two frames, although one has to bear in mind that experimental data (redshift, distances, CMB temperature, etc.) have their usual interpretation in the Jordan frame.

Einstein frame and dark energy In the Einstein frame, the line element is

$$ds^2 = -\mathrm{d}t^2 + a^2(t)\delta^{ij}\mathrm{d}x^i\mathrm{d}x^j \,. \tag{1.81}$$

The Friedmann equations (1.22) take exactly the same form as in GR, but the scalar field contribution $T^{(\phi)}_{\mu\nu}$ [Eq. (1.44)] to the total energy-momentum tensor has to be taken into account through the density and pressure terms. From the expression of the stress-energy tensor of a perfect fluid Eq. (1.19), we readily extract $\rho = u^\mu u^\nu T_{\mu\nu}$

and $3p = (g^{\mu\nu} + u^\mu u^\nu)T_{\mu\nu}$. With the assumption that $\partial_i\phi = 0$, we get

$$\rho_\phi = u^\mu u^\nu T^{(\phi)}_{\mu\nu} = \big(u^0\big)^2(\partial_t\phi)^2 + \frac{1}{2}g^{00}(\partial_t\phi)^2 + V(\phi) = \frac{1}{2}(\partial_t\phi)^2 + V(\phi)\,, \tag{1.82a}$$

$$p_\phi = \frac{1}{3}(g^{\mu\nu} + u^\mu u^\nu)T^{(\phi)}_{\mu\nu} = \frac{1}{3}\big[-g^{00}(\partial_t\phi)^2 - 4V(\phi) + \rho_\phi\big] = \frac{1}{2}(\partial_t\phi)^2 - V(\phi)\,, \tag{1.82b}$$

where we have used the fact that the 'scalar fluid' is at rest in comoving coordinates i.e. $u^\mu = \delta^\mu_0$, $g^{00} = -1$ and $g_{\mu\nu}u^\mu u^\nu = -1$. Therefore, the scalar field behaves as a perfect fluid whose equation of state of the scalar field is simply

$$w_\phi = \frac{(\partial_t\phi)^2 - 2V(\phi)}{(\partial_t\phi)^2 + 2V(\phi)} \tag{1.83}$$

and the Friedmann equations are obtained by substituting $\rho \to \rho + \rho_\phi$, $p \to p + p_\phi$ in Eq. (1.22). Likewise, the equation of motion for the scalar field — the Klein–Gordon equation Eq. (1.45) — becomes

$$\partial_t^2\phi + 3H\partial_t\phi + \frac{\mathrm{d}V}{\mathrm{d}\phi} = \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}(3p - \rho)\,, \tag{1.84}$$

where of course $\partial_t^2 = \partial_t\partial_t$. The continuity equation (1.23) is no longer applicable in its original form since $\nabla_\mu T^{\mu\nu} \neq 0$ in the Einstein frame. However, it is easily recovered if we substitute $\rho \to \rho + \rho_\phi$, $p \to p + p_\phi$ as before.

While these Einstein frame Friedmann equations are easy to work with, the usual quantities derived from them do not correspond to the measured ones. For instance, the measured redshift z of a distant object would be

$$1 + z = \frac{\Omega_{\mathrm{rec}}}{\Omega_{\mathrm{em}}}\frac{a_{\mathrm{rec}}}{a_{\mathrm{em}}} \neq \frac{a_{\mathrm{rec}}}{a_{\mathrm{em}}}\,, \tag{1.85}$$

where the subscript 'em' and 'rec' refer to the spacetime events of emission and reception of the light respectively. From Eq. (1.85) we see that, if the conformal factor Ω remained nearly constant (close to unity) between these two events, the ratio $a_{\mathrm{rec}}/a_{\mathrm{em}}$ can dominate the redshift contribution. This turns out to be the case when the scalar field ϕ is weakly coupled to matter. Interestingly, in the limit of no coupling at all $\Omega(\phi) \equiv 1$ at all times, the distinction between the two frames becomes irrelevant and our scalar-tensor model boils down to a quintessence model (see Ref. [87]), which is completely described by the potential function V. The rhs of Eq. (1.84) vanishes and we recover the usual dark energy tale: the scalar field tends to roll down the potential, with the Hubble parameter H acting as a friction term. Therefore, if the potential is shallow enough, the scalar field will roll very slowly, leading to a kinetic term $(\partial_t\phi)^2$ much smaller than the potential term $V(\phi)$. Plugging this approximation into the equation of state Eq. (1.83) yields $w_\phi \sim -1$, which is reminiscent of the equation of state of the cosmological constant. Actually, Friedmann equations (1.22b) show that any $w_\phi < -1/3$ provides an accelerated expansion. Among the most studied potentials are power-law functions — specifically the Ratra–Peebles potential [88] — of the form

$$V(\phi) = \frac{M^{n+4}}{\phi^n}\,, \tag{1.86}$$

where M is some constant energy scale and n is generally a positive number called the slop of the potential. On the one hand, this type of potential is appealing as the energy scale can be compatible with the one from particle physics, meaning that the embarrassing *fine-tuning* issue of the cosmological constant is greatly alleviated [87]. On the other hand, we have yet to find a potential of the form (1.86) in particle physics. In contrast, the simple example of a quadratic potential

$$V(\phi) = \frac{1}{2}m_\phi^2\phi^2\,, \tag{1.87}$$

where m_ϕ is called the *mass* of the scalar field,[18] is way more conventional in quantum field theory. But while this potential has the advantage of being 'particle physics friendly', the fine-tuning issue strikes back as the field's mass would have to be tiny in front of the familiar masses of elementary particles (by dozens of orders of magnitude) [12].

As a side note, we recall that according to the criterion presented in Ref. [62], quintessence is not modified gravity in the sense that the SEP is not violated.

[18]The reason for using the term 'mass' comes from the fact that, upon quantization of the field, momentum eigenstates are collection of particles, each with a mass m_ϕ. At the classical level (to which we are limiting ourselves here), the mass of the field can simply be thought of as a convenient way to characterize the dynamics of the field.

Jordan frame and self-acceleration In the general case, the assumption that the conformal factor temporal evolution has negligible effects on observables does not hold. Indeed, $\Omega(\phi)$ can *a priori* have any shape. In the Jordan frame, the line element is

$$d\tilde{s}^2 = -\mathrm{d}\tilde{t}^{\,2} + \tilde{a}^2(\tilde{t})\delta^{ij}\mathrm{d}x^i\mathrm{d}x^j\,. \tag{1.88}$$

Recalling that $d\tilde{s}^2 = \Omega^2(\phi)\,ds^2$, we get the scaling relations

$$\mathrm{d}\tilde{t} = \Omega\,\mathrm{d}t \qquad \text{and} \qquad \tilde{a} = \Omega\,a \tag{1.89}$$

The universe is still of the FLRW type in the Jordan frame. The background equations follow from the field equations (1.53, 1.56), reading

$$3M_{\rm Pl}^2\tilde{H}\left(F\tilde{H} + \partial_{\tilde{t}}F\right) = \tilde{\rho} + \frac{1}{2}Z\big(\partial_{\tilde{t}}\varphi\big)^2 + U\,, \tag{1.90a}$$

$$-2M_{\rm Pl}^2F\,\partial_{\tilde{t}}\tilde{H} = \tilde{\rho} + \tilde{p} + Z\big(\partial_{\tilde{t}}\varphi\big)^2 + M_{\rm Pl}^2\left(\partial_{\tilde{t}}^2F - \tilde{H}\,\partial_{\tilde{t}}F\right), \tag{1.90b}$$

$$Z\left(\partial_{\tilde{t}}^2\varphi + 3\tilde{H}\,\partial_{\tilde{t}}\varphi\right) = 3M_{\rm Pl}^2\frac{\mathrm{d}F}{\mathrm{d}\varphi}\left(\partial_{\tilde{t}}\tilde{H} + 2\tilde{H}^2\right) - \frac{1}{2}\frac{\mathrm{d}Z}{\mathrm{d}\varphi}\big(\partial_{\tilde{t}}\varphi\big)^2 - \frac{\mathrm{d}U}{\mathrm{d}\varphi}\,, \tag{1.90c}$$

where $\tilde{H} = \mathrm{d}\ln\tilde{a}/\mathrm{d}\tilde{t}$. Thanks to $\tilde{\nabla}_\mu\tilde{T}^{\mu\nu} = 0$, the continuity equation takes its usual form $\partial_{\tilde{t}}\tilde{\rho} + 3\tilde{H}(\tilde{\rho} + \tilde{p}) = 0$.

Let us take the time to discuss Eqs. (1.88–1.90). Eqs. (1.90a, 1.90b) correspond to the first and second Friedmann equations, while Eq. (1.90c) is the scalar field equation in the FLRW background. As stated above, it is easy to derive cosmological observables from Jordan frame quantities. For example, the measured redshift from the light emitted by a distant object is simply

$$1 + z = \frac{\tilde{a}_{\rm rec}}{\tilde{a}_{\rm em}}\,, \tag{1.91}$$

which is consistent with Eq. (1.85) given the fact that the scale factor transforms as $\tilde{a} = \Omega\,a$ [Eq. (1.89)]. A proper discussion of the background evolution of the universe therefore requires to solve the coupled equations (1.90). Such a discussion is undertaken in the illuminating paper by Esposito-Farèse and Polarski [83].

If our scalar-tensor model is to describe our universe, which we *know* is undergoing a phase of accelerated expansion [89, 90], we must have $\mathrm{d}^2\tilde{a}/\mathrm{d}\tilde{t}^2 > 0$ at the present time. Interestingly, this is not enough to fix the sign of $\mathrm{d}^2a/\mathrm{d}t^2$. To give a concrete example, the case of a vanishing potential $V(\phi) \equiv 0$ results in

$$\frac{1}{a}\frac{\mathrm{d}^2a}{\mathrm{d}t^2} = -\frac{1}{6M_{\rm Pl}^2}\left[\rho + 3p + 2(\partial_t\phi)^2\right], \tag{1.92}$$

so that the universe is clearly decelerating in the Einstein frame. In the Jordan frame however, the relation $\tilde{a} = \Omega\,a$ [see Eq. (1.89)] prevents us from drawing that same conclusion. Whether the universe is decelerating in the Jordan frame is a question that cannot be answered unless the conformal factor function is specified. In this perspective, cosmologists have coined the term 'self-acceleration' [91–94] to designate scalar-tensor models which simultaneously exhibit an accelerated expansion in the Jordan frame and a decelerated expansion in the Einstein frame. A model that *self-accelerates* can be appealing. Indeed, the cosmic acceleration stems entirely from the Weyl rescaling [Eq. (1.89)] and not from a putative dark energy fluid which, as discussed above, can suffer from fine-tuning issues or non-standard potentials in particle physics.

Fifth forces

One inescapable topic when discussing scalar-tensor models is the motion of massive and massless particles.

Timelike geodesics As already said, matter test particles follow geodesics of the Jordan-frame metric $\tilde{g}_{\mu\nu}$. As such, it is straightforward to write the geodesic equation in that frame, as all we have to do is translate Eq. (1.7) in terms of Jordan frame quantities. Denoting $\tilde{u}^\mu = \mathrm{d}x^\mu/\mathrm{d}\tilde{\tau}$ the 4-velocity of such a test particle in free fall, we have

$$\tilde{u}^\alpha\tilde{\nabla}_\alpha\tilde{u}^\mu = 0\,, \qquad \text{with} \quad \tilde{g}_{\alpha\beta}\tilde{u}^\alpha\tilde{u}^\beta = -1\,. \tag{1.93}$$

Now let us have a look at what happens in the Einstein frame. Precisely, that means finding an expression for $u^\alpha\nabla_\alpha u^\mu$, where $u^\mu = \mathrm{d}x^\mu/\mathrm{d}\tau = \Omega^{-1}\tilde{u}^\mu$ so that $g_{\alpha\beta}u^\alpha u^\beta = -1$ (u^μ is still tangent to the Jordan frame geodesic). For any vector A_μ, the covariant derivatives are related through

$$\nabla_\alpha A_\mu = \tilde{\nabla}_\alpha A_\mu + \big(\tilde{\Gamma}^\lambda_{\alpha\mu} - \Gamma^\lambda_{\alpha\mu}\big)A_\lambda\,, \qquad \text{with} \quad \tilde{\Gamma}^\lambda_{\alpha\mu} - \Gamma^\lambda_{\alpha\mu} = \Omega^{-1}\left(\delta^\lambda_\alpha\partial_\mu\Omega + \delta^\lambda_\mu\partial_\alpha\Omega - g_{\alpha\mu}g^{\lambda\sigma}\partial_\sigma\Omega\right) \tag{1.94}$$

Therefore, we get

$$\begin{aligned}\tilde{\nabla}_\alpha \tilde{u}^\mu &= \tilde{g}^{\mu\nu}\tilde{\nabla}_\alpha \tilde{u}_\nu \\ &= \tilde{g}^{\mu\nu}\left[\nabla_\alpha \tilde{u}_\nu - \big(\tilde{\Gamma}^\lambda_{\alpha\nu} - \Gamma^\lambda_{\alpha\nu}\big)\tilde{u}_\lambda\right] \\ &= \Omega^{-2} g^{\mu\nu}\left[u_\nu \partial_\alpha \Omega + \Omega \nabla_\alpha u_\nu - \big(\delta^\lambda_\alpha \partial_\nu \Omega + \delta^\lambda_\nu \partial_\alpha \Omega - g_{\alpha\nu} g^{\lambda\sigma}\partial_\sigma \Omega\big) u_\lambda\right] \\ &= \Omega^{-1}\Big\{\partial_\alpha(\ln\Omega) u^\mu + \nabla_\alpha u^\mu - \big[\delta^\lambda_\alpha \partial_\nu(\ln\Omega) + \delta^\lambda_\nu \partial_\alpha(\ln\Omega) - g_{\alpha\nu} g^{\lambda\sigma}\partial_\sigma(\ln\Omega)\big] g^{\mu\nu} u_\lambda\Big\},\end{aligned}$$

where we have made use of the fact that $\nabla_\alpha g^{\mu\nu} = \tilde{\nabla}_\alpha \tilde{g}^{\mu\nu} = 0$. Finally, contracting this expression by $\tilde{u}^\alpha$ yields

$$\tilde{u}^\alpha \tilde{\nabla}_\alpha \tilde{u}^\mu = \Omega^{-1} u^\alpha \tilde{\nabla}_\alpha \tilde{u}^\alpha = \Omega^{-2}\left[u^\alpha \nabla_\alpha u^\mu + \big(g^{\mu\nu} + u^\mu u^\nu\big)\partial_\nu(\ln\Omega)\right] = 0\,, \tag{1.95}$$

where we have used the fact $u_\lambda u^\lambda = -1$. What we find is that geodesics of the Jordan-frame metric do not coincide with the Einstein-frame metric ones, since

$$u^\alpha \nabla_\alpha u^\mu = -\perp^{\mu\nu}\partial_\nu(\ln\Omega) = -\perp^{\mu\nu}\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\partial_\nu\phi\,, \qquad \text{with} \quad \perp^{\mu\nu} = g^{\mu\nu} + u^\mu u^\nu\,. \tag{1.96}$$

Note that the $\perp^{\mu\nu}$ can be interpreted as the projector on the 3-space normal to u^μ. In plain language, Eq. (1.96) tells us that test particles do not follow geodesics of $g_{\mu\nu}$ (otherwise, the rhs would be zero), their trajectory being perturbed by a term that depends on the conformal factor and the gradient of the scalar field. From the Einstein frame perspective, everything happens as if the particles were subjected to a force along their trajectory, hence their non-geodesic motion. This force — called the fifth force in the literature[19] — can be readily expressed as

$$a^\rho_\phi = -\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\perp^{\mu\rho}\partial_\mu\phi\,. \tag{1.97}$$

It is worth mentioning that this derivation is valid for any two metric conformally related as in Eq. (1.35) and thus does not depend at all on the action of the theory at hand (as long as the coupling remains universal).

To gain further insight into this fifth force, we proceed as in Sec. 1.1.1 and compute the Newtonian limit of the modified geodesic equation (1.96) in the Einstein frame. Putting the metric $g_{\mu\nu}$ in the Newtonian gauge (1.46), we obtain

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} = -\partial_i\Phi - \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\partial_i\phi \qquad \text{i.e.} \quad \frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}t^2} = -\boldsymbol{\nabla}\Phi - \frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\boldsymbol{\nabla}\phi \tag{1.98}$$

We recover the acceleration due to the gravitational potential Φ, as in GR, plus the 3-force $\mathbf{a}_\phi = -\boldsymbol{\nabla}[\ln\Omega(\phi)]$.[20] For the case of an extended body of volume $\mathcal{V}$ and mass M rather than a point-like particle, the total fifth force experienced by that body is obtained through the integration

$$\mathbf{a}_\phi = \frac{1}{M}\int_{\mathcal{V}} \rho(\mathbf{x})\boldsymbol{\nabla}\big[\ln\Omega(\phi)\big]\,\mathrm{d}\mathbf{x}\,. \tag{1.99}$$

Here, it should be stressed that, while the WEP holds for point-like particles (as they do not disturb the background field ϕ), the same cannot be said about extended test bodies. Indeed, there is no obvious reason why the fifth force (1.99) should be the same for all bodies with matter distribution $\rho(\mathbf{x})$ and scalar field profile $\phi(\mathbf{x})$. A concrete example of such an *apparent* WEP violation is given later in Sec. 1.2.2.

Null geodesics It is obvious that conformal transformations leave null *curves* invariant, since $\tilde{g}_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = 0$ is equivalent to $g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu = 0$. Actually, we can show that they further leave null *geodesics* invariant. Let $k^\mu = \mathrm{d}x^\mu/\mathrm{d}\lambda$ be the tangent vector to a null geodesic of the Einstein-frame metric, affinely parameterized by λ. Then $k^\alpha\nabla_\alpha k^\mu = 0$, by definition. In particular k^μ is a null vector, which implies that (see e.g. Ref. [13], Appendix G)

$$k^\alpha\tilde{\nabla}_\alpha k^\mu = \big[2k^\alpha\partial_\alpha(\ln\Omega)\big]k^\mu\,.$$

This is the the equation of a non-affinely parameterized geodesic of the Jordan-frame metric — which ends the proof. Note that we can further define a new affine parameter $\tilde{\lambda}$, related to λ via $\mathrm{d}\tilde{\lambda} = \Omega^2\mathrm{d}\lambda$, so that

$$\tilde{k}^\alpha\tilde{\nabla}_\alpha\tilde{k}^\mu = 0\,, \qquad \text{with} \quad \tilde{k}^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tilde{\lambda}} = \Omega^{-2}k^\mu\,.$$

[19] The name comes from the fact it would be a new force, mediated by the scalar field, alongside the four known fundamental interactions in nature.

[20] We acknowledge being rather sloppy name-wise as 'fifth force' should sometimes be understood as 'fifth acceleration', but that is clear enough from context.

Given the conformal invariance of null geodesics, massless particles are not affected by the presence of the scalar field and do not 'feel' any fifth force. This property can be leveraged in gravity tests that involve the bending of light by a massive body as the *lensing mass* (inferred from lensing) is different from the dynamical mass (inferred from the dynamics of other objects around that body) — see e.g. Ref. [71] for a nice discussion of this effect.

Massless and massive scalar fields explained through two examples: Brans–Dicke theory and the Yukawa approximation

Brans–Dicke theory This theory is considered to be the very first scalar-tensor model, proposed by Brans and Dicke in 1962 [79, 95] (building on top of the earlier work of Jordan) as a modification of GR that respects Mach's principle [96]. It is the prototypical example of a massless scalar-tensor theory whose action, in the Jordan frame, corresponds to setting

$$F(\varphi) = \varphi \,, \qquad Z(\varphi) = M_{\rm Pl}^2 \frac{\omega}{\varphi} \,, \qquad U(\varphi) = 0 \,, \tag{1.100}$$

where $\omega > 0$ is a constant parameter of the theory. In this theory, the scalar field φ is said to be massless because the potential function U is set to zero. Applying the Newtonian approximation in the Jordan frame [Eqs. (1.64, 1.66)] to this special case yields

$$2M_{\rm Pl}^2 \varphi \tilde{\Delta}\tilde{\Phi} = \frac{2\omega+4}{2\omega+3}\tilde{\rho} \qquad \text{and} \qquad 2M_{\rm Pl}^2 \tilde{\Delta}\varphi \left(\omega + \frac{3}{2}\right) = -\tilde{\rho} \,. \tag{1.101}$$

Interestingly, the scalar field obeys a linear Poisson equation (just like the gravitational potential in Newtonian gravity). The equation for the potential $\tilde{\Phi}$ can be interpreted as having an effective gravitational constant G_* instead of the 'bare' one G embedded in $M_{\rm Pl}^2$ [see Eq. (1.1)], reading

$$G_*(\varphi) = \frac{G}{\varphi}\frac{2\omega+4}{2\omega+3} \simeq \frac{G}{\varphi_0}\frac{2\omega+4}{2\omega+3}\big(1 - \delta\varphi(\mathbf{x})\big) \,, \tag{1.102}$$

where we have further made the assumption that φ can be expanded around a nonzero background $\varphi = \varphi_0(1+\delta\varphi)$ with $|\delta\varphi| \ll 1$. The background value of the scalar field φ_0 can be interpreted as the value of φ today far from the system being studied, which is therefore determined by an appropriate cosmological boundary condition[21] [3, 97]. The gravitational "constant" G_* corresponds to what one would measure in a Cavendish-like experiment (see e.g. Refs. [98, 99] for the description of such experiments in the laboratory). In this theory, it ceases to be a true constant for

1. the background value φ_0 may change as a result of cosmic evolution (see Ref. [73] for a comprehensive review of '$\dot{G}/G$' constraints);

2. small fluctuations around this background $\delta\varphi$ arise, meaning G_* could also vary in space.

It is often said in the literature that GR is recovered in the limit $\omega \to \infty$.[22] A somewhat hand-wavy argument to see this is the following: it is clear that the field equations (1.53) reduce to the Einstein field equations when φ cease to be a dynamical quantity. Roughly speaking, in the limit of big ω, one finds from Eq. (1.56) $\tilde{\Box}\varphi \sim 0$ which admits constants as solutions.

Actual constraints on ω are best brought out using the PPN formalism. Specifically, the γ parameter (which is related to the bending of light by massive objects and the Shapiro time delay, see Table 1.2) can be computed within the Brans–Dicke theory, yielding $\gamma = (1+\omega)/(2+\omega)$. The Cassini–Huygens experiment puts the best lower bound to date $\omega > 4 \times 10^4$ [36]. To get a sense of why the PPN parameter γ can be used to put constraints on the model at stake, it is insightful to consider light deflection measurements. On the one hand, we have seen above that the trajectory of massless particles is not affected by the presence of the scalar field. In this respect, the deflection angle must be proportional to the bare gravitational constant G appearing in the action of the theory through the reduced Planck mass $M_{\rm Pl}$. On the other hand, the strength of the force between massive bodies is proportional to the effective gravitational 'constant' G_* given by Eq. (1.102) and which is measured in Cavendish-like experiments. The latter being dependent of the scalar field and the value of model's parameter(s), it is different from G. Therefore, the Brans–Dicke model can be tested by checking the consistency

[21] Indeed, it is clear that φ_0 has to be specified, for otherwise Eq. (1.101) is obviously not a well-posed PDE problem! Note that the notion of well-posedness will be discussed more in depth in Chapt. 2.

[22] This statement is of course not very rigorous, and giving a proper mathematical meaning involves all the subtleties associated with the topology of functional spaces. Nonetheless, it is true that as far as Solar system experiment are concerned, the predictions of the Brans–Dicke theory given by Eqs. (1.41, 1.100) approach those of GR as $\omega \to \infty$ (see Brans' own historical perspective in Refs. [100, 101]).

of the inferred value of the gravitational constant from the deflection angle of light by a massive body and the dynamics of other bodies around it. This test principle is relevant not only for the Brans–Dicke model but for any massless scalar-tensor theory.

As a consequence of the large lower bound on ω, the Brans–Dicke theory as introduced here with Eq. (1.101) is no longer of much interest as its predictions are basically those of GR.[23] A relevant extension is to promote the constant ω to a function of the scalar field $\omega(\varphi)$. In particular, Damour and Esposito-Farèse [102] provide a framework for studying general massless scalar-tensor theories, which is more concisely laid out in Ref. [70] Sec. 4.

Finally, this discussion could have equally been made in the Einstein frame, see e.g. Ref. [103].

Massive scalar field and the Yukawa potential Another example worth developing is the case of a massive scalar field ϕ. The canonical example of a scalar field with constant mass is when the potential V is quadratic as in Eq. (1.87), $V(\phi) = m_\phi^2\phi^2/2$. For the model to be fully specified, we set $\Omega(\phi) = \exp(\beta\phi/M_{\rm Pl})$ where $\beta > 0$ is some dimensionless coupling constant. In this framework, m_ϕ is called the mass of the scalar field[18]. For a more general potential exhibiting a minimum at $\phi_{\rm min}$, the concept of mass can still be defined as $m_\phi = \sqrt{V''(\phi_{\rm min})}$. Note that with this choice of functions V and Ω, the 'effective mass' defined in Box C coincides with standard mass defined from the potential alone.

Box C: Effective potential & Effective mass of the scalar field

Looking at the field equation (1.45) for the scalar field ϕ in the Einstein frame, it is natural to define an *effective potential*

$$V_{\rm eff}(\phi) = V(\phi) - T\ln\Omega\,, \tag{1.103}$$

so that the dynamics is described by $\Box\phi = V'_{\rm eff}(\phi)$.

The concept of the scalar field having an *effective mass* m_ϕ can be defined as long as the effective potential exhibits a minimum at some $\phi_{\rm min}$. In that case, one sets

$$m_\phi^2 = \frac{{\rm d}^2V_{\rm eff}}{{\rm d}\phi^2}(\phi_{\rm min})\,. \tag{1.104}$$

This barely comes from the fact that, when the field oscillates close to the minimum of its effective potential, we have the following Taylor expansion

$$\begin{aligned} V_{\rm eff}(\phi) &\simeq V_{\rm eff}(\phi_{\rm min}) + (\phi-\phi_{\rm min})V'_{\rm eff}(\phi_{\rm min}) + \frac{1}{2}(\phi-\phi_{\rm min})^2V''_{\rm eff}(\phi_{\rm min}) \\ &\simeq V_{\rm eff}(\phi_{\rm min}) + \frac{1}{2}m_\phi^2(\phi-\phi_{\rm min})^2\,, \end{aligned} \tag{1.105}$$

the constant $V_{\rm eff}(\phi_{\rm min})$ having no effect on the dynamics, see Eq. (1.45).

Assuming $|\beta\phi/M_{\rm Pl}| \ll 1$ leads to $\tilde\rho \sim \rho$, and the scalar field equation (1.45) in the weak field, static case, reduces to

$$\Delta\phi = m_\phi^2\phi + \frac{\beta}{M_{\rm Pl}}\rho\,. \tag{1.106}$$

This Klein–Gordon equation is linear, which greatly simplify its study. In particular, we know the free-space Green function of the $(\Delta - m_\phi^2)$ operator to be

$$\mathcal{G}_Y(r) = -\frac{\exp(-m_\phi r)}{4\pi r} \qquad \text{so that} \quad \phi(\mathbf{r}) = -\frac{\beta}{4\pi M_{\rm Pl}}\int_{\mathcal{V}}\frac{{\rm e}^{-m_\phi\|\mathbf{r}-\mathbf{r}'\|}}{\|\mathbf{r}-\mathbf{r}'\|}\rho(\mathbf{r}')\,{\rm d}^3\mathbf{r}' \tag{1.107}$$

for the field generated by an extended body of volume $\mathcal{V}$. The total acceleration created by a point-particle of mass M [i.e. $\rho(\mathbf{r}') = M\delta(\mathbf{r}')$] is the sum of the Newtonian acceleration and the scalar acceleration [Eq. (1.98)], that is

$$\frac{{\rm d}\mathbf{x}^2}{{\rm d}t^2} = \boldsymbol{\nabla}\left[\frac{GM}{r}\left(1+2\beta^2{\rm e}^{-m_\phi r}\right)\right] = -\frac{GM}{r^2}\left[1+2\beta^2(1+m_\phi r){\rm e}^{-m_\phi r}\right]\mathbf{e}_r\,. \tag{1.108}$$

The presence of the scalar fifth force can thus be interpreted as a modification of the standard Newtonian potential $\Phi = -GM/r$ from which the acceleration due to gravity is computed. This modified potential is called

[23] This is an example of the application of Occam's razor; or to quote from Einstein *"It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."* https://www.nature.com/articles/d41586-018-05004-4.

the Yukawa potential V_Y and reads (still for a point mass)

$$V_Y(r) = -\frac{GM}{r}\left(1 + 2\beta^2 \mathrm{e}^{-m_\phi r}\right) = -\frac{GM}{r}\left[1 + \alpha \exp\left(-\frac{r}{\lambda}\right)\right], \tag{1.109}$$

where we have set $\alpha = 2\beta^2$ and $\lambda = 1/m_\phi$ in order to recover the canonical form of the Yukawa potential often found in the literature — see e.g. Refs. [104–108]. We see from Eq. (1.108) that the fifth force gets exponentially suppressed for distances $r > \lambda$. Therefore, it makes sense to refer to λ as a 'Compton wavelength', the typical length scale of the interaction mediated by the scalar field.

Let us emphasize the fact that this canonical example of a massive scalar-tensor theory is of particular importance. The phenomenology we just derived with the Yukawa potential and the associated gravitational acceleration [Eqs. (1.108, 1.109)] is *universal* to all scalar-tensor models for which the scalar field has an effective mass — see Box C. Of course, this is to be understood in a qualitative sense since Eqs. (1.108, 1.109) are only exact for the case of a quadratic potential [Eq. (1.87)] and conformal factor $\Omega = \beta\phi/M_{\rm Pl}$. Nevertheless, any (local) minimum of the effective potential (provided it has some) can be approximated as a quadratic function, see Eq. (1.105), from which an effective mass can be extracted. When the scalar field is close to that minimum, its field equation becomes well-approximated by Eq. (1.106) and the Yukawa potential is a good approximation of the 'total' gravitational potential. In particular, the finite range of the fifth force is typical of massive scalar-tensor models. In contrast, massless scalars mediate infinite-range interactions, as illustrated with the prime example of massless Brans–Dicke theory just above.

The fact that the Yukawa approximation applies to a wide range of scalar-tensor models makes it a good choice for parameterizing deviations from GR, as it most often captures the essence of the model at hand. In that respect, it is useful to try and put constraints on the parameter space given by the pair of model-independent parameters (α, λ) — see e.g. Figs. 11 & 12 from Ref. [63] — as these constraints can sometimes be mapped to physically-motivated models with a massive scalar.

The need to hide light scalars

As already stated, scalar fields are candidates for explaining astrophysical effects associated with dark matter [109, 110] and the apparent cosmic expansion acceleration [87]. In many of such models, the scalar field needs to be *light* enough to fulfill its number one *raison d'être*, that is to fit observations, with masses sometimes as low as the Hubble constant i.e. $\sim 10^{-33}\,\mathrm{eV}$ [88].

This should immediately be put into perspective with our discussion of fifth forces in scalar-tensor models just above. As a matter of fact, light scalar fields mediate long-range interactions which constitute deviations from GR through, among other things, a modified *apparent* gravitational constant. This is obviously problematic as GR is very well tested, especially at Solar system scales (see Sec. 1.1.1). The moral of the story is that, within the framework of scalar-tensor theories, modifying gravity is a tricky game to play: we want the scalar field to explain astrophysical conundra (dark matter or dark energy related) — which most often requires the scalar field to be light — and, at the same time, be consistent with basically all tests of gravity. Unfortunately, these two constraints are often incompatible and many models can be *killed* using the powerful PPN-formalism together with the tight PPN bounds (see Table 1.2) — see Refs. [1, 3].

The first workaround that comes to mind is to willingly decouple the scalar field from the matter sector by setting $\Omega(\phi) \equiv 1$. The coupling vanishes and so does the fifth force (recovering a quintessence model). Problem solved? Not really since from a quantum mechanical point of view, the introduction of a scalar field in the gravitational sector *always* generates a coupling between this scalar field and matter. Additionally, setting $\Omega \equiv 1$ reduces the space of possible models (our only freedom is to specify the potential function V), thereby lessening the scope of phenomenologies.

1.2 Screening mechanisms

The apparent incompatibility between light scalar fields having relevant astrophysical phenomenology and the fact that they mediate long-range fifth forces puts us in a dead-end. Fortunately, some theoretical physicists are quite stubborn and discovered some ways out: *screening mechanisms*.

Screening mechanisms are theoretical constructs designed, or better said engineered, to hide or "screen" the effects of the scalar field in Earth-based and Solar system experiments, while allowing for deviations from GR at astrophysical and cosmological scales. In fact, some clever choices of Lagrangians can dynamically suppress fifth forces at scales for which gravity is well-tested. As we are going to see, this can be achieved in a variety of ways, though all screening mechanisms have in common the fact that they are enabled through nonlinearities.

In this section, we briefly review the existing mechanisms in the literature, with a focus on the chameleon mechanism. In particular, we will see that despite being advertised as convenient ways to hide scalars in the

laboratory and in the Solar system, scalar-tensor models equipped with screening mechanisms do nonetheless predict deviations from GR, however small they might be.

1.2.1 A convenient classification of screening mechanisms

There is of course not a unique way to classify screening mechanisms. The way we proceed here is in the spirit of the review Ref. [111]. One can adopt either of the two following perspectives:

1. *Mathematical perspective* — The focus is put on the Lagrangian of the model and the equation of motion of the scalar field that results from it. Specifically, we look at where the nonlinearity manifests itself in the partial differential equation (PDE). Depending on the mechanism at stake, it can appear on ϕ-terms, $\partial\phi$-terms or $\partial^2\phi$-terms, corresponding to semi-linear, quasi-linear and fully nonlinear PDE. This classification of nonlinear PDEs is explained in more details in Chap. 2, Sec. 2.2.

2. *Force-law perspective* — The idea is to underline the physical reasons that lead to a suppression of the fifth force in the local environment. The available levers are *(i)* the coupling strength of the scalar field matter, *(ii)* the mass of the scalar field and *(iii)* its kinetic term.

These two classifications are not aligned in the sense that there is no one-to-one map between physical levers and PDE type. We start with the screening mechanisms that can be implemented within the traditional scalar-tensor theory subclass that we introduced in Sec. 1.1.2 and then extend the framework to Horndeski's theories. A summary of the discussed mechanisms can be found in Table 1.3 below.

Screening mechanisms designed from the traditional scalar-tensor theory subclass

It is natural to begin with the traditional scalar-tensor theory subclass — the one introduced in Sec. 1.1.2 and more specifically in Eqs. (1.33–1.41) — as we already have gained insight into this model. From the Einstein frame perspective, we only have two degrees of freedom to produce screening, namely the conformal factor function $\Omega(\phi)$ and the potential $V(\phi)$.

Weak coupling At the end of Sec.1.1.3, we suggested that suppressing the fifth force was as simple as turning the coupling off (i.e. set $\Omega \equiv 1$). For reasons already touched upon, it is not always desirable for the scalar force to be universally weak. However, making this coupling environmentally weak is an idea worth keeping. Indeed, local tests of gravity can be passed provided that the coupling becomes sufficiently small in regions of high density (or equivalently, of high Newtonian potential). This is called the weak coupling principle and underlies at least two screening mechanisms.

The first one is the *symmetron mechanism* [112, 113], for which the coupling of the scalar to matter is proportional to the vacuum expectation value (VEV) of the scalar field. The effective potential (combination of the two functions Ω and V, see Box C) is chosen such that *(i)* this VEV is nonzero in low-density environment, and *(ii)* the $\mathbb{Z}_2$-symmetry $\phi \to -\phi$ is restored in high-density regions, so that the field have a zero VEV in such regions and does not couple to matter. A choice of functions that produces this behavior is

$$\Omega(\phi) = 1 + \frac{\phi^2}{2M^2} \qquad \text{and} \qquad V(\phi) = -\frac{\mu^2}{2}\phi^2 + \frac{\lambda}{4}\phi^4\,, \tag{1.110}$$

where M is some high mass scale ($\phi \ll M$) and μ, λ are also model parameters.

The *Damour–Polyakov* mechanism (or *least coupling principle*) [74, 114], in the same vein as the symmetron, is another density-dependent screening mechanism with

$$\Omega(\phi) = 1 + \frac{1}{2M}(\phi - \phi_\star)^2 \qquad \text{and} \qquad V(\phi) = V_0 \exp\left(-\frac{\phi}{M_{\rm Pl}}\right). \tag{1.111}$$

Again, M, $\phi_\star$, V_0 are model parameters.

Large mass The other 'knob' we have at our disposal is the mass of the scalar field. As we have seen in Sec. 1.1.2, the concept of mass for a scalar field is a good indicator of the range of the interaction mediated by that scalar field — the Compton wavelength being inversely proportional to the mass. Therefore, another way to hide the scalar field is to make its effective mass environment-dependent: heavy in high-density environment and light in low-density regions (as in deep space). In this way, the field would have barely detectable effects in the Solar system while playing its intended role of a light scalar at astrophysical and cosmological scales.

Mechanism	Type of Equation	Rule of thumb
Weak coupling – Symmetron [112, 113] – Damour–Polyakov [74, 114] *Large mass* – Chameleon [115, 116, 124]	$\Box\phi = \dfrac{\mathrm{d}V}{\mathrm{d}\phi} - \dfrac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}T$	Occurs in regions of high Newtonian potential $\lvert\Phi\rvert$
Large inertia – Kinetic screening [117–120]	$\Box\phi + A_1\partial_\mu\Big[\big(\partial\phi\big)^2\partial^\mu\phi\Big] + A_2 T = 0$	Occurs in regions where the gravitational acceleration $\lvert\boldsymbol{\nabla}\Phi\rvert$ is large
– Vainshtein [121–123]	$6\Box\phi + B_1\Big[\big(\Box\phi\big)^2 - \big(\partial_\mu\partial_\nu\phi\big)^2\Big] = B_2 T$	Occurs in regions where the spatial curvature $\lvert\Delta\Phi\rvert$ is large

Table 1.3: Classification of the most popular screening mechanisms found in the literature — based on Ref. [111]. A_1, A_2, B_1, B_2 are model-dependent constants which are irrelevant here.

This idea is best illustrated by the *chameleon mechanism*, first introduced by Khoury and Weltman in Refs. [115, 116]. In this model, the conformal factor and 'quintessence-inspired' potential are of the form

$$\Omega(\phi) = \exp\left(\frac{\beta\phi}{M_{\rm Pl}}\right) \qquad \text{and} \qquad V(\phi) = \Lambda^4\left(\frac{\Lambda}{\phi}\right)^n, \tag{1.112}$$

where β is some coupling constant, Λ is some energy scale and n is the slope of the potential. In Sec. 1.2.2 below, we will delve further into this particular model. Besides, Appendix C deals with the well-posedness of the nonlinear Klein–Gordon equations arising in the symmetron and chameleon models.

Large inertia

Other 'screening phenomenologies' can be developed in the context of more general scalar-tensor theories. As a matter of fact, the mechanisms introduced so far were all specific cases of the action Eqs. (1.33–1.34) and could therefore only rely on the scalar self-coupling and/or coupling to matter. Quite naturally, we get more room for engineering screening mechanisms if we allow terms involving the derivatives of the scalar field to be part of the Lagrangian (aside from the canonical kinetic term which is what makes the field dynamical). In this respect, the models we are about to introduce belong to the wider class of Horndeski's theories — see Fig. 1.2 and Eqs. (1.31–1.32). These are slightly beyond the scope of the PhD work so we do not delve into them.

Kinetic screening Kinetic screening relies on the introduction of higher powers of the kinetic term $X = -g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi$ in the Lagrangian on the Einstein-frame action, gathered in $P(X)$. From a physical point of view, doing so is well-motivated in the context of K-inflation models, K-essence models, Dirac–Born–Infeld models, etc. (see e.g. Ref. [117]). The simplest example is $P(X) = X - (X/\Lambda^2)^2$, for which the force mediated by the scalar field is suppressed in regions where the gradient of the Newtonian potential is high. See Refs. [117–120] for detailed examples.

Vainshtein mechanism The Vainshtein mechanism [121–123] goes a step further as it requires the introduction of second-order derivatives of the scalar field in the Lagrangian. The latter must be well-chosen to avoid having derivatives beyond second-order in the equation of motion [which is granted in the context of Horndeski theories given by Eqs. (1.31–1.32)]. Nonlinear effects suppress deviation from GR below the so-called Vainshtein radius nearby massive bodies. They effectively kick in when the spatial curvature $\sim \Delta\Phi$ becomes large. In particular, the Vainshtein mechanism can manifest itself in bimetric gravity, galileon models and massive gravity.

Of course this short review is by no means exhaustive. Other mechanisms in the framework of scalar-tensor theories are discussed in the literature — see e.g. the pressuron [125, 126], the runaway-dilaton [127, 128] — but they always rely on the principles set out above (weak coupling, large mass or large inertia). It is also worth mentioning that the phenomenology of screening mechanism can be broaden once we relax the hypothesis of having only one scalar field. For instance, Refs. [129–131] show that the nonlinear interplay between a light axion and a dilaton can effectively screen the latter.

1.2.2 Focus on the chameleon mechanism

The chameleon mechanism was shortly discussed above. It operates when the effective mass of the scalar field is designed to be environment-dependent (hence the name): light in low-density regions of the universe, and heavy in high-density regions. Here, we discuss this mechanism more in depth. It is to be noted that many articles in the literature (including our own) refer to the chameleon as a 'model' rather than a 'mechanism'. Clearly, different combinations of conformal factor and potential functions (Ω, V) can exhibit the chameleon screening mechanism. Be that as it may, the way it was first showcased by Khoury and Weltman in Refs. [115, 116], with an exponential coupling Ω and power-law potential V specified by Eq. (1.112), became to be known as the chameleon model. For instance, let us mention that certain $f(R)$ models can screen using the chameleon mechanism — see Hu and Sawicki models [132] and Appendix B.1 for the mapping of $f(R)$ theories to scalar-tensor theories.

Tutorial: how to make the effective mass environment-dependent

Following Box C, the effective mass of the scalar field is given by Eq. (1.104), where

$$\frac{\mathrm{d}^2 V_{\rm eff}}{\mathrm{d}\phi^2} = \frac{\mathrm{d}^2 V}{\mathrm{d}\phi^2} - \frac{\mathrm{d}^2 \ln\Omega}{\mathrm{d}\phi^2} T \simeq \frac{\mathrm{d}^2 V}{\mathrm{d}\phi^2} + \frac{\mathrm{d}^2 \ln\Omega}{\mathrm{d}\phi^2}\rho\,. \tag{1.113}$$

The specifications are the following:

1. First of all, the effective potential should be such that a minimum actually exists (otherwise this whole discussion is pointless). Looking at the form of $V_{\rm eff}$ [Eq. (1.103)], we see that a necessary condition for this to be true is to have $V' \times \mathrm{d}\ln\Omega/\mathrm{d}\phi < 0$ on a given interval of ϕ-values. A simple choice that fulfills this requirement is to take $V' < 0$ and $\mathrm{d}\ln\Omega/\mathrm{d}\phi > 0$ globally. Specifically, $\Omega(\phi) = \exp(\beta\phi/M_{\rm Pl})$ with $\beta > 0$ does the job and is quite natural from a theoretical point of view.[24]

2. With this specific choice of conformal factor function, Eq. (1.113) simplifies to $V''_{\rm eff} = V''$. Therefore, the condition that the extremum be a minimum implies that V'' must be positive around it. A simple way to ensure this property is to have $V''(\phi) > 0$ over the whole range of ϕ-values.

3. Finally, for the model to be a chameleon, the crucial part is for the effective mass to be an increasing function of the density. The value $\phi_{\rm min}$ that minimizes the effective potential is such that

$$V'_{\rm eff}(\phi_{\rm min}) = 0 \iff V'(\phi_{\rm min}) = -\beta\rho/M_{\rm Pl}\,, \tag{1.114}$$

   highlighting the fact that $\phi_{\rm min}$ is a function of the density, and so is the mass. Taking the derivative of (the square of) the latter with respect to ρ yields

$$\frac{\mathrm{d}m_\phi^2}{\mathrm{d}\rho} = \frac{\mathrm{d}}{\mathrm{d}\rho}\Big[V''\big(\phi_{\rm min}(\rho)\big)\Big] = V'''\big(\phi_{\rm min}(\rho)\big)\frac{\mathrm{d}\phi_{\rm min}}{\mathrm{d}\rho} > 0\,. \tag{1.115}$$

   The way $\phi_{\rm min}$ depends on ρ is obtained by computing the derivative of Eq. (1.114) with respect to ρ, yielding

$$V''\big(\phi_{\rm min}(\rho)\big)\frac{\mathrm{d}\phi_{\rm min}}{\mathrm{d}\rho} = -\frac{\beta}{M_{\rm Pl}} < 0\,. \tag{1.116}$$

   Having taken $V'' > 0$, this readily implies that $\phi_{\rm min}$ is a decreasing function of the density, and so Eq. (1.115) provides the condition $V''' < 0$.

In particular, the functions proposed in Eq. (1.112) can be shown to satisfy all the above conditions[25] and exhibit the chameleon mechanism when $n \in 2\mathbb{Z}^*_- \setminus \{-2\} \cup \mathbb{R}^*_+$, where $2\mathbb{Z}^*_-$ designates the set of all strictly negative even numbers.

This whole discussion is perhaps best summarized by Fig. 1.3 which shows how density shapes the effective potential function for $n > 0$. The behavior for $n \in 2\mathbb{Z}^*_- \setminus \{-2\}$ is illustrated in Ref. [71], Fig. 1. In particular, this shows that the chameleon field has a nonzero effective mass (which is moreover density-dependent). Therefore, the phenomenology of massive scalar fields — discussed through the example of the Yukawa potential in Sec. 1.1.3 — also applies, to some extent, to the chameleon model and can (should!) be leveraged to gain insight into its own phenomenology.

[24] Writing the conformal factor function Ω as an exponential is often considered a 'natural' choice. Indeed, it arises in massless Brans–Dicke theory [see Eq. (1.100)] and in other more fundamental theoretical contexts such as string theory and other higher-dimensional theories. As an example, we show how this happens in $(4+d)$-dimensional Kaluza–Klein theory in Appendix B.2. Recasting $f(R)$ theories into scalar-tensor theories (see Appendix B.1) in the Einstein frame also yields an exponential conformal factor.

[25] One should bear in mind that the conditions derived above are *necessary* conditions for the model to exhibit the chameleon mechanism. We have not actually shown that they were *sufficient* conditions, although one can easily check that the functions V and Ω given by Eq. (1.112) indeed produce the desired behavior.

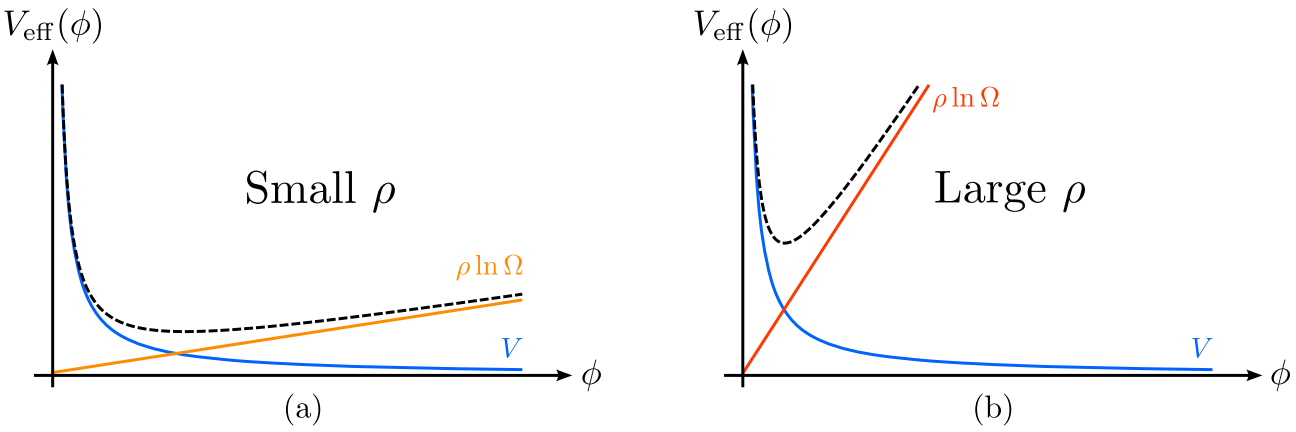


Figure 1.3: Sketch of the chameleon effective potential for strictly positive exponent n in a low density region (left panel) and in a higher density region (left panel). Adapted from Ref. [71].

Equations and phenomenology of the chameleon model

In the chameleon model given by Eq. (1.112), the field's equation (1.45) takes the following form

$$\Box\phi = -n\frac{\Lambda^{n+4}}{\phi^{n+1}} - \frac{\beta}{M_{\rm Pl}}T \xrightarrow{\text{Eq. (1.52)}} \Delta\phi \simeq \frac{\beta}{M_{\rm Pl}}\rho - n\frac{\Lambda^{n+4}}{\phi^{n+1}}\,. \tag{1.117}$$

In particular, the field's value that minimizes the effective potential together with the effective mass can be readily computed (see Box C) as

$$\phi_{\rm min}(\rho) = \left(M_{\rm Pl}\frac{n\Lambda^{n+4}}{\beta\rho}\right)^{\frac{1}{n+1}}\,, \tag{1.118a}$$

$$m_\phi^2 = n(n+1)\Lambda^{n+4}\left(\frac{\beta\rho}{nM_{\rm Pl}\Lambda^{n+4}}\right)^{\frac{n+2}{n+1}}\,. \tag{1.118b}$$

In the quasistatic, weak field regime, the chameleon field is governed by a nonlinear Klein–Gordon equation (1.117), where the nonlinearity stems from the potential function power-law form. The equation being nonetheless linear in the field's derivatives, it is classified as a semi-linear PDE (see Chap. 2, Sec. 2.2). The existence of solutions to that equation on a bounded domain with Dirichlet boundary conditions is examined in Appendix C.1. To get a grasp on the chameleon field behavior, let us first imagine that all space is being filled with a fluid of density $\rho_{\rm vac}$. In this idealized situation, a trivial solution to Eq. (1.117) is $\phi = \phi_{\rm min}(\rho_{\rm vac})$, i.e. the constant field value that minimizes the effective potential $V_{\rm eff}$ in this medium.

Things become less trivial when we study more realistic configurations for which the density varies across space. It is convenient to start with the simple example of a homogeneous ball of density $\rho_{\rm in}$ immersed in the vacuum. If the ball has an extremely large radius, one gets the intuition that deep inside that ball, the field should go to the value that minimizes the effective potential $\phi_{\rm min}(\rho_{\rm in})$ as earlier. In this situation, the field is expected to be roughly constant both very far away from the ball [where it goes to $\phi_{\rm min}(\rho_{\rm vac})$] and deep inside the ball [where it stays close to $\phi_{\rm min}(\rho_{\rm in})$]. In other words, the field is attracted towards the value that minimizes $V_{\rm eff}$ in the presence of media with different densities. This picture can be altered if we start shrinking down the ball's size. In the limit where the radius goes to zero, the ball shall leave the field unperturbed. Therefore, a sufficiently small ball would not be a big enough perturbation for the field to reach $\phi_{\rm min}(\rho_{\rm in})$. The decisive criteria for turning this qualitative explanation into a quantitative one is given by the field's effective mass Eq. (1.118b). Indeed, the inverse of the effective mass m_ϕ is the Compton wavelength of the field λ. As seen in our discussion of the Yukawa approximation in Sec. 1.1.3, the latter provides an estimate of the relaxation length scale of the field when going from one medium to the other. Consequently, we have the following criteria:

- if $\lambda(\rho_{\rm in})$ is much smaller than the ball's radius (or more generally, the size of the object at stake) we can expect the field to reach $\phi_{\rm min}(\rho_{\rm in})$ — this is the *screened* scenario;

- if on the other hand $\lambda(\rho_{\rm in})$ is greater than or equal to the size of the object, ϕ will not 'have enough space' to reach $\phi_{\rm min}(\rho_{\rm in})$ — this is the *unscreened* scenario.

To go beyond these general remarks and gain further insight into the behavior of the chameleon field, it is useful to analytically study the case of the perfect sphere with radius $R_{\rm b}$. In the spherically symmetric case,

Eq. (1.117) can be written in terms of the radial coordinate r as a mere ordinary differential equation (ODE)

$$\forall r > 0\,,\ \frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}\phi}{\mathrm{d}r}\right) = \frac{\beta}{M_{\rm Pl}}\rho(r) - n\frac{\Lambda^{n+4}}{\phi^{n+1}}\,, \qquad \text{with} \quad \rho(r) = \begin{cases}\rho_{\rm in} & \text{if } r \leq R_{\rm b}\\ 0 & \text{if } r > R_{\rm b}\end{cases} \tag{1.119}$$

Before embarking on any calculation, this nonlinear Klein–Gordon equation (1.119) needs to be supplemented with relevant boundary conditions. The spherical symmetry immediately imposes $\phi'(r=0)=0$. Far away from the ball, the discussion above indicates that the field should relax towards the value that minimizes the effective potential in the exterior medium (with infinite spatial extension). We therefore impose the *asymptotic boundary condition*

$$\phi(r) \underset{r\to+\infty}{\longrightarrow} \phi_{\rm vac} \equiv \phi_{\rm min}(\rho_{\rm vac})\,. \tag{1.120}$$

Intuitively, these two conditions are enough for the ODE (1.119) to have a unique solution.[26] Many articles dealing with this ODE further impose a vanishing gradient condition at infinity, i.e. $\phi'(r)\to 0$ as $r\to+\infty$, see e.g. Refs. [133–136]. As a matter of fact, we proved in Ref. [137] that this additional asymptotic condition is redundant with Eq. (1.120), given the structure of the ODE at stake.[27] This proof is reproduced in Appendix D.

Analytical approximations to this problem are derived in many studies (see e.g. Refs. [71, 72, 111, 115, 138–140]) and reported here. Let us denote by $R_{\rm TS} \in [0, R_{\rm b}[$ the radius below which the field stays frozen at $\phi_{\rm min}(\rho_{\rm in})$ in the screened scenario depicted above (which occurs for a sufficiently large body). There are three distinct regions: *(i)* $r < R_{\rm TS}$ where the field is frozen, *(ii)* $R_{\rm TS} \leq r \leq R_{\rm b}$ which is often called the *thin shell*, and *(iii)* $r > R_{\rm b}$ which is the exterior of the ball. An analytical approximation might be derived by approximating the rhs of Eq. (1.119) in each of the three regions, and then smoothly reconnecting the piece-wise solutions.

- In the first region, the field is constant $\phi = \phi_{\rm min}(\rho_{\rm in})$. In the rhs of Eq. (1.119), the density term and the $\phi^{-(n+1)}$ term balance each other so that $\Delta\phi \simeq 0$.

- In the second region, $\phi_{\rm in} < \phi < \phi_{\rm vac}$ due to the hierarchy between the minima of the two effective potentials (in the object and in the surrounding medium). Therefore, the $\phi^{-(n+1)}$ term decreases while the density term remains constant, and starts to dominate the rhs of Eq. (1.119).

- Going from the center of the ball to the exterior region, the density undergoes a jump $\rho_{\rm in} \to \rho_{\rm vac}$. If $\rho_{\rm vac} \ll \rho_{\rm in}$, $\Delta\phi \sim 0$ for $r < \lambda(\rho_{\rm vac})$. As r becomes greater than the field's Compton wavelength in the exterior medium, ϕ gets closer to its asymptotic value $\phi_{\rm vac}$ so that the effective potential is well-approximated by Eq. (1.105). Therefore, the Yukawa approximation applies and the Klein–Gordon equation boils down to $\Delta\phi \simeq m_\phi^2(\rho_{\rm vac})(\phi - \phi_{\rm vac})$.

Given the above, we get

$$\text{[Screened case]} \quad \phi(r) \simeq \begin{cases}\phi_{\rm in} & \text{for } r < R_{\rm TS}\,,\\ \phi_{\rm in} + \dfrac{\beta\rho_{\rm in}}{6M_{\rm Pl}}\left(r^2 + 2\dfrac{R_{\rm TS}^3}{r} - 3R_{\rm TS}^2\right) & \text{for } R_{\rm TS} \leq r \leq R_{\rm b}\,,\\ \phi_{\rm vac} - \left(\dfrac{3\beta}{4\pi M_{\rm Pl}}\right)\left(\dfrac{\Delta R}{R}\right)\dfrac{GM_{\rm b}}{r}\mathrm{e}^{-m_\phi(\rho_{\rm vac})(r-R_{\rm b})} & \text{for } r > R_{\rm b}\,,\end{cases} \tag{1.121}$$

where $M_{\rm b}$ denotes the mass of the ball and $(\Delta R/R)$ is an estimate of the relative thickness of the thin shell, called the *thin shell parameter*, reading

$$\frac{\Delta R}{R} = \frac{\phi_{\rm vac} - \phi_{\rm in}}{6\beta M_{\rm Pl}|\Phi_N|} \simeq \frac{R_{\rm b} - R_{\rm TS}}{R_{\rm b}}\,, \qquad \text{with} \quad |\Phi_N| = \frac{GM_{\rm b}}{R_{\rm b}}\,. \tag{1.122}$$

Note that we have assumed $(\Delta R/R) \ll 1$ in the above expression. We observe that in the exterior domain, the chameleon field is well-described by a Yukawa potential except its coupling constant β has been multiplied by the thin shell parameter. As a direct consequence, the fifth force experienced by a test particle (not perturbing the chameleon field) is similar to a Yukawa interaction [see e.g. Eq. (1.108)], but depleted by the factor $(\Delta R/R) \ll 1$ given by Eq. (1.122). Everything happens as if the scalar fifth force was sourced only by the mass contained in the thin shell. A sketch of the radial chameleon field profile in the screened case is illustrated in the left panel of Fig. 1.4.

[26] Note that for ODEs, the question of existence and uniqueness of the solution is usually addressed using the Cauchy–Lipschitz theorem. However, Eq. (1.119) together with the boundary conditions $\phi'(0)=0$ and $\phi(r)\to\phi_{\rm vac}$ as $r\to+\infty$ does not constitute a Cauchy problem and so the theorem is not applicable in the present form.

[27] This statement was already claimed but unproved in Ref. [115]. Yet it turns out that the proof is not immediate.

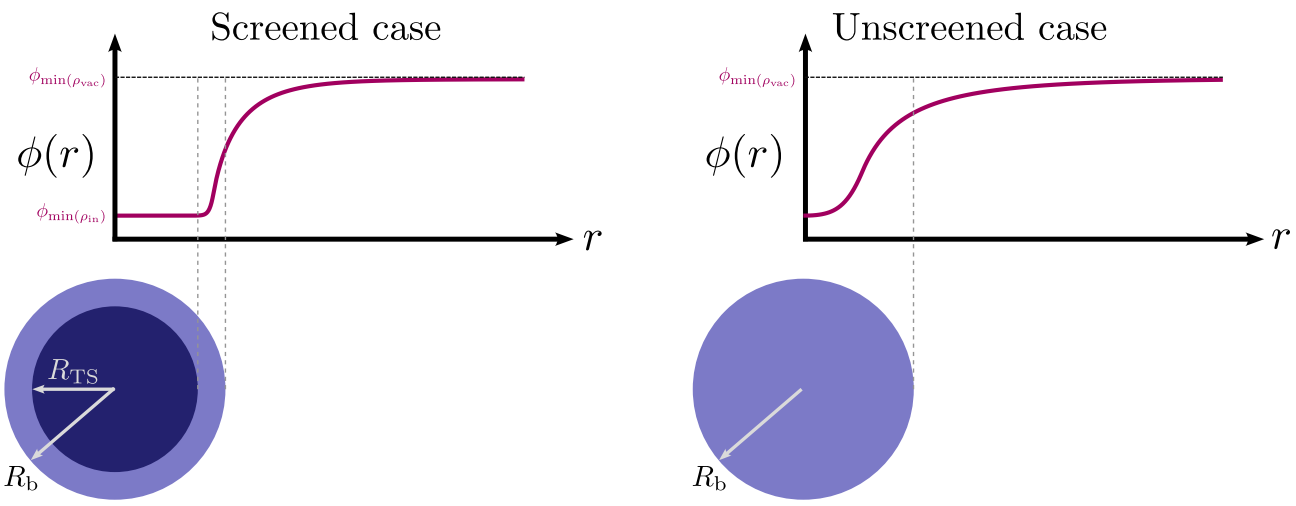


Figure 1.4: Radial chameleon field profiles in the presence of an homogeneous solid sphere of radius $R_{\rm b}$ and density $\rho_{\rm in}$ immersed in a lower density medium $\rho_{\rm vac}$. The left panel corresponds to the screened case: inside the ball (dark blue), the field is frozen to the value that minimizes its effective potential $\phi_{\rm min}(\rho_{\rm in})$ everywhere but in a *thin shell* beneath the surface (light blue) where it starts to depart from this equilibrium. Outside the sphere, the chameleon's behavior resembles that of a massive scalar field and is well-approximated with a Yukawa exponential factor, with $\phi_{\rm min}(\rho_{\rm vac})$ as its asymptotic value. The right panel corresponds to the unscreened case: here the chameleon does not have enough space to reach $\phi_{\rm min}(\rho_{\rm in})$ inside the ball, where it behaves more or less like the Newtonian potential. Outside the ball, it is well-approximated by a Yukawa potential.

On the other hand, in the unscreened regime, the first region where the field used to be frozen does not exist anymore, so that $R_{\rm TS} = 0$. Using the same arguments as for the screened case, we get

$$\text{[Unscreened case]} \quad \phi(r) \simeq \begin{cases} \phi_{\rm vac} - \dfrac{\beta\rho_{\rm in}}{6M_{\rm Pl}}(R_{\rm b}^2 - r^2) & \text{for } r < R_{\rm b}\,, \\ \phi_{\rm vac} - \dfrac{\beta}{4\pi M_{\rm Pl}}\dfrac{GM_{\rm b}}{r}\mathrm{e}^{-m_\phi(\rho_{\rm vac})(r-R_{\rm b})} & \text{for } r \geq R_{\rm b}\,. \end{cases} \tag{1.123}$$

This situation occurs when the mass of the field inside the ball is smaller than its radius. The absence of screening means that the ratio between the fifth force and the usual Newtonian force is roughly equal to $2\beta^2$, see Eq. (1.108). Again, a sketch of the radial chameleon profile in the unscreened case in illustrated in the right panel of Fig. 1.4.

Of course, there exists an intermediate case where the field reaches $\phi_{\rm min}(\rho_{\rm in})$ at the center of the ball but $(\Delta R/R)$ is not very small compared to one. This situation is sometimes referred to as the *partially screened* regime, for which the analytical approximation (1.121) has to be slightly adapted (see e.g. Refs. [137, 138]), but the behavior is essentially the same. In practice, it can be verified numerically — for lack of other alternatives — that the transition from the screened regime to the unscreened regime occurs over a very narrow region of the chameleon parameter space, see Ref. [141] Sec. IV A 2.

Finally, note that in Ref. [137] we used an analytical approximation which slightly differs from Eqs. (1.121–1.123). The idea of simplifying the equation in different non-overlapping regions still applies, but extra care is given when imposing the smooth connection of the field across neighboring regions — in particular $R_{\rm TS}$ is obtained by computing the real roots of a third-degree polynomial. Other analytical approximations exist for other simple geometries including the cylinder [142], ellipsoids [143] or parallel plates [144].

Critical discussion of the asymptotic condition Eq. (1.120)

In the above, we supplemented the Klein–Gordon equation governing the chameleon field with the asymptotic boundary condition [Eq. (1.120)] so that the problem has a unique solution (although we did not prove it mathematically). However, while this asymptotic condition is perfectly valid for the idealized case of a body immersed in a background medium with infinite spatial extension, our universe looks quite different from this picture.

In reality, between the body of interest and spatial infinity, there lies stars, galaxies, dust, etc. In that sense, the scalar field never really relaxes toward a constant value, however far away from the region of interest we go: there is no isolated system, which is at odds with what one usually does in physics.

Yet, let us assume that we can go sufficiently far away from the studied body, where the density becomes

homogeneous, without encountering any of the parasitic sources of density mentioned above. Demanding that the scalar field reaches the value that minimizes its effective potential in this far out region of the universe is also questionable. Indeed, as was the case when discussing the massless Brans–Dicke theory in Sec. 1.1.3, the background value of the scalar field is determined by an appropriate cosmological boundary condition. In this regard, it is shown in Ref. [145] Sec. III D that throughout its cosmological evolution in a FLRW universe, the scalar field does not necessarily remain at the minimum of its effective potential. This means we could have

$$\phi(r) \underset{r\to+\infty}{\longrightarrow} \phi_\infty \neq \phi_{\min}(\rho_{\rm vac})\,, \qquad \text{with} \quad \phi_\infty < \phi_{\min}(\rho_{\rm vac}) \tag{1.124}$$

As advocated in Ref. [145] and to be more consistent, one should solve Eq. (1.84) which governs the cosmological evolution of the field to get a good estimate of ϕ_∞ when applying the asymptotic boundary condition for Eq. (1.117). This careful treatment is beyond the scope of the present PhD work, but had to be mentioned nonetheless.

Physical relevance of the chameleon model

The bare potential for the chameleon model specified by Eq. (1.112) is nothing but the Ratra–Peebles potential [88], which was already introduced in a cosmological context in Eq. (1.86). As a matter of fact, the chameleon model was introduced as a candidate for explaining cosmic acceleration [115, 116]. However, it has been shown that 'chameleon-like' models could not at the same time *(i) self-accelerate* the cosmic expansion (this notion is given a meaning in Sec. 1.1.3) and *(ii)* ensure that the Sun (and the Milky Way) is screened, which seems like a reasonable assumption for otherwise Solar system tests would not be satisfied — see Refs. [94, 145]. However, it can still act as a dark energy quintessence field [87, 124] if we accept to add a cosmological constant on top of the Ratra–Peebles potential Eq. (1.86) as

$$V(\phi) = \Lambda_{\rm DE}^4 + \frac{\Lambda^{n+4}}{\phi^n}\,, \tag{1.125}$$

where $\Lambda_{\rm DE} \simeq 2.4\,\text{meV}$ is the so-called dark energy scale.[28] Note that in spite of this modification of the bare potential, our discussion of the chameleon model above remains valid as the potential only comes into play through its derivatives. The only quantity introduced so far that depends on $V(\phi)$ itself is the equation of state of the scalar field in the Einstein frame [Eq. (1.83)].

Let us stress once again that the chameleon model discussed above is just one example of realization of the chameleon screening mechanism. The latter may be leveraged to hide scalars in other physical contexts. For example, Refs. [146, 147] consider a 'chameleonic' dark matter candidate, while Ref. [148] propose to stabilize the Higgs potential in the early universe by regarding the Higgs field as a chameleon field coupled to the inflaton alone.

Violations of the weak equivalence principle

In practice there are two ways in which the WEP can be violated in the framework of the chameleon model.

Explicit violation The first is to have different coupling constant β_i for different particle species: this is the case of non-universal coupling which we presented in Sec. 1.1.2 through the matter action [Eq. (1.80)]. This composition dependent coupling is expected from a theoretical view point as quantum corrections would generally produce slightly different coupling constants for different matter species. A direct consequence of this non-universal coupling is that different particle species follow different geodesics in spacetime (and the variation of fundamental constants [73]). From the Einstein frame perspective, they experience different fifth forces. This first option can thus be regarded as a violation of the WEP at the microscopic level, i.e. at the level of the Lagrangian of the theory.

Macroscopic violation The second possibility is more subtle as it applies even in the case of a universal β, but only for extended bodies (as opposed to test particles). It relies on the fact that macroscopic objects with different densities do not necessarily possess the same thin shell radius $R_{\rm TS}$ in the screening regime. As such, their thin shell parameter $(\Delta R/R)$ given by Eq. (1.122) would be different, and so would be the fifth force they experience. The WEP thereby strictly holds for idealized test particles, yet extended bodies do not fall in the exact same way. This is the reason why this phenomenon is sometimes referred to as an *apparent* WEP violation, see e.g. Refs. [138, 149]. Note that this violation is different from a violation of the SEP since macroscopic bodies do not necessarily have a dominating gravitational binding energy.

[28] Note that $\Lambda_{\rm DE}$ is not to be confused with the cosmological constant Λ reported in Table 1.1 (they do not even share the same physical dimension). The given value of the dark energy scale is computed as $(3\Omega_{0,\Lambda} H_0^2 M_{\rm Pl}^2)^{1/4}$ [see Eq. (1.26)].

Current constraints

In the above, we have seen through the example of the chameleon model that, despite screening in regions of high Newtonian potential, the chameleon field is never completely *invisible*. The fifth force it mediates, no matter how tiny it may be, is nonzero. This means we can look for it, at scales ranging from the laboratory on Earth up to astrophysical scales.

The chameleon model, as introduced here with the functions Ω and V defined by Eq. (1.112), has three parameters: n, β, and Λ. For any triplet of such parameters, one can compute the fifth force amplitude in a given setup and compare the model's prediction against measurements. If the two outcomes — the prediction and the observation — fail to be consistent with each other, this specific triplet is ruled out.[29] The chameleon model can thereby be tested via astrophysical observations (rotation-curve observations, lensing by galaxy clusters, redshift-space distortions, etc.) or laboratory experiments (precision atomic tests, atom interferometry, torsion balance experiments, Casimir-force tests, precision neutron tests, levitated force sensor, etc.), see reviews by Burrage and Sakstein [71, 150] and Ref. [151]. Currently, laboratory experiments provide the most stringent constraints, astrophysical observations not being competitive enough. At the time of writing and to the best of our knowledge, the current status of the chameleon parameter space can be consulted in Ref. [152], Fig. 4.

It is to be noted that the PPN formalism, discussed in Sec. 1.1.1, is not efficient for constraining the chameleon, and more generally screened scalar-tensor theories. The procedure laid out in Ref. [1] Chapt. 5.3.2, which shows how to derive the PPN parameters γ and β (see Table 1.2), relies on the assumption that the scalar field can be expanded around some background as $\phi = \phi_0(1 + \delta\phi)$, with $|\delta\phi| \ll 1$. This does not hold true in general for theories with screening, as the nonlinearity allows $\delta\phi \sim O(1)$ in some situations — see e.g. the examples provided throughout Chapt. 5 or Refs. [149, 153]. A derivation can nonetheless be found based on the approximations (1.121, 1.123). In particular, bounds on the PPN parameter are satisfied for a wide range of coupling constants β thanks to the thin shell parameter $(\Delta R/R)$ that gives a small "effective" coupling constant — see e.g. Ref. [71, 145].

1.3 Space-based tests: the legacy of the MICROSCOPE space mission

In the above, we have mentioned the possibility of testing screened scalar-tensor models of gravity using Earth-based experimental setups or astrophysical observations. Somewhere in between these two scales lies the Solar system scale. While it is true that screening mechanisms were introduced precisely to recover GR in the Solar system (and thereby satisfying PPN bounds, see Table 1.2), in no way does that preclude the possibility of using spacecraft for probing gravity and getting state-of-the-art bounds.

From a historical perspective, the advent of the space era in the second half of the XX$^{\text{th}}$ century broadened the landscape of gravity tests within our reach. The rapid development of interplanetary space program and Earth observation satellites went hand in hand with the equally rapid development of the underlying technology (such as radar ranging and accurate clocks for instance). Launched in 1976, Gravity Probe A — also known as the Vessot–Levine experiment [25] — was the first space mission dedicated to testing gravity (through the measurement of the gravitational redshift), whose successful outcome paved the way for other space-based experiments: Gravity Probe B [154], LAGEOS 1 and 2, LARES 1 and 2 [155], MICROSCOPE [156]...

In this section, we discuss the benefits and disadvantages of performing gravity tests in outer space, with a focus on the MICROSCOPE space mission. In particular, we review its main scientific results — which are mainly related to the WEP, although not exclusively — and see to what extent the mission's data can be used to constrain screened modified gravity.

1.3.1 Testing gravity with spacecraft

Space-based experiments are particularly well-suited for testing gravity due to their ability to operate in low-gravity environments and thanks to the precision they can achieve in measuring gravitational effects. Indeed, space offers a very stable environment compared to that of the Earth in terms of seismic noise and thermal stability (principally in Sun-synchronous orbit), which are known to be limiting factors when it comes to increasing the sensitivity of Earth-based experiments.[30] Additionally, although ground-based experiments, such as those using torsion balances [19, 20, 98, 157] or atom interferometers [158], have achieved remarkable precision, they are confined to testing gravity at small scales ($\sim \mu$m to $\sim$ m scales) and in the relatively strong gravitational field of the Earth (compared to outer space). Space-based experiments, on the other hand, can measure gravitational effects over hundreds to millions of kilometers and can operate in low-gravity environments for extended periods of time.

[29] This is of course a very schematic view. In practice this must be done on a statistical basis, using the uncertainty in the measurement.

[30] In this regard, it is insightful to look at how seismic isolation is performed in the gravitational wave observatory LIGO.

Beyond these fairly general considerations, space-based experiments must also be discussed from the perspective of testing screened scalar-tensor models. In both the chameleon and symmetron models discussed above, the screening mechanism is at work in high density environments, because the scalar field acquires a large mass or decouples from matter namely. In Earth orbit, the density of the residual atmosphere is already many orders of magnitude smaller than at the Earth surface, neighboring $\sim 10^{-15}\,\mathrm{kg/m^3}$ at $10^3\,\mathrm{km}$ altitude. In the case of the chameleon model specifically, whether an object develops a thin shell not only depends on the objects properties (size and density), but also on the background density. An object that exhibits a thin shell in the laboratory here on Earth may be depleted from it in outer space, just because there the background density is very low [see e.g. Eq. (1.122) which shows that the thin shell parameter grows with $\phi_{\rm vac}$, which itself grows as the background density is decreased, see Eq. (1.118a)]. This leads to at least two important conclusions: *(i)* a Cavendish-like experiment could measure a value for the gravitational constant G_* on Earth but $G_*(1+2\beta^2)$ in space, and *(ii)* WEP violations could arise at levels higher than what is allowed on Earth (in the case of non-universal coupling) [115, 116, 149, 153]. Of course, there are some strong hypotheses underlying these bold statements:

1. They assume that the test bodies used to perform the experiment are screened down on Earth and unscreened in outer space.

2. They assume that the satellite that carries the test masses is also unscreened in space.

These assumptions will be discussed in Sec. 1.3.4 and re-examined later in Chapt. 5.

To sum up, space-based experiments partially bridge the length-scale gap between laboratory tests and astrophysical measurements. Claims such as the ones reported above in Refs. [115, 116, 149, 153] generated interest, and perhaps sparked hope, among the community for fundamental physics space mission that were planned in the early 2000's. Among them, let us mention the SEE project [159, 160] which aimed, among other things, at measuring the gravitational constant; STEP [161–163], Galileo Galilei (GG) [164] and MICROSCOPE [156] whose goals were to test the WEP at a precision of 10^{-18}, 10^{-17} and 10^{-15} on η respectively [the Eötvös parameter, see Sec. 1.1.1, Eq. (1.28)]. At the present date, MICROSCOPE is the only one that flew. In the following, we take a step back from the now-over MICROSCOPE space mission and present the lessons drawn from it, especially regarding the exciting claims that were made prior to launch for testing screened modified gravity.

As final remarks, it must be reminded that space inevitably leads to multiple limitations to which ground-based experiments are usually not subject. The satellite payload is constrained not only in terms of size and weight, but also in terms of power consumption. A space mission is hardly modular and any failure can be fatal to its smooth operation. Last but not least, performing an experiment in space is much more expensive than performing that same experiment in the laboratory.

1.3.2 A brief description of the MICROSCOPE experiment and its result on the weak equivalence principle

The MICROSCOPE experiment, standing for "MICRO-Satellite à Compensation de traînée pour l'Observation du Principe d'Équivalence" is a French space mission led by a CNES-ESA-ONERA-CNRS-OCA-DLR-ZARM collaboration. The satellite was launched in April 2016 and collected data for more than two years, before being decommissioned in October 2018. As mentioned above, the mission's goal was to test one of GR's cornerstone — the weak equivalence principle — with a targeted precision $\eta \lesssim 10^{-15}$. The full description of the mission together with its scientific return are provided in a special issue of *Classical and Quantum Gravity*, see Ref. [156]. In parallel, the raw mission data has been made available at https://cmsm-ds.onera.fr/user/microscope.

Mission design

The simplest test of the universality of free fall one can imagine is to drop objects — with different masses and different compositions — from an altitude and check whether they hit the ground at the same instant in time. The popular belief is that this simple thought experiment was performed by Galileo, back in the XVI$^{\rm th}$ century, by dropping unequal weights of the same material from the leaning tower of Pisa. This popularized picture, however, is by no means a good way to accurately test the universality of free fall, mainly due to air drag and short integration time. Instead, verified sources indicate that the Italian scientist proceeded by using inclined planes [165], which is arguably a cleverer experimental concept. The history of WEP tests, starting from the XX$^{\rm th}$ century, can be found in Fig. 1.1 (left panel).

Nonetheless, the primary thought experiment is perhaps the right starting place to think of the MICROSCOPE experiment. It extends on that idea by comparing the free fall of test masses, not dropped from a tower but in orbit. The benefits of space are manifest: *(i)* atmospheric drag, despite not being quite zero, is reduced by many orders of magnitude compared to air drag down on Earth, and *(ii)* the duration of free fall is virtually infinite since the satellite follows a (perturbed) Keplerian orbit around the Earth. Monitoring the free fall of test masses for an extended amount of time constitute a simple yet powerful way of testing the WEP. In practice, the test

masses used to perform this experiment are not exactly in free fall but are instead constrained to remain in equilibrium with respect to each other and to the satellite platform by means of electrostatic forces. This is achieved using differential ultrasensitive electrostatic accelerometers consisting of two coaxial and concentric hollow cylinders — the test masses — made of different materials. The electrostatic forces undergone by the cylinders are an image of the electrostatic potentials applied on the various controlling electrodes. Therefore, if the WEP holds, applied potentials should be such that concentric cylinders are subject to the same electrostatic acceleration. If not, that would be a smoking gun for a violation of the universality of free fall. The accelerometers are designed such that the measurement is most sensitive along the cylinders' axis, hereby labeled the X-axis. For reasons including the minimization of instrumental noise, the satellite is put into a spinning mode,[31] with angular velocity $2\pi f_{\rm spin}$. By doing so, the frequency of the potential WEP violation $f_{\rm EP}$ is equal to

$$f_{\rm EP} = f_{\rm orb} + f_{\rm spin}\,, \tag{1.126}$$

where $f_{\rm orb}$ is the satellite's orbital frequency.

The satellite was put on a Sun-synchronous orbit at an altitude of 710 km. It was oriented so as to always show the same face to the Sun, ensuring a good thermal stability throughout the mission. The chosen altitude results from a trade-off. On the one hand, it is desirable to reduce atmospheric drag as much as possible by going to higher altitudes. On the other hand, the expected WEP violation signal being proportional to the gravity acceleration sourced by the Earth, aiming for too high altitudes is not desirable either. For the test masses to be as close to free fall as possible — i.e. only subject to the gravitational force (Newtonian physics interpretation), or equivalently following a timelike geodesic (GR's interpretation) — non-gravitational perturbations have to be counteracted. Among those, atmospheric drag is unsurprisingly the dominant perturbation, followed by Solar radiation pressure and electromagnetic forces. The spacecraft was thus equipped with a Drag-Free and Attitude Control System (DFACS) using cold gas thrusters to counteract these non-gravitational perturbations at the level of tens of μN.

The payload

Outlined above, MICROSCOPE's payload (T-SAGE) inherits from ONERA's expertise in the field of ultrasensitive electrostatic accelerometry. T-SAGE consists of two differential accelerometers or sensor units (SU). Each SU is composed of two concentric cylindrical test masses, each of which is partly surrounded by two electrode-bearing gold coated silica cylinders in charge of both sensing the proof mass position and acting on it via electrostatic forces in case it gets displaced from its equilibrium position. In this way all four test masses are controlled along the six degrees of freedom, where as mentioned above translation along the X-axis is the most sensitive of all and is the one used to perform the WEP test.

The two sensor units serve different purposes:

- One is called the reference sensor unit — SUREF. The two test masses composing SUREF are made of the same material (a platinum alloy) and should, in principle, be insensitive to composition-dependent WEP violations.
- The other one, called SUEP, is used to measure the WEP. Its two test masses are made of different materials (a platinum alloy and a titanium alloy) that were chosen partly so as to maximize the WEP violation from a light dilaton [74, 76].

A new upper bound on the WEP

While the mere principle of the experiment may seem straightforward enough, the actual mission design and data analysis phase that took place subsequently are far from being simple. Speaking of the latter phase, translating raw voltages applied to the various electrodes into an upper bound on the Eötvös parameter η requires to account for diverse physical effects entering the stage inside MICROSCOPE's payload, and are called *errors*. Errors are the answer to the question: *"what are the sources of non-null differential acceleration, aside from a true violation of the WEP?"*. Caused by the instrument's imperfections, they can be categorized into stochastic errors and systematic errors.

On the one hand, systematic errors can be estimated and minimized. They include the Earth gravity gradient together with the off-centering of the test masses, temperature variations (dominating systematics), electrical bias, etc. See Ref. [167] for a comprehensive review of all systematic errors. Most of these systematic effects can be estimated thanks to dedicated calibration sessions in orbit.

On the other hand, statistical error can be attributed to all sources of noise. The main ones are the electronic noise, the thermal noise and the noise coming from the gold wire that is employed to control the electrical potential of the test mass it is attached to. The spinning frequency of the satellite was chosen so that the total

[31] See Ref. [166] for a complete discussion of how spinning frequencies were chosen.

noise is minimum at $f_{\rm EP}$ [see Eq. (1.126)]. The long duration of measurement sessions, allowed by the mere fact of being in orbit around the Earth, is such that the statistical error can be brought down to the targeted $\sim 10^{-15}$ level on the Eötvös parameter.

Final results from the MICROSCOPE mission were published in 2022 [23, 24]. In particular, the result for SUEP, the differential accelerometer with two test masses made of different materials, is

$$\eta(\mathrm{Ti},\,\mathrm{Pt}) = \left[-1.5 \pm 2.3\,(\mathrm{stat}) \pm 1.5\,(\mathrm{syst})\right] \times 10^{-15}\,, \tag{1.127}$$

where the statistical error is given at 1σ. This result is consistent with $\eta = 0$ given the errors. Besides, the reference sensor unit, SUREF, provided a null result $\eta(\mathrm{Pt},\,\mathrm{Pt}) = [0.0 \pm 1.1\,(\mathrm{stat}) \pm 2.3\,(\mathrm{syst})] \times 10^{-15}$, showing no sign of unaccounted systematic errors in Eq. (1.127).

1.3.3 Implications beyond the weak equivalence principle

To a certain extent, MICROSCOPE is a *one-number* mission: a new upper bound on the Eötvös parameter, roughly 10^{-15} as hoped. However, despite being a substantial technical achievement, there is a lot more to be said about the scientific return of this successful space mission, and treating it as 'just' a new bound on η is reductive. First, such a tight constraint on WEP violation *automatically* results in constraints on alternative theories predicting one. Moreover, the raw data might also be analyzed in the framework of models which do not directly translate into a non-zero η. Here, we report on all these scientific results beyond the weak equivalence principle.

Constraints on Yukawa gravity

In Sec. 1.1.3, we introduced the Yukawa potential [Eq. (1.109)]. Although it was first presented in the specific framework of scalar-tensor theories with a quadratic potential and an exponential conformal function, we saw while discussing the chameleon model in Sec. 1.2.2 that the Yukawa potential could constitute a good approximation for estimating the fifth force mediated by massive scalars coupled to matter. In that sense, the Yukawa potential has a generic character and, as such, is a sensible way of parameterizing deviations from the Newtonian inverse-square law. As a matter of fact, it is arguably the most common parameterization found in the literature. We remind that the only two parameters for this model are a coupling constant α, representing the strength of the deviation from Newtonian gravity, and a Compton wavelength λ representing the range of the fifth force.

Long range Yukawa fifth force MICROSCOPE becomes relevant for putting constraints as soon as one considers a composition-dependent coupling α_{ij}, where each material is associated with a scalar dimensionless 'Yukawa charge' q so that

$$\alpha_{ij} = \alpha \left(\frac{q}{\mu}\right)_i \left(\frac{q}{\mu}\right)_j . \tag{1.128}$$

In this expression, α (without subscripts) plays the role of a universal dimensionless coupling constant while μ_i is the atomic mass in atomic units of the constituent element. Note that there is no single way of defining the Yukawa charge q (since this depends on the explicit coupling of ϕ to the standard matter fields). Some works take the material's baryon and/or lepton numbers to play that role, see e.g. Refs. [168, 169]. Regardless of the adopted definition for q, this composition dependent coupling can be expressed in terms of the Eötvös parameter as

$$\eta = \alpha \left[\left(\frac{q}{\mu}\right)_{\rm Pt} - \left(\frac{q}{\mu}\right)_{\rm Ti}\right] \left(\frac{q}{\mu}\right)_{\oplus} \left(1 + \frac{r}{\lambda}\right) \mathrm{e}^{-\frac{r}{\lambda}} \tag{1.129}$$

in the case of MICROSCOPE test masses. Equating the Yukawa charge q with the baryon number of the various elements, one can translate the constraint on η into constraints on $(\alpha,\,\lambda)$. In particular, the satellite's altitude is such that the experiment is sensitive to a Yukawa fifth force in the range $\lambda \in [\sim 10^5\,\mathrm{m},\,+\infty[$. The constraints on such a light Yukawa field can be found in Ref. [170].

Short range Yukawa fifth force Although MICROSCOPE was not designed for testing putative short range interactions, it is nonetheless possible to leverage the data acquired through characterization sessions in this direction. Characterization sessions are dedicated to the estimation of various parameters related to the instrument. During some of those sessions, the tests masses are subjected to a sinusoidal excitation in position by means of electrostatic forces applied by the electrodes surrounding them. At first order, the electrostatic force acts as a stiffness. In Ref. [171], the authors analyzed all possible classical sources of discrepancy between the measured and expected electrostatic stiffness, before considering the observed discrepancy as the result of a short range Yukawa fifth force. The latter also acts as a stiffness when the cylindrical test mass is displaced from its centered position, owing to the interaction with the various parts of the apparatus. The discrepancy-budget

can then be translated into constraints in the (α, λ)-plane. However, the constraints derived in Ref. [171] are not competitive with the state-of-art — roughly eight orders of magnitude looser than the current best upper bounds. This was expected as MICROSCOPE was not designed for probing gravity at centimetric length scales. Nonetheless, it does not preclude the possibility that a highly-optimized MICROSCOPE-like experiment could provide competitive constraints on short range interaction between several bodies.

Constraints on dilaton models

Massless dilaton Massless scalar-tensor models have been illustrated in Sec. 1.1.3 with the prototypical example of the massless Brans–Dicke model [see Eqs. (1.100–1.102)]. A light dilaton scalar field interacts with the standard model fields — it couples to the electromagnetic and gluonic tensors, and to fermion spinors — with different coupling strengths, see Refs. [172, 173]. A result of these interactions is that the dilaton effectively couples differently to each atom, and we are brought back to the discussion of non-universal coupling in scalar-tensor theories we had in Sec. 1.1.2. In particular, a direct consequence of non-universal coupling is WEP violations. Constraints on massless dilaton-like scalar field are derived from preliminary MICROSCOPE results in Ref. [170].

It should also be stressed that the framework of light dilaton-like scalar fields laid out in Refs. [172, 173] is prototypical of several *ultra-light dark matter* (ULDM) scenarios (see e.g. Ref. [174]). As such, MICROSCOPE's bound on WEP violations also provide constraints on ULDM models.

Massive dilaton The idea is more or less the same as above except the dilaton scalar is massive and so the force it mediates gets Yukawa suppressed for distances larger than $\sim 1/m_\phi$. Again, see Ref. [170] for the detailed results.

Miscellaneous

Aside from the aforementioned results, MICROSCOPE's bound on η has also been used to put constraints on the runaway dilaton model [175], local Lorentz invariance [176], and spin-1 U boson [177, 178]. In the following, we report on the attempt made to constrain the chameleon model introduced in Sec. 1.2.2.

1.3.4 Attempts to look for a chameleon fifth force

As part of his PhD thesis, M. Pernot-Borràs studied the possibility of testing chameleon gravity with MICROSCOPE [138]. There are three ways in which the chameleon model could in principle be tested, and they need to be clearly distinguished:

1. *WEP violation due to non-universal coupling* — This amounts to having different coupling constants β_i for different matter species, and thus a WEP violation at the level of the Lagrangian. The discussion of this scenario is more or less the same as the case of the massive dilaton (see above and Ref. [170]) except screening comes into play. This possibility has not been investigated in this thesis.

2. *Apparent WEP violation* — We saw earlier that, even in the universal coupling case, chameleon gravity can feature apparent WEP violations, see Sec. 1.2.2 or Refs. [138, 149]. This phenomenon happens when macroscopic objects develop different thin shells and therefore do not experience the same chameleon acceleration. While MICROSCOPE's test masses could indeed play the role of such extended objects, we will see below that this effect is unlikely to be measured by MICROSCOPE

3. *Chameleon stiffness* — The principle is the same as testing a short range Yukawa fifth force (see above). For small enough displacements of the test mass with respect to its centered position, the fifth force mediated by the chameleon scalar field behaves as a stiffness. Therefore, constraints can be set on the model parameters (n, β, Λ) by comparing the discrepancy between the measured MICROSCOPE's stiffness and the expected stiffness when accounting for all *classical* known effects at stake.

In the following, we discuss in more detail the second and third ways respectively.

Apparent WEP violation

A necessary condition for apparent WEP violations (in the sense given above) to be observable is that the test masses should be at least partially screened. Otherwise, in the unscreened regime, the chameleon field couples to all the object's mass and the latter therefore follows the same trajectory as a test particle located at its center of mass would: all unscreened test masses follow the same geodesics and no WEP violation is expected. When at least one of the two masses starts being partially screened, the chameleon accelerations they undergo have no particular reason to be equal, resulting in an apparent WEP violation — even if the two masses share the same composition.

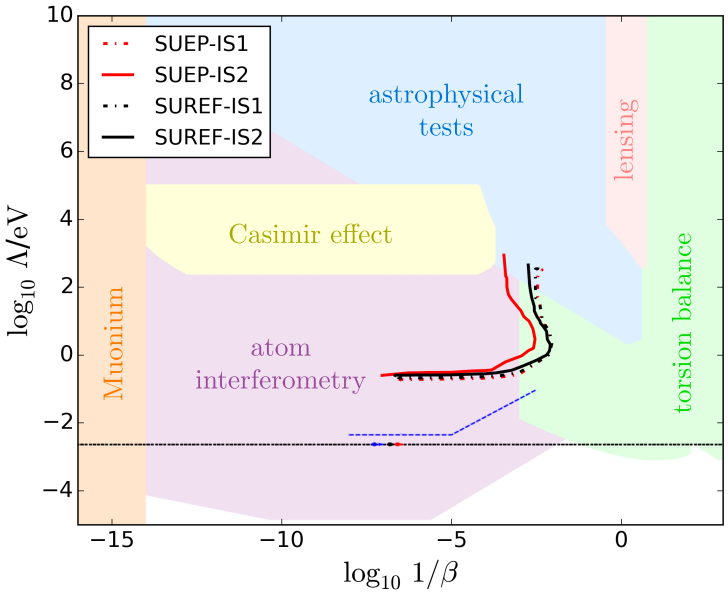


Figure 1.5: Constraints on the chameleon model for $n = 1$ from the MICROSCOPE experiment using stiffness measurement sessions. The test masses are labeled IS1 (internal test mass) and IS2 (external test mass), while SUEP and SUREF denote the two sensor units presented in Sec. 1.3.2. This figure is reproduced from Ref. [179]. Note that the other constraints (colored areas) are those from 2021 [71, 150] — see Fig. 4 from Ref. [152] for updated constraints.

Let us assume that the two concentric test masses check this first necessary condition. Of course, they are not floating around in space but are embedded within a satellite, whose density and size cannot but affect the chameleon field profile around the test masses. In order to get a WEP violation signal sourced by the Earth at f_{EP}, the satellite itself must not be screened, for otherwise the field inside the satellite becomes completely decoupled from the outside. Criteria for determining whether or not the satellite is screened are proposed in Refs. [116, 136]. Ref. [116] provides a criterion based on an estimate of the spacecraft Newtonian potential [in accordance with Eq. (1.122)], whereas Ref. [136] proposes a criterion based on the Compton wavelength of the chameleon field inside the satellite's walls, backed up with numerical computations. Specifically, the latter work concludes that MICROSCOPE is screened in most of the relevant part of the model's parameter space i.e. in unconstrained regions, see Fig. 21 from Ref. [136]. This precludes the possibility of setting competitive constraints with MICROSCOPE's WEP test. Aside, let us remark that it would be rather difficult to be in a situation where the satellite is unscreened while the test masses are not. Here, 'difficult' is to be understood in the sense that this situation would only occur in a narrow region of the parameter space, if any. Incidentally, this crucial 'satellite's screening' question is re-examined in more depth in Chapt. 5.

This possibility of an apparent WEP violation by a chameleon scalar field is not considered in Pernot-Borràs PhD work.

Chameleon stiffness

As for the test of a short range Yukawa fifth force, one can take advantage of MICROSCOPE's technical sessions aimed at characterizing the electrostatic stiffness of the instrument for testing the chameleon model. Again, the principle of this test consists in computing the chameleon's stiffness and checking whether it fits into the discrepancy-budget between the measured stiffness and the expected one (from classical effects only, mostly electrostatics).

This study was performed in Ref. [179] and heavily relies on two former work [136, 180] for computing the chameleon field inside the nested cylinders and deriving the chameleon stiffness. The constraints obtained are represented in Fig. 1.5, and turn out not to be competitive with the current state-of-art. Again, this is not overly surprising as MICROSCOPE was not designed for looking for such fifth forces, and technical sessions did not primarily serve that purpose either. A suggested route to improve these constraints would be to increase asymmetries in the device [138].

As a closing remark, it is to be noted that, the actual geometry of the experimental apparatus being complex and given the techniques developed in Refs. [136, 180], a lot of simplifying assumptions were necessary to get the results from Ref. [179]. As a matter of fact, different techniques had to be used depending of the chameleon field regime (namely screened, deeply screened and unscreened), with associated hypotheses. In the screened regime,

a semi-analytical 2D model could be applied while in the other regimes, 1D numerical simulations had to be employed. All in all, this makes the computation of the field in the relatively simple setup of nested cylinders quite complex!

1.4 The need for new numerical tools

Section 1.3.4 illustrated how MICROSCOPE's data — both collected from sessions dedicated to the WEP test and from technical sessions aiming at characterizing the instrument — can be used to test alternative theories of gravity. Some of these analyses provided state-of-the-art bounds on given models (mainly those that come with a WEP violation at the Lagrangian level), some resulted in non-competitive constraints on other models (mainly those predicting the existence of a putative short range fifth force). In the light of these results, it appeared that the testability of models with nonlinearities at the PDE level is crucially dependent on the development of new numerical methods for a realistic modeling of their feature. This path for future development is advocated in several work, including Refs. [104, 138, 181]. While valuable qualitative insights can be gained using analytical approximations, numerical computations remain essential to establish a quantitative connection with real-world observations and experimental data.

Being able to numerically compute scalar field profiles, in the context of scalar-tensor theories exhibiting screening mechanisms, is indeed desirable is several respects:

1. *Validation of analytical approximations* — As seen in Sec. 1.2.2 for the case of the chameleon model, approximate solutions to the nonlinear Klein–Gordon equation (1.117) can be derived analytically.[32] Essentially, such approximations are found by adequately simplifying the PDE in specific regions, based on physical and mathematical intuition. It is therefore crucial to compare them against numerical solution, not only to validate them but also to clearly identify their limits. Indeed, analytical approximations are generally expected to be valid in a certain regime, and numerical computations can prove to be valuable for identifying this regime's 'boundary'.

2. *Going beyond analytical approximations* — Numerical computations can be viewed as the way to free ourselves from the limitations of analytical approximations. The latter are derived in the case of highly symmetrical geometries (sphere, cylinder, infinite walls, etc.). In general, experimental setups are much more complex in terms of geometry and distribution of mass within that geometry, challenging the relevance of analytical approximations for obtaining faithful scalar field profiles. Specifically, deriving accurate constraints on screened scalar-tensor models from a given experiment must sometimes be done on the basis of numerical approximations, for otherwise the risk of underestimating or overestimating bounds on the model's parameters is present.

3. *Designing new tests (prospective)* — Having a numerical tool for computing fifth force effects also proves to be useful when it comes to exploring new ways of testing the model at stake. Such a tool could indeed help design relevant new tests, and in that sense, guide the hunt for screened scalar fields.

In this section, we begin by reviewing the existing available numerical tools and shed light on their limitations. From there, we establish the overall specifications of the numerical tool developed in this PhD work.

1.4.1 Existing solutions and limitations

1D solver with shooting method

Static three-dimensional PDEs can be reduced to one-dimensional ODEs when one of the following symmetry conditions applies:

- translation invariance along two axes (infinite parallel walls),
- rotational invariance along two axes (spherical symmetry),
- translation and rotational invariance along one axis (infinite cylinder).

From a numerical view point, nonlinear ODEs are easier to solve than nonlinear PDEs. There are plenty of options to choose from in the special case of initial value problems, Runge–Kutta schemes being the most commonly employed class of methods. However, the scalar field and its derivative values are, in general, unknown in the numerical domain. In the spherically symmetric case, all we know is that $\phi'(r = 0)$ should be zero, while the asymptotic behavior of the field is given by the asymptotic boundary condition (1.120).

[32] For instance, Ref. [149] compiles a great number of such approximations.

Because asymptotic boundary conditions cannot be imposed 'by hand' in traditional ODE solvers, Refs. [136, 145, 182] implement a *shooting method* for properly accounting for the boundary condition at infinity — this careful treatment of boundary condition is overlooked in Refs. [183, 184] and semi-analytically accounted for in Ref. [185] by imposing an asymptotic Yukawa profile. This technique is used in the subsequent work Ref. [179]. To the best of our knowledge, this was the only work properly accounting for the asymptotic boundary condition (1.120) numerically prior to the present PhD work.

While this shooting method works well for cases for which the scalar field PDE boils down to an ODE, it does not generalizes to the 2D and 3D cases. Likewise, numerical techniques for solving PDEs are drastically different from those used to solve ODEs.

2D semi-analytical model for nested cylinders

With a view to compute the chameleon field profile inside MICROSCOPE's SUEP and SUREF (see Sec. 1.3), Ref. [180] proposes a semi-analytical technique for solving the scalar field nonlinear Klein–Gordon equation (1.117) in the presence of not perfectly coaxial nested cylinders. There, the three-dimensional PDE does not reduce to an ODE but rather to a two-dimensional PDE (due to translation invariance along the infinite cylinders' axes). Note that the semi-analytical prescription developed in this article, based on mode decomposition, works for small off-centering of the cylinders and is able to account for the asymptotic boundary condition (1.120) through the nullity of the monopole at infinity.

Finite Difference codes

The finite difference method (FDM) is one commonly employed numerical technique to numerically solve PDEs in dimension greater than or equal to two, where differential operators are approximated with finite differences. FDM has been used in the context of screened scalar-tensor models. In a nutshell, the numerical procedure consists in a finite difference method in which, from an initial guess, the algorithm iteratively converges towards the correct solution — see e.g. the Appendix of Ref. [186] for a more detailed description of this algorithm (under-relaxed Gauss-Siedel scheme). Ref. [180] obtains the chameleon field profile for the case of off-centered, infinitely long, nested cylinders. Likewise, Ref. [187] solves the Klein–Gordon equation where ρ mimics the density distribution in the Eöt-Wash (torsion pendulum) experiment. Ref. [186] calculates the chameleonic fifth force in a cylindrical vacuum chamber with a source mass inside of it in the context of atom interferometry experiments.

In all the aforementioned references, the asymptotic boundary condition (1.120) has to be abandoned. Instead, a standard Dirichlet boundary condition[33] is set at a finite distance from the examined setup, thereby assuming that the scalar field reaches its asymptotic value not too far away from the latter. As already underlined, this assumption is not always valid. As a matter of fact, it is plainly wrong when the Compton wavelength of the field in the ambient medium is much larger than the typical size of the numerical domain, and imposing $\phi_\infty = \phi(\|\mathbf{x}\| = +\infty)$ at a finite distance can result in significant errors (see Chapt. 3). To get a flavor of why this is problematic, Fig. 1.6 shows what happens to the Newtonian potential for a homogeneous solid sphere of radius R_b when the Dirichlet boundary condition $\Phi(R_c) = 0$ is applied at some radius R_c (colored solid curves) instead of applying the asymptotic condition $\Phi(r) \to 0$ when $r \to +\infty$ (gray dashed line). In particular, the bottom panel, which represent the relative error with respect to the latter benchmark, shows that even when the truncation radius is set at $R_c = 100R_\mathrm{b}$, the relative error is no better than $\sim 1\%$.

Additionally, let us note that FDM is usually not used for irregular computer-aided design (CAD) geometries. Instead, the use of rectangular grids works well for rectangular or block-shaped models, which are unfortunately not necessarily adapted to complex shapes found in any realistic experimental setup.

Finite Element codes

More recently, the finite element method (FEM) has been leveraged to compute scalar field profiles in various models with screening. FEM is quite straightforward to apply for linear elliptic second-order PDEs, and can be extended to the nonlinear case, although this is sometimes challenging from the perspective of numerical convergence. Among the method's advantages, let us stress that it can be used to model virtually any given geometry, as complex as it may be, and can be adapted to handle time-dependent problems. Chapt. 2 is dedicated to a more in depth presentation of FEM and covers all the aforementioned points.

A succinct review of existing FEM codes dedicated to the study of screening mechanisms arising in scalar-tensor theories can be found in Ref. [188]. To the best of our knowledge and in line with Ref. [188], it started being used by a research group from the University of Nottingham led by C. Burrage for studying the shape dependence of the chameleon screening mechanism [189]. Building on top of Ref. [189], C. Briddon developed a user-friendly code called SELCIE as part of his PhD work for investigating the chameleon model using FEM [190].

[33] The various types of boundary conditions that can be applied on the boundary of a bounded open set are discussed in Chapt. 2.

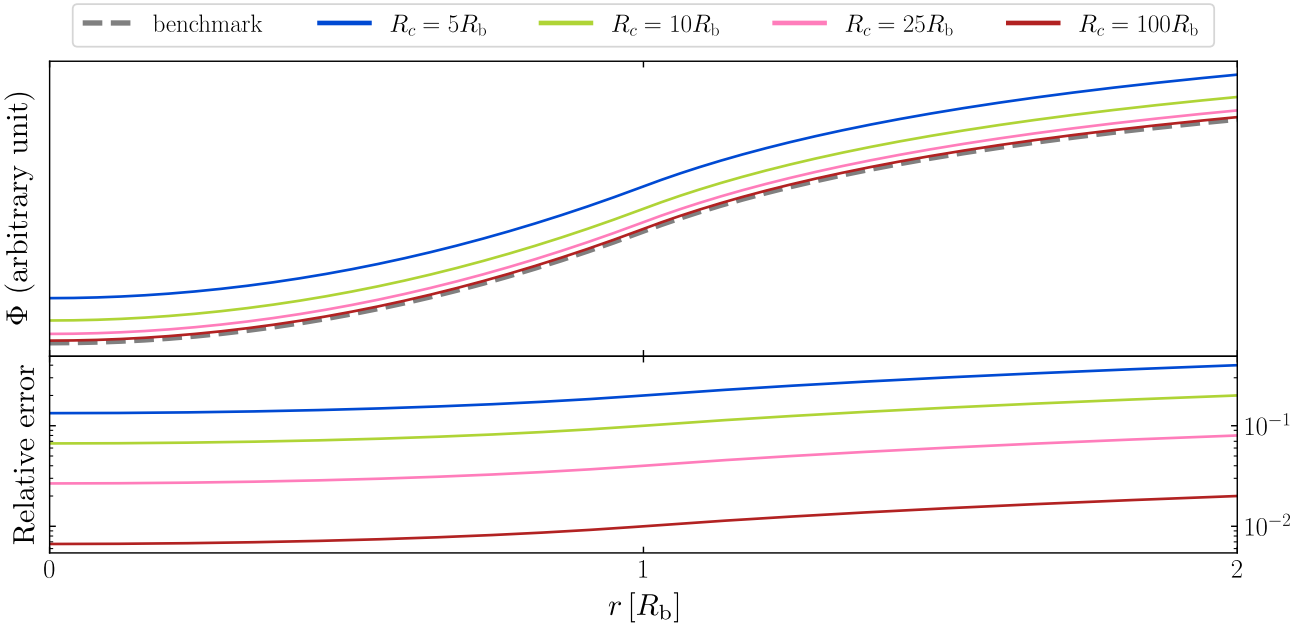


Figure 1.6: Influence of the truncation radius R_c on the computation of the Newtonian potential of a perfect sphere with radius R_b. The truncation radius $R_c \in \{5R_\mathrm{b}, 10R_\mathrm{b}, 25R_\mathrm{b}, 100R_\mathrm{b}\}$ corresponds to the radius where the Dirichlet boundary condition $\Phi(R_c) = 0$ is applied. The top panel shows the resulting Newtonian potential profiles, where the gray dashed curve labeled 'benchmark' corresponds to the solution for which $\Phi(r) \to 0$ as $r \to +\infty$. The bottom panel shows the relative error with respect to the benchmark solution. As intuition suggests, the larger the truncation radius R_c, the smaller the relative error.

FEM has also been used to gain insight into the symmetron model [191, 192]. There is also the φ-`enics` code which offers a one-dimensional implementation of Vainshtein models [193].

The main limitation of all existing FEM code is that computation can only be performed on bounded domains and thus cannot account for the asymptotic boundary condition (1.120). Indeed, as we will see in the next chapter, FEM relies on the meshing of the numerical domain and cannot — in its standard implementation — encompass a region of infinite spatial extension. As a result, all the aforementioned work are limited to finite simulation boxes and have to use Dirichlet or Neumann boundary conditions on the boundary of the numerical domain. This can be justified in some situations — let us take the example of laboratory experiments performed in vacuum chamber (e.g. atom interferometry experiments [142, 158]) for testing the chameleon model. In some parts of the parameter space, the walls of the vacuum chamber can legitimately be considered as screened (when they are thick and dense enough), and one can thus apply the Dirichlet boundary condition $\phi = \phi_\mathrm{min}(\rho_\mathrm{wall})$ deep inside the wall (i.e. at the boundary of the numerical domain) or a vanishing Neumann boundary condition $\partial\phi/\partial n = 0$ as done in Ref. [190]. However, this assumption — that the scalar field reaches the value that minimizes the effective potential at the boundary of the simulation box — does not hold in the general case. For instance, the simple example of an unscreened satellite in space cannot be handled that way. More generally, the case of a dense object embedded in a less dense medium of density ρ_vac is problematic because there is a priori no reason for the field to reach $\phi_\mathrm{min}(\rho_\mathrm{vac})$ anywhere near the dense object. This is physically well-understood by looking at the expression for the effective mass of the scalar field [Eq. (1.118b)], which shows that the field's mass decreases with ρ, and can therefore be very light in low-density environment. More precisely, the Compton wavelength in such a low density medium can be several orders of magnitude larger than the typical size of the region of interest (where the fifth force is to be evaluated for instance). In this kind of situations, the numerical domain is required to be very large and the simulation can therefore become computationally prohibitive.

N-body simulations in modified gravity

Modified gravity models have also been considered in the context of astrophysical N-body simulations. N-body codes for modified gravity are usually based on standard gravity codes, see Refs. [188, 194] for reviews of those. The scalar field profile (and hence the scalar fifth force) is usually obtained under the quasi-static hypothesis using multigrid techniques — see e.g. Ref. [194] for a good overview of how they work building on top of finite differences. State-of-the-art codes are `ECOSMOG` [195], `ISIS` [196] and `MG-GADGET` [197]. Generally speaking, regular grids are useful for very-large-scale simulations (meteorological, seismological and astrophysical simulations).

Specification	ODE solver [184, 185]	1D shooting method [136]	Semi-analytical nested cylinders [180]	FDM / multigrid [180, 186, 187, 194]	FEM codes [189–193]
Asymptotic boundary condition	✗	✓	✓	✗	✗
Complex geometries	✗	✗	✗	rectangular grid	✓
Spatial dimensions	1D	1D	2D	1D, 2D, 3D	1D, 2D, 3D
Coordinate system	—	—	polar	Cartesian	Cartesian, cylindrical
Time-dependence	✗	✗	✗	✗	✗
Multi-scale simulations	✗	✗	✗	✓	possible through h-adaptivity

Table 1.4: Summary of the specifications of existing numerical tools presented in Sec. 1.4.1.

Time-dependent solvers

The picture would not be complete without mentioning several work that attempt to go beyond the quasi-static approximation. This is a path that needs to be taken if one is interested in scalar radiation (also known as scalar waves), or in knowing whether accounting for the time dependency in the Klein–Gordon equation can affect the efficiency of the screening mechanism. Ref. [198] studies the chameleon field profile around a radially pulsating mass (one time dimension + one spatial dimension), although the exact numerical scheme employed is not stated. Ref. [199] performs the first fully relativistic simulations of binary neutron stars in scalar-tensor theories with kinetic screening (see Sec. 1.2.1) using FDM on grids with six refinement levels. To the best of our knowledge, there is no FEM code (e.g. similar to SELCIE [190]) for handling the time-dependence in the chameleon or symmetron models, for which the equation of motion is a nonlinear wave equation.

1.4.2 Outline of the tool's overall specifications

A non-exhaustive summary of the technical specifications of existing numerical tools for studying screened scalar-tensor gravity is provided in Table 1.4. Here, we list the features we would like to have in a numerical tool for studying screened scalar-tensor gravity in space and take M. Pernot-Borràs' work [138] a step further. These desired functionalities are matched by a choice a technical specifications — Table 1.4 is useful in this respect.

The list of desired features, which will shape the development of a new numerical tool, roughly follows the entries of Table 1.4 but are reordered in terms of priority:

- *Handling of arbitrary shapes* — From our discussion above in Sec. 1.4.1, it appears that handling complex geometries is key to a realistic modeling of modified gravity in both experimental setups and Solar system / astrophysical environments. In this regard, the finite element method is perhaps the best suited numerical method as it is possible in principle, through finite element meshes, to represent any given geometry, however complex it may be. Moreover, it is particularly well-suited to solve elliptic second-order PDE, a category of PDEs under which most scalar-field equations fall. As the chosen method for our tool, FEM is presented in its own chapter, Chapt. 2. In particular, it will be shown how one can address nonlinear PDE, which are at the heart of screening mechanisms.

- *Implementation of asymptotic boundary conditions* — As stressed above, having the possibility to account for asymptotic boundary conditions is more than desirable as it would allow us to model modified gravity in very general cases and avoid having to make assumptions about the behavior of the scalar field within a bounded numerical domain as done in Refs. [189–193]. The way they are handled in Ref. [136], namely by means of a shooting method in 1D, does not generalize well to the higher dimensional case (2D and 3D). Nor is it possible to mesh the whole unbounded domain as this would require the storage of an infinite number of elements on a finite-memory computer. Nonetheless, standard FEM can be extended to account for the behavior of the unknown at infinity in satisfactory ways. We delve into some of these advanced techniques in Chapt. 3.

- *Spatial dimensions* — While highly symmetrical geometry allows for dimensional reduction — for instance, only one coordinate is needed to deal with the spherically symmetric case, see our discussion at the beginning of Sec. 1.4.1 — computations involving arbitrary shapes cannot be undertaken but in 3D. In the light of this remark, our numerical tool should be able perform 1D and 2D simulations for cases which enjoy a certain symmetry, and 3D simulations to accommodate for the most general geometries. Again, FEM can be formulated in either of those three distinct dimensional cases.

- *Ability to perform time-dependent simulations* — In Sec. 1.2.2, we considered the time-independent version of the nonlinear Klein–Gordon equation governing the dynamics of the chameleon field, for which the

d'Alembertian is replaced by the Laplacian. Doing so is valid as long a the typical timescale of the field's dynamics is much shorter than the typical time-variation of the matter that is being modeled. The former scales as L_0/c, where L_0 denotes the length scale of the problem at stake, whereas the latter scales as L_0/v, where v is the typical velocity of matter in this problem. Therefore, this condition, better known as the *quasi-static* regime, holds in the non-relativistic limit. When the quasi-static approximation does not hold, keeping the $\partial^2/\partial t^2$ term of the d'Alembert operator is necessary. This turns out to be the case for all physical phenomena involving scalar waves. Our aim being to develop a general-purpose numerical code, it should be sufficiently flexible to implement time dependency. This is possible within the framework of FEM, the ins and outs being discussed in Chapt. 2.

– *Possibility to keep track of physical effects on a large range of scales* — Tests of gravity in space involve various length scales: from the smallest, such as MICROSCOPE's cylinders, larger ones, such as the Earth-satellite distance (the Earth being the main source of gravity in orbit), up to the astronomical unit scale. Therefore, the tool should be able to handle a wide range of length scales, $\sim 10^{-3}$–10^{7} m. There are several ways in which FEM can deal with such a wide range of length scales. The first idea that comes to mind is to employ h-adaptivity, which consists in locally adapting the mesh size in the numerical domain: fine in regions where fine details are present or where high gradient occurs (e.g. nearby the satellite in our example), and coarse elsewhere. That way, computing resources are focused where most needed, thereby avoiding the computation cost of resolving the smallest scales throughout the entire domain. Other more advanced methods include domain decomposition techniques [200, 201].

– *Implementation of several coordinate systems* — For practicality of use, it is desirable to be able to conduct numerical computation using different coordinate systems depending on the problem at stake. For instance, the PDEs of problems enjoying invariance by rotation along an axis might be written either in polar coordinates or cylindrical coordinates (with vanishing partial derivatives with respect to the corresponding angle in both cases). One may be better suited than the other given a specific problem.

Given the above list, it appears that among the existing tools reported in Table 1.4, none of them meets all our needs. This observation led us to develop a new numerical tool called *femtoscope*, which will be presented in Chapt. 4. Before that, the next two chapters will lay out its mathematical foundations. Specifically, Chapt. 2 introduces the finite element method while Chapt. 3 tackles the question of dealing with PDE problems posed on unbounded regions of $\mathbb{R}^n$, $n \in \{1, 2, 3\}$.

Chapter summary

This chapter was dedicated to the introduction of scalar-tensor theories of gravity, one of the most natural alternatives to General Relativity where gravity is mediated by both a rank-2 tensor field and a scalar field. While scalar fields are key players in many extensions beyond the standard models, they must somehow comply with all the numerous tests of gravity accumulated throughout the past decades. In particular, the fifth force they mediate must not betray their presence in Solar system tests or laboratory experiments. We saw that this can be achieved by means of screening mechanisms which dynamically suppress deviation from General Relativity in classical fifth force searches, with a focus on the chameleon model.
Although space-based experiments were long-expected to provide new constraints on the chameleon model, the legacy of the MICROSCOPE mission has shown that experimental tests of the weak equivalence principle in space are not as straightforward and groundbreaking as initially hoped in this regard. One of the lessons learnt is that the testability of screened scalar-tensor models is crucially dependent on the development of new numerical methods, for a realistic modeling of their effects. In this perspective, we outlined the specifications of a FEM-based numerical tool to be developed in this PhD work. The next chapter therefore introduces the finite element method in more depth.

Chapter 2

The Finite Element framework

Outline of the current chapter

In this chapter, we describe, in a rather concrete manner, how PDE problems posed on bounded domains can be solved with the Finite Element Method (FEM). Starting with the linear elliptic case, we introduce the so-called *weak formulation* of the problem which, under an appropriate functional framework, can be given a precise meaning and constitutes a well-posed mathematical problem. From there, things become a little less abstract as we discretize this weak formulation according to a mesh of the domain. This procedure results in a mere linear system which, upon solving, yields a numerical approximation of the solution to the original PDE problem. In a second stage, we deal with nonlinear problems and review the most commonly used techniques in the literature. Most of them build on top of standard FEM by iteratively solving a sequence of linearized problems, which in turn requires the definition of suitable stopping criteria. The main difficulty with such iterative techniques is to ensure their convergence. In that respect, we provide a list of (mostly empirical) common practices that were put into use in this PhD work in order to enhance their robustness. We also give insights into *(i)* dimensional reduction of the PDE problem in the presence of global continuous symmetries, and *(ii)* how time-dependent problems may be addressed in the future.

In contrast to what can be found in several FEM textbooks, the philosophy of this chapter is to expose the very basics of FEM on the basis of examples, sometimes at the expense of mathematical rigor. In particular, techniques to deal with nonlinear PDEs are illustrated on the nonlinear Klein–Gordon equation (1.117) arising as the Newtonian limit of the scalar field equation in the chameleon model that we discussed in the previous chapter. Finally, this chapter sets the stage for Chapt. 3 in which we will deal with problems posed on unbounded domains.

2.1 Overview of the Finite Element Method

The Finite Element Method is a general numerical method for solving PDEs together with a set of constraints imposed on the boundary of the domain. The latter constraints are referred to as *boundary conditions* and have to be specified in order to ensure the uniqueness of the solution (provided it exists). The main idea behind FEM is to mesh a continuous spatial domain into a finite set of non-overlapping subdomains — the finite elements — over which the problem takes a simpler form. One of the key contributions in FEM comes from the analysis of aircraft structures back in the 1950s [202], which is why it is often associated with elasticity and structural analysis problems in aeronautical engineering. Since then, the method has been widely adopted in many other engineering disciplines, including heat transfer, electromagnetism, acoustics, and fluid dynamics (see e.g. Ref. [203]).

2.1.1 Problem definition

Before diving straight into the ins and outs of FEM, it is important to recall its scope: what type of problems can it address? To that extent, we are going to consider partial differential equations of the form

$$\mathcal{L}u = f \text{ on } \Omega\,, \tag{2.1}$$

where Ω is an open connected bounded subset of $\mathbb{R}^n$, $n \in \mathbb{N}^*$ being the dimension of the problem, $f\colon \Omega \to \mathbb{R}$ is a given function, $\mathcal{L}$ is some linear partial differential operator and $u\colon \Omega \to \mathbb{R}$ is the unknown. In this PhD work, the PDEs we are trying to solve are all second-order equations. Under the assumption that u is of class $\mathcal{C}^2(\Omega)$, the $\mathcal{L}$ operator can then be given a more explicit form

$$\mathcal{L}u = -\mathbf{C}\colon \mathbf{H}_u^T + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u + du = -\sum_{i,j=1}^{n} c_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n}\beta_i\frac{\partial u}{\partial x_i} + du\,. \tag{2.2}$$

In the above expression, coefficients $\mathbf{C} = (c_{ij})_{1\leq i,j\leq n}$, $\boldsymbol{\beta} = (\beta_i)_{1\leq i\leq n}$ and d can be any given real functions of $\mathbf{x} \in \Omega$ only.[1] $\mathbf{H}_u$ denotes the Hessian matrix of u while the colon operator ':' represents the Frobenius inner product (also known as the double-dot product). Due to the symmetry of second derivatives, the matrix $\mathbf{C}$ can always be assumed to be symmetric without loss of generality. Alternatively, if coefficients c_{ij} are $\mathcal{C}^1(\Omega)$ functions, it is possible to rewrite the $\mathcal{L}$ operator as

$$\mathcal{L}u = -\mathrm{div}\,(\mathbf{C}\boldsymbol{\nabla}u) + \mathbf{b}\cdot\boldsymbol{\nabla}u + du = -\sum_{i,j=1}^{n}\frac{\partial}{\partial x_i}\left(c_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n} b_i\frac{\partial u}{\partial x_i} + du\,. \tag{2.3}$$

The PDE (2.1) is said to be in the *divergence form* if $\mathcal{L}$ is given by Eq. (2.3) and in the non-divergence form if it is given by Eq. (2.2) instead. Note that these two forms are simply related though $\beta_j = b_j - \sum_{i=1}^{n}\partial_{x_i}c_{ij}$. We will see that, for our discussion of FEM, the divergence form is the most natural of the two.

Second-order linear PDEs in physics usually come in three different flavors:

1. *Elliptic* if for all $\mathbf{x} \in \Omega$, $\mathbf{C}(\mathbf{x})$ is positive definite. This is the case, for example, of the Poisson equation $-\Delta u = f$ governing the potential in Newtonian gravity for which $\mathbf{C} \equiv \mathbf{I}_n$.

2. *Parabolic* if for all $\mathbf{x} \in \Omega$, the eigenvalues of $\mathbf{C}(\mathbf{x})$ are all strictly positive except exactly one that is zero. The most famous example of a parabolic PDE is the heat equation $\partial_t u = \alpha\Delta u$.

3. *Hyperbolic* if for all $\mathbf{x} \in \Omega$, $\mathbf{C}(\mathbf{x})$ has 1 strictly negative and $n-1$ strictly positive eigenvalues. For instance, the wave equation $\partial_{tt}u - c^2\Delta u = 0$ is a hyperbolic PDE.

Note that, so far, we have not assigned a specific physical role to the components of $\mathbf{x} = (x_1, \ldots, x_n)$, and so nothing prevents us from having one time dimension alongside $n-1$ spatial dimensions as is the case for the heat equation or wave equation. Nonetheless, for the time being, we restrict ourselves to stationary PDE problems of elliptic nature, for which FEM is well-suited (this idea will be revived when discussing spacetime FEM). For this type of PDEs, each term in Eq. (2.3) can be given a physical interpretation: $-\mathrm{div}(\mathbf{C}\boldsymbol{\nabla}u)$ represents the *diffusion* of u within Ω, $\mathbf{b}\cdot\boldsymbol{\nabla}u$ is a *transport* term (picturing $\mathbf{b}$ as some velocity field), also referred to as an *advection* term, and du might be interpreted as a *creation/annihilation* term. We defer the discussion of time-dependent problems to Sec. 2.1.4.

[1] If the coefficients also depend on u and/or its partial derivatives, the partial differential operator $\mathcal{L}$ becomes nonlinear.

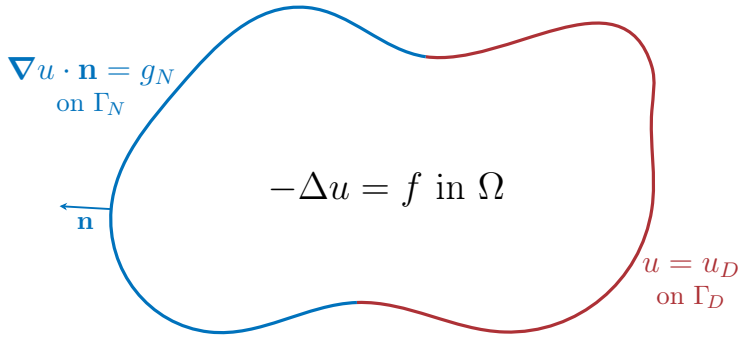


Figure 2.1: Poisson equation posed on some domain Ω together with Dirichlet boundary condition on Γ_D (in red) and Neumann boundary condition on Γ_N (in blue). The outward normal vector $\mathbf{n}$ is represented by the blue arrow.

Box D: Definition of well-posed problems (in the sense of Hadamard)

The notion of well-posedness in PDE problems was introduced by French mathematician Jacques Hadamard [204]. A problem is said to be well-posed if all the following conditions are met:

1. *a solution exists,*
2. *this solution is unique,*
3. *the solution depends continuously on the data given in the problem.*

Otherwise, it is ill-posed. The meaning of the first two points is clear. Essentially, the last requirement reflects the fact that "small" changes in either initial conditions, boundary conditions or parameters' value should result in "small" changes in the solution. It is thus particularly important for problems arising from physical applications.

For the problem Eq. (2.1) to be well-posed — see Box D —, it is necessary to supplement the PDE with boundary conditions imposed at the border Γ of Ω. Indeed, without such additional constraints, the uniqueness of the solution is likely not to hold, leading to an ill-posed problem. The specification of boundary conditions generally takes different forms depending on the physical problem at stake. Here, we partition the border Γ into $\Gamma = \Gamma_D \cup \Gamma_N$ and focus on the two main types:

1. *Dirichlet* or *essential* boundary conditions, for which the value of the unknown on Γ_D is imposed

$$u = u_D \text{ on } \Gamma_D\,, \tag{2.4}$$

   where $u_D : \Gamma_D \to \mathbb{R}$ is part of the problem's data.

2. *Neumann* boundary conditions, which relates to the gradient of the unknown on Γ_N. Traditionally, it consists in setting the normal derivative $\partial u/\partial n := \nabla u \cdot \mathbf{n}$, where $\mathbf{n}$ denotes the outward normal vector to Γ_N. Here however, we provide a slightly different definition by specifying the co-normal derivative

$$\frac{\partial u}{\partial \nu} := (\mathbf{C}\nabla u) \cdot \mathbf{n} = \nabla u \cdot (\mathbf{C}^T \mathbf{n}) = g_N \text{ on } \Gamma_N\,, \tag{2.5}$$

   where $g_N : \Gamma_N \to \mathbb{R}$ is part of the problem's data. Indeed, using this oblique derivative instead of the usual normal derivative will make things easier when deriving the weak formulation of the problem. While the knowledge of the normal derivative usually does not imply the knowledge of the co-normal derivative (and *vice-versa*), we still refer to Eq. (2.5) as a Neumann boundary condition. Moreover, this distinction becomes irrelevant when $\mathbf{C} \equiv \mathbf{I}_n$.

These two types of boundary conditions — Dirichlet and Neumann — are illustrated in Fig. 2.1 for a Poisson equation. It should be noted that, even though they model the situations most often encountered in real problems, there exist other types of boundary conditions.[2]

[2]We can mention three more: Robin, mixed, Cauchy and periodic boundary conditions.

Our model second-order PDE problem finally reads

$$\mathcal{L}u = f \text{ in } \Omega \quad \text{with} \quad \begin{cases} u = u_D & \text{on } \Gamma_D \\ \partial u/\partial \nu = g_N & \text{on } \Gamma_N \end{cases} . \tag{2.6}$$

2.1.2 Variational formulation

We now transform the so-called *strong* PDE problem (2.6) by writing it in its *weak* form (also known as *variational* form). This step is not only the starting point of FEM, but also leads to a convenient framework for the theoretical study of PDEs, where many concepts and properties from functional analysis can be applied. We first show in a pragmatic way how to derive the weak formulation. Only then do we add a layer of rigor by introducing an adequate functional framework for it to make sense.

Heuristic derivation of the weak formulation

Schematically, the weak formulation is obtained by *(i)* multiplying the strong PDE (2.1) by some *test function* v belonging to some functional space V to be further specified, *(ii)* integrating over the whole space Ω, and *(iii)* integrating by parts the highest derivative term. Let us illustrate this process on the model problem (2.6). Anticipating the fact that we will have to perform an integration by parts in step *(iii)*, we choose to write the differential operator $\mathcal{L}$ in its divergence form (2.3). After completing the first two steps, we obtain the expression

$$-\int_\Omega \operatorname{div}(\mathbf{C}\boldsymbol{\nabla} u)\, v \,\mathrm{d}\mathbf{x} + \int_\Omega (\mathbf{b}\cdot\boldsymbol{\nabla} u)\, v \,\mathrm{d}\mathbf{x} + \int_\Omega duv \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} . \tag{2.7}$$

Before going any further, let us clarify the role of the test function $v \in V$ appearing in the above equation. Requiring Eq. (2.7) to hold only for one given test function is not restrictive enough. Instead, we demand it to hold true for *any* function $v \in V$. In that sense, Eq. (2.7) actually conceals a non-countable infinite number of equations![3] Back to the algebra, performing an integration by parts[4] in the first integral yields

$$\int_\Omega (\mathbf{C}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} - \int_\Gamma \big[(\mathbf{C}\boldsymbol{\nabla} u)\cdot\mathbf{n}\big] v \,\mathrm{d}\gamma + \int_\Omega (\mathbf{b}\cdot\boldsymbol{\nabla} u)\, v \,\mathrm{d}\mathbf{x} + \int_\Omega duv \,\mathrm{d}\mathbf{x} = \int_\Omega fv \,\mathrm{d}\mathbf{x} . \tag{2.8}$$

This has had two consequences worth of notice:

1. First, this has led to the appearance of a boundary term where the integral is carried over Γ which is a lower-dimensional topological entity with respect to Ω. We can thus account for the boundary conditions stated in Eq. (2.6) through this term. For practical reasons justified later, the test function is chosen to be zero on Γ_D (the part of the boundary where the Dirichlet boundary condition is prescribed). As a result, there remains
$$\int_\Gamma \big[(\mathbf{C}\boldsymbol{\nabla} u)\cdot\mathbf{n}\big] v \,\mathrm{d}\gamma = \int_{\Gamma_N} g_N v \,\mathrm{d}\gamma . \tag{2.9}$$

2. Eq. (2.8) only involves at most first-order derivatives of u through the gradient term while the original strong-form PDE (2.1) was dependent on second-order derivatives. In other words, the latter strong formulation only makes sense for at least twice-differentiable functions, yet Eq. (2.8) could be satisfied by only-once-differentiable functions.

In order to end-up with a somewhat more *canonical* form, we assume V to be a vector space and define the (bi-)linear maps

$$\begin{aligned} a &: V \times V \to \mathbb{R} \\ &(u, v) \mapsto \int_\Omega (\mathbf{C}\boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + \int_\Omega (\mathbf{b}\cdot\boldsymbol{\nabla} u)\, v \,\mathrm{d}\mathbf{x} + \int_\Omega duv \,\mathrm{d}\mathbf{x} \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} l &: V \to \mathbb{R} \\ &v \mapsto \int_\Omega fv \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v \,\mathrm{d}\gamma . \end{aligned} \tag{2.11}$$

[3] The function space V is indeed an uncountable set.

[4] Some work invoke the "divergence theorem" or the "Green's formula". This is just a matter of terminology.

In this way, Eq. (2.8) can be concisely written as $a(u,v) = l(v)$ and the weak form loosely reads

$$\text{Find } u \text{ such that for all } v \in V \,/\, v \equiv 0 \text{ on } \Gamma_D\,,\ a(u,v) = l(v)\,. \tag{2.12}$$

There are nonetheless several shortcomings with the weak formulation (2.12) that need to be clarified. First, we have to make sure that all the above integrals actually exist. This is going to translate into conditions on the space V, on the coefficients of the PDE $\mathbf{C}$, $\mathbf{b}$, d, and on the data f. Furthermore, the way Dirichlet boundary conditions are taken into account is not clear yet. And last but not least, how do we turn Eq. (2.12) into a well-posed problem? How can well-posedness, as defined in Box D, be checked?

The adequate functional framework: Sobolev spaces

Let us now be a little more rigorous mathematically speaking. While the discussion could be carried out within the general framework of second-order elliptic PDEs — see Evans' book [205], Chapt. 6 —, we narrow it down to the case of the Poisson equation $-\Delta u = f$. Indeed, processing this example is sufficient for introducing all the main concepts, other cases being handled in the same way but for a few technical details that are of minor importance to us. We thus have

$$a(u,v) = \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} \quad \text{and} \quad l(v) = \int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N v \,\mathrm{d}\gamma\,. \tag{2.13}$$

Our choice of an adequate functional space V is greatly facilitated by the Lax–Milgram theorem (reported in Box E below), which is an extremely powerful tool for assessing the well-posedness of weak forms. Until further notice, we assume *homogeneous* Dirichlet conditions on Γ_D, that is $u_D \equiv 0$. The non-homogeneous case will be dealt with the end of this discussion.

Box E: Lax–Milgram theorem

Let V be a Hilbert space with norm $\|\cdot\|_V$, and V' its dual space with norm $\|\cdot\|_{V'}$. Let a be a real bilinear mapping defined over $V \times V$ and l be a real linear mapping defined over V. Assume that

1. a is continuous over $V \times V$, i.e. there exists a constant $M \geq 0$ such that for all $u, v \in V$

$$|a(u,v)| \leq M \,\|u\|_V \,\|v\|_V\,; \tag{2.14}$$

2. a is coercive over V, i.e. there exists a constant $\alpha > 0$ such that $\forall v \in V$

$$a(v,v) \geq \alpha \,\|v\|_V^2\,; \tag{2.15}$$

3. l is continuous over V, i.e. there exists a constant $L \geq 0$ such that for all $v \in V$

$$|l(v)| \leq L \,\|v\|_V\,. \tag{2.16}$$

Then, there exists a unique element $u \in V$ such that

$$a(u,v) = l(v)\,,\ \forall v \in V\,.$$

Additionally, the solution u depends continuously on l, i.e.

$$\|u\|_V \leq \frac{1}{\alpha}\,\|l\|_{V'}\,,\ \text{where } \|l\|_{V'} \coloneqq \sup_{v \neq 0} \frac{l(v)}{\|v\|_V}\,. \tag{2.17}$$

The first idea that comes to mind is to use $V = L^2(\Omega)$, the space of square-integrable functions which is defined as

$$L^2(\Omega) \coloneqq \left\{ u\colon \Omega \to \mathbb{R} \text{ such that } \int_\Omega |u(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} < +\infty \right\}. \tag{2.18}$$

The reason being that, when equipped with the inner product

$$\langle u, v \rangle_{L^2} \coloneqq \int_\Omega u\, v \,\mathrm{d}\mathbf{x}\,, \tag{2.19}$$

$L^2(\Omega)$ forms a Hilbert space,[5] which is a first step in the direction of the Lax–Milgram theorem. Moreover, it becomes possible to give a meaning to the integral of the product of two functions by merely requiring them to be in $L^2(\Omega)$. Indeed, in view of the Cauchy–Schwarz inequality

$$\int_\Omega u(\mathbf{x})\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} \leq \left(\int_\Omega |u(\mathbf{x})|^2\right)^{1/2} \left(\int_\Omega |v(\mathbf{x})|^2\right)^{1/2} = \|u\|_{L^2}\, \|v\|_{L^2} < +\infty\,,\ \forall u, v \in L^2(\Omega)\,, \tag{2.20}$$

where $\|\cdot\|_{L^2}$ is the norm associated with the inner product $\langle \cdot, \cdot \rangle_{L^2}$. Therefore, it seems natural to ask for the test function v and the rhs function f appearing in Eq. (2.13) to be in $L^2(\Omega)$. Likewise, it seems reasonable to ask for the partial derivatives of u and v to be in $L^2(\Omega)$ so that the bilinear form $a(\cdot, \cdot)$ in Eq. (2.13) is well-defined. Yet, for technical reasons not worth delving into here, one has to abandon the notion of smooth functions and turn to *distribution* theory, where functions are replaced by distributions (also called generalized functions) and partial derivatives in the usual sense are replaced by *weak derivatives*. Luckily, while this transition represents a big leap conceptually speaking, most of the operations and properties that hold in the *classical* sense happen to also hold in the *weak* sense. To that extent, we keep exactly the same notations and define the Sobolev space

$$H^1(\Omega) \coloneqq \left\{ u \in L^2(\Omega) \text{ such that } \frac{\partial u}{\partial x_i} \in L^2(\Omega)\,,\ \forall i \in \{1, \dots, n\} \right\}. \tag{2.21}$$

In this definition, u is a distribution and the notation "$\partial u / \partial x_i$" now refers to the weak derivative of u with respect to the coordinate x_i. This space can be equipped with a new inner product that reads

$$\langle u, v \rangle_{H^1} \coloneqq \int_\Omega u\, v\, \mathrm{d}\mathbf{x} + \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v\, \mathrm{d}\mathbf{x}\,, \tag{2.22}$$

where again, "$\boldsymbol{\nabla}$" is to be understood in the weak sense. One can show that $H^1(\Omega)$ is a Hilbert space, and now $a(u, v)$ in Eq. (2.13) is well-defined for any $u, v \in H^1(\Omega)$. The norm associated with the inner product Eq. (2.22) can be decomposed as $\|u\|_{H^1}^2 = \|u\|_{L^2}^2 + |u|_{H^1}^2$, where $|\cdot|_{H^1}$ defines a semi-norm in $H^1(\Omega)$.[6]

Remark 2.1. The definition of Sobolev spaces can be generalized to any order $m \in \mathbb{N}$:

$$H^m(\Omega) \coloneqq \left\{ u \in L^2(\Omega) \text{ such that } D^\alpha u \in L^2(\Omega)\,,\ \forall |\alpha| \leq m \right\}, \tag{2.23}$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ is used to denote a multi-index and $D^\alpha u = (\partial^{\alpha_1} / \partial x_1^{\alpha_1}) \dots (\partial^{\alpha_n} / \partial x_n^{\alpha_n}) u$ (in the weak sense).

We are getting closer to our objective, that is to check that Lax–Milgram theorem applies to our model problem. Before tackling the theorem's hypotheses, there remains to update our somewhat vague notion of boundary conditions. The problem being that the border Γ has a zero Lebesgue measure in $\mathbb{R}^n$. Therefore, the condition "$u = u_D$ on Γ_D" makes no sense whatsoever if u is sought in $L^2(\Omega)$. Fortunately, there is a way around this issue when u belongs to $H^1(\Omega)$. To see it, let us consider the trace operator

$$\begin{aligned} \gamma_0 \colon \mathcal{C}^1(\bar{\Omega}) &\to \mathcal{C}^0(\Gamma) \quad \text{such that} \quad \forall \mathbf{x} \in \Gamma\,,\ (\gamma_0 v)(\mathbf{x}) = v(\mathbf{x})\,. \\ v &\mapsto \gamma_0 v \end{aligned} \tag{2.24}$$

Eq. (2.24) defines a linear application from $\mathcal{C}^1(\bar{\Omega})$ equipped with the $H^1(\Omega)$-norm to $\mathcal{C}^0(\Gamma)$ equipped with the $L^2(\Gamma)$-norm. Provided that the boundary Γ is *sufficiently regular*,[7] the density of $\mathcal{C}^1(\bar{\Omega})$ in $H^1(\Omega)$ allows one to extend γ_0 to a bounded linear operator from $H^1(\Omega)$ to $L^2(\Gamma)$ — this is the *trace theorem* — so that

$$\text{there exists a constant } C > 0 \text{ such that } \forall v \in H^1(\Omega)\,,\ \|\gamma_0 v\|_{L^2(\Gamma)} \leq C\, \|v\|_{H^1}\,. \tag{2.25}$$

This concept of trace operator can be adapted to our subset $\Gamma_D \subset \Gamma$ (we still denote it γ_0) and allows us to define

$$H_0^1(\Omega) \coloneqq \left\{ u \in H^1(\Omega) \text{ such that } \gamma_0 u = 0 \text{ on } \Gamma_D \right\}, \tag{2.26}$$

which is a Hilbert space when equipped with the norm inherited from $H^1(\Omega)$. Moreover, the homogeneous Dirichlet boundary condition is now easily satisfied just by requiring the unknown generalized function u to itself belong to $H_0^1(\Omega)$. As for the Neumann boundary term in Eq. (2.13), it is enough to demand that g_N be in

[5] Note that $L^2(\Omega)$ plays a special role among the family of $L^p(\Omega)$ Banach spaces, $p \in [1, \infty]$, as it is the only one to be a Hilbert space.

[6] $|\cdot|_{H^1}$ is not a norm because it is not positive definite. Indeed, $|u|_{H^1}$ implies $\boldsymbol{\nabla} u = 0$ in $L^2(\Omega)$ but not $u = 0$.

[7] Here, Ω needs to be at least a Lipschitz domain.

$L^2(\Gamma_N)$ since then

$$\int_{\Gamma_N} g_N \, \gamma_0 v \, \mathrm{d}\gamma \leq \|g_N\|_{L^2(\Gamma_N)} \, \|v\|_{L^2(\Gamma_N)} < +\infty \,, \ \forall v \in H^1_0(\Omega) \,. \tag{2.27}$$

We now have all the tools in our hand to make use of the Lax–Milgram theorem recalled in Box E. We have found a good candidate for the role of Hilbert space, namely $H^1_0(\Omega)$, and we are left to review the theorem's three hypotheses. Hypotheses 1 and 3 [Eqs. (2.14) and (2.16) respectively] are easily checked by applying Cauchy–Schwarz inequalities in $L^2(\Omega)$ and $L^2(\Gamma_N)$ as seen before, together with the trivial inequality $\|u\|_{L^2} \leq \|u\|_{H^1}$ and the (less trivial) trace inequality Eq. (2.25) respectively. The point requiring most attention is the second one, namely the coerciveness of the bilinear form $a(\cdot, \cdot)$ defined by Eq. (2.15). Poincaré's inequality, reported in Box F below, comes to help in proving this last point. Indeed, it provides us with a constant $C > 0$ such that for all $u \in H^1_0(\Omega)$,

$$a(u,u) = |u|^2_{H^1} \geq \frac{1}{2}\left(\frac{1}{C^2}\|u\|^2_{L^2} + |u|^2_{H^1}\right) \geq \frac{1}{2}\min\left(\frac{1}{C^2}, 1\right) \|u\|^2_{H^1} \,. \tag{2.28}$$

Box F: Poincaré's inequality

Let Ω be an open connected bounded subset of $\mathbb{R}^n$ with a Lipschitz boundary and $\Gamma_D \subset \Gamma =: \partial\Omega$ with non-vanishing Lebesgue measure. There exists a constant $C > 0$ such that

$$\text{for all } u \in H^1_0(\Omega) \,, \quad \|u\|_{L^2} \leq C \, |u|_{H^1} \,. \tag{2.29}$$

All assumptions of the Lax–Milgram theorem hold, and so the weak formulation

$$\text{Find } u \in H^1_0(\Omega) \text{ such that for all } v \in H^1_0(\Omega) \,, \ a(u,v) = l(v) \tag{2.30}$$

has a unique solution, whose solution depends continuously on the problem's data encapsulated in $l(\cdot)$. Eq. (2.30) thus constitutes a well-posed problem.

While this proof sketch was dealing specifically with the Poisson equation, several results set out above can be readily re-applied to other elliptic PDEs, e.g. Laplace equation $\Delta u = 0$ or linear Klein–Gordon equation $-\Delta u - \lambda u = 0$ with $\lambda < 0$. When the coefficient matrix $\mathbf{C}$ does not boil down to the identy, additional conditions are to be met. In order to preserve the coerciveness of the bilinear form $a(\cdot, \cdot)$, one usually requires the operator $\mathcal{L}$ to be uniformly elliptic, i.e. there exists a constant $\theta > 0$ such that $\mathbf{C}(\mathbf{x})$ is positive definite with smallest eigenvalue greater than or equal to θ, almost everywhere in Ω. Also, for the integrals to continue to exist without having to change the functional framework, it is natural to demand that the coefficients of the PDE be bounded, i.e. $c_{ij}, b_i, d \in L^\infty(\Omega)$.

Dealing with non-homogeneous boundary conditions

Now what happens if we relax the assumption that $u_D \equiv 0$? It is tempting to slightly adapt the definition of $H^1_0(\Omega)$ as the space of all generalized function u belonging to $H^1(\Omega)$ such that $\gamma_0 u = u_D$ on Γ_D. Unfortunately, such a space is not a vector space and so the whole development carried out in the above fails. One breaks this deadlock by decomposing the unknown into $u = u_0 + \tilde{u}_D$ where $u_0 \in H^1_0(\Omega)$ and $\tilde{u}_D \in H^1(\Omega)$ is constructed so that $\gamma_0 \tilde{u}_D = u_D$ on Γ_D. This way, u_0 satisfies Eq. (2.30) where the rhs f is replaced by $\tilde{f} = f - \mathcal{L}\tilde{u}_D$ in the linear form $l(\cdot)$. Back to the terminology, we see that Neumann boundary conditions appear directly in the linear form $l(\cdot)$: they do not have to be imposed by hand which is why they are sometimes described as *natural*. On the other hand, Dirichlet boundary conditions are *essential* in the sense that they are enforced as a constraint on the function space.[8]

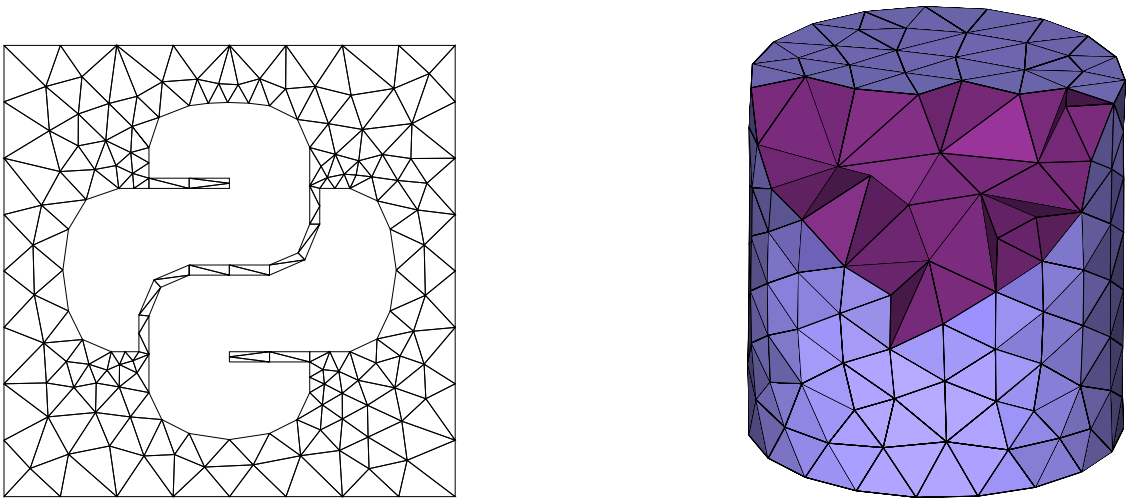
2.1.3 The Finite Element approximation

As pointed out before, the weak formulation Eq. (2.30), despite being well-posed, is an infinite-dimensional problem. The whole point of the following is to turn it into a finite-dimensional problem, so that it can be solved numerically on a finite-memory machine.

Discrete weak formulation

Discretizing the weak form is straightforward enough. The idea consists in approximating the infinite dimensional space V (in which we look for the solution) by a smaller, finite dimensional space V^h. Let $(w_i)_{1 \leq i \leq N}$ be a basis

[8] Additionally, note that Poincaré's inequality (Box F) does not hold if no Dirichlet conditions are prescribed on the boundary.



(a) 2D mesh (b) 3D mesh

Figure 2.2: Illustrations of meshes. The 2D mesh (left panel) showcases the power of mesh refinement to approximate complex shapes. The 3D cylinder mesh (right panel) has been purposely clipped to make tetrahedral elements apparent (shown in purple) while facet elements are depicted in blue.

of V^h, where N represent its dimension. Then, any function $u^h \in V^h$ can be decomposed equivocally as

$$u^h = \sum_{i=1}^{N} U_i \, w_i \quad \text{with} \quad \mathbf{U} = (U_1, \dots, U_N)^T \in \mathbb{R}^N \tag{2.31}$$

As a consequence, testing against all $v \in V$ in the weak form Eq. (2.30) is equivalent to testing against all basis functions of V^h. Using the bilinearity of $a(\cdot, \cdot)$ and leaving Dirichlet boundary conditions aside for the moment, the discrete weak formulation reads

$$\text{Find } \mathbf{U} \in \mathbb{R}^N \text{ such that for all } i \in \{1, \dots, N\}, \; \sum_{j=1}^{N} U_j \, a(w_j, w_i) = l(w_i) \,, \tag{2.32}$$

which is nothing but a linear system of unknown $\mathbf{U}$, with matrix $\mathbf{A} = (a(w_j, w_i))_{1 \le i,j \le N}$ (often called the *stiffness matrix*) and rhs vector $\mathbf{L} = (l(w_i))_{1 \le i \le N}$ (also known as the *load vector*).

Eq. (2.32) is the discrete counterpart of the weak form. If the latter is well-posed, the linear system $\mathbf{AU} = \mathbf{L}$ is invertible on the sole condition that $V^h \subset V$ — the coercivity of $a(\cdot, \cdot)$ implying that the stiffness matrix is positive definite.

Mesh and $\mathbb{P}_k$ elements

The remaining ingredient of FEM is the mesh which serves as the basis for the definition of a finite dimensional space $V^h \subset V$. As we are going to see, there is in fact a tight interplay between the discretization of space [geometry] on the one hand, and the discretization of the function space [analysis] on the other hand.

A mesh of Ω is a tessellation composed of simple cells, such as triangles in 2D or tetrahedra in 3D.[9] We denote by $\mathcal{T}^h$ such a collection of cells. All cells do not have to be the same shape or size, which means we can use them to approximate virtually any given geometry — see Fig. 2.2. The boundary Γ ends up being approximated by polytopes[10] of dimension stricly less than n. The resulting elements of dimension $n-1$ (exactly) are called *facet* elements and denoted by F. The collection of all facet elements is referred to as Σ^h.

For the discretization of the function space, we provide $\mathcal{T}^h$ with N *degrees of freedom*, each of which being associated with a basis function w_i introduced above. In order to make this somewhat abstract notion clearer, let us consider a simple 2D triangular mesh (as in Fig. 2.2a) where the triangles' vertices are numbered from 1 to N. These vertices (which are of course shared between several triangles) are going to act as anchors for the basis functions. A common choice is to demand that *(i)* basis functions are first-order polynomials on each triangle (i.e. in a piecewise manner), and *(ii)* for all $i \in \{1, \dots, N\}$, w_i is equal to one at the i^{th} vertex and equal to zero at all other vertices. These two conditions, which can be interpreted as Lagrangian interpolation

[9] Other types of elements exist, e.g. quadrangle elements in 2D, hexahedron, pentahedron or pyramid elements in 3D. Complex meshes can even combine several types of elements.

[10] A polytope is the generalization of the notion of two-dimensional polygons to arbitrary dimensions.

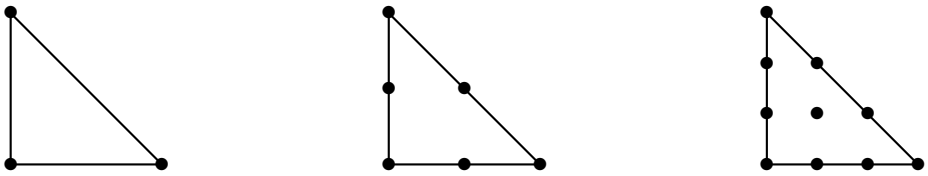


Figure 2.3: Illustration of $\mathbb{P}_k$ triangles for $k = 1, 2, 3$. The bullet points represent the dofs location for each type. There are as many as polynomial coefficients in 2D, namely 3, 6, 10 for polynomials of degree 1, 2, 3 respectively.

basically, are enough for $(w_i)_{1\leq i\leq N}$ to define a basis of the space

$$V^h = \left\{ w \in \mathcal{C}^0\big(\mathcal{T}^h\big) \text{ such that } w \in \mathbb{P}_1(K)\,,\ \forall K \in \mathcal{T}^h \right\}. \tag{2.33}$$

In Eq. (2.33), $\mathbb{P}_1(K)$ denotes the space of all polynomials of degree 1 on the element K of the mesh. This illustrates the notion of first-order Lagrange elements on a triangular mesh, or $\mathbb{P}_1$ elements for short.

However, the exact solution to our PDE problem is a priori more regular than piecewise polynomial functions of degree 1. If so, the *distance* from the exact solution to the subspace V^h defined by Eq. (2.33) might be large. A way to remedy this issue to some extent is to employ higher-degree polynomials. In the same fashion as before, we ask our basis functions w_i to be k^{th}-order polynomials on each triangles, be equal to one for exactly one DOF and be equal to zero for all the others. For this procedure to be well-defined, we need to add new physical DOFs on the mesh elements. Indeed, higher-degree polynomials having more coefficients, their value has to be specified on more points for them to be completely determined. Fig. 2.3 helps visualizing how such additional DOFs are distributed on a reference triangle in practice for $\mathbb{P}_k$ elements, $k \in \{1, 2, 3\}$. We can thus generalize the definition of V^h to

$$V^h = \left\{ w \in \mathcal{C}^0\big(\mathcal{T}^h\big) \text{ such that } w \in \mathbb{P}_k(K)\,,\ \forall K \in \mathcal{T}^h \right\}, \text{ for some } k \in \mathbb{N}^*\,. \tag{2.34}$$

Finally, one can show that the finite-dimensional space V^h as defined in Eq. (2.34) is indeed a subset of the infinite-dimensional space V. As mentioned previously, this ensures the well-posedness of the arising linear system.

Linear system assembly and solution

As highlighted by Eq. (2.32), the computation of stiffness matrix $\mathbf{A}$ and rhs vector $\mathbf{L}$ involves evaluating the quantities $a(w_j, w_i)$ and $l(w_i)$, for all $(i,\, j) \in [\![1, N]\!]^2$. In the case of the Poisson equation, they read

$$\begin{aligned} a(w_j, w_i) &= \int_\Omega \boldsymbol{\nabla} w_j \cdot \boldsymbol{\nabla} w_i \,\mathrm{d}\mathbf{x} & l(w_i) &= \int_\Omega f w_i \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} g_N w_i \,\mathrm{d}\gamma \\ &= \sum_{K\in\mathcal{T}^h} \int_K \boldsymbol{\nabla} w_j \cdot \boldsymbol{\nabla} w_i \,\mathrm{d}\mathbf{x}\,, & &= \sum_{K\in\mathcal{T}^h} \int_K f w_i \,\mathrm{d}\mathbf{x} + \sum_{F\in\Sigma^h} \int_F g_N w_i \,\mathrm{d}\gamma\,. \end{aligned} \tag{2.35}$$

These integrals over cell and facet elements can be evaluated through the use of suitable *quadrature rules*. Schematically, the integral of some function g over some cell element $K \in \mathcal{T}^h$ is computed as

$$\int_K g(\mathbf{x}) \,\mathrm{d}\mathbf{x} \simeq |K| \sum_{l=1}^{N_q} \omega_l \, g(\mathbf{x}_l)\,, \tag{2.36}$$

where $|K|$ denotes the surface/volume of K, $(\mathbf{x}_l)_{1\leq l\leq N_q}$ are the N_q quadrature points and $(\omega_l)_{1\leq l\leq N_q}$ are weights. In particular, the Gauss quadrature rule is implemented in most FEM codes as it allows for the exact integration of polynomials.[11]

This assembly step results in a linear system $\mathbf{AU} = \mathbf{L}$ that is *sparse* thanks to the small support of basis functions w_i. The solving stage — which dominates the time-complexity budget of FEM — is generally carried out by direct solvers (e.g. *LU* and Cholesky factorizations) for relatively small problems ($\sim$ less than one million DOFs) and by iterative solvers (e.g. conjugate gradient method) for the larger ones.

[11] More precisely, an m-point Gauss quadrature rule will exactly integrate a polynomial of degree $2m-1$ [206].

Dealing with Dirichlet boundary conditions

So far, we have not discussed the numerical implementation of Dirichlet boundary conditions — in fact, they do not appear in the definition of V^h in Eq. (2.34). Yet in Sec. 2.1.2, we saw that they played a key role in making the weak form problem well-posed. Accounting for them at the level of the discrete problem is equally important as, more often than not, they guarantee the invertibility of the matrix $\mathbf{A}$. The fact that the solution vector $\mathbf{U}$ is supposedly known at some DOFs means that the corresponding w_i functions should not have been taken into account at the assembly stage. Here we present two techniques for dealing with non-homogeneous Dirichlet boundary conditions. Denoting $(\mathbf{x}_i)_{1\leq i\leq N}$ the coordinates of all DOFs, we need to introduce the vector

$$\mathbf{U}_D \in \mathbb{R}^N \text{ such that } (U_D)_i = \begin{cases} u_D(\mathbf{x}_i) & \text{if the } i^{\text{th}} \text{ DOF belongs to } \Sigma_D^h \\ 0 & \text{otherwise} \end{cases}, \tag{2.37}$$

where Σ_D^h refers to the part of Σ^h where Dirichlet boundary conditions are applied. For convenience, we denote by I_D the index-set of DOFs belonging to Σ_D^h.

The first method is described in Algorithm 1. First, the rhs vector is modified in a similar fashion as we did with the continuous weak form in Sec. 2.1.2 where $l(v) \leftarrow l(v) - a(\tilde{u}_D, v)$ in order to account for non-homogeneous boundary conditions. Then, entries of $\mathbf{A}$ and $\mathbf{L}$ associated with fixed DOFs are set *by hand* in order to account for the fact the corresponding entries of the solution vector $\mathbf{U}$ are already known.

Algorithm 1 A first implementation of non-homogeneous Dirichlet boundary conditions

```
Assemble matrix A and rhs vector L
L ← L − A U_D
for i ∈ I_D do
    for j ∈ ⟦1, N⟧ do          ▷ Matrix entries corresponding to fixed DOFs are set to zero
        A_ij ← 0
        A_ji ← 0
    end for
    A_ii ← 1                   ▷ Diagonal coefficients are set to one to avoid matrix singularity
    L_i ← (U_D)_i
end for
Obtain U by solving the linear system AU = L
```

The method described above involves solving a linear system that is larger than necessary. This is because the value of the solution vector is already known for all entries $i \in I_D$, since it is imposed by the essential boundary condition. Therefore from a *performance* point of view (linear system solving, memory space), it may be worthwhile to keep only the "true" unknowns of the system, even if the management of indices is a little more cumbersome. Let $N_r = N - \text{card}(I_D)$ be the reduced number of active DOFs. One can construct a matrix $\mathbf{T}$ of size $(N \times N_r)$ — with zeros and ones only — that allows to go from a reduced size unknown vector to the full unknown vector. Precisely,

$$\mathbf{U} = \mathbf{T}\,\mathbf{U}_r + \mathbf{U}_D\,, \tag{2.38}$$

where $\mathbf{U}_r \in \mathbb{R}^{N_r}$ is called the reduced unknown vector. Substituting $\mathbf{U}$ by its expression in Eq. (2.38) and left multiplying by $\mathbf{T}^T$ yields the reduced square linear system of size N_r

$$\mathbf{T}^T\mathbf{A}\mathbf{T}\,\mathbf{U}_r = \mathbf{T}^T\left(\mathbf{L} - \mathbf{A}\mathbf{U}_D\right). \tag{2.39}$$

This second method is summarized by Algorithm 2.

Algorithm 2 A second implementation of non-homogeneous Dirichlet boundary conditions

```
Assemble matrix A and rhs vector L
Assemble the T matrix defined by Eq. (2.38)
Compute matrix Ã = T^T A T and vector L̃ = T^T (L − A U_D)
Solve linear system Ã U_r = L̃
Reconstruct full solution vector U with Eq. (2.38)
```

Error estimation

So far, we pretended that is was normal for the finite element approximation to get closer to the actual PDE solution when *(i)* decreasing the typical size of cells (i.e. going from a coarse mesh to a fine mesh) and *(ii)*

increasing the order of polynomial approximation. These statements, put in those terms, are not very precise, nor are they obvious. In the following development, we consciously leave aside all the problems that have to do with numerical calculations, e.g. badly conditioned stiffness matrix, floating-point errors, inexact quadrature rules for the evaluation of integrals, etc. In plain language, all errors inherent in machine computation are assumed to be zero.

We start with a clarification of the notion of error. For that, we need to introduce the Π operator, which acts on continuous functions on the mesh $\mathcal{T}^h$ as

$$\begin{aligned} \Pi\colon \mathcal{C}^0(\mathcal{T}^h) &\to V^h \\ u &\mapsto \Pi u = \sum_{i=1}^{N} u(\mathbf{x}_i)\, w_i \,, \end{aligned} \tag{2.40}$$

where $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ corresponds to DOFs coordinates, and Πu is called the interpolation of u into the finite element space V^h. Now, if $u \in \mathcal{C}^2(\Omega)$ is taken to be the solution of the strong form PDE, and $u^h \in V^h$ the approximation obtained via FEM, u and u^h are solutions to the variational problems

$$\forall v \in V \,,\ a(u,v) = l(v) \qquad \text{and} \qquad \forall v^h \in V^h \,,\ a(u^h, v^h) = l(v^h) \,. \tag{2.41}$$

Letting $\alpha > 0$ be a coercivity constant and $M > 0$ a continuity constant of the bilinear form $a(\cdot\,,\cdot)$, we get, with the help of Eq. (2.41), that for any $v^h \in V^h$

$$\alpha \|u - u^h\|_V^2 \leq a(u-u^h, u-u^h) = a(u-u^h, u) = a(u-u^h, u-v^h) \leq M\, \|u-u^h\|_V\, \|u-v^h\|_V \,, \tag{2.42}$$

where we have used the a-orthogonality of $u - u^h$ with respect to V^h. In other words,

$$\|u-u^h\|_V \leq \frac{M}{\alpha} \inf_{v^h \in V^h} \|u - v^h\|_V \leq \frac{M}{\alpha} \|u - \Pi u\|_V \,. \tag{2.43}$$

In this expression $\|u-u^h\|_V$ represents the finite element error and $\|u-\Pi u\|_V$ represents the interpolation error, both in V-norm. This result is known as Céa's lemma [207], and means that the accuracy of the finite element solution is primarily determined by how well the finite element space can approximate the exact solution within that space (note that the constants α and M only depend on properties of the bilinear form, not on the sub-space V^h).

Céa's lemma paves the way for further obtaining *a priori*[12] error estimates, as the task now boils down to knowing how the interpolation error relates to the actual solution u and the finite element space parameters. We consider only two "macro" parameters:

1. $k \in \mathbb{N}^*$ the polynomial approximation order appearing in Eq. (2.34).

2. $h > 0$, which reflects the degree of fineness of the mesh, often called the mesh size. To be more precise, we first define
$$h_K \coloneqq \operatorname{diam}(K) = \sup\big\{\|\mathbf{x}-\mathbf{y}\|_2 \,,\ \mathbf{x},\mathbf{y} \in K\big\} \quad \text{and then} \quad h \coloneqq \max_{K \in \mathcal{T}^h} h_k \,. \tag{2.44}$$

From there, a lot of useful inequalities can be derived, with varying assumptions and varying norms. In fact, if the best approximation error goes to zero, then so does the finite element error (at the same rate) thanks to Céa's lemma. We report on one well-known *a priori* error estimate in Box G. The proof of a more general statement is given in Brenner and Scott's book [208], Chapt. 4. A similar estimate of the error in L^2-norm can be derived using the "Aubin–Nitsche duality argument", under a few additional assumptions. The rule of thumb is that, if $u \in H^{k+1}(\Omega)$, then a Galerkin approximation using degree k Lagrange finite elements converges at $\mathcal{O}(h^k)$ in H^1-norm and at $\mathcal{O}(h^{k+1})$ in L^2-norm.

[12] An *a priori* error estimate, as opposed to an *a posteriori* error estimate, can be obtained without having to actually compute u or u^h.

Box G: An *a priori* error estimate

Let $k \in \mathbb{N}^*$ be the order of the Lagrange elements. Suppose that u is regular enough, i.e. it belongs to the space $H^{k+1}(\Omega)$ defined by Eq. (2.23). Then, there exists a positive constant C, which does not depend on u nor h, such that

$$\|u - u^h\|_{H^1} \leq C\, h^k\, |u|_{H^{k+1}}\,. \tag{2.45}$$

In this expression, $|u|_{H^{k+1}}$ denotes the Sobolev semi-norm, which extends the definition given for $|u|_{H^1}$ to

$$|u|_{H^m} = \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{L^2}^2 \right)^{\frac{1}{2}}\,, \quad \forall m \in \mathbb{N}^*\,. \tag{2.46}$$

2.1.4 Time-dependent problems in FEM

Throughout this PhD work, we assume that the various fields at stake vary slowly with respect to the characteristic time of the phenomena being studied. This legitimates the use of the so-called *quasi-static approximation*. Bluntly speaking, terms of the fields' equation involving derivatives with respect to time are all set to zero, resulting in simpler PDEs. Yet, performing time-dependent simulations of screened scalar fields was at some point envisioned. For the sake of completeness, we here provide a short guide explaining how to take FEM a step further for capturing *dynamical* (i.e. time-dependent) physical effects and review the most mainstream approaches.

It is worth noting that the stationarity assumption generally alters the nature of a given PDE (elliptic, parabolic or hyperbolic, see Sec. 2.1.1). In the case of wave-like PDEs, such as the Klein–Gordon equation (1.45) governing the scalar field in scalar-tensor models, setting partial derivatives involving time to zero changes the nature of the PDE from hyperbolic to elliptic. Moreover, regardless of the chosen approach, the unknown is now time-dependent $u = u(\mathbf{x}, t)$. That means new extra "boundary conditions" have to be provided for the time component, namely *initial conditions*.

The model problem we consider here is a linear wave-equation, reading

$$\frac{\partial^2 u}{\partial t^2} - c^2 \Delta u = f \qquad \text{in } \Omega \times [0,\, T]\,, \tag{2.47a}$$

$$u(\mathbf{x}, t) = u_D(\mathbf{x}, t) \quad \text{on } \Gamma_D \times [0,\, T]\,, \tag{2.47b}$$

$$\frac{\partial u}{\partial n}(\mathbf{x}, t) = g_N(\mathbf{x}, t) \quad \text{on } \Gamma_N \times [0,\, T]\,, \tag{2.47c}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}) \qquad \text{in } \Omega\,, \tag{2.47d}$$

$$\frac{\partial u}{\partial t}(\mathbf{x}, 0) = v_0(\mathbf{x}) \qquad \text{in } \Omega\,, \tag{2.47e}$$

for some time $T > 0$. In the above, Eqs. (2.47b–2.47c) represent the boundary conditions (Dirichlet and Neumann respectively) while Eqs. (2.47d–2.47e) correspond to the initial conditions. The latter are twofold since the PDE (2.47a) is second-order in time. In the following, we present two approaches for numerically solving problem (2.47). In Eq. (2.47a), c need not necessarily be interpreted as the speed of light.

Finite differences for time, finite elements for space

The first approach, probably the most commonly used one, consists in combining the finite element method for space discretization on the one hand, and the finite difference method (FDM) for time discretization on the other hand. Because the discretization is twofold, notations need to be made slightly more precise:

- N_s is used to denote the total number of DOFs, h refers to the mesh size (see Eq. 2.44) and finite element matrices/vectors are labeled with subscripts i, j.
- N_t denotes the total number of discrete time steps, $(t_n)_{0 \leq n \leq N_t}$ is the sequence of points in $[0,\, T]$ such that $t_n = n\Delta t$ with $\Delta t \coloneqq T/N_t$.

From the outset, one seemingly inconsequential question arises: should we *(i)* discretize time and then space, or conversely *(ii)* discretize space first, and then time? Option *(i)* — referred to as Rothe method — leads to a stationary PDE at each time step, that is then solved via FEM. With option *(ii)* — called the method of lines —

we get a system of N coupled ordinary differential equations (ODEs) which can be solved with a relevant finite difference scheme.

Actually, the difference between the two techniques is quite subtle. Indeed, one can show that they lead to the same fully-discretized (time and space) set of equations, provided that the mesh remains the same across all time steps. With the method of lines, we construct a mesh $\mathcal{T}^h$ once and for all, and spatially discretize the unknown as

$$u^h(\mathbf{x},t) = \sum_{i=1}^{N} U_i(t)\, w_i(\mathbf{x}) \;, \quad \forall (\mathbf{x},t) \in \mathcal{T}^h \times [0,\, T] \,. \tag{2.48}$$

Pluging the decomposition Eq. (2.48) into the wave equation Eq. (2.47a) yields the N-dimensional ODE

$$\mathbf{M}\,\ddot{\mathbf{U}}(t) + \mathbf{A}\,\mathbf{U}(t) = \mathbf{L}(t) \,, \tag{2.49}$$

where dots refer to time derivatives. This intermediate expression is called a semi-discretized equation owing to the fact that space has been made discrete while time still flows continuously. Then, the second-order time derivative appearing in Eq. (2.49) is approximated through traditional finite differences. Cutting short this discussion, we lay emphasis on the fact that this method is not particularly suited to our potential future needs. The reason being that, with the mesh $\mathcal{T}^h$ set in concrete, it is not possible to dynamically keep track of moving parts in the simulation, nor is it possible to adapt the mesh refinement to the region of interest (which is likely to travel across the numerical domain). In contrast, Rothe method can accommodate these features which is why it is given the focus in the following.

We start with an optional step, that consists in re-writing problem (2.47) as two first-order-in-time PDEs. This fairly common trick has the effect of reducing the number of time steps to be kept in computer's memory, and brings us back to the 'canonical' framework of first-order ODEs. Setting $v \coloneqq \partial_t u$, we get the set of equations

$$\begin{aligned}
\partial_t v - c^2 \Delta u &= f && \text{(2.50a)} & \qquad \partial_t u &= v && \text{(2.50e)} & \quad &\text{in } \Omega \times [0,\, T] \\
u(\mathbf{x},t) &= u_D(\mathbf{x},t) && \text{(2.50b)} & \qquad v(\mathbf{x},t) &= \partial_t u_D(\mathbf{x},t) && \text{(2.50f)} & \quad &\text{on } \Gamma_D \times [0,\, T] \\
\partial_n u(\mathbf{x},t) &= g_N(\mathbf{x},t) && \text{(2.50c)} & \qquad \partial_n v(\mathbf{x},t) &= \partial_t g_N(\mathbf{x},t) && \text{(2.50g)} & \quad &\text{on } \Gamma_N \times [0,\, T] \\
u(\mathbf{x},0) &= u_0(\mathbf{x}) && \text{(2.50d)} & \qquad v(\mathbf{x},t) &= v_0(\mathbf{x}) && \text{(2.50h)} & \quad &\text{in } \Omega \,,
\end{aligned}$$

where we have added new boundary conditions (2.50f–2.50g) for the new unknown v. Following Rothe method, we discretize Eqs. (2.50a, 2.50e) with respect to time first, using finite difference approximations. In order to remain as general as possible, we do so by employing the θ-scheme, reading for all $\mathbf{x} \in \Omega$, $n \in [\![0, N_t - 1]\!]$

$$\frac{u^{n+1} - u^n}{\Delta t} = \theta v^{n+1} + (1-\theta) v^n \,, \tag{2.51a}$$

$$\frac{v^{n+1} - v^n}{\Delta t} = c^2 \theta \Delta u^{n+1} + c^2 (1-\theta) \Delta u^n + \theta f^{n+1}(\mathbf{x}) + (1-\theta) f^n(\mathbf{x}) \,. \tag{2.51b}$$

In above, we use the notation $f^n(\mathbf{x}) \equiv f(\mathbf{x}, t_n)$, and the same goes for all other functions. This scheme is explicit only when $\theta = 0$, which corresponds to the forward Euler scheme. It becomes implicit as soon as $\theta > 0$. In particular, $\theta = 1$ corresponds to backward Euler scheme while $\theta = 1/2$ is the well-known Crank-Nicolson scheme. The latter has the advantage of being a second-order method, compared to Euler schemes which constitute first-order methods. We then rearrange the terms in Eqs. (2.51a–2.51b) so as to be able to first determine u^{n+1}, and then v^{n+1}:

$$\left[1 - (\Delta t\, \theta)^2 \Delta\right] u^{n+1} = \left[1 + \Delta t^2\, \theta(1-\theta)\Delta\right] u^n + \Delta t v^n + \theta \Delta t^2 \left[\theta f^{n+1}(\mathbf{x}) + (1-\theta) f^n(\mathbf{x})\right] \,, \tag{2.52a}$$

$$v^{n+1} = v^n + \Delta t \left[c^2 \theta \Delta u^{n+1} + c^2 (1-\theta) \Delta u^n + \theta f^{n+1}(\mathbf{x}) + (1-\theta) f^n(\mathbf{x})\right] \,. \tag{2.52b}$$

From there, we apply the finite element procedure, which has been discussed at length in Sec. 2.1.2 and 2.1.3. Deriving the weak formulations from Eqs. (2.52a–2.52b) does not present any particular difficulty. Things become a little trickier when turning to the finite element discretization. Indeed, we willingly allow the underlying mesh to evolve from one time step to the next. As a result, the basis decomposition Eq. (2.48), which assumed a single mesh for all time steps, is no longer valid. Instead, it should be made time-dependent as well, reading for all $n \in [\![0,\, N_t]\!]$

$$\left(u^n\right)^h (\mathbf{x}) = \sum_{i=1}^{N_s^n} U_i^n\, w_i^n(\mathbf{x}) \;, \quad \forall \mathbf{x} \in \mathcal{T}_n^h \,, \tag{2.53}$$

where it should be noted that the number of DOFs, the basis functions, and the tessellation are indexed by n as they are now time step-dependent. We can already see the issues ahead: the assembly of some FE terms is going

to involve the computation of integrals of the form

$$\int_\Omega (v^n)^h(\mathbf{x})\, w_j^{n+1}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{i=1}^{N_s^n} V_i^n \int_\Omega w_i^n(\mathbf{x})\, w_j^{n+1}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,. \tag{2.54}$$

The problem with these type of integrals is that w_i^n and w_j^{n+1} are not defined on the same mesh, making such computations quite messy in appearance. One way possible around this awkwardness would be to pre-evaluate known terms (obtained at the n^{th} step) at the quadrature points of each cell element of $\mathcal{T}_{n+1}^h$ (see Eq. 2.36) — which is doable in practice without too much hassle.

The only point we have not touched upon so far is the *stability* of the method. One can show that explicit methods will typically exhibit a CFL[13] stability condition of the form $\Delta t < (2h)/(c\pi)$ — h being the mesh size [see Eq. (2.44)], while implicit methods are generally unconditionally stable (for any step size). Either way, *accuracy* cannot be expected unless $\Delta t < h/c$. Combining these two arguments, it is reasonable in that case to use explicit time stepping methods, as checking the accuracy condition matches the stability limit. This is important matter in terms of computational cost since, unlike for implicit methods, explicit methods do not require solving a linear system at each time step (set $\theta = 0$ in Eqs. 2.52a, 2.52b), making them a cheaper alternative.

To sum up, the method of lines with explicit time stepping is probably a good place to start implementing a time-dependent solver. If, however, the ultimate goal is to be able to perform a dynamical simulation with moving parts (e.g. time-dependent density distribution $\rho(\mathbf{x}, t)$, with back reaction from the scalar force), Rothe method might be more suitable and worth the extra implementation-effort. An illustration of this method is given in Fig. 2.4a. Moreover, the model problem we considered in Eq. (2.47) is that of a linear wave-equation. In Sec. (2.2.4), we provide a short extra note on how to deal with the time-dependent nonlinear Klein–Gordon equation.

Spacetime FEM

As highlighted in Sec. 2.1.1, we could in principle ask some component x_i of the $\mathbf{x}$ vector to represent time, and the discussion would remain the same. This naive observation underlies the slightly more exotic *spacetime finite element method*. It treats space and time as a unified domain, allowing for the discretization of both dimensions using finite element basis functions. The idea extending the finite element framework to time-dependent problems by encompassing the time-dimension into the numerical domain Ω was first proposed in Refs. [209–211]. The method has then been successfully applied to second-order hyperbolic PDEs as soon as the early 1990's, see e.g. Ref. [212]. Compared to traditional methods that separate spatial and temporal discretizations, spacetime FEM can offer more efficient and accurate solutions, especially for problems with strong coupling between space and time. Note that problems with three spatial dimensions require the use of 4-dimensional meshes. This added layer of abstraction does not alter the roots of FEM (in the sense that all computation can still be fairly well automated), but data visualization may require extra post-processing tools. An illustration of this method is given in Fig. 2.4b.

2.2 Dealing with nonlinear problems

In this second section, we relax one of the first assumptions we made at the beginning of Sec. 2.1.1, namely that the differential operator $\mathcal{L}$ be linear — the motivation being to be able to solve the nonlinear Klein–Gordon equation Eq. (1.117). There are three main types of nonlinear PDEs:

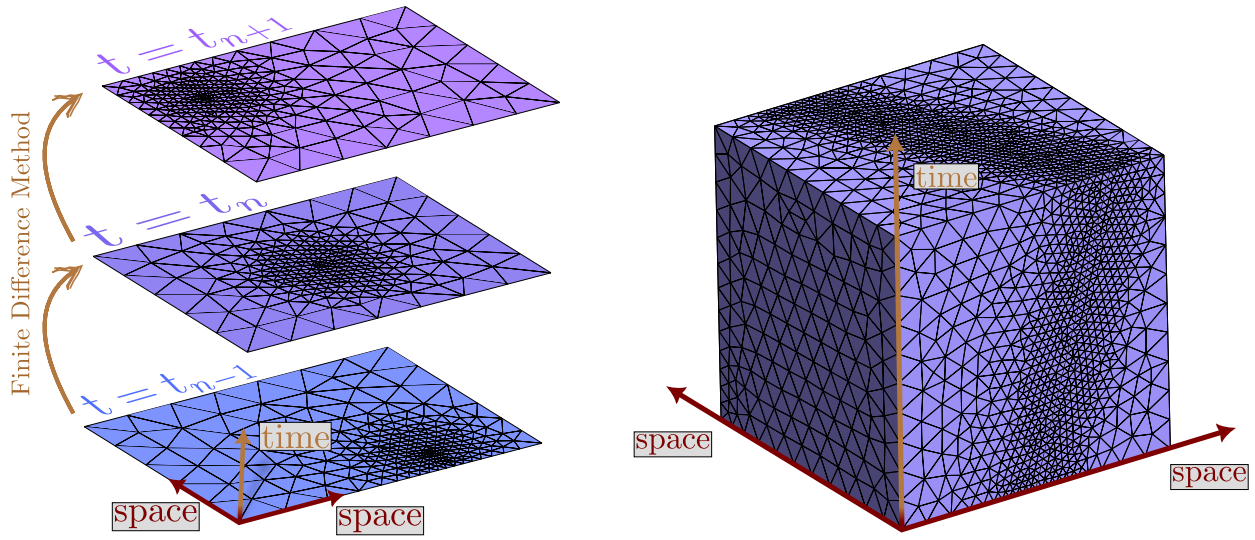
1. *semi-linear* PDEs, for which only the highest order derivatives appear as linear terms. For the specific case of second-order PDEs [Eqs. (2.2–2.3)], that means that the coefficient matrix $\mathbf{C}$ only depends on $\mathbf{x}$ while the other coefficients $\mathbf{b}$ and d are allowed to depend on u and its partial derivatives.

2. *quasi-linear* PDEs, where the coefficients of the highest order derivatives terms are allowed to be functions of lower-order derivatives. In the case of second-order PDEs, this translates to $\mathbf{C} = \mathbf{C}(\mathbf{x}, \{\partial^\alpha u\})$ for all the multi-indices α satisfying $|\alpha| \leq 1$. Einstein field equations Eq. (1.12) are a concrete example of a system of quasi-linear PDEs.

3. *fully nonlinear* if none of the above linearity properties hold.

We see that the Klein–Gordon equation (1.117) falls into the first category of semi-linear PDEs, which is the form closest to linear equations.

[13] Courant-Friedrichs-Lewy



(a) Combining FDM and FEM discretizations

(b) Spacetime Finite Element Method

Figure 2.4: Illustration of the two proposed ways to handle time-dependent PDE problems. In order to facilitate the understanding through visualization, the time dimension (represented by the gold vertical arrow) is supplemented with only two spatial dimensions (represented by the red arrows). On the left-hand-side, (a) depicts the use of FEM for space discretization while FDM is employed for time stepping. Note that this sketch corresponds to the Rothe method as the mesh evolves from one time step to the next — allowing to keep track of some moving region of interest where the mesh needs to be refined (adaptivity). On the right-hand-side, (b) represents a spacetime FEM mesh. Mesh refinement is also allowed, both in space and time directions. While time is still discretized, basis functions make it possible to retrieve the data at any given instant.

In the following, we show how all the FE frameworks laid out in the previous section can be re-invested in the context of nonlinear PDEs. We do not deal with nonlinearities due to boundary conditions, which are another source of nonlinearity.[14]

2.2.1 Iterative techniques

On the surface, the principle of iterative techniques is fairly simple:

1. linearize the PDE around some *initial guess* u_0;
2. solve the resulting linear PDE with the techniques laid out in Sec. 2.1;
3. update the initial guess thanks to the previous solution and go back to step 2 until convergence is met.

While this is a not-so-inaccurate outline of how iterative techniques work, it shadows a great number difficulties. As we are going to see, there is not a single way of linearizing a given PDE. What makes a *good* initial guess and how to choose it accordingly? How do we actually update the solution from one iteration to the next? How do we assess the convergence of the algorithm? How fast does it converge? More fundamentally, how do we know for sure it is going to converge?

In order to address these questions, we decide to proceed as in Sec. 2.1. For the sake of remaining sufficiently general, the main principles are first introduced on a generic quasi-linear PDE with essential boundary conditions whose weak formulation is assumed to take the form[15]

$$\int_\Omega \mathbf{C}(\mathbf{x},\{D^\alpha u\})\boldsymbol{\nabla} u\cdot\boldsymbol{\nabla} v\,\mathrm{d}\mathbf{x} - \int_\Gamma \big[(\mathbf{C}(\mathbf{x},\{D^\alpha u\})\boldsymbol{\nabla} u)\cdot\mathbf{n}\big]v\,\mathrm{d}\gamma + \int_\Omega \mathbf{b}(\mathbf{x},\{D^\alpha u\})\cdot\boldsymbol{\nabla} u\,v\,\mathrm{d}\mathbf{x} + \int_\Omega d(\mathbf{x},\{D^\alpha u\})\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega f(\mathbf{x})v\,\mathrm{d}\mathbf{x}\,. \quad (2.55)$$

[14] Nonlinear boundary condition are common in structural mechanics for problems where the boundary constraints depend explicitely on the deformation state of the system.

[15] The techniques introduced thereafter could also be applied to fully nonlinear PDEs to some extent. Yet, we restrict the general discussion to weak formulations of the form Eq. (2.55) because *(i)* their handling does not require the introduction of additional tools/objects and *(ii)* we do not go beyond semi-linear PDEs in this PhD work.

In this expression, all coefficients are allowed to depend both on the coordinates $\mathbf{x}$, and on the unknown u and all its first-order weak derivatives, shortened to $\{D^\alpha u\}$. It is also worth noting that the integral over Γ cannot be taken to be zero (yet)! Indeed, the *trick* presented at the end of Sec. 2.1.2 to deal with non-homogeneous Dirichlet boundary conditions cannot be applied here due to the nonlinearity of $\mathcal{L}$. The ideas are first exposed in the general case, before being applied to the semi-linear Klein–Gordon equation (1.117) which we reproduce here (in a simplified fashion) for the sake of convenience:

$$\alpha \Delta u = \rho - u^{-(n+1)} \quad \text{on } \Omega \subset \mathbb{R}^3 \,. \tag{2.56}$$

We denote the unknown by u (instead of ϕ) for consistency with the previous section and draw attention to the fact that n now denotes the potential index of the chameleon model, the dimension being set to 3. Following Sec. 2.1.2, we can derive the weak form of Eq. (2.56), for which we have

$$a(u,v) = \alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, \mathrm{d}\mathbf{x} - \alpha \int_\Gamma (\boldsymbol{\nabla} u \cdot \mathbf{n}) v \, \mathrm{d}\gamma - \int_\Omega u^{-(n+1)} v \, \mathrm{d}\mathbf{x} \quad \text{and} \quad l(v) = -\int_\Omega \rho \, v \, \mathrm{d}\mathbf{x} \,. \tag{2.57}$$

The issue with the weak forms (2.55, 2.57) is that they are not linear in u, and so the basis decomposition (2.31) will fail to produce a linear system.

We start by giving a sense to what it means to *linearize* a nonlinear PDE with two distinct procedures, namely Picard iterations and Newton's method. In both cases, our starting point is Eq. (2.57). In other words, the linearization process occurs at the stage of the continuous weak form. This is a choice that we make, as linearizing at earlier stages (e.g. in the strong form PDE) or later stages (e.g. after discretization) may sometimes lead to different problem formulations. Moreover, such procedures are of course not restricted to the FEM framework and can be applied in very diverse contexts.

Picard iterations

Picard's method — which is also known as fixed-point iteration, successive substitution or even nonlinear Richardson iteration — is perhaps the simplest way of linearizing nonlinear PDEs put in the weak form. In nonlinear terms, the unknown u is replaced *by hand* by some already known approximation u_* of u. This procedure is best illustrated on algebraic equations. Consider the simple example of a second-order algebraic equation

$$a u^2 + b u + c = 0 \,. \tag{2.58}$$

Here, the nonlinear term can be approximated through $u^2 \simeq u_* u$, resulting in a linear algebraic equation $a u_* u + b u + c = 0$. From there the algorithm is elementary: solve the linearized equation with respect to u, set $u_* \leftarrow u$ and repeat. This procedure is guaranteed to converge under the assumptions of the Banach fixed-point theorem [213]. Back to PDEs, the very same idea can be applied except that now u lives in an infinite-dimensional space. Eq. (2.55) can be made linear in the unknown by inputting u_* instead of u in all coefficients, yielding

$$\int_\Omega \mathbf{C}(\mathbf{x}, \{D^\alpha u_*\}) \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, \mathrm{d}\mathbf{x} + \int_\Omega \mathbf{b}(\mathbf{x}, \{D^\alpha u_*\}) \cdot \boldsymbol{\nabla} u \, v \, \mathrm{d}\mathbf{x} + \int_\Omega d(\mathbf{x}, \{D^\alpha u_*\}) \, uv \, \mathrm{d}\mathbf{x} = \int_\Omega f(\mathbf{x}) v \, \mathrm{d}\mathbf{x} \,. \tag{2.59}$$

Applying the same treatment to Eq. (2.57) results in

$$\alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, \mathrm{d}\mathbf{x} - \int_\Omega u_*^{-(n+2)} u \, v \, \mathrm{d}\mathbf{x} = -\int_\Omega \rho \, v \, \mathrm{d}\mathbf{x} \,. \tag{2.60}$$

Standard FEM can then be employed to solve Eqs. (2.59–2.60) whatever the known function u_*. It is important to point out that, unlike their nonlinear counterparts, these weak forms are now exempt from boundary terms. In fact, the linearity of Eqs. (2.59–2.60) allows for the use of null test functions v on the boundary, i.e. $v_{|\Gamma} \equiv 0$.

Note that the above derivation of a linearized weak form may not seem quite *algorithmic*, in the sense that other choices could have been made. Indeed, Picard's method can be viewed as solving iteratively a fixed-point problem of the form $u = g(u)$ — where the function g is generally not unique. Back to the algebraic equation (2.58), one can write a fixed-point scheme in either of the following forms

$$u = -\frac{a u^2 + c}{b} \tag{2.61a}$$

$$u = -\frac{c}{a u + b} \tag{2.61b}$$

$$u = \sqrt{-\frac{b u + c}{a}} \tag{2.61c}$$

provided that the coefficients $\{a, b, c\}$ make these expressions well-defined for u in some given interval. The linearization $u^2 \simeq u_* u$ employed above corresponds to the fixed-point scheme given by Eq. (2.61b) while Eq. (2.61a) is equivalent to approximating $u^2 \simeq u_*^2$. By analogy, an alternative way of linearizing the weak

formulation of the Klein–Gordon equation (2.57) is

$$\alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \, \mathrm{d}\mathbf{x} = \int_\Omega u_*^{-(n+1)} v \, \mathrm{d}\mathbf{x} - \int_\Omega \rho \, v \, \mathrm{d}\mathbf{x} \,. \tag{2.62}$$

Looking at the two linearized weak forms (2.60, 2.62), we can see that they correspond to different (bi)linear forms. Most importantly, the coercivity assumption listed in the Lax–Milgram theorem (see Box E) may or may not hold depending on the linearization process. Even worse, the coerciveness of the bilinear form defined by the lhs of Eq. (2.60) may depend on the current iterate u_*. In comparison, the bilinear form defined by the lhs of Eq. (2.62) is coercive for the H^1-norm, under the assumptions of Sec. 2.1 ($\alpha > 0$). And still, we stress that having a well-posed linearized weak formulation at each iteration does not ensure convergence.

For the sake of completeness, let us give the discrete version of Picard iterations. At any given iteration k, the discrete approximation $u_k^h \in V^h$ is defined according to Eq. (2.31), that is

$$u_k^h = \sum_{i=1}^N U_i^k \, w_i \quad \text{with} \quad \mathbf{U}_k = (U_1^k, \dots, U_N^k)^T \,. \tag{2.63}$$

Consequently, weak formulations Eqs. (2.60, 2.62) result in a linear system reading

$$\mathbf{A}\big(u_k^h\big) \, \mathbf{U}_{k+1} = \mathbf{L}\big(u_k^h\big) \quad \text{and so} \quad \mathbf{U}_{k+1} = \mathbf{A}\big(u_k^h\big)^{-1} \mathbf{L}\big(u_k^h\big) \,. \tag{2.64}$$

Recalling that u_k^h only depends on $\mathbf{U}_k \in \mathbb{R}^N$, we recover the classic form of fixed-point iterations in dimension N, that is

$$\mathbf{U}_{k+1} = \mathbf{K}(\mathbf{U}_k) \,. \tag{2.65}$$

A few comments are in order:

- The invertibility of matrix $\mathbf{A}(\mathbf{U}_k)$ follows directly from the coerciveness of $a(\cdot, \cdot)$ assumed here. And yet, one must bear in mind that coercivity is only a sufficient condition for invertibility.

- The final form Eq. (2.65) is the multi-dimensional counterpart of the algebraic fixed-point problem "$u = g(u)$" discussed above.

- Compared to the continuous weak formulation [e.g. Eqs. (2.59, 2.60, 2.62)], the convergence of iterations of the N-dimensional version given by Eq. (2.65) is somewhat easier to assess, at least theoretically. Given some assumptions on the fixed-point map $\mathbf{K}$, one can find several relevant convergence results in Kelley's book [214], Chapt. 4.

Newton iterations

Newton's method is another way of linearizing weak form PDEs. In comparison with Picard iterations, for which we saw that there was not a single way of proceeding, it is somewhat more algorithmic. Here, the foundations of the method are properly laid since it ended up being used for nearly all numerical computations conducted in this PhD work.

Newton's method is a root finding algorithm in numerical analysis. Its starting point consists in writing the nonlinear equation in the residual form "$f(u) = 0$", which is a more general form than the fixed-point iteration "$u = g(u)$" seen above. From there, the equation is linearized by approximating $f(u)$ by its first-order Taylor series expansion around a known guess u_*. As before, it is enlightening to start with the one-dimensional algebraic case. Taking Eq. (2.58) as an example, we have $f(u) = au^2 + bu + c$ and so

$$f(u) = f\left[u_* + (u - u_*)\right] = f(u_*) + (u - u_*) f'(u_*) + o(u - u_*) = u(2au_* + b) - au_*^2 + c + o(u - u_*) \,. \tag{2.66}$$

Then u is chosen so as to cancel this expansion up to first order, i.e.

$$u = u_* - \frac{f(u_*)}{f'(u_*)} = \frac{au_*^2 - c}{2au_* + b} \,, \tag{2.67}$$

The solution given by Eq. (2.67) is used as the new guess $u_* \leftarrow u$ and so on until convergence is met. Note that Eq. (2.67) does not correspond to any of the Picard linearizations attempted in Eq. (2.61).

The very same ideas can be applied to the case of nonlinear weak forms. This requires nonetheless adapting several notions. First of all, f is no longer a real function but a *functional* instead. In the case of the Klein–Gordon

equation, we set for $v \in V$

$$f_v \colon V \to \mathbb{R} \qquad (2.68)$$
$$u \mapsto \alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}x - \alpha \int_\Gamma (\boldsymbol{\nabla} u \cdot \mathbf{n}) v \,\mathrm{d}\gamma - \int_\Omega u^{-(n+1)} v \,\mathrm{d}x + \int_\Omega \rho v \,\mathrm{d}x \,.$$

The recursion formula (2.67) cannot be used directly as "f_v'" is undefined. The appropriate form of differentiation is the Gateaux derivative. Given a direction $z \in V$, it is defined as

$$J_v(u, z) \coloneqq \lim_{\epsilon \to 0} \frac{f_v(u + \epsilon z) - f_v(u)}{\epsilon} \,. \qquad (2.69)$$

Our goal being to cancel $f_v(u)$, we set $\delta u \coloneqq u - u_*$ and write

$$f_v(u) = f_v(u_* + \delta u) = f_v(u_*) + J_v(u_*, \delta u) + \text{higher order terms} = 0 \,. \qquad (2.70)$$

Dropping the higher order terms, we merely end up with $J_v(u_*, \delta u) = -f_v(u_*)$, where $u_* \in V$ is known and the Gateaux derivative is linear in δu. This is of course reminiscent of Eq. (2.67) when rewritten as $f'(u_*)\delta u = -f(u_*)$.

We illustrate this procedure on f_v given by Eq. (2.68), we have

$$f_v(u + \epsilon z) - f_v(u) = \epsilon\alpha \int_\Omega \boldsymbol{\nabla} z \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} - \epsilon\alpha \int_\Gamma (\boldsymbol{\nabla} z \cdot \mathbf{n}) v \,\mathrm{d}\gamma - \int_\Omega \left[(u + \epsilon z)^{-(n+1)} - u^{-(n+1)} \right] v \,\mathrm{d}\mathbf{x} \qquad (2.71)$$
$$= \epsilon\alpha \int_\Omega \boldsymbol{\nabla} z \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} - \epsilon\alpha \int_\Gamma (\boldsymbol{\nabla} z \cdot \mathbf{n}) v \,\mathrm{d}\gamma + \epsilon(n+1) \int_\Omega u^{-(n+2)} z v \,\mathrm{d}\mathbf{x} + o(\epsilon) \,, \qquad (2.72)$$

so that

$$J_v(u, z) = \alpha \int_\Omega \boldsymbol{\nabla} z \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} - \alpha \int_\Gamma (\boldsymbol{\nabla} z \cdot \mathbf{n}) v \,\mathrm{d}\gamma + (n+1) \int_\Omega u^{-(n+2)} z v \,\mathrm{d}\mathbf{x} \,. \qquad (2.73)$$

Therefore, at the k^{th} iteration, $\delta u_k = u_{k+1} - u_k$ satisfies $J_v(u_k, \delta u_k) = -f_v(u_k)$ i.e.

$$\alpha \int_\Omega \boldsymbol{\nabla} u_{k+1} \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + (n+1) \int_\Omega u_k^{-(n+2)} u_{k+1} v \,\mathrm{d}\mathbf{x} = (n+2) \int_\Omega u_k^{-(n+1)} v \,\mathrm{d}\mathbf{x} - \int_\Omega \rho v \,\mathrm{d}\mathbf{x} \,. \qquad (2.74)$$

Let us pause here and make a few remarks:

- Again, the boundary term has been dropped in Eq. (2.74) because the tests functions v are taken to be zero on Γ when solving this sequence of linear problems.
- The linearized weak form (2.74) is different from the previous two weak forms Eqs. (2.60, 2.62) derived in the context of Picard iterations. Note that the Newton linearization produces more terms than Picard does.
- At this stage (continuous linearized weak form), it is equivalent to work the u_{k+1} or δu_k as the unknown,[16] since Eq. (2.74) is linear. Most importantly, this change of variable does not affect the bilinear form defined by the lhs of Eq. (2.74), only its rhs gets modified.
- The coerciveness of the bilinear form is easy to study. From physical insights relating to the chameleon model (see Sec. 1.2.2), we know that *(i)* there exists a constant $\phi_{\max} > 0$ such that for any $k \in \mathbb{N}$ and for all $\mathbf{x} \in \Omega$, $u_k(\mathbf{x}) \leq \phi_{\max}$ and *(ii)*, $\alpha > 0$. Consequently, for all $u \in V$,

$$\alpha \int_\Omega \|\boldsymbol{\nabla} u\|^2 \,\mathrm{d}\mathbf{x} + (n+1) \int_\Omega u_k^{-(n+2)} |u|^2 \,\mathrm{d}\mathbf{x} \geq \min\left(\alpha, (n+1)\phi_{\max}^{-(n+2)}\right) \|u\|_{H^1}^2 \,. \qquad (2.75)$$

  Note that we have not made use of Poincaré's inequality (see Box F), unlike for the Poisson problem where it plays a crucial role in proving the coercivity of the bilinear form. It follows that Eq. (2.75) holds even in the absence of essential boundary conditions on the boundary Γ, which is remarkable.

The next step is the discretization of Eq. (2.74), already discussed at length in Sec. 2.1.3 and when presenting Picard's method, which is why we skip it here. Note however that, regardless of the linearization technique employed, we inevitably end up with terms that depend on u_k^h and terms that do not. Using the subscript 'mod'

[16] In this respect, Ref. [190] uses the wrong terminology by stating that Eq. (2.74) corresponds to the Picard iteration while the same equation but written out with the δu_k variable corresponds to the Newton iteration.

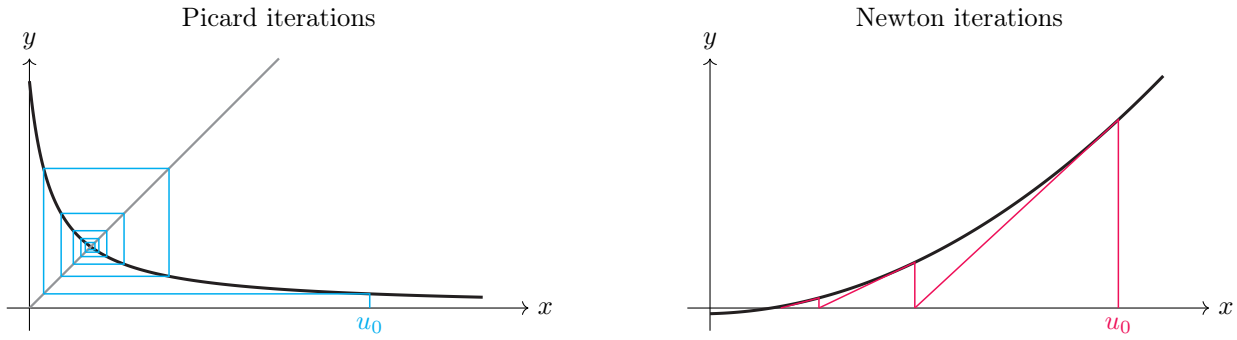


Figure 2.5: Illustration of Picard and Newton methods on the scalar case Eq. (2.58), with $\{a = 10,\, b = 1,\, c = -1\}$. For the Picard iterations, we chose the fixed-point scheme given by Eq. (2.61b). In this example, Newton's iterations converge much faster than Picard's iterations. Note that the two pairs of axes do not share the same scaling.

to denote the former and 'cst' to denote the latter, we have, schematically,

$$\mathbf{A} = \mathbf{A}_{\text{cst}} + \mathbf{A}^k_{\text{mod}} \quad \text{and} \quad \mathbf{L} = \mathbf{L}_{\text{cst}} + \mathbf{L}^k_{\text{mod}}\,. \tag{2.76}$$

An optimized solver would thus only have to re-assemble the terms $\mathbf{A}^k_{\text{mod}}$ and $\mathbf{L}^k_{\text{mod}}$ from one iteration to the next. Again, let us make closing remarks regarding Newton's method:

- After the discretization step, the process boils down to an N-dimensional Newton's method. Its convergence is studied from a theoretical point of view again in Kelley's book [214], Chapt. 4, 5. In practice, convergence issues are commonplace. We review a number of techniques designed to help convergence in Sec. 2.2.3.
- One of the take-home messages from the theory is that, provided the initial guess $\mathbf{U}_0$ is not "too far away from the root", Newton's method converges *quadratically* (that is much faster the Picard iterations, see e.g. Fig. 2.5 for an illustration on the scalar case). A good rule of thumb is that the "less" we linearize f_v, the faster the convergence is likely to be.
- As a side remark, the distinction we draw between Picard and Newton methods is not in line with Refs. [190, 191]. The definitions given in those references thus clashes with ours.

2.2.2 Stopping criteria and inspection of the residual

A numerical iterative technique cannot go but hand in hand with stopping criteria, i.e. conditions to be checked after each iteration to determine whether or not the procedure has converged. We consider roughly two categories of criteria: *(i)* the ones based on *how much* the current iterate changes from one iteration to the next, and *(ii)* the ones based on the residual, which is given a precise meaning thereafter. Unless specified otherwise, we use the notation $\|\cdot\|_2$ to refer to the two-norm in $\mathbb{R}^N$.

Criteria based on the stalling of the evolution of the current iterate $\mathbf{U}_k$ can be either

- *absolute*, e.g. checking the condition $\|\mathbf{U}_k - \mathbf{U}_{k-1}\|_2 \overset{?}{<} \epsilon_{\text{abs}}$, or
- *relative*, e.g. checking the condition $\|\mathbf{U}_k - \mathbf{U}_{k-1}\|_2 \,/\, \|\mathbf{U}_{k-1}\|_2 \overset{?}{<} \epsilon_{\text{rel}}$.

It is even possible to combine them both as $\|\mathbf{U}_k - \mathbf{U}_{k-1}\|_2 \overset{?}{<} \epsilon_{\text{abs}} + \|\mathbf{U}_{k-1}\|_2\, \epsilon_{\text{rel}}$. The tuning of ϵ_{rel}, ϵ_{abs} is left to the user's appreciation (especially ϵ_{abs} as it directly linked to the characteristic scales of the problem at stake, while ϵ_{rel} can be set to a constant value regardless of the underlying problem, e.g. $\epsilon_{\text{rel}} = 10^{-6}$).

These above criteria are a good way to assess the convergence part of the algorithm. What they do not do is provide some sort of feedback on how accurate the current iterate $\mathbf{U}_k$ is. For this, one very important tool is the residual. Any given nonlinear PDE can be cast into the so-called residual form, that is $f(u) = 0$, where $f(u)$ is referred to as the *residual*. The closer the residual $f(u_k)$ is to zero, the better the approximated solution u_k. Here however, it is not possible to perform this test as is, simply because we only have $u^h_k = \sum_i U^k_i\, w_i$ at our disposal, which in general does not have the sufficient regularity to even be inputted in the original strong form PDE.[17] A workaround is to define a discrete residual vector $\mathbf{R}_k = F(\mathbf{U}_k)$ using the non-linearized weak form

[17] In this regard, let us recall that using $\mathbb{P}_2$ Lagrange triangles only result in continuous functions (i.e. not even $\mathcal{C}^1$).

Eq. (2.55):

$$
\begin{aligned}
&F\colon \mathbb{R}^N \to \mathbb{R}^N \\
&\mathbf{U} \mapsto \left(\sum_{i=1}^{N} U_i \left\{ \int_\Omega \Big[\mathbf{C}(\mathbf{x}, \{D^\alpha u^h\}) \boldsymbol{\nabla} w_i \cdot \boldsymbol{\nabla} w_j + (\mathbf{b}(\mathbf{x}, \{D^\alpha u^h\}) \cdot \boldsymbol{\nabla} w_i)\, w_j \right. \right. \\
&\qquad \left. \left. + d(\mathbf{x}, \{D^\alpha u^h\})\, w_i w_j \Big] \mathrm{d}\mathbf{x} - \int_\Gamma \big(\mathbf{C}(\mathbf{x}, \{D^\alpha u^h\}) \boldsymbol{\nabla} w_i \cdot \mathbf{n} \big) w_j \, \mathrm{d}\gamma \right\} - \int_\Omega f(\mathbf{x}) w_j \, \mathrm{d}\mathbf{x} \right)_{1 \le j \le N} .
\end{aligned}
\tag{2.77}
$$

For the Klein–Gordon equation (2.56), we have the much more compact expression

$$
R_j^k = \underbrace{\alpha \sum_{i=1}^{N} U_i^k \left\{ \int_\Omega \boldsymbol{\nabla} w_i \cdot \boldsymbol{\nabla} w_j \, \mathrm{d}\mathbf{x} - \int_\Gamma (\boldsymbol{\nabla} w_i \cdot \mathbf{n}) w_j \, \mathrm{d}\gamma \right\}}_{=(t_1)_j^k} \quad \underbrace{- \int_\Omega \left(u_k^h \right)^{-(n+1)} w_j \, \mathrm{d}\mathbf{x}}_{=(t_2)_j^k} \quad \underbrace{+ \int_\Omega \rho(\mathbf{x}) w_j \, \mathrm{d}\mathbf{x}}_{=(t_3)_j^k} . \tag{2.78}
$$

This residual vector $\mathbf{R}_k$ can be used in several ways to assess convergence towards the problem's actual solution — see Fig. 2.6 for a concrete example of such an analysis. We decide to use $\|\mathbf{R}_k\|_2$ as a *global* monitoring quantity. The issue is that $\|\mathbf{R}_k\|_2$ is an absolute quantity, and thus needs to be compared against some reference for it to be useful in practice — otherwise what does it mean for it to be 'small'? To that extent, several ideas have been considered in this work, such as

- the relative decrease for one iteration to the next $(\|\mathbf{R}_k\|_2 - \|\mathbf{R}_{k-1}\|_2) \,/\, \|\mathbf{R}_{k-1}\|_2$, or
- the improvement with respect to the initial guess, that is $\|\mathbf{R}_k\|_2 \,/\, \|\mathbf{R}_0\|_2$.

The drawback from using the 2-norm of the residual vector is that local information is ineluctably lost. The numerical approximation may end up being very accurate in some regions of the numerical domain while being quite poor in others. This is difficult to estimate with the residual vector alone. Even if we assume a constant relative error throughout the domain, the residual vector is allowed to vary by several orders of magnitude. The trick we employ in this work consists in comparing (locally) the residual vector against each of the terms that make it up. These terms are denoted by $(t_i)_{1 \le i \le 3}$ in Eq. (2.78) for the chameleon case. For a numerical approximation to be deemed *good*, we demand that the residual vector be locally several orders of magnitude smaller than the dominant term (in absolute value). This additional check can be performed "off-line", when the iterations are over, as post-processing. It is used in particular in Chapt. 5.

Finally, we mentioned earlier that we may encounter convergence issues when using iterative techniques. In such situations, none of the above criteria are likely to be fulfilled. In order to ensure the termination of the program and thus avoid falling into an infinite loop, it is common practice to additionally set up a maximum number of iterations $k_{\max}$.

2.2.3 Resolving convergence issues

Ultimately, the most critical point in iterative techniques is convergence. As mentioned above, whether the chosen method — Picard or Newton — converges, depends on a number of factors which are not necessarily easy to assess in actual computations. As a result, there are no miracle techniques to address convergence issues but rather recipes and good practices, that are reported here.

Starting from a good initial guess

When solving nonlinear PDEs with iterative techniques, finding a good initial guess is crucial for a number of reasons: prevent divergence or oscillations in the iterative process, avoid local minima/maxima [which corresponds to the case when the Jacobian given by Eq. (2.69) is singular], reduce the computational cost by reducing the number of iterations to be taken, and increase the robustness of the method against variations in the problem's data. All tricks are thus fair game.

The most convenient case is when an analytical approximation to the solution is available. That, however, may require a fair amount of work, yielding an approximation applicable only to a restricted family of cases. Another typical trick consists in starting by solving a *simpler* problem whose solution is expected to be close to the solution of the actual, more complex, problem. A simpler problem might be a closely related linear problem if the nonlinear term is relatively small, or a lower-dimensional one in cases where the physics almost has symmetrical properties.

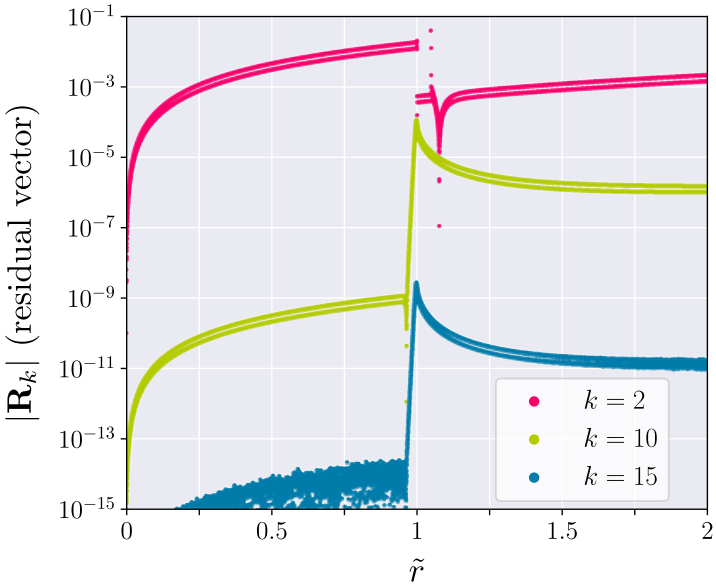


Figure 2.6: Residual vector as a monitoring tool. Absolute value of the residual vector $|\mathbf{R}_k|$ at iterations $k \in \{2, 10, 15\}$ of the Newton method for the radial Klein–Gordon equation (2.56), expressed as a function of $\tilde{r}$. A Dirichlet boundary condition is set at $\tilde{r} = 5$. At $\tilde{r} = 1$, the density term, which acts as a source term for this semi-linear PDE, drops by five orders of magnitude. As a result, the residual tends to be large in this localized region where the solution undergoes rapid variations. We observe that the pointwise residual is uniformly decreased over the computational domain by several orders of magnitude as iterations are carried out. Eventually, the residual starts stagnating and the algorithm terminates.

Relaxation

Sometimes, convergence issues arise because the newly computed approximation $\mathbf{U}_{k+1}$ is "too far away" from the previously computed one $\mathbf{U}_k$. That can be mitigated through the introduction of a so-called *relaxation parameter* $\omega \in]0, 1]$, which allows one to take smaller steps. With this parameter at hand, the update procedure is generalized to a mere convex combination of the previous approximation $\mathbf{U}_k$ and $\mathbf{U}_{k+\omega}$ — the solution of the relevant linearized problem with current guess $\mathbf{U}_k$ —, reading

$$\mathbf{U}_{k+1} \leftarrow \omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k \,. \tag{2.79}$$

The price to pay for this added stability to the algorithm is a potentially slower convergence.

Line search algorithm

The linear search algorithm extends the idea of using a relaxation parameter by turning the choice of its value into an optimization problem. It thereby becomes an iteration-dependent parameter $(\omega_k)_{1 \le k \le k_{\max}}$. A convenient choice is to define this sequence of relaxation parameters with respect to the residual vector as

$$\omega_{k+1} = \arg\min_{\omega} \|\mathbf{R}_k\|_2^2 = \arg\min_{\omega} \|F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big)\|_2^2 \,. \tag{2.80}$$

See Fig. 2.7 to get a flavor of the meaning of Eq. (2.80) on the example of the nonlinear Klein-Gordon problem (2.56–2.57) being discussed.

This optimization problem is easily solved by first computing the derivative with respect to ω. Using the chain rule, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\omega}\Big[\|F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big)\|_2^2\Big] &= \frac{\mathrm{d}}{\mathrm{d}\omega} \sum_{i=1}^{N} \Big[F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big)_i\Big]^2 \\
&= 2 \sum_{i=1}^{N} F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big)_i \times (\mathbf{U}_{k+\omega} - \mathbf{U}_k) \cdot \mathbf{J}_{F_i}\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big) \\
&= 2F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big) \cdot \Big[\mathbf{J}_F\big(\omega \mathbf{U}_{k+\omega} + (1 - \omega)\mathbf{U}_k\big)(\mathbf{U}_{k+\omega} - \mathbf{U}_k)\Big] \,,
\end{aligned}$$

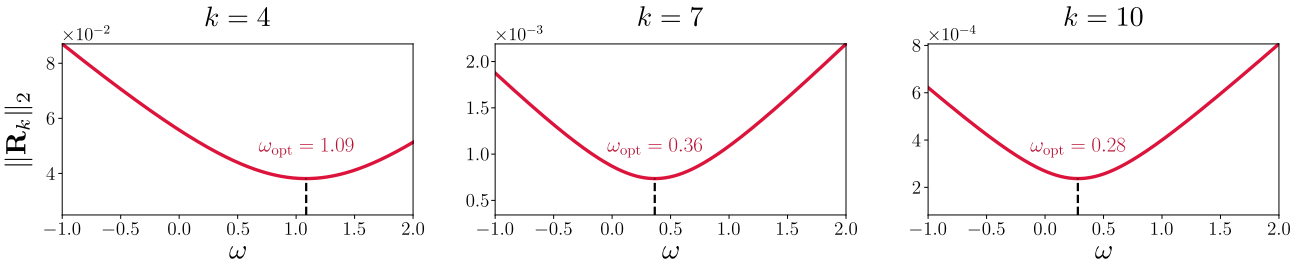


Figure 2.7: Illustration of the line search process on the radial nonlinear Klein-Gordon equation (2.56). The three panels represent the 2-norm of the residual vector $\mathbf{R}_k$ given by Eq. (2.78) at iteration $k \in \{4, 7, 10\}$ of the Newton solver as a function of the relaxation parameter ω. The line search algorithm consists in finding the optimal parameter ω_{opt} that minimizes the residual in the sense given by Eqs. (2.80).

where $\mathbf{J}_{F_i} \colon \mathbb{R}^N \to \mathbb{R}^N$ denotes the Jacobian of the i^{th} component of F and $\mathbf{J}_F \colon \mathbb{R}^N \to \mathbb{R}^N \times \mathbb{R}^N$ is the full Jacobian matrix of F defined by Eq. (2.68). Then, we find out for which ω the derivative cancels out, that is we solve

$$F\big(\omega \mathbf{U}_{k+\omega} + (1-\omega)\mathbf{U}_k\big) \cdot \Big[\mathbf{J}_F\big(\omega \mathbf{U}_{k+\omega} + (1-\omega)\mathbf{U}_k\big)(\mathbf{U}_{k+\omega} - \mathbf{U}_k)\Big] = 0\,. \tag{2.81}$$

This last step cannot be carried out analytically, and one has to rely on some root-finding algorithm. The subsequent multiple evaluations of F together with its Jacobian matrix to determine the best ω is what makes up the computational cost of the line search algorithm overall.

The form of Eq. (2.81) can be made explicit in the case of the Klein–Gordon equation. The vector F has components given by Eq. (2.78) while the Jacobian matrix $\mathbf{J}_F$ has entries

$$\begin{aligned}
(\mathbf{J}_F)_{j,l}(\mathbf{U}) &= \big(\mathbf{J}_{F_j}\big)_l(\mathbf{U}) = \frac{\partial F_j}{\partial U_l}(\mathbf{U}) \\
&= \alpha \sum_{i=1}^{N} \frac{\partial U_i}{\partial U_l} \left\{ \int_\Omega \boldsymbol{\nabla} w_i \cdot \boldsymbol{\nabla} w_j \,\mathrm{d}\mathbf{x} - \int_\Gamma (\boldsymbol{\nabla} w_j \cdot \mathbf{n}) w_i \,\mathrm{d}\gamma \right\} - \int_\Omega w_j \frac{\partial}{\partial U_l} \left(\sum_{i=1}^{N} U_i w_i \right)^{-(n+1)} \mathrm{d}\mathbf{x} + \frac{\partial}{\partial U_l} \int_\Omega \rho(\mathbf{x}) w_j \,\mathrm{d}\mathbf{x} \\
&= \alpha \sum_{i=1}^{N} \delta_i^l \left\{ \int_\Omega \boldsymbol{\nabla} w_i \cdot \boldsymbol{\nabla} w_j \,\mathrm{d}\mathbf{x} - \int_\Gamma (\boldsymbol{\nabla} w_j \cdot \mathbf{n}) w_i \,\mathrm{d}\gamma \right\} + (n+1) \int_\Omega (u^h)^{-(n+2)} \sum_{i=1}^{N} \delta_i^l w_i \,\mathrm{d}\mathbf{x} \\
&= \alpha \left\{ \int_\Omega \boldsymbol{\nabla} w_i \cdot \boldsymbol{\nabla} w_j \,\mathrm{d}\mathbf{x} - \int_\Gamma (\boldsymbol{\nabla} w_j \cdot \mathbf{n}) w_i \,\mathrm{d}\gamma \right\} + (n+1) \int_\Omega (u^h)^{-(n+2)} w_l \, w_j \,\mathrm{d}\mathbf{x}\,.
\end{aligned} \tag{2.82}$$

We observe that, when using Newton's method, most finite element matrices/vectors needed to carry out the present computations are already available (as they were also needed for the computation of $\mathbf{U}_{k+\omega}$). What dominate the root finding stage are thus the various scalar products and matrix-vector multiplications of Eq. (2.81).

Algorithm 3 Nonlinear FEM iterative techniques — summary

Inputs, initialization:
  Set a value to meta-parameters ϵ_{abs}, ϵ_{rel}, $k_{\max}$, ω, etc. ▷ *empirical, physical insights*
  Linearize the PDE according to Picard or Newton method
  Assemble matrix $\mathbf{A}_0 = \mathbf{A}_{\text{cst}} + \mathbf{A}^0_{\text{mod}}$ and rhs vector $\mathbf{L}_0 = \mathbf{L}_{\text{cst}} + \mathbf{L}^0_{\text{mod}}$
  Select an initial guess $\mathbf{U}_0$ ▷ *physical insight, solution of a simpler problem*
for $k = 0$ to $k_{\max}$ **do**
  Solve linear system $\mathbf{A}_k \mathbf{U}_{k+\omega} = \mathbf{L}_k$
  if employ a line search algorithm **then**
    Determine ω_{k+1} as the root of Eq. (2.81)
  else
    $\omega_{k+1} = \omega$ (constant)
  end if
  $\mathbf{U}_{k+1} \leftarrow \omega_{k+1}\mathbf{U}_{k+\omega} + (1-\omega_{k+1})\mathbf{U}_{k-1}$
  Criteria checks:
    Evaluate the residual vector, evaluate the criteria introduced in Sec. 2.2.2
    if stop is true: **break**
  Compute $\mathbf{A}^{k+1}_{\text{mod}}$ and $\mathbf{L}^{k+1}_{\text{mod}}$ and Update $\mathbf{A}_{k+1} \leftarrow \mathbf{A}_{\text{cst}} + \mathbf{A}^{k+1}_{\text{mod}}$, $\mathbf{L}_{k+1} \leftarrow \mathbf{L}_{\text{cst}} + \mathbf{L}^{k+1}_{\text{mod}}$
end for

A summary of how nonlinear problems are handled with Picard or Newton method (see Sec. 2.2.1), together with the useful practices laid out in Sec. 2.2.2 and 2.2.3, is given in Algorithm 3.

Continuation techniques

When all the above prescriptions fail, one can resort to so-called *ramping* or *numerical continuation* techniques [215–218]. Say Newton's iterations fail to converge on the problem of interest (the *target* problem), but are successful for solving a somewhat modified version of this problem (the *entry* problem) nonetheless. Very broadly speaking, the idea of continuation techniques is to go from the entry problem to the target one, gradually, through a sequence of sub-problems that bridge the gap.

The typical application is to progressively turn on the nonlinearity in a given PDE by weighting the nonlinear term with a parameter that goes from 0 to 1. In our case, we regularly faced the situation where Newton's method would successfully converge when solving Eq. (2.56) with a given value of $\alpha = \alpha_{\text{c}}$, but diverge for $\alpha = \alpha_{\text{d}}$. In that case, the continuation parameter would be $\alpha \in [\alpha_{\text{c}}, \alpha_{\text{d}}]$. Specifically, we would create a sequence $(\alpha_i)_{1 \leq i \leq M}$ such that $\alpha_1 = \alpha_{\text{c}}$, $\alpha_M = \alpha_{\text{d}}$, and use the solution of the $(i-1)^{\text{th}}$ problem as the initial guess for the i^{th} problem — see Chapt. 5. This approach was also adapted to the density profile $\rho(\mathbf{x})$, which proved to be a valuable aid in reaching convergence for certain atmospheric profiles, see Chapt. 5.

2.2.4 A word about the time-dependent nonlinear Klein–Gordon equation

Back in Sec. 2.1.4, we provided some leads on how hyperbolic problems could be addressed numerically by combining the finite element method (for space discretization) together with the finite difference method (for time discretization). The toy model we used to showcase this technique was a linear wave-equation. Yet, the Klein–Gordon equation governing the dynamics of the chameleon model is only semi-linear. Fortunately, the concepts introduced in the present section for nonlinear problems can be readily reinvested to supplement Sec. 2.1.4. With little regard to the actual coefficients of the chameleon field's equation, we are interested in the prototypical PDE

$$\frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) - c^2 \Delta u(\mathbf{x}, t) = \rho(\mathbf{x}, t) - u(\mathbf{x}, t)^m \,, \tag{2.83}$$

for some $m \in \mathbb{Z}^-$. Taking the same steps as in Sec. 2.1.4, Eqs. (2.52a–2.52b) become, for $\theta \in [0, 1]$ and $n \in [\![0, N_t - 1]\!]$,

$$\begin{aligned}\left[1 - (\Delta t\, \theta)^2 \Delta\right] u^{n+1} = \left[1 + \Delta t^2\, \theta(1-\theta)\Delta\right] u^n + \Delta t v^n \\ + \theta \Delta t^2 \left[\theta \rho^{n+1} + (1-\theta)\rho^n\right] - \theta \Delta t^2 \left[\theta u^{n+1} + (1-\theta)u^n\right]^m \,,\end{aligned} \tag{2.84a}$$

$$v^{n+1} = v^n + \Delta t \left\{ c^2 \theta \Delta u^{n+1} + c^2(1-\theta)\Delta u^n + \theta \rho^{n+1} + (1-\theta)\rho^n - \left[\theta u^{n+1} + (1-\theta)u^n\right]^m \right\} . \tag{2.84b}$$

As expected, Eq. (2.84) is not linear in u^{n+1} [Eq. (2.84) is not subject to this issue in comparison]. Forgetting about time stepping, Eq. (2.84) is nothing but a semi-linear elliptic PDE of unknown u^{n+1}. Therefore, it is quite possible to apply Newton's method on its corresponding weak form, as seen in Sec. 2.2.1.

Symmetry		Name	dim	Remark
independent of position	$\mathbf{x} \to \mathbf{x} + \mathbf{a}$ $\forall \mathbf{a} \in \mathbb{R}^3$	maximally symmetric	0	uninteresting in stationary problems
rotations w.r.t. two axes	$\mathbf{x} \to \mathbf{R}\mathbf{x}$ $\forall \mathbf{R} \in \mathrm{SO}(3)$	spherical symmetry	1	nested spheres, spherical coordinates
translations along two axes (e.g. y and z)	$\mathbf{x} \to \mathbf{T}\mathbf{x}$ $\forall \mathbf{T} \in \mathbb{T}_{yz}$	—	1	parallel infinite walls, x-coordinate
translations & rotations along one axis (e.g. z)	$\mathbf{x} \to \mathbf{T}\mathbf{x}$, $\mathbf{x} \to \mathbf{R}\mathbf{x}$ $\forall (\mathbf{T}, \mathbf{R}) \in \mathbb{T}_z \times \mathrm{SO}(2, z)$	—	1	infinite cylinder, cylindrical coordinates
translations along one axis (e.g. z)	$\mathbf{x} \to \mathbf{T}\mathbf{x}$ $\forall \mathbf{T} \in \mathbb{T}_z$	planar symmetry	2	infinitely long objects with constant section
rotations w.r.t. one axis (e.g. z)	$\mathbf{x} \to \mathbf{R}\mathbf{x}$ $\forall \mathbf{R} \in \mathrm{SO}(2, z)$	cylindrical symmetry	2	cylindrical coordinates or spherical coordinates

Table 2.1: Global continuous symmetries in $\mathbb{R}^3$ that lead to dimensional-reduction. SO(3) denotes the special orthogonal group in three dimension, or the 3D rotation group; $\mathrm{SO}(2, z)$ denotes the group of rotations with respect to the z-axis, which is a subgroup of SO(3); $\mathbb{T}_z$ denotes the group of translations along the z-axis; $\mathbb{T}_{yz}$ denotes the group of translations in the yz-plane. We give the most suitable coordinate systems to adopt in the last column on an indicative basis.

While the numerical implementation would grow in complexity, there are no major anticipated stumbling blocks in tackling time-dependent chameleons. Nonetheless, having to solve for a nonlinear problem at each time step severely increases the computational cost of the method. The real challenge may thus lie in code optimization and the implementation of high-performance computing (HPC) techniques.

2.3 Taking advantage of problem symmetries

Another important aspect to discuss is when the problem at stake exhibits a global continuous symmetry. As it was mentioned at the beginning of Sec. 1.4.1, such symmetries allow for a dimensional reduction of the PDE problem. This section is restricted to stationary PDE problems, with three spatial dimensions. In that respect, Table 2.1 reports the rotational and translation symmetries in $\mathbb{R}^3$ that effectively lead to a reduction of the problem's dimension.

When taken into account in numerical computations, these continuous symmetries can drastically reduce the complexity of the underlying FEM calculations. Going from a 3D mesh to a lower-dimensional mesh greatly lowers the number of DOFs that has to be employed for a given precision on the numerical approximation. For instance, it was observed empirically that for an axisymmetric setup, FEM computations conducted using a 2D mesh were roughly ~ 500 times faster than their three-dimensional counterparts (at fixed relative error) [137]. The moral is that symmetries should be leveraged in numerical computations whenever possible.

In this section, we show how one can adapt the framework laid out in Sec. 2.1 in order to account for the aforementioned symmetries (when relevant). In particular, this process requires abandoning Cartesian coordinates $\mathbf{x}$ in favor of better suited coordinate systems in the presence of rotational symmetries. We illustrate how this is performed in the spherically symmetric and axisymmetric cases. To stay in line with Sec. 2.1, we consider the case of the Poisson equation in a bounded open set $\Omega \subset \mathbb{R}^3$ with homogeneous Dirichlet boundary conditions on $\Gamma \equiv \partial\Omega$, reading

$$-\Delta u = f \text{ in } \Omega \qquad \text{with} \quad u = 0 \text{ on } \Gamma\,. \tag{2.85}$$

2.3.1 Spherical symmetry

As reported in Table 2.1, a PDE problem is said to be spherically symmetric in 3D when the following properties hold simultaneously:

1. The domain Ω is invariant under rotations, i.e. $\forall \mathbf{R} \in \mathrm{SO}(3)$ and $\forall \mathbf{x} \in \Omega$, $\mathbf{R}\mathbf{x} \in \Omega$. For instance, Ω can be a ball centered at the origin or the whole space $\mathbb{R}^3$. Because we work (for now) under the assumption that

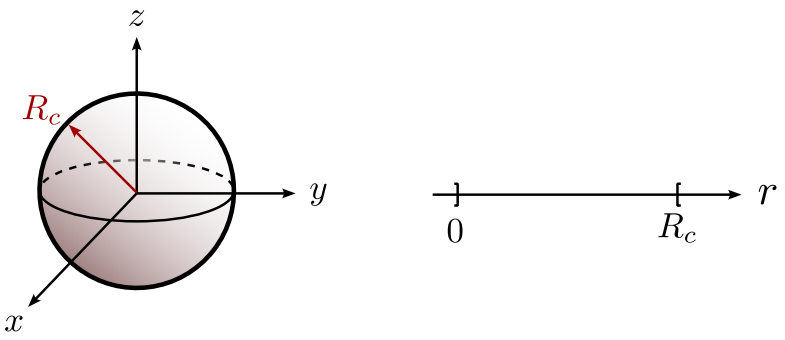


Figure 2.8: 3D spherically symmetric problems can be treated as 1D radial problems.

Ω is bounded, we assume that there exists a radius R_c such that $\Omega = \mathcal{B}(R_c)$, the open ball of radius R_c centered at the origin.

2. The rhs function f depends only on the distance of the coordinate $\mathbf{x}$ from the origin, i.e. $f(\mathbf{x}) = f(\|\mathbf{x}\|)$.

3. The boundary conditions expressed on Γ (which is necessarily a sphere given the first condition) must be invariant under rotations. Homogeneous Dirichlet conditions considered in Eq. (2.85) trivially satisfy this condition.

The first condition expresses that the domain Ω must be spherically symmetric while the two others signify that the problem's data is also invariant under rotations. Ergo one looks for radial solutions to Eq. (2.85).

Simplified PDE and corresponding weak form

Dimensional reduction in this case is best brought out by rewriting the Poisson equation (2.85) in spherical coordinates

$$\mathbf{s} = (r,\, \theta,\, \varphi) \in \Pi \qquad \text{where} \quad \Pi :=]0,\, R_c[\times]0,\, \pi[\times]0,\, 2\pi[\,. \tag{2.86}$$

To be rigorous in doing so, one should define $\mathcal{M}_{\text{sph}}$ the bijective mapping from Cartesian coordinates $\mathbf{x}$ to spherical coordinates $\mathbf{s}$, and for any $v\colon \Omega \to \mathbb{R}$, let

$$\begin{aligned} \tilde{v}\colon \Pi &\to \mathbb{R} \\ \mathbf{s} &\mapsto \tilde{v}(\mathbf{s}) = v\big(\mathcal{M}_{\text{sph}}^{-1}(\mathbf{s})\big) \end{aligned} \tag{2.87}$$

Under this change of coordinates, the Poisson PDE (2.85) thereby becomes

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \tilde{u}}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial \tilde{u}}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 \tilde{u}}{\partial \varphi^2} = -\tilde{f}\,. \tag{2.88}$$

For spherically symmetric problems (see the three conditions listed above), it makes sense to look for solutions $\tilde{u}$ that only depend on the radial coordinate r, and so treat $\tilde{u}$ as a function of r only, which we denote u_r — see Fig. 2.8. Likewise, $\tilde{f}(r, \theta, \varphi)$ boils down to $f_r(r)$. This results in the following ODE put in the divergence form

$$-\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}u_r}{\mathrm{d}r}\right) = r^2 f_r \qquad \text{with} \quad u_r(R_c) = 0 \text{ and } u_r'(0) = 0\,, \tag{2.89}$$

where the latter homogeneous Neumann boundary condition at the origin is called a *compatibility condition* and is imposed to ensure that the solution is physically meaningful and smooth at $r = 0$.

Let us denote W_r a suitable functional space — to be specified later — for writing the weak formulation associated with Eq. (2.89). The (bi)linear forms a_r and l_r for this radial problem read, for all $u_r,\, v_r \in W_r$

$$a_r(u_r, v_r) = \int_0^{R_c} r^2\, u_r'\, v_r' \mathrm{d}r \qquad \text{and} \qquad l_r(v_r) = \int_0^{R_c} r^2\, f_r\, v_r\, \mathrm{d}r\,. \tag{2.90}$$

Well-posedness

To find a suitable definition for W_r, it is useful to start from a weak formulation we know to be well-posed. As a matter of fact, we have shown in Sec. 2.1.2 that the problem "find $u \in H_0^1\big(\mathcal{B}(R_c)\big)$ such that for all $v \in H_0^1\big(\mathcal{B}(R_c)\big)$, a(u, v) = l(v)" — with $a(\cdot,\,\cdot)$ and $l(\cdot)$ given by Eq. (2.13) — satisfies all the assumptions of the Lax–Milgram theorem (see Box E).

Therefore, let

$$H^1_r\big(\mathcal{B}(R_c)\big) = \Big\{ u \in H^1_0\big(\mathcal{B}(R_c)\big) \text{ such that for all } \mathbf{x} \in \mathcal{B}(R_c) \text{ and for all } \mathbf{R} \in \mathrm{SO}(3),\, u(\mathbf{x}) = u(\mathbf{R}\mathbf{x}) \Big\}. \quad (2.91)$$

In plain English, H^1_r is nothing but the subset of radial functions in H^1_0.

Lemma 2.1. H^1_r is a closed linear subspace of H^1_0, and therefore a Hilbert space in its own right.

Proof. The only non-trivial point to show is that H^1_r is closed in H^1_0 for the norm $\|\cdot\|_{H^1}$. Let $(g_n)_{n\in\mathbb{N}}$ be a sequence of elements in H^1_r that converges to $g \in H^1_0$. To show that g belongs to H^1_r, let $\mathbf{R} \in \mathrm{SO}(3)$ and define the operator

$$\begin{aligned} T_{\mathbf{R}} \colon H^1_0 &\to H^1_0 \\ v \mapsto \colon \mathcal{B}(R_c) &\to \mathbb{R} \\ \mathbf{x} &\mapsto v(\mathbf{R}\mathbf{x})\,. \end{aligned} \quad (2.92)$$

The goal then consists in showing that $\|g - T_{\mathbf{R}}(g)\|_{H^1} = 0$. We have for $n \in \mathbb{N}$,

$$\begin{aligned} \|g - T_{\mathbf{R}}(g)\|^2_{H^1} &= \big\|(g - g_n) - \big(T_{\mathbf{R}}(g) - g_n\big)\big\|^2_{H^1} \\ &\le \|g - g_n\|^2_{H^1} + \|T_{\mathbf{R}}(g - g_n)\|^2_{H^1}\,, \end{aligned} \quad (2.93)$$

where we have made use of the fact that $T_{\mathbf{R}}(g_n) = g_n$ since $g_n \in H^1_r$. The fist term of Eq. (2.93) goes to zero when $n \to +\infty$ by definition. Moreover, $T_{\mathbf{R}}$ is a linear operator and for $v \in H^1_0$,

$$\begin{aligned} \|T_{\mathbf{R}}(v)\|^2_{H^1} &= \int_{\mathcal{B}(R_c)} |v(\mathbf{R}\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} + \int_{\mathcal{B}(R_c)} \big\|\boldsymbol{\nabla}\big(v(\mathbf{R}\mathbf{x})\big)\big\|^2 \,\mathrm{d}\mathbf{x} \\ &= \int_{\mathcal{B}(R_c)} |v(\mathbf{y})|^2 \,|\det(\mathbf{R})|\,\mathrm{d}\mathbf{y} + \int_{\mathcal{B}(R_c)} \big\|\mathbf{R}^T \boldsymbol{\nabla} v(\mathbf{y})\big\|^2 \,|\det(\mathbf{R})|\,\mathrm{d}\mathbf{y} \\ &= \|v\|^2_{H^1}\,. \end{aligned}$$

Therefore, $T_{\mathbf{R}}$ is continuous and so the second term of Eq. (2.93) also goes to zero as $n \to +\infty$, which ends the proof. □

The space H^1_r is a Hilbert subspace of H^1_0. In particular, the problem "find $u \in H^1_r$ such that for all $v \in H^1_r$, $a(u,v) = l(v)$" is well-posed. In fact, the assumptions of the Lax–Milgram theorem are automatically satisfied in H^1_r since the latter space inherits from the H^1-norm. Looking at Eq. (2.90) motivates the definition

$$W_r = \left\{ v \colon]0,\, R_c[\to \mathbb{R} \text{ such that } \int_0^{R_c} r^2 |v(r)|^2 \,\mathrm{d}r < +\infty,\ \int_0^{R_c} r^2 |v'(r)|^2 \,\mathrm{d}r < +\infty \text{ and } v(R_c) = 0 \right\}. \quad (2.94)$$

We also define the inner product

$$\begin{aligned} \langle \cdot,\, \cdot \rangle_{W_r} \colon W_r \times W_r &\to \mathbb{R} \\ u,\, v &\mapsto \int_0^{R_c} r^2 \Big[u(r)\, v(r) + u'(r)\, v'(r) \Big] \mathrm{d}r \end{aligned} \quad (2.95)$$

and denote by $\|\cdot\|_{W_r}$ the associated norm.

Lemma 2.2. H^1_r is isomorphic to W_r

Proof. Let us define the linear mapping between W_r and H^1_r

$$\begin{aligned} \Phi \colon W_r &\to H^1_r \\ v_r \mapsto \colon \mathcal{B}(R_c) &\to \mathbb{R} \\ \mathbf{x} &\mapsto v_r(\|\mathbf{x}\|)/4\pi\,. \end{aligned} \quad (2.96)$$

This map is well-defined, since for $v_r \in W_r$, $\Phi(v_r) \equiv 0$ on Γ and

$$\int_{\mathcal{B}(R_c)} |\Phi(v_r)(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} = \frac{1}{4\pi} \int_{\mathcal{B}(R_c)} |v_r(\|\mathbf{x}\|)|^2 \,\mathrm{d}\mathbf{x} = \frac{1}{4\pi} \int_{\Pi} |v_r(r)|^2 r^2 \sin(\theta) \,\mathrm{d}r\mathrm{d}\theta\mathrm{d}\varphi = \int_0^{R_c} r^2 |v_r(r)|^2 \,\mathrm{d}r < +\infty \,,$$

$$\int_{\mathcal{B}(R_c)} \left\|\boldsymbol{\nabla}\big(\Phi(v_r)\big)\right\|^2 \mathrm{d}\mathbf{x} = \frac{1}{4\pi} \int_{\mathcal{B}(R_c)} |v_r'(\|\mathbf{x}\|)|^2 \,\mathrm{d}\mathbf{x} = \int_0^{R_c} r^2 |v_r'(r)|^2 \,\mathrm{d}r < +\infty \,.$$

Moreover, it is bijective:

- *injectivity* — If $\Phi(v_1) = \Phi(v_2)$, v_1, $v_2 \in W_r$, then for all $\mathbf{x} \in \mathcal{B}(R_c)$, $v_1(\|\mathbf{x}\|) = v_2(\|\mathbf{x}\|)$ i.e. for all $r \in]0,\, R_c[$, $v_1(r) = v_2(r)$. Of course, these equalities hold *almost everywhere*.
- *surjectivity* — Let $\mathbf{n} \in \mathbb{R}^3$ with $\|\mathbf{n}\| = 1$. For $v \in H^1_r$, we can define $v_r \colon]0,\, R_c[\to \mathbb{R}$ such that for all $r \in]0,\, R_c[$, $v_r(r) = v(r\mathbf{n})$. Then $v_r \in W_r$ and $\Phi(v_r) = v$.

Finally, the above calculations show that for $v_r \in W_r$,

$$\|\Phi(v_r)\|^2_{H^1} = \int_0^{R_c} r^2 |v_r(r)|^2 \,\mathrm{d}r + \int_0^{R_c} r^2 |v_r'(r)|^2 \,\mathrm{d}r = \|v_r\|^2_{W_r} \,, \tag{2.97}$$

so that Φ is indeed an isomorphism between W_r and H^1_r. □

Therefore, W_r inherits from the topological properties of H^1_r — in particular, W_r equipped with the norm $\|\cdot\|_{W_r}$ is complete. Thus, W_r equipped with the inner product $\langle \cdot\,, \cdot \rangle_{W_r}$ given by Eq. (2.95) is a Hilbert space.

The last step in showing that the radial weak formulation is well-posed consists in examining the (bi)linear forms $a_r(\cdot,\cdot)$ and $l_r(\cdot)$ given by Eq. (2.90). The continuity conditions are proved using the Cauchy–Schwarz inequality. Here, let us remark that in the original problem, $f \in L^2\big(\mathcal{B}(R_c)\big)$. This condition translates into

$$\|f_r\|_{L^2\left(]0,\, R_c[,\, r^2\right)} \coloneqq \int_0^{R_c} r^2 \, |f_r(r)|^2 \,\mathrm{d}r < +\infty \,.$$

Finally, the coercivity of $a_r(\cdot,\,\cdot)$ in W_r is obtained thanks to the coercivity of $a(\cdot,\,\cdot)$ in H^1, since for all $u_r \in W_r$ we have

$$a_r(u_r, u_r) = a\big(\Phi(u_r), \Phi(u_r)\big) \geq \alpha \, \|\Phi(u_r)\|^2_{H^1} = \alpha \, \|u_r\|^2_{W_r} \,, \tag{2.98}$$

where $\alpha > 0$ is a coercivity constant of the bilinear form $a(\cdot,\,\cdot)$ (see Sec. 2.1.2). Besides, the functional space $L^2\big(]0,\, R_c[,\, r^2\big)$ is a *weighted space*. This notion shall be formalized in Chapt. 3.

All in all, we have shown that the weak formulation " find $u_r \in W_r$ such that for all $v_r \in W_r$, $a_r(u_r, v_r) = l_r(v_r)$" is well-posed. This legitimates the use of radial FEM computations conducted at several stages in this PhD work. Of course, this approach is generalizable to other elliptic PDEs than the Poisson equation.

2.3.2 Cylindrical symmetry

In cylindrical symmetry, the problem is invariant under rotations with respect to an axis. Without loss of generality, we can have the z-axis play that role. We also denote by $\mathrm{SO}(2, z)$ the subgroup of $\mathrm{SO}(3)$ containing the rotations about the z-axis. As for the spherically symmetric case discussed above, both the domain Ω and the problem's data must be invariant under such rotations.

Simplified PDE and corresponding weak form

Dimensional reduction is made clear when writing the PDE (2.85) in cylindrical or spherical coordinates — see Fig. 2.9.

Cylindrical coordinates Let us denote cylindrical coordinates by $\mathbf{c} = (\rho,\, \varphi,\, z)$ and $\mathcal{M}_{\mathrm{cyl}}$ the mapping from Cartesian coordinates to cylindrical coordinates. The domain Ω is mapped to $\Xi_{\mathrm{3D}} = \mathcal{M}_{\mathrm{cyl}}(\Omega)$. Rewriting Eq. (2.85) in cylindrical coordinates yields

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\tilde{u}}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2\tilde{u}}{\partial\varphi^2} + \frac{\partial^2\tilde{u}}{\partial z^2} = -\tilde{f} \,, \tag{2.99}$$

where for any $v \colon \Omega \to \mathbb{R}$ and for any $\mathbf{c} \in \Xi_{\mathrm{3D}}$, $\tilde{v}(\mathbf{c}) = v\big(\mathcal{M}^{-1}_{\mathrm{cyl}}(\mathbf{c})\big)$. Because of axisymmetry, we look for solutions to Eq. (2.99) that do not depend on the angle φ, i.e. $\partial_\varphi \tilde{u} \equiv 0$.[18] Denoting Ξ_{2D} a slice of Ξ_{3D} at a fixed angle φ,

[18] In Cartesian coordinates, one can show that this condition translates to $x\,\partial_y u - y\,\partial_x u \equiv 0$.

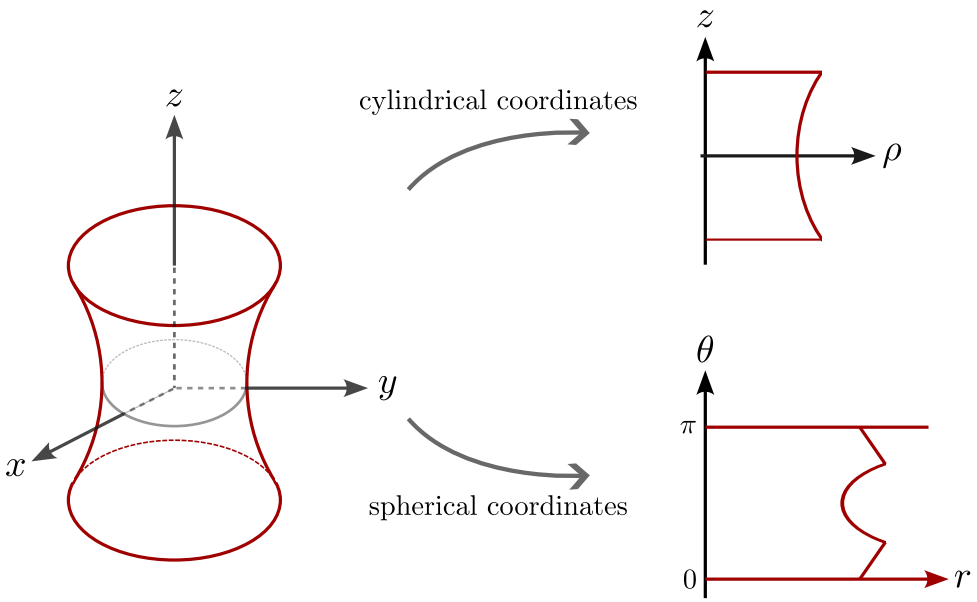


Figure 2.9: 3D axisymmetric problems can be treated as 2D problems using either cylindrical or spherical coordinates.

the strong problem boils down to finding $u_{\text{cyl}} : \Xi_{\text{2D}} \to \mathbb{R}$ such that

$$\frac{\partial}{\partial \rho}\left(\rho \frac{\partial u_{\text{cyl}}}{\partial \rho}\right) + \rho \frac{\partial^2 u_{\text{cyl}}}{\partial z^2} = -\rho f_{\text{cyl}} \,. \tag{2.100}$$

Denoting W_{cyl} a suitable functional space to be specified later, the weak formulation associated with Eq. (2.100) involves the (bi)linear forms a_{cyl} and l_{cyl}, reading for all u_{cyl}, $v_{\text{cyl}} \in W_{\text{cyl}}$

$$a_{\text{cyl}}(u_{\text{cyl}}, v_{\text{cyl}}) = \int_{\Xi_{\text{2D}}} \rho \boldsymbol{\nabla}_{\rho,z} u_{\text{cyl}} \cdot \boldsymbol{\nabla}_{\rho,z} v_{\text{cyl}} \, \mathrm{d}\rho \mathrm{d}z \qquad \text{and} \qquad l_{\rho,z}(v_{\text{cyl}}) = \int_{\Xi_{\text{2D}}} \rho f_{\text{cyl}} v_{\text{cyl}} \, \mathrm{d}\rho \mathrm{d}z \,. \tag{2.101}$$

Here, $\boldsymbol{\nabla}_{\rho,z} = (\partial_\rho, \partial_z)^T$. The homogeneous Dirichlet boundary condition is to be applied on $\Gamma_{\text{cyl}} \subset \partial\Xi_{\text{2D}}$, where $\rho \neq 0$ (see Fig. 2.9).

Spherical coordinates Axisymmetry can also be accounted for using spherical coordinates, where partial derivatives with respect to φ are dropped. The resulting strong form PDE problem is to find $u_{\text{pol}} : \Pi_{\text{2D}} \to \mathbb{R}$ such that

$$\frac{\partial}{\partial r}\left(r^2 \sin\theta \frac{\partial u_{\text{pol}}}{\partial r}\right) + \frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_{\text{pol}}}{\partial \theta}\right) = -r^2 \sin\theta f_{\text{pol}} \,, \tag{2.102}$$

where Π_{2D} is a slice of Π at a fixed angle φ.

Denoting W_{pol} a suitable functional space to be specified later, the weak formulation associated with Eq. (2.102) involves the (bi)linear forms a_{pol} and l_{pol}, reading for all u_{pol}, $v_{\text{pol}} \in W_{\text{pol}}$

$$\begin{aligned} a_{\text{pol}}(u_{\text{pol}}, v_{\text{pol}}) &= \int_{\Pi_{\text{2D}}} \left[\begin{pmatrix} r^2 \sin\theta & 0 \\ 0 & \sin\theta \end{pmatrix} \boldsymbol{\nabla}_{r,\theta} u_{\text{pol}} \right] \cdot \boldsymbol{\nabla}_{r,\theta} v_{\text{pol}} \, \mathrm{d}r\mathrm{d}\theta \,, \\ l_{\text{pol}}(v_{\text{pol}}) &= \int_{\Pi_{\text{2D}}} r^2 \sin\theta f_{\text{pol}} v_{\text{pol}} \, \mathrm{d}r\mathrm{d}\theta \,. \end{aligned} \tag{2.103}$$

Here, $\boldsymbol{\nabla}_{r,\theta} = (\partial_r, \partial_\theta)^T$ and we set $\mathbf{A}(r,\theta) = \mathrm{diag}(r^2 \sin\theta, \sin\theta)$. The homogeneous Dirichlet boundary condition is to be applied on $\Gamma_{\text{pol}} \subset \partial\Pi_{\text{2D}}$, where $r \neq 0$ (see Fig. 2.9).

Well-posedness

Following what we did in Sec. 2.3.1, we introduce the subspace

$$H^1_\varphi(\Omega) = \left\{ u \in H^1_0(\Omega) \text{ such that for all } \mathbf{x} \in \Omega \text{ and for all } \mathbf{R} \in \mathrm{SO}(2,z),\ u(\mathbf{x}) = u(\mathbf{R}\mathbf{x}) \right\}. \tag{2.104}$$

One can show, in a similar way to the proof of Lemma 2.1, that H^1_φ is closed in H^1_0 for the H^1-norm, and is therefore a Hilbert space.

We then define the functional spaces

$$W_{\text{cyl}} = \left\{ v\colon \Xi_{\text{2D}} \to \mathbb{R} \text{ such that } \int_{\Xi_{\text{2D}}} \rho|v|^2 \,\mathrm{d}\rho\mathrm{d}z < \infty,\ \int_{\Xi_{\text{2D}}} \rho\|\boldsymbol{\nabla}_{\rho,z} v\|^2 \,\mathrm{d}\rho\mathrm{d}z < \infty \text{ and } v = 0 \text{ on } \Gamma_{\text{cyl}} \right\},$$
$$W_{\text{pol}} = \left\{ v\colon \Pi_{\text{2D}} \to \mathbb{R} \text{ such that } \int_{\Pi_{\text{2D}}} r^2 \sin\theta|v|^2 \,\mathrm{d}r\mathrm{d}\theta < \infty,\ \int_{\Pi_{\text{2D}}} (\boldsymbol{\nabla}_{r,\theta} v)^T \mathbf{A} (\boldsymbol{\nabla}_{r,\theta} v) \mathrm{d}r\mathrm{d}\theta < \infty \text{ and } v = 0 \text{ on } \Gamma_{\text{pol}} \right\},$$

and the inner products

$$\begin{aligned} \langle u, v\rangle_{\text{cyl}} &= \int_{\Xi_{2D}} \rho\, u\, v \,\mathrm{d}\rho\mathrm{d}z + \int_{\Xi_{2D}} \rho\big(\boldsymbol{\nabla}_{\rho,z} u\big) \cdot \big(\boldsymbol{\nabla}_{\rho,z} v\big) \,\mathrm{d}\rho\mathrm{d}z\,,\ \forall (u,v) \in W^2_{\text{cyl}} \\ \langle u, v\rangle_{\text{pol}} &= \int_{\Pi_{2D}} r^2 \sin\theta\, u\, v \mathrm{d}r\mathrm{d}\theta + \int_{\Pi_{2D}} \big[\mathbf{A}\boldsymbol{\nabla}_{r,\theta} u\big] \cdot \boldsymbol{\nabla}_{r,\theta} v \,\mathrm{d}r\mathrm{d}\theta\,,\ \forall (u,v) \in W^2_{\text{pol}} \end{aligned} \tag{2.105}$$

Following the proof of Lemma 2.2, one can show that W_{cyl} and W_{pol} are isomorphic to H^1_φ using the mappings

$$\begin{aligned} \Phi_{\text{cyl}}\colon W_{\text{cyl}} &\to H^1_\varphi \\ v_{\text{cyl}} &\mapsto\colon \Omega \to \mathbb{R} \\ \mathbf{x} &\mapsto v_{\text{cyl}}\big(\rho(\mathbf{x}),\, z(\mathbf{x})\big) \end{aligned} \qquad \text{and} \qquad \begin{aligned} \Phi_{\text{pol}}\colon W_{\text{pol}} &\to H^1_\varphi \\ v_{\text{pol}} &\mapsto\colon \Omega \to \mathbb{R} \\ \mathbf{x} &\mapsto v_{\text{pol}}\big(r(\mathbf{x}),\, \theta(\mathbf{x})\big)\,. \end{aligned}$$

In particular, one inherits from the interesting properties of H^1_φ: W_{cyl}, W_{pol} equipped with the inner products (2.105) are Hilbert spaces, the (bi)linear forms are all continuous, a_{cyl} and a_{pol} are coercive over W_{cyl} and W_{pol} respectively.

All in all, the weak formulations

$$\text{“ Find } u \in W_{\text{cyl}} \text{ such that for all } v \in W_{\text{cyl}},\ a_{\text{cyl}}(u,v) = l_{\text{cyl}}(v) \text{ ”},$$
$$\text{“ Find } u \in W_{\text{pol}} \text{ such that for all } v \in W_{\text{pol}},\ a_{\text{pol}}(u,v) = l_{\text{pol}}(v) \text{ ”},$$

are both well-posed.

Chapter summary

This chapter was dedicated to the presentation of the finite element method, which was identified as an adequate numerical method for studying scalar-tensor models with screening mechanisms at the end of Chapt. 1. FEM is indeed well-suited for solving the second-order elliptic PDEs that arise in such models. We showed how the method can readily be adapted to deal with time-dependent problems and nonlinearities in the PDE — the latter point being a common feature of screened scalar-tensor models.

However, the mathematical grounds and numerical techniques discussed in this chapter only apply to the case where the PDE problem is formulated on a bounded domain. This translates into the inability of standard FEM to deal with problems posed on unbounded sets, i.e. PDEs that come with asymptotic boundary conditions, which was one the required specifications of the numerical tool under development. In the following chapter, we study ways into which FEM can be extended to account for those.

Chapter 3

Problems posed on unbounded domains

Outline of the current chapter

One of the very first assumptions we made when presenting the finite element method in Chapt. 2 was that the domain Ω over which the PDE is solved had to be *bounded*. The goal of the present chapter is to relax this restrictive hypothesis, as several PDEs arising in the context of modified gravity theories are posed on unbounded spatial regions. In particular, we strive to maintain the same level of mathematical rigor as in the previous chapter in the presentation of the explored numerical techniques.

3.1 Problem statement and state of the art

3.1.1 Motivations

PDE problems are sometimes formulated on unbounded regions of space Ω, in which case their well-posedness generally hinges on the specification of asymptotic boundary conditions. To give an example, the latter sometimes take the form

$$u(\mathbf{x}) \underset{\|\mathbf{x}\|\to+\infty}{\longrightarrow} u_\infty \in \mathbb{R}\,, \tag{3.1}$$

although many other behaviors may be sought. This situation is of much interest to us in the context of modified theories of gravity. For instance, the Poisson equation (1.4) governing the Newtonian potential Φ is usually supplemented with condition that Φ decays to zero far from the sources (in the non-cosmological case). Likewise in scalar-tensor models, the stationary Klein–Gordon equation (1.52) driving the dynamics of the scalar field ϕ cannot always be solved on bounded domains for the lack of physical conditions prescribed at their boundary. In such situations, one cannot but resort to using an asymptotic prescription, namely that the field goes to the value that minimizes its effective potential at infinity [see e.g. Eq. (1.120) and refer to the discussion we had in Sec. 1.2.2]. Therefore, imposing asymptotic boundary conditions is dictated by both the necessity to deal with well-posed problems and the desire to obtain physically meaningful solutions. This point is illustrated on a simple ODE toy-problem in Box H.

Box H: Asymptotic condition and uniqueness of the solution [example]

Consider the following ODE problem posed on $\mathbb{R}_+$

$$\begin{cases} u''(x) - u(x) = 0 \\ u(0) = 1 \end{cases}, \tag{3.2}$$

whose solutions are of the form $u(x) = c\mathrm{e}^x + (1-c)\mathrm{e}^{-x}$, $c \in \mathbb{R}$. The set of solutions can be reduced upon specifying an asymptotic behavior:

- the only solution satisfying $u(x) \to 0$ as $x \to +\infty$ is $u(x) = \mathrm{e}^{-x}$;
- there is no solution satisfying $u(x) \to \alpha \in \mathbb{R}^*$ as $x \to +\infty$;
- there is an infinity of solutions satisfying $u(x) \to \pm\infty$ as $x \to +\infty$;
- there is only one solution satisfying $u(x)/\mathrm{e}^x \to 1$ as $x \to +\infty$.

In particular, this example highlights the fact that, for problems posed on unbounded domains (here $\mathbb{R}_+$), reducing the set of solutions to a singleton depends on the specific form of the asymptotic boundary condition employed.

From a numerical view point however, and more specifically in the framework of FEM, enforcing asymptotic boundary conditions is not as straightforward as imposing 'standard' boundary conditions (e.g. Dirichlet, see Sec. 2.1.3). Indeed, meshing an unbounded domain would require an infinite number of elements, which obviously cannot fit into finite-memory computers. The naive workaround for this issue is to *(i)* truncate the unbounded domain Ω at a finite distance $R_c > 0$, i.e. $\Omega \to \Omega \cap \mathcal{B}(R_c)$ and *(ii)* replace the asymptotic condition (3.1) by a Dirichlet boundary condition $u = u_\infty$ at the boundary of the cropped domain. This procedure has several issues:

1. While the resulting boundary value problem fits into the standard FEM framework laid out in Chapt. 2, it is only an approximation to the original unbounded problem. The exact solution of the former can therefore be quite different from the exact solution of the latter — see Fig. 1.6 for an illustration of this phenomenon at play on a radial Poisson equation.

2. For this 'truncation error' to be small, the truncation radius R_c must sufficiently large, which a priori translates to a higher computational cost.

3. In practice, this error cannot be estimated quantitatively without additional tricks. In that sense, the process of selecting an adequate size for the truncated domain is a blind experiment.

4. If the unbounded domain is indeed truncated to a ball, setting $u = u_\infty$ on $\mathcal{S}(R_c) = \partial\mathcal{B}(R_c)$ wantonly imposes spherical symmetry in a direct neighborhood of the boundary. This might be detrimental to the study of problems that only slightly deviate from spherical symmetry — see Chapt. 5 for a concrete example.

3.1.2 The landscape of proposed solutions

In the face of the aforementioned issues, several approaches have been developed in the past decades to take better account of asymptotic conditions than the naive truncation idea outlined above. These approaches can be roughly divided into two categories.

The first one consists of those techniques based on the artificial truncation of the computational domain at some finite distance. Let us mention artificial boundary conditions (see e.g. Refs. [219–224]), absorbing boundary conditions or perfectly matched layers (see e.g. Refs. [225–227]) for wave-like equations.

The second category encompasses the techniques that manage to preserve the unbounded nature of the domain. It includes boundary element techniques (see e.g. Refs. [228, 229]) for exterior problems, infinite elements (see e.g. Refs. [224, 230–232]), spectral methods (see e.g. Refs. [233–236]), and techniques based on the mapping of the unbounded domain into a bounded one. Among the latter are compactification-based techniques: they include compactifications of the whole domain (see e.g. Refs. [237–239]) and approaches based on the inversion of some exterior unbounded domain (see e.g. Refs. [137, 240–248]).

3.1.3 Organization of the present chapter

This chapter presents the various approaches explored during this PhD work, all of which are based on compactification transforms — i.e. the mapping of the unbounded domain into a bounded one. While the ideas to be outlined here are quite general and could probably be applied to a wide range of PDE problems, there are multiple problem-dependent subtleties which prevent us from directly discussing the generic case of a second-order elliptic PDE [see Eq. (2.3)]. Instead, we focus our attention to two specific examples which are most relevant to this PhD work.

Example 3.1. The n-dimensional linear Klein–Gordon equation reads

$$-\Delta u + d(\mathbf{x})u = f(\mathbf{x}) \text{ in } \Omega = \mathbb{R}^n \,. \tag{3.3}$$

The Yukawa potential Eq. (1.106) obeys a linear Klein–Gordon equation with $d(\mathbf{x}) = m_\phi^2$ and $f(\mathbf{x}) = -\beta\rho(\mathbf{x})/M_{\rm Pl}$. Moreover, the linearized chameleon field equation at the $k^{\rm th}$ iteration of the Newton method [see e.g. Eq. (2.74)] can be recast in the form of Eq. (3.3) with

$$d(\mathbf{x}) = \frac{n+1}{\alpha} u_k^{-(n+2)}(\mathbf{x}) \qquad \text{and} \qquad f(\mathbf{x}) = \frac{n+2}{\alpha} u_k^{-(n+1)}(\mathbf{x}) - \rho(\mathbf{x}) \,.$$

Example 3.2. The Poisson equation reads

$$-\Delta u = f(\mathbf{x}) \text{ in } \Omega = \mathbb{R}^n \,. \tag{3.4}$$

The Newtonian potential obeys a Poisson equation [see Eq. (1.4)] with $f(\mathbf{x}) = -4\pi G\rho(\mathbf{x})$.

Due to their relevance in the context of studying scalar-tensor theories of gravity, we have deliberately chosen these two examples to showcase the numerical techniques involved in this chapter. In particular, the various proofs provided thereafter sometimes rely on assumptions that would not be valid in the general case — we will endeavor to specify when this is the case.

Let us explain in more details the way the present chapter is organized. In the same spirit as Chapt. 2, we first start by looking for suitable functional spaces for writing the weak formulations associated with Eqs. (3.3–3.4). Note that at this stage, the choice of such spaces is guided by the sole aim of obtaining well-posed weak problems, not by numerical considerations. Schematically, we will specify for each problem *(i)* a space W, *(ii)* a bilinear form $a(\cdot,\cdot)$ defined over $W \times W$ and *(iii)* a linear form $l(\cdot)$ defined over W, such that the weak formulation

$$\text{“ Find } u \in W \text{ such that for all } v \in W \,,\; a(u,v) = l(v) \text{”}$$

is well-posed in the sense of Hadamard (see Box D). Only then do we start addressing the practical issue raised in Sec. 3.1.1, namely that one cannot apply an essentially finite numerical process to an infinite domain. In that perspective, Sec. 3.3 illustrates several approaches based on compactification transforms, whereby Ω is mapped to a bounded domain by means of suitable coordinate transforms. This change of coordinates is to be performed in the integrals appearing in the definition of the (bi)linear forms $a(\cdot,\cdot)$ and $l(\cdot)$. In Sec. 3.3.2, we single out a particularly nice transformation — called the *Kelvin inversion* — which we adopt for the remainder of this chapter. The next important step is the definition of W^h, a discrete counterpart of W. Although there is no single way to proceed, the choice of W^h is not a blinded one but is rather guided by two conditions:

1. the inclusion $W^h \subset W$ (*conformal approximation*);

2. the possibility of returning to the FE framework presented in Sec. 2.1.3, in particular, the possibility of using usual polygonal meshes and $\mathbb{P}_k$ elements, so that one can build on top of virtually any already-existing FEM code.

The second condition is mostly motivated by practical considerations: we do not want to resort to the internal modifications of an existing FEM code.[1] This process — which is mostly inspired by the *inverted finite elements method* (*ifem*) introduced by T. Boulmezaoud in Ref. [242] — is thoroughly explained in Sec. 3.4. In Sec. 3.5, we present a novel approach developed during this PhD called the *alternate inverted finite elements method* (*a-ifem*) which builds on top of *ifem* and domain decomposition techniques. Finally, Sec. 3.6 is dedicated to the presentation of results from numerical experiments we conducted: efficiency of the methods, comparison of *a-ifem* against *ifem*, error estimates and influence of auxiliary parameters appear on the list of issues addressed.

3.2 Functional framework

Given a PDE problem, regardless of the underlying numerical technique employed, it is crucial to specify the functional space in which we look for solutions. Back in Chapt. 2.1, we saw that the choice of a suitable Sobolev space was key to obtaining well-posed problems in bounded domains. In particular, Dirichlet boundary conditions — also called *essential* boundary conditions — are encoded in the definition of H^1_0 [see Eqs. (2.21, 2.26)] and turn out to be, indeed, essential to the derivation of Poincaré's inequality (see Box F). When $\Omega = \mathbb{R}^n$, the domain has no boundary and the concept of Dirichlet boundary conditions no longer makes sense. Instead, they have to be replaced by asymptotic conditions (see the example discussed in Box H). In a similar way to the bounded case, these asymptotic conditions are going to be encoded in the very definition of functional spaces.

To illustrate our point, we delve into examples 3.1 and 3.2. The latter requires the introduction of so-called *weighted Sobolev spaces*, which we have already encountered Sec. 2.3 showing how to take advantage of problems' symmetries, yet barely took the time to discuss.

3.2.1 (Why) do we need new function spaces?

The Klein–Gordon example

Let us look at example 3.1 first and consider W an abstract functional space in which all the expressions we are about to write make sense. The weak formulation of Eq. (3.3) involves the (bi)linear forms $a_{\text{KG}}(\cdot,\cdot)$ and $l_{\text{KG}}(\cdot)$, defined for all u, $v \in W$ by

$$a_{\text{KG}}(u,v) = \int_{\mathbb{R}^n} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^n} d(\mathbf{x}) uv \,\mathrm{d}\mathbf{x} \qquad \text{and} \qquad l_{\text{KG}}(v) = \int_{\mathbb{R}^n} f(\mathbf{x}) v \,\mathrm{d}\mathbf{x}\,. \tag{3.5}$$

For these integrals to exist, a sensible choice for W would be $H^1(\mathbb{R}^n)$, i.e. the space of all generalized functions such that

$$\int_{\mathbb{R}^n} |u(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} < +\infty \qquad \text{and} \qquad \int_{\mathbb{R}^n} \|\boldsymbol{\nabla} u(\mathbf{x})\|^2 \,\mathrm{d}\mathbf{x} < +\infty\,, \tag{3.6}$$

which is a Hilbert space when equipped with the inner product given by Eq. (2.22) with $\Omega = \mathbb{R}^n$. Indeed, under the further minimal assumptions that $d \in L^\infty(\mathbb{R}^n)$ and $f \in L^2(\mathbb{R}^n)$, setting $W = H^1(\mathbb{R}^n)$ is sufficient for the (bi)linear forms given by Eq. (3.5) to be well-defined and continuous with respect to the H^1-norm.[2] In the following, we shall make use of a constant $d_\infty \geq d(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$

The only property that we have not examined yet is the coercivity of a_{KG}. In the bounded case, the Poincaré inequality (see Box F) was key in proving the coerciveness of the bilinear form associated with the Poisson equation. Here, if we further assume that there exists $d_0 > 0$ such that for all $\mathbf{x} \in \mathbb{R}^n$, $d(\mathbf{x}) > d_0$, we simply no longer need such inequality. As a matter of fact, under this additional assumption, one has

$$a_{\text{KG}}(u,\, u) = \int_{\mathbb{R}^n} \|\boldsymbol{\nabla} u\|^2 \,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^n} d(\mathbf{x}) |u|^2 \,\mathrm{d}\mathbf{x} \geq \min(1,\, d_0)\, \|u\|^2_{H^1}\,. \tag{3.7}$$

Therefore, Lax–Milgram theorem applies and the problem is well-posed. Additionally, we will need the following lemma.

Lemma 3.1. The space of test functions $\mathscr{D}(\mathbb{R}^n)$ is dense in $H^1(\mathbb{R}^n)$.

[1] One could imagine implementing new basis functions to comply with the condition $W^h \subset W$. This is far from being impossible but if alternative solutions not requiring such technical tweaks exist, they shall be preferred here.

[2] This is indeed straightforward to prove using the Cauchy–Schwarz inequality.

The Poisson example

As done above, suppose that W is a suitable functional space for writing the weak formulation of the Poisson problem. Then, corresponding (bi)linear forms read, for all u, $v \in W$,

$$a_{\mathrm{P}}(u,v) = \int_{\mathbb{R}^n} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} \qquad \text{and} \qquad l_{\mathrm{P}}(v) = \int_{\mathbb{R}^n} f(\mathbf{x}) v \,\mathrm{d}\mathbf{x}\,. \tag{3.8}$$

The Poisson example 3.2 requires a different functional framework, since $H^1(\mathbb{R}^n)$ fails to capture the behavior of the physical solution at infinity. Indeed, consider the case of the Newtonian potential sourced by a solid sphere immersed in vacuum in three dimensions. Outside the sphere, the potential behaves as $\Phi \sim -A/r$ (for some constant $A \in \mathbb{R}$) and so Φ does not even belong to $L^2(\mathbb{R}^3)$. Consequently, it makes no sense to look for physical solutions in $H^1(\mathbb{R}^3)$, and new functional spaces should be used instead.

Beyond this issue, it is well-known that Poincaré's inequality (see Box F) does not hold in $\mathbb{R}^n$. Consequently, the proof of the coercivity of the bilinear form in bounded domains cannot be adapted to the present case.

These two issues are addressed by introducing the notion of *weighted Sobolev spaces*, which is a generalization of the usual Sobolev spaces [Eq. (2.23)].

3.2.2 Weighted Sobolev spaces

Definitions, properties

A few definitions are in order. First, we give a meaning to the notion of weight.

Definition 3.1. We say that a real-valued function on $\Omega \subseteq \mathbb{R}^n$ is a weight if it is locally integrable on Ω and strictly positive almost everywhere.

Definition 3.2. Given $\omega = (\omega_0, \omega_1)$ a pair of two weight functions, we denote by $L^2(\Omega, \omega_0)$ the set of all functions $v\colon \Omega \to \mathbb{R}$ such that

$$+\infty > \int_\Omega |v(\mathbf{x})|^2 \,\omega_0(\mathbf{x}) \,\mathrm{d}\mathbf{x} =: \|v\|^2_{L^2(\Omega,\,\omega_0)}\,;$$

and by $W_\omega(\Omega)$ the set of all functions $v \in L^2(\Omega, \omega_0)$ such that for all $i \in \{1, \cdots, n\}$

$$+\infty > \int_\Omega |D_{x_i} v(\mathbf{x})|^2 \,\omega_1(\mathbf{x}) \,\mathrm{d}\mathbf{x}\,,$$

where $D_{x_i} v$ denotes the weak partial derivative of v in the direction i.

The space $W_\omega(\Omega)$ is called a weighted Sobolev space. It can readily be equipped with the following norm [249, 250]

$$\|v\|_{W_\omega(\Omega)} \coloneqq \left(\int_\Omega |v(\mathbf{x})|^2 \,\omega_0(\mathbf{x}) \,\mathrm{d}\mathbf{x} + \sum_{i=1}^{n} \int_\Omega |D_{x_i} v(\mathbf{x})|^2 \,\omega_1(\mathbf{x}) \,\mathrm{d}\mathbf{x} \right)^{1/2}, \tag{3.9}$$

and semi-norm

$$|v|_{W_\omega(\Omega)} \coloneqq \left(\int_\Omega \omega_1(\mathbf{x}) \,\|\boldsymbol{\nabla} v\|^2 \right)^{1/2}. \tag{3.10}$$

The weights we use in the remainder of this work are all such that for all $\mathbf{x} \in \Omega$, $0 < \omega_k(\mathbf{x}) \leq 1$, $k \in \{0, 1\}$. The k subscript is dropped when the two weights are chosen as equals. Let us also report a useful lemma that will be used thereafter.

Lemma 3.2. Suppose that for $k \in \{0, 1\}$, ω_k and ω_k^{-1} are locally integrable functions on Ω. Then, $W_\omega(\Omega)$ equipped with the norm $\|\cdot\|_{W_\omega}$ [see Eq. (3.9)] is a uniformly convex Banach space — see e.g. Refs. [250, 251].

Provided that the weight functions satisfy the hypothesis of Lemma 3.2, W_ω can be promoted to rank of Hilbert space when equipped with the inner product

$$\begin{aligned} \langle \cdot, \cdot \rangle_{W_\omega} \colon W_\omega \times W_\omega &\to \mathbb{R} \\ (u,\, v) &\mapsto \int_\Omega \omega_0(\mathbf{x}) uv \,\mathrm{d}\mathbf{x} + \int_\Omega \omega_1(\mathbf{x}) \,\boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x}\,. \end{aligned} \tag{3.11}$$

Remark 3.1. We note that for any function $v \in W_\omega$ and for any compact set $K \in \Omega$, $v_{|K} \in H^1(K)$, where $H^1(K)$ is the usual Sobolev space defined over K. In other words, the role of the weights is to specify the asymptotic behavior of functions belonging to W_ω at infinity. They have no influence on their local properties, which are identical to those of H^1.

Remark 3.2. $H^1(\mathbb{R}^n)$ is a particular case of weighted Sobolev space for which the weights are unitary, i.e. $\omega_0 \equiv 1 \equiv \omega_1$.

Back to the Poisson problem

There is an extended literature devoted to the study of weighted Sobolev spaces for PDEs involving the Laplace operator and defined over unbounded sets of $\mathbb{R}^n$ [252–257]. The weight functions employed in these references are constructed using powers of the two functions

$$\rho(\mathbf{x}) \coloneqq \sqrt{1+\|\mathbf{x}\|^2} \qquad \text{and} \qquad \lg(\mathbf{x}) \coloneqq \ln(2+\|\mathbf{x}\|^2)\,. \tag{3.12}$$

In line with Definition 3.2, the adequate weighted Sobolev space for studying the Poisson problem in $\mathbb{R}^n$ makes use of $\omega = (\omega_0, \omega_1)$ with

$$\omega_0 = \rho^{-2}\,, \qquad \omega_1 \equiv 1 \tag{3.13}$$

for the case $n \in \{1, 3\}$, and

$$\omega_0 = (\rho \lg)^{-2}\,, \qquad \omega_1 \equiv 1 \tag{3.14}$$

for the case $n = 2$ (see e.g. Refs. [252, 257]). We set W^{P}_{ω} the weighted Sobolev space defined with the pair of weights given by Eq. (3.13) or Eq. (3.14) depending on the dimension of the problem. As explained in Ref. [257], the choice of weight exponents in the definition of W^{P}_{ω}, and in particular the introduction of a logarithmic weight $\lg(\mathbf{x})$ in the critical case $n = 2$ are dictated by the generalized Hardy inequalities [258]. Precisely, one can show that there exists a constant $C > 0$ such that for all $u \in W^{\mathrm{P}}_{\omega}$ [254, 255],

$$|u|^2_{W_\omega} \geq C\,\|u\|^2_{L^2(\mathbb{R}^n,\,\omega_0)}\,, \qquad \text{i.e.} \quad \int_{\mathbb{R}^n} \|\boldsymbol{\nabla} u\|^2\,\mathrm{d}\mathbf{x} \geq C \int_{\mathbb{R}^n} \omega_0(\mathbf{x})\,|u|^2\,\mathrm{d}\mathbf{x}\,. \tag{3.15}$$

This inequality is reminiscent of the traditional Poincaré inequality (2.29), except the norms involved in Eq. (3.15) are different. In that sense, inequalities of the form Eq. (3.15) go under the same of 'Poincaré-type inequalities' or 'generalized Poincaré inequalities'.

Going back to Example 3.2, let us show that W^{P}_{ω} as defined above ticks all the boxes. First, the bilinear form a_{P} defined over $W^{\mathrm{P}}_{\omega} \times W^{\mathrm{P}}_{\omega}$ is well-defined and continuous. Moreover, it is coercive over W^{P}_{ω} thanks to the Poincaré-type inequality (3.15). Second, for the linear form l_{P} to be well-defined, we demand that the data f be in $W'_{\omega}(\mathbb{R}^n)$, the dual space of $W_{\omega}(\mathbb{R}^n)$. To be more explicit (but a little more restrictive), we can demand that

$$f \in L^2\big(\mathbb{R}^n,\,\omega_0^{-1}\big)\,, \quad \text{i.e.} \quad \int_{\mathbb{R}^n} \omega_0^{-1}(\mathbf{x})\,|f(\mathbf{x})|^2\,\mathrm{d}\mathbf{x} < +\infty\,. \tag{3.16}$$

Under such an assumption, l_{P} is well-defined and continuous on W^{P}_{ω}. Therefore, Lax–Milgram theorem applies and the problem is well-posed. Additionally, one can check that Newtonian potentials of the form $\Phi \sim -A/r$ (for sufficiently large r) do indeed belong to W^{P}_{ω} thanks to the weight functions (ω_0, ω_1).

Remark 3.3. Different notations are used in the literature to refer to the space W^{P}_{ω} with ω given by Eqs. (3.13–3.14), such as $\mathcal{H}^1_{-1,0}$ in Ref. [242] or $W^{1,2}_{0,0}$ in Refs. [252, 255, 256]. In the case $n = 2$, it is sometimes referred to as $W^1_{\log}$, see e.g. Ref. [245].

Additionally, in the subsequent discussion, we will need the following lemma.

Lemma 3.3. The space of test functions $(\mathbb{R}^n)$ is dense in W^{P}_{ω}, with ω given by Eqs. (3.13–3.14) — see e.g. Refs. [252, 256].

3.2.3 A word about the integration by parts in $\mathbb{R}^n$

One important aspect that has been completely overlooked in the above examples is the derivation of the weak formulations from the strong problem. As a matter of fact, we assumed without even questioning that it was legitimate to integrate by parts the divergence term in $\mathbb{R}^n$. Precisely, we implicitly assumed that

$$\forall u \in \mathcal{C}^2(\mathbb{R}^n) \cap W_\omega \text{ such that } \Delta u \in L^2\big(\mathbb{R}^n,\,\omega_0^{-1}\big) \text{ and } \forall v \in W_\omega\,,\ \int_{\mathbb{R}^n} \Delta u\, v\,\mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^n} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v\,\mathrm{d}\mathbf{x}\,, \tag{3.17}$$

where W_ω denotes the adequate weighted Sobolev space for either the linear Klein–Gordon equation or the Poisson equation. Note that this equality is sometimes referred to as Green's formula.

In order to examine the validity of the identity (3.17), let us write the usual integration by parts formula on a ball of radius R, which we know to be valid. For $u \in \mathcal{C}^2(\mathbb{R}^n) \cap W_\omega$ such that $\Delta u \in L^2\big(\mathbb{R}^n,\,\omega_0^{-1}\big)$ and for $v \in W_\omega$, we have

$$\int_{\mathcal{B}(R)} \Delta u\, v\,\mathrm{d}\mathbf{x} = -\int_{\mathcal{B}(R)} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v\,\mathrm{d}\mathbf{x} + \int_{\mathcal{S}(R)} \frac{\partial u}{\partial n} v\,\mathrm{d}\gamma\,. \tag{3.18}$$

Given the spaces in which u and v live, the integrals over $\mathcal{B}(R)$ converge to the integrals over the whole space as

$R \to +\infty$. Consequently, the surface term in the rhs of Eq. (3.18) also converges as $R \to +\infty$, and its limit has to be zero for equality (3.17) to hold.

This last point can be proven using a classical density argument. Consider the function

$$\begin{aligned} F\colon W_\omega(\mathbb{R}^n) &\to \mathbb{R} \\ v &\mapsto \lim_{R\to+\infty} \int_{\mathcal{S}(R)} \frac{\partial u}{\partial n}\, v \,\mathrm{d}\gamma\,. \end{aligned} \tag{3.19}$$

Our goal is to show that $F \equiv 0$. In that perspective, let us note that

1. for any $v \in \mathscr{D}(\mathbb{R}^n)$, $F(v) = 0$ since v is compactly supported;
2. in virtue of Lemmas 3.1 and 3.3, $\mathscr{D}(\mathbb{R}^n)$ is dense in $W_\omega(\mathbb{R}^n)$.

Consequently, it is sufficient to show that the linear form F is continuous with respect to the norm $\|\cdot\|_{W_\omega}$. Let $v \in W_\omega(\mathbb{R}^n)$ and $R > 0$, we have

$$\begin{aligned} \left| \int_{\mathcal{S}(R)} \frac{\partial u}{\partial n}\, v \,\mathrm{d}\gamma \right| &\leq \left| \int_{\mathcal{B}(R)} \Delta u\, v \,\mathrm{d}\mathbf{x} \right| + \left| \int_{\mathcal{B}(R)} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} \right| \\ &\leq \|\Delta u\|_{L^2\left(\mathcal{B}(R),\, \omega_0^{-1}\right)} \|v\|_{L^2\left(\mathcal{B}(R),\, \omega_0\right)} + \|\boldsymbol{\nabla} u\|_{L^2\left(\mathcal{B}(R)\right)} \|\boldsymbol{\nabla} v\|_{L^2\left(\mathcal{B}(R)\right)} \\ &\leq \|\Delta u\|_{L^2\left(\mathbb{R}^n,\, \omega_0^{-1}\right)} \|v\|_{L^2\left(\mathbb{R}^n,\, \omega_0^{-1}\right)} + |u|_{W_\omega}\, |v|_{W_\omega} \\ &\leq \left(\|\Delta u\|_{L^2\left(\mathbb{R}^n,\, \omega_0^{-1}\right)} + |u|_{W_\omega} \right) \|v\|_{W_\omega} \end{aligned} \tag{3.20}$$

Inequality (3.20) remains true in the limit $R \to +\infty$, so that F is indeed continuous. This ends the proof.

3.3 Approaches based on compactification transforms

Now that we have established a clear functional framework, let us return to practical considerations. As mentioned above, a convenient way of preserving the 'unboundedness' of the computational domain is to *compactify* it by means of an adequate mapping. In this section, we present two examples that were considered in this PhD work:

1. Compactification of the whole space using a mapping of the form $\mathbf{x} \in \mathbb{R}^n \mapsto \mathbf{x}/(1 + \|\mathbf{x}\|)$.
2. Splitting the domain into an interior part and an exterior part $\bar{\Omega} = \bar{\Omega}_{\text{int}} \cup \bar{\Omega}_{\text{ext}}$ and then applying an *inversion* transform to the exterior domain Ω_{ext}. This process results in two bounded domains which have to somehow exchange information at their shared boundary.

At the level of the weak form, we perform the corresponding change of variable in the integrals associated with the (bi)linear forms. Doing so generally leads to the appearance of singular coefficients in the integrands. This caveat is discussed (from a numerical perspective) in Sec. 3.3.3.

3.3.1 Compactification of the whole domain

Example of a compactification transform

Compactifying the whole space — i.e. applying a global coordinate transform to turn $\mathbb{R}^n$ into a bounded domain — is perhaps the most intuitive idea one can have. This technique is very useful in physics, one of the most famous illustration of which being Penrose diagrams, for capturing the causal relations between different points in spacetime through a conformal treatment of infinity [259].[3] It can also be leveraged in numerical simulations. For instance, Ref. [260] solves hyperbolic equations on unbounded domains with a compactification transform involving both time and space coordinates. Similarly, we solve a non-linear Klein–Gordon equation in Ref. [137] using a compactification transform in spherical coordinates for an axisymmetric configuration (see Sec. 2.3.2).

In order to get a clearer representation of what a compactification of the whole space may look like, Fig. 3.1 illustrates the action of

$$\begin{aligned} \mathcal{T}\colon \mathbb{R}^n &\to \hat{\Omega} \\ \mathbf{x} &\mapsto \frac{R_c}{1 + \|\mathbf{x}\|}\mathbf{x} \end{aligned} \tag{3.21}$$

[3] The advantage of this class of transformations is that they leave the light-like geodesics invariant for $n = 4$.

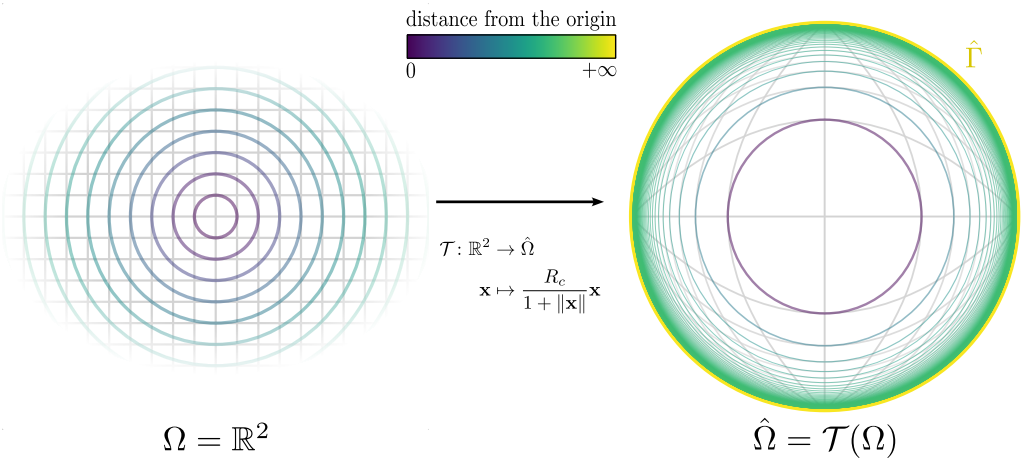


Figure 3.1: Compactification of the whole space $\Omega = \mathbb{R}^2$ by means of the $\mathcal{T}$ transform. The resulting domain $\hat{\Omega}$ is the open disk of radius R_c, with boundary $\hat{\Gamma}$ representing spatial infinity. In $\hat{\Omega}$, the angles are not preserved ($\mathcal{T}$ is not a conformal transformation) and the distances are altered: lines of iso-radius get squashed as one approaches the boundary $\hat{\Gamma}$.

on $\mathbb{R}^2$, for some $R_c > 0$. The 2D plane is complemented by a Cartesian grid and concentric circles spanning from the origin to infinity (whose radii grow at an arithmetic rate). The latter are mapped to circles again, with R_c as the upper limit on the biggest radius, while the former gets non-conformally distorted. The overall resulting domain is the open disk of radius R_c.

Illustration on the Klein–Gordon problem

Following what we did in Ref. [137], let us apply the transform (3.21) to the (bi)linear forms featured in the Klein–Gordon problem a_{KG} and l_{KG} [see Eq. (3.5)]. Doing this in Cartesian coordinates is not handy though, because the resulting expressions involve complicated terms. If anything, that would obscure our point. Instead, it is much more natural to consider the use of spherical coordinates, for which the compactification only alters the radial coordinate r as

$$\mathcal{T}_r : \begin{array}{c} \Pi_\infty \to \Pi_{R_c} \\ (r,\, \theta,\, \varphi) \mapsto (\eta,\, \theta,\, \varphi) \end{array}, \qquad \text{with } \eta = \frac{r R_c}{1+r} \tag{3.22}$$

and $\Pi_\infty :=]0, +\infty[\times]0,\, \pi[\times]0,\, 2\pi[$ and $\Pi_{R_c} :=]0, R_c[\times]0,\, \pi[\times]0,\, 2\pi[$. Under the further assumption of spherical symmetry (see Sec. 2.3.1), this yields

$$\frac{a_{\text{KG}}(u,\, v)}{4\pi} = \int_0^{+\infty} r^2 (\partial_r u)(\partial_r v)\, \mathrm{d}r + \int_0^{+\infty} r^2 d(r) uv\, \mathrm{d}r = \int_0^{R_c} \frac{\eta^2}{R_c} (\partial_\eta \hat{u})(\partial_\eta \hat{v})\, \mathrm{d}\eta + \int_0^{+\infty} \frac{R_c \eta^2}{(R_c - \eta)^4} \hat{d}(\eta) \hat{u}\hat{v}\, \mathrm{d}\eta \tag{3.23}$$

$$\frac{l_{\text{KG}}(v)}{4\pi} = \int_0^{+\infty} r^2 f(r) v\, \mathrm{d}r = \int_0^{R_c} \frac{R_c \eta^2}{(R_c - \eta)^4} \hat{f}(\eta) \hat{v}\, \mathrm{d}\eta\,, \tag{3.24}$$

where we use the hat-notation to distinguish functions expressed in the new coordinate, e.g. for spherical coordinates $\mathbf{s}$, $u(\mathbf{s}) = \hat{u}\big(\mathcal{T}_r(\mathbf{s})\big)$. As may have been anticipated, the compactification $r \to \eta$ yields coefficients in the above integrals that are not bounded in the neighborhood of $\eta = R_c$. Put another way, the mapping is singular on the boundary representing spatial infinity. Nonetheless, we know for a fact that these integrals are all perfectly well-defined as long as $u,\, v \in H^1(\mathbb{R}^3)$, $d \in L^\infty(\mathbb{R}^3)$ and $f \in L^2(\mathbb{R}^n)$ — refer to Sec. 3.2.1. The difficult part consists in finding a suitable finite-dimensional subspace of $H^1(\mathbb{R}^3)$. However, we do not delve further into this topic at this stage as it will be covered in details in Sec. 3.4.

Finally, note that the compactification transform [Eqs. (3.21, 3.22)] discussed here is just one example among others. For instance, one could imagine compactifying each individual coordinate separately as in, e.g., $(x,\, y,\, z) \mapsto [\arctan(x), \arctan(y), \arctan(z)]$. In such a scenario, $\mathbb{R}^3$ would be mapped to a finite cube instead of a ball as was the case for the $\mathcal{T}$ transform (3.21). The main drawback from compactifications of the whole space is that the geometry in the resulting bounded domain is not intuitive (see e.g. Fig. 3.1). On the one hand, there is no denying that the $\mathcal{T}$ transform is a bijection, allowing one to go back and forth between the two domains without any loss of information. On the other hand, the process of constructing suitable meshes and the whole post-processing stage is much more cumbersome to carry out in the compactified space. This practical issue can be circumvented by applying the compactification only to the exterior of a ball centered at the origin instead of

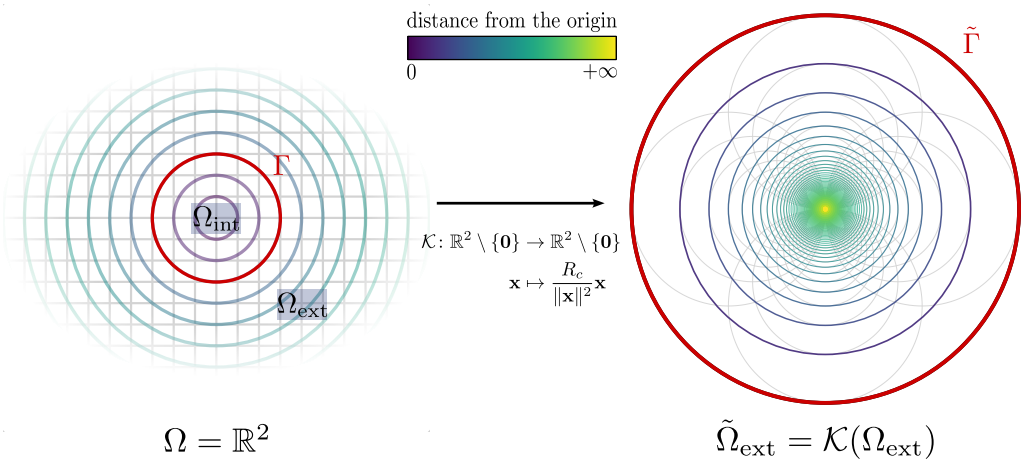


Figure 3.2: Compactification of the exterior domain $\Omega_{\text{ext}} = \mathbb{R}^n \setminus \overline{\mathcal{B}}(R_c)$ by means of the Kelvin inversion $\mathcal{K}$. The resulting *inverted exterior domain* $\tilde{\Omega}_{\text{ext}}$ is $\mathcal{B}(R_c) \setminus \{\mathbf{0}\}$, where the closure point $\mathbf{0}$ represents spatial infinity in the real space. Note that $\mathcal{K}$ is a conformal mapping so that the angles are preserved.

the whole space, as suggested in Ref. [260]. The next technique we present also avoids this issue.

3.3.2 Domain splitting and Kelvin inversion

Inversion transform

We start by partitioning the whole domain Ω into two subdomains. Let $R_c > 0$ be a truncation radius. We define $\Omega_{\text{int}} \coloneqq \Omega \cap \mathcal{B}(R_c)$ and $\Omega_{\text{ext}} \coloneqq \Omega \setminus \bar{\Omega}_{\text{int}}$ so that *(i)* $\bar{\Omega} = \bar{\Omega}_{\text{int}} \cup \bar{\Omega}_{\text{int}}$ and *(ii)* Ω_{int} encapsulates the various sources of physical interest. R_c is chosen large enough so that $\mathbb{R}^n \setminus \mathcal{B}(R_c) = \bar{\Omega}_{\text{ext}}$. The resulting interface $\Gamma \coloneqq \partial\bar{\Omega}_{\text{int}} \cap \partial\bar{\Omega}_{\text{ext}}$ is nothing but $\mathcal{S}(R_c)$, the sphere of radius R_c and centered at the origin in $\mathbb{R}^n$. The *interior domain* Ω_{int} is bounded, the *exterior domain* Ω_{ext} is not.

The exterior domain Ω_{ext} can be mapped to a bounded domain using the *Kelvin inversion* $\mathcal{K}$, reading

$$\begin{aligned} \mathcal{K}\colon \mathbb{R}^n \setminus \{\mathbf{0}\} &\to \mathbb{R}^n \setminus \{\mathbf{0}\} \\ \mathbf{x} &\mapsto \frac{R_c^2}{\|\mathbf{x}\|^2}\mathbf{x} =: \boldsymbol{\xi}\,. \end{aligned} \tag{3.25}$$

This involution maps Ω_{ext} to $\tilde{\Omega}_{\text{ext}} = \mathcal{K}(\Omega_{\text{ext}}) = \mathcal{B}(R_c) \setminus \{\mathbf{0}\}$. The resulting bounded domain $\tilde{\Omega}_{\text{ext}}$ is referred to as the *inverted exterior domain* [261] (or *fictitious domain* in Refs. [242, 245, 246]). We further notice that the boundary Γ is invariant under $\mathcal{K}$, i.e. $\tilde{\Gamma} = \mathcal{K}(\Gamma) = \Gamma$. Fig. 3.2 illustrates the application of this inversion on $\mathbb{R}^2 \setminus \overline{\mathcal{B}}(R_c)$ fitted with a Cartesian grid and concentric circles growing at an arithmetic rate. Fig. 3.3 provides a somewhat clearer view of the main notations associated with this inversion transform.

The generic (bi)linear forms $a(\cdot,\cdot)$ and $l(\cdot)$ can be split into

$$a = a_{\text{int}} + a_{\text{ext}} \quad \text{and} \quad l = l_{\text{int}} + l_{\text{ext}}\,, \tag{3.26}$$

where, by virtue of Chasles' relation, a_{int} and l_{int} feature integrals over Ω_{int}, while a_{ext} and l_{ext} feature integrals over Ω_{ext}. The next logical step consists in applying the coordinates change $\mathbf{x} \mapsto \mathcal{K}(\mathbf{x}) = \boldsymbol{\xi}$ in the integrals associated with a_{ext} and l_{ext}. In that perspective, we extend the tilde notation to functions w as

$$\forall \boldsymbol{\xi} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\,,\ \widetilde{w}(\boldsymbol{\xi}) = w\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big)\,. \tag{3.27}$$

Illustration on the Klein–Gordon problem

Let us illustrate this procedure on the linear Klein–Gordon equation, Example 3.1, for which a_{KG} and l_{KG} are given by Eq. (3.5). For $u, v \in H^1(\mathbb{R}^n)$, one has

$$\begin{aligned} a_{\text{KG,ext}}(u,v) &= \int_{\Omega_{\text{ext}}} \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + \int_{\Omega_{\text{ext}}} d(\mathbf{x}) uv \,\mathrm{d}\mathbf{x} \\ &= \int_{\tilde{\Omega}_{\text{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2(n-2)} \widetilde{\boldsymbol{\nabla}}\tilde{u} \cdot \widetilde{\boldsymbol{\nabla}}\tilde{v} \,\mathrm{d}\boldsymbol{\xi} + \int_{\tilde{\Omega}_{\text{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \tilde{d}(\boldsymbol{\xi})\, \tilde{u}\, \tilde{v} \,\mathrm{d}\boldsymbol{\xi}\,, \end{aligned} \tag{3.28}$$

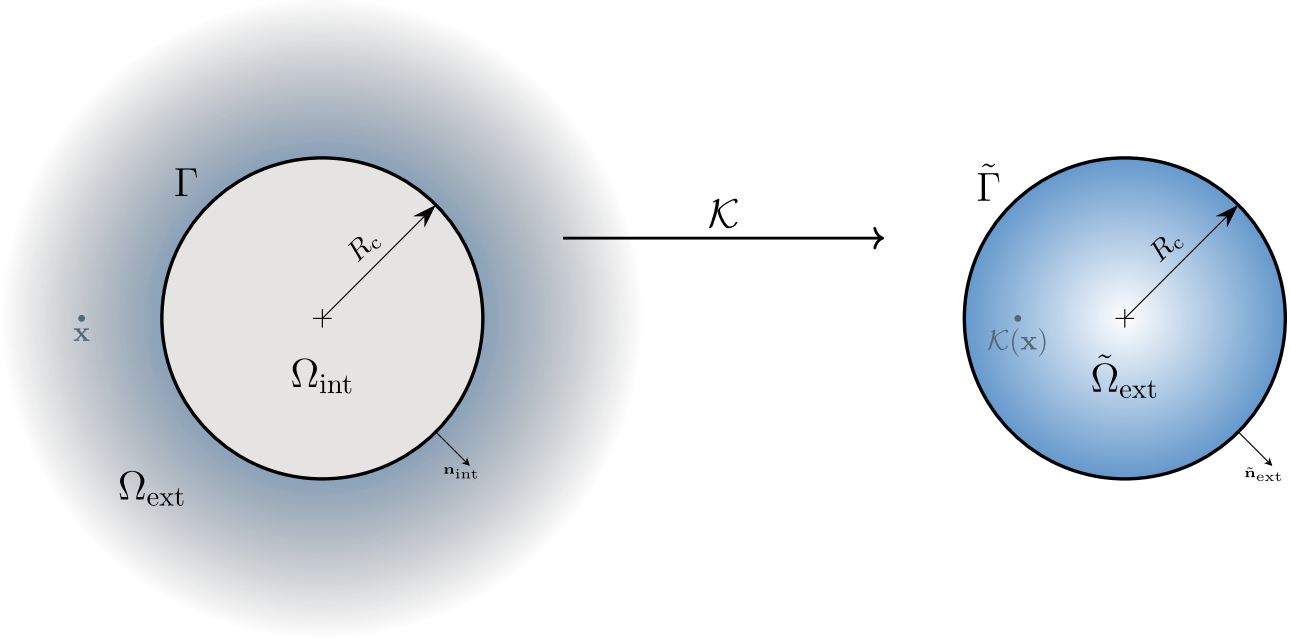


Figure 3.3: Kelvin inversion of the exterior domain.

and

$$l_{\text{KG,ext}}(v) = \int_{\Omega_{\text{ext}}} f(\mathbf{x}) v \, \mathrm{d}\mathbf{x} = \int_{\tilde{\Omega}_{\text{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \tilde{f}(\boldsymbol{\xi})\, \tilde{v} \, \mathrm{d}\boldsymbol{\xi}\,, \tag{3.29}$$

where $\widetilde{\boldsymbol{\nabla}} = (\partial_{\xi_1}, \ldots, \partial_{\xi_n})^T$ is the gradient operator in the new coordinate system. The change of coordinates carried out in these integrals is detailed in Box I below.

Box I: Kelvin inversion — change of coordinates

Performing the change of coordinates $\mathbf{x} \mapsto \mathcal{K}(\mathbf{x}) = \boldsymbol{\xi}$ in the integrals appearing in a_{ext} and l_{ext} requires, among other things, the computation of the Jacobian of the Kelvin inversion (3.25). We denote by $\mathbf{J}_{\mathcal{K}}$ and $\mathbf{J}_{\mathcal{K}^{-1}}$ the Jacobian matrix of $\mathcal{K}$ and $\mathcal{K}^{-1}$ respectively. For $\boldsymbol{\xi} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, one has

$$\mathbf{J}_{\mathcal{K}}(\mathcal{K}^{-1}(\boldsymbol{\xi})) = -\frac{1}{R_c^2}\begin{bmatrix} \xi_1^2 - \sum_{i\neq 1}\xi_i^2 & 2\xi_1\xi_2 & \cdots & 2\xi_1\xi_n \\ & \xi_2^2 - \sum_{i\neq 2}\xi_i^2 & \cdots & 2\xi_2\xi_n \\ & & \ddots & \vdots \\ & \text{SYM} & & \xi_n^2 - \sum_{i\neq n}\xi_i^2 \end{bmatrix} \quad \text{and} \quad \mathbf{J}_{\mathcal{K}^{-1}}(\boldsymbol{\xi}) = \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^4 \mathbf{J}_{\mathcal{K}}(\mathcal{K}^{-1}(\boldsymbol{\xi}))\,.$$

For u, $v \in \mathcal{C}^1(\mathbb{R}^n)$, one has

$$\begin{aligned} \boldsymbol{\nabla} u\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big) \cdot \boldsymbol{\nabla} v\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big) &= \widetilde{\boldsymbol{\nabla}}\tilde{u}(\boldsymbol{\xi})^T\, \mathbf{J}_{\mathcal{K}}\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big)\, \mathbf{J}_{\mathcal{K}}\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big)^T\, \widetilde{\boldsymbol{\nabla}}\tilde{v}(\boldsymbol{\xi}) \\ &= \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^4 \widetilde{\boldsymbol{\nabla}}\tilde{u}(\boldsymbol{\xi}) \cdot \widetilde{\boldsymbol{\nabla}}\tilde{v}(\boldsymbol{\xi})\,. \end{aligned}$$

Finally, the volume elements $\mathrm{d}\mathbf{x}$ and $\mathrm{d}\boldsymbol{\xi}$ are related through the Jacobian determinant of the mapping, reading

$$\mathrm{d}\mathbf{x} = \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \mathrm{d}\boldsymbol{\xi}\,.$$

We notice that, as was the case for the compactification transform (3.22), the change of coordinates $\mathbf{x} \mapsto \boldsymbol{\xi}$ in Eqs. (3.28–3.29) leads to the appearance of coefficients proportional to negative powers of $\|\boldsymbol{\xi}\|$ in the integrals. From a mathematical viewpoint, the choice of an adequate functional space for u and v, $H^1(\mathbb{R}^n)$ in this particular case (see Sec. 3.2.1), mitigates such singularities in the neighborhood of $\boldsymbol{\xi} = 0$. In actual FEM computations,

these singularities are 'killed' by a suitable choice of discrete space W^h and associated basis functions. In that perspective, the following section explores two ways for ensuring that the problem is numerically free of singularities:

1. The first idea is to choose a finite-dimensional space with carefully selected basis functions. More precisely, instead of using piecewise polynomials as in standard FEM (see Sec. 2.1.3), we weight the latter by $\|\boldsymbol{\xi}\|^\beta$, where $\beta \in \mathbb{R}$ is chosen *(i)* large enough to cancel the singularity in the integrands, and *(ii)* so that the inclusion $W^h \subset W$ holds. This approach is most notably leveraged in a series of articles by T. Boulmezaoud *et al.* (see e.g. Refs. [242–247]) in the framework of *ifem*. For this reason, we will refer to this approach as 'à la Boulmezaoud'.

2. The second idea we investigate is somehow more convoluted. Instead of working with the weak forms featuring singular coefficients in the integrals [e.g. Eqs. (3.23, 3.24, 3.28, 3.29)], we introduce a weight ω directly in the strong formulation of the PDE problem so as to make such singular coefficients disappear. We will therefore refer to this second approach as the 'explicit weight regularization technique'. It is implemented in, e.g., Refs. [137, 240, 241, 262].

3.3.3 Dealing with arising unbounded coefficients

The two aforementioned approaches are showcased using Examples 3.1 and 3.2. Specifically:

- the *weight regularization technique* is applied to the linear Klein–Gordon equation (3.3), Example 3.1, for $\Omega = \mathbb{R}^2$;
- the Poisson problem, Example 3.2, is handled *à la Boulmezaoud* for $\Omega = \mathbb{R}^2$.

For the sake of generality, the computations are conducted in an arbitrary dimension n as far as possible. Ultimately though, we specify $n = 2$ to obtain explicit expressions.

First approach: à la Boulmezaoud

We illustrate this technique on the Poisson equation (3.4), Example 3.2. Paving the way for Sec. 3.4, we introduce the *hat* operator acting on generalized functions w as

$$\forall \boldsymbol{\xi} \in \mathbb{R}^n \setminus \{\mathbf{0}\}\,,\ \widehat{w}(\boldsymbol{\xi}) = \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^\beta w\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big) = \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^\beta \widetilde{w}(\boldsymbol{\xi})\,, \tag{3.30}$$

where $\widetilde{w}$ is given by Eq. (3.27) and $\beta \in \mathbb{R}$ is to be fixed subsequently. For u, $v \in W^{\mathrm{P}}_\omega$, rewriting $a_{\mathrm{P,ext}}$ and $l_{\mathrm{P,ext}}$ in terms of $\widehat{u}$ and $\widehat{v}$ reads

$$\begin{aligned} a_{\mathrm{P,ext}}(u,v) &= \int_{\tilde{\Omega}_{\mathrm{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2(n-2)} \widetilde{\boldsymbol{\nabla}}\tilde{u}\cdot\widetilde{\boldsymbol{\nabla}}\tilde{v}\,\mathrm{d}\boldsymbol{\xi} \\ &= \int_{\tilde{\Omega}_{\mathrm{ext}}} \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^{2(\beta-n+1)} \left\{ \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^2 \widetilde{\boldsymbol{\nabla}}\widehat{u}\cdot\widetilde{\boldsymbol{\nabla}}\widehat{v} + \frac{\beta}{R_c^2}\Big[\widehat{u}\,\widetilde{\boldsymbol{\nabla}}\widehat{v} + \widehat{v}\,\widetilde{\boldsymbol{\nabla}}\widehat{u}\Big]\cdot\boldsymbol{\xi} + \left(\frac{\beta}{R_c}\right)^2 \widehat{u}\,\widehat{v}\right\}\mathrm{d}\boldsymbol{\xi}\,, \end{aligned} \tag{3.31}$$

$$l_{\mathrm{P,ext}}(v) = \int_{\tilde{\Omega}_{\mathrm{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \tilde{f}(\boldsymbol{\xi})\,\tilde{v}\,\mathrm{d}\boldsymbol{\xi} = \int_{\tilde{\Omega}_{\mathrm{ext}}} \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^{\beta-2n} \tilde{f}(\boldsymbol{\xi})\,\widehat{v}\,\mathrm{d}\boldsymbol{\xi}\,. \tag{3.32}$$

Above all, the choice of a suitable value for β must be guided by the adopted discrete space W^h in which we approximate functions of W^{P}_ω. We will see in Sec. 3.4 that demanding the inclusion $W^h \subset W^{\mathrm{P}}_\omega$ is the same as demanding the convergence of the integrals (3.31–3.32).

Second approach: explicit weight regularization technique

We illustrate this technique on the linear Klein–Gordon equation (3.3), Example 3.1, for $\Omega = \mathbb{R}^2$. It is worth noting this technique is the same as the *method of auxiliary mapping* implemented in the work of Oh *et al.* [240, 241], except they use an exponential weight whereas we use a polynomial weight.

The weighted weak form Given a single weight $\omega\colon \mathbb{R}^2 \to \mathbb{R}_+^*$ of class $\mathcal{C}^1$, multiplying both sides of Eq. (3.3) by ω and rearranging the terms yields

$$-\mathrm{div}\big[\omega(\mathbf{x})\boldsymbol{\nabla}u\big] + \boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u + \omega(\mathbf{x})d(\mathbf{x})u = \omega(\mathbf{x})f(\mathbf{x})\,,\quad \forall \mathbf{x}\in\mathbb{R}^2\,. \tag{3.33}$$

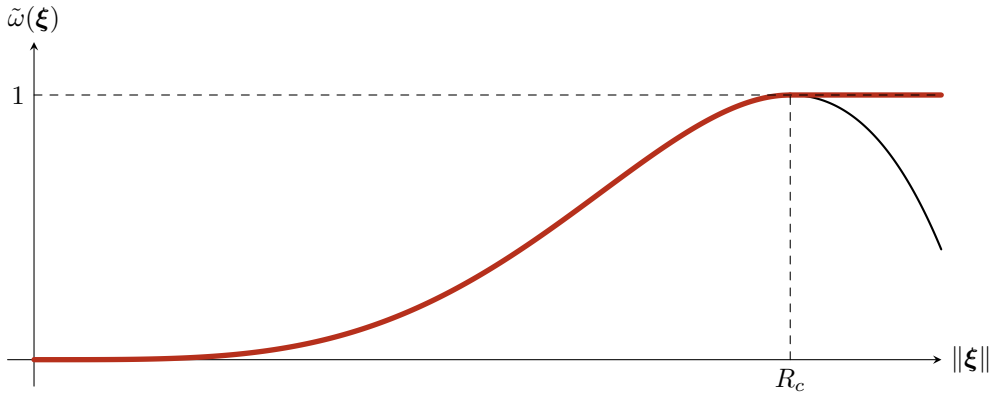


Figure 3.4: Chosen weight expressed as a function of $\|\boldsymbol{\xi}\|$ — see Eq. (3.36).

This *weighted* PDE, despite being equivalent to the original one [Eq. (3.3)], has a different weak form. The (bi)linear forms a_{KG} and l_{KG} are indeed replaced by a^*_{KG} and l^*_{KG}. In particular, Eqs. (3.28–3.29) become

$$a^*_{\text{KG,ext}}(u,v) = \int_{\tilde{\Omega}_{\text{ext}}} \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2(n-2)} \widetilde{\boldsymbol{\nabla}}\tilde{u}\cdot\left[\widetilde{\boldsymbol{\nabla}}\tilde{v}\,\omega + \widetilde{\boldsymbol{\nabla}}\tilde{\omega}\,\tilde{v}\right]\mathrm{d}\boldsymbol{\xi} + \int_{\tilde{\Omega}_{\text{ext}}} \tilde{\omega}\left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \tilde{d}(\boldsymbol{\xi})\,\tilde{u}\,\tilde{v}\,\mathrm{d}\boldsymbol{\xi}\,, \tag{3.34}$$

$$l^*_{\text{KG,ext}}(v) = \int_{\tilde{\Omega}_{\text{ext}}} \tilde{\omega}\left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^{2n} \tilde{f}(\boldsymbol{\xi})\,\tilde{v}\,\mathrm{d}\boldsymbol{\xi}\,, \tag{3.35}$$

for u, v in some adequate space to be specified. The weight function ω, that we introduced *manu militari*, should fulfill the following properties:

1. $\omega(\mathbf{x}) = 1$ in Ω_{int} [the weight has no effect in the interior domain];
2. $\tilde{\omega}(\boldsymbol{\xi}) \sim \|\boldsymbol{\xi}\|^{2n}$ in the neighborhood of $\boldsymbol{\xi} = 0$ [the weight removes the singularities in Eqs. (3.34–3.35)];
3. $\tilde{\omega}(R_c) = 1$ and $\widetilde{\boldsymbol{\nabla}}\tilde{\omega}(\boldsymbol{\xi}) \underset{\|\boldsymbol{\xi}\|\to R_c}{\longrightarrow} \mathbf{0}$ [smooth connection at the interface between Ω_{int} and Ω_{ext}].

We now select $n = 2$. In regards to the three above conditions, an admissible weight ω is a polynomial in the variable $\|\boldsymbol{\xi}\|$, reading

$$\begin{cases} \tilde{\omega}(\boldsymbol{\xi}) = -\dfrac{4}{R_c^5}\|\boldsymbol{\xi}\|^5 + \dfrac{5}{R_c^4}\|\boldsymbol{\xi}\|^4 \\ \widetilde{\boldsymbol{\nabla}}\tilde{\omega}(\boldsymbol{\xi}) = \dfrac{20}{R_c^5}\left(R_c\|\boldsymbol{\xi}\|^2 - \|\boldsymbol{\xi}\|^3\right)\boldsymbol{\xi} \end{cases}, \ \forall \boldsymbol{\xi} \in \tilde{\Omega}_{\text{ext}} \quad \text{and} \quad \tilde{\omega}(\boldsymbol{\xi}) = 1 \text{ for } \|\boldsymbol{\xi}\| \geq R_c \tag{3.36}$$

in the $\boldsymbol{\xi}$ coordinate, and

$$\begin{cases} \omega(\mathbf{x}) = \dfrac{5R_c^4\|\mathbf{x}\| - 4R_c^5}{\|\mathbf{x}\|^5} \\ \boldsymbol{\nabla}\omega(\mathbf{x}) = -4\dfrac{5R_c^4\|\mathbf{x}\| - 5R_c^5}{\|\mathbf{x}\|^7}\mathbf{x} \end{cases}, \ \forall \mathbf{x} \in \Omega_{\text{ext}} \quad \text{and} \quad \omega(\mathbf{x}) = 1 \text{ for } \|\mathbf{x}\| \leq R_c \tag{3.37}$$

in the original Cartesian coordinate system. This weight function $\tilde{\omega}$ is represented in Fig. 3.4 (red curve).

Weighted Sobolev space and well-posedness A suitable functional space is the weighted Sobolev space $W^{\text{kg}}_{\omega}(\mathbb{R}^2)$, where the weight ω is given by Eqs. (3.36–3.37) — see Sec. 3.2.2. The following two lemmas serve the purpose of proving this statement.

Lemma 3.4. With the weight defined through Eqs. (3.36–3.37), $W^{\text{KG}}_{\omega}(\mathbb{R}^2)$ is a uniformly convex Banach space. Moreover, the space of test functions $\mathscr{D}(\mathbb{R}^2)$ is dense in $W^{\text{KG}}_{\omega}(\mathbb{R}^2)$.

Proof. One has ω, $\omega^{-1} \in L^1_{\text{loc}}(\mathbb{R}^2)$ so W^{KG}_{ω} is a uniformly convex Banach space and $\mathscr{D}(\mathbb{R}^2)$ is a subset of W^{KG}_{ω} — see Lemma 3.2. The density property can be proven in a similar fashion as done in Ref. [253], that is in two stages:

1. given an element u of W_ω^{KG}, one can construct a sequence of elements in W_ω^{KG} with compact support that converges to u;

2. any element of that sequence having a compact support, one can choose an element of $\mathscr{D}(\mathbb{R}^2)$ arbitrarily close to it.

Let $u \in W_\omega^{\text{KG}}$ and $\varepsilon > 0$. Let $\varphi \in \mathscr{D}(\mathbb{R}^2)$ such that $0 \leq \varphi \leq 1$, $\varphi(\mathbf{x}) = 1$ for $\|\mathbf{x}\| \leq 1$, $\varphi(\mathbf{x}) = 0$ for $\|\mathbf{x}\| \geq 2$ and define

$$\text{for any } k \in \mathbb{N}^*\,,\ \varphi_k(\mathbf{x}) = \varphi\left(\frac{\mathbf{x}}{k}\right)\ \forall \mathbf{x} \in \mathbb{R}^2 \text{ and } u_k = \varphi_k u\,,$$

so that φ_k and u_k are compactly supported in $\mathcal{B}(2k)$. We then show that

$$\lim_{k\to+\infty} \|u - u_k\|_{W_\omega^{\text{KG}}} = 0\,.$$

On the one hand,

$$\int_{\mathbb{R}^2} \omega|u-u_k|^2\,\mathrm{d}\mathbf{x} = \int\limits_{k\leq\|\mathbf{x}\|\leq 2k} \omega|u-u_k|^2\,\mathrm{d}\mathbf{x} + \int\limits_{\|\mathbf{x}\|\geq 2k} \omega u^2\,\mathrm{d}\mathbf{x} \leq 2\int\limits_{\|\mathbf{x}\|\geq k} \omega u^2\,\mathrm{d}\mathbf{x} \underset{k\to+\infty}{\longrightarrow} 0$$

where the limit is justified by the fact that $u \in W_\omega^{\text{KG}}$. On the other hand, using $\boldsymbol{\nabla}\varphi_k(\mathbf{x}) = k^{-1}\boldsymbol{\nabla}\varphi(\mathbf{x}/k)$ and the fact that there exists a constant $C > 0$ such that for all $\mathbf{x} \in \mathbb{R}^2$, $\|\boldsymbol{\nabla}\varphi(\mathbf{x})\|^2 \leq C$,

$$\begin{aligned}\int_{\mathbb{R}^2} \omega\|\boldsymbol{\nabla}(u-u_k)\|^2\,\mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^2} \omega\|(1-\varphi_k)\boldsymbol{\nabla}u - u\boldsymbol{\nabla}\varphi_k\|^2\,\mathrm{d}\mathbf{x}\\ &\leq 2\int_{\mathbb{R}^2} \omega|1-\varphi_k|^2\|\boldsymbol{\nabla}u\|^2\,\mathrm{d}\mathbf{x} + 2\int_{\mathbb{R}^2} \omega|u|^2\|\boldsymbol{\nabla}\varphi_k\|^2\,\mathrm{d}\mathbf{x}\\ &\leq 2\int\limits_{\|\mathbf{x}\|\geq k} \omega\|\boldsymbol{\nabla}u\|^2\,\mathrm{d}\mathbf{x} + \frac{2C}{k^2}\int\limits_{\|\mathbf{x}\|\geq k} \omega|u|^2\,\mathrm{d}\mathbf{x} \underset{k\to+\infty}{\longrightarrow} 0\,,\end{aligned}$$

because again $u \in W_\omega^{\text{KG}}$. Therefore, there exists $M_1 \geq 1$ such that for all $k \geq M_1$, $\|u - u_k\|_{W_\omega^{\text{KG}}} < \varepsilon/2$. Let $k \geq M_1$; u_k being compactly supported, it belongs to $H^1(\mathbb{R}^2)$. $\mathscr{D}(\mathbb{R}^2)$ being dense in $H^1(\mathbb{R}^2)$, we can find a sequence $(\Psi_\ell)_{\ell\in\mathbb{N}} \in \mathscr{D}(\mathbb{R}^2)^{\mathbb{N}}$ such that

$$\lim_{\ell\to+\infty} \|\Psi_\ell - u_k\|_{H^1} = 0\,.$$

Recalling that $\forall \mathbf{x} \in \mathbb{R}^2, 0 < \omega(\mathbf{x}) \leq 1$, we get

$$\lim_{\ell\to+\infty} \|\Psi_\ell - u_k\|_{W_\omega^{\text{KG}}} = 0\,,\ \text{therefore, there exists } M_2 \in \mathbb{N}^* \text{ such that } \forall\,\ell \geq M_2,\ \|\Psi_\ell - u_k\|_{W_\omega^{\text{KG}}} < \varepsilon/2\,.$$

To conclude, for $k \geq M_1$ and $\ell \geq M_2$, we have

$$\|u - \Psi_\ell\|_{W_\omega^{\text{KG}}} \leq \|u - u_k\|_{W_\omega^{\text{KG}}} + \|u_k - \Psi_\ell\|_{W_\omega^{\text{KG}}} < \varepsilon\,.$$

□

Lemma 3.5. For u, $v \in W_\omega^{\text{KG}}(\mathbb{R}^2)$, one has

$$\int_{\mathbb{R}^2} \big[\boldsymbol{\nabla}\omega \cdot \boldsymbol{\nabla}u\big]\,v\,\mathrm{d}\mathbf{x} \leq \frac{4}{R_c}\,\|v\|_{L^2(\mathbb{R}^2,\,\omega)}\,|u|_{W_\omega^{\text{KG}}}\,, \tag{3.38}$$

where $|\cdot|_{W_\omega^{\text{KG}}}$ denotes the semi-norm in W_ω^{KG} — see Eq. (3.10).

Proof. Let u, $v \in W_\omega^{\mathrm{KG}}(\mathbb{R}^2)$. Recalling that ω is constant in Ω_{int}, we get

$$\begin{aligned}
\left|\int_\Omega [\boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u]\,v\,\mathrm{d}\mathbf{x}\right| &\leq \int_\Omega \left|\,[\boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u]\,v\right|\mathrm{d}\mathbf{x} = \int_{\Omega_{\mathrm{ext}}} \left|\,[\boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u]\,v\right|\mathrm{d}\mathbf{x} && \text{(triangle inequality and } \omega\equiv 1 \text{ in } \Omega_{\mathrm{int}}\text{)}\\
&\leq 4\int_{\Omega_{\mathrm{ext}}} \frac{5R_c^4\|\mathbf{x}\| - 5R_c^5}{\|\mathbf{x}\|^5}\left|\frac{\mathbf{x}}{\|\mathbf{x}\|^2}\cdot\boldsymbol{\nabla}u\right||v|\,\mathrm{d}\mathbf{x} && (\|\mathbf{x}\|\geq R_c \text{ in } \Omega_{\mathrm{ext}})\\
&\leq 4\int_{\Omega_{\mathrm{ext}}} \omega(\mathbf{x})\left|\frac{\mathbf{x}}{\|\mathbf{x}\|^2}\cdot\boldsymbol{\nabla}u\right||v|\,\mathrm{d}\mathbf{x} && (0\leq 5R_c^4\|\mathbf{x}\| - 5R_c^5 \leq 5R_c^4\|\mathbf{x}\| - 4R_c^5)\\
&= 4\int_{\Omega_{\mathrm{ext}}} \sqrt{\omega(\mathbf{x})}\,|v|\sum_{i=1}^{2}\sqrt{\omega(\mathbf{x})}\frac{\partial u}{\partial x_i}\frac{x_i}{\|\mathbf{x}\|^2}\,\mathrm{d}\mathbf{x}\\
&\leq 4\int_{\Omega_{\mathrm{ext}}} \sqrt{\omega(\mathbf{x})}\,|v|\left[\sum_{i=1}^{2}\omega(\mathbf{x})\left(\frac{\partial u}{\partial x_i}\right)^2\right]^{\frac{1}{2}}\left[\sum_{i=1}^{2}\frac{x_i^2}{\|\mathbf{x}\|^4}\right]^{\frac{1}{2}}\mathrm{d}\mathbf{x} && \text{(Hölder's inequality)}\\
&\leq \frac{4}{R_c}\left(\int_{\Omega_{\mathrm{ext}}}\omega(\mathbf{x})\,v^2\,\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}}\left(\int_{\Omega_{\mathrm{ext}}}\omega(\mathbf{x})\,\|\boldsymbol{\nabla}u\|^2\,\mathrm{d}\mathbf{x}\right)^{\frac{1}{2}} && \text{(Cauchy–Schwarz inequality)}
\end{aligned}$$

□

Lemma 3.4 justifies the integration by parts in $\mathbb{R}^2$ underlying the definition of a_{KG}^* — see Sec. 3.2.3. Lemma 3.5 is useful in several respects for further proving the well-posedness of the corresponding weak formulation. Let u, $v \in W_\omega^{\mathrm{KG}}$, the well-defined character and continuity of a_{KG}^* follow from inequality (3.38) as

$$\begin{aligned}
|a_{\mathrm{KG}}^*(u,v)| &\leq \left|\int_{\mathbb{R}^2}\omega(\mathbf{x})\big[\boldsymbol{\nabla}u\cdot\boldsymbol{\nabla}v\big]\,\mathrm{d}\mathbf{x}\right| + \left|\int_{\mathbb{R}^2}\big[\boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u\big]\,v\,\mathrm{d}\mathbf{x}\right| + \left|\int_{\mathbb{R}^2} d(\mathbf{x})uv\,\mathrm{d}\mathbf{x}\right|\\
&\leq |u|_{W_\omega^{\mathrm{KG}}}\,|v|_{W_\omega^{\mathrm{KG}}} + \frac{4}{R_c}\,\|v\|_{L^2(\mathbb{R}^2,\,\omega)}\,|u|_{W_\omega^{\mathrm{KG}}} + d_\infty\,\|u\|_{L^2(\mathbb{R}^2,\,\omega)}\,\|v\|_{L^2(\mathbb{R}^2,\,\omega)}\\
&\leq \max(1,\,4/R_c,\,d_\infty)\,\|u\|_{W_\omega^{\mathrm{KG}}}\,\|v\|_{W_\omega^{\mathrm{KG}}}\,.
\end{aligned}$$

The continuity of l_{KG}^* in W_ω^{KG}, under the assumption that $f \in L^2(\mathbb{R}^2,\,\omega)$, is obtained similarly by the Cauchy–Schwarz inequality. Ultimately, the coercivity of a_{KG}^* over W_ω^{KG} is examined in the following lemma.

Lemma 3.6. If $R_c\,\min(1,\,d_0) > 2$, the bilinear for $a_{\mathrm{KG}}^*(\cdot,\,\cdot)$ is coercive over W_ω^{KG}.

Proof. Let $u \in W_\omega^{\mathrm{KG}}$. On the one hand,

$$\int_\Omega \omega(\mathbf{x})\,\|\boldsymbol{\nabla}u\|^2\,\mathrm{d}\mathbf{x} + \int_\Omega \omega(\mathbf{x})\,d(\mathbf{x})u^2\,\mathrm{d}\mathbf{x} \geq \int_\Omega \omega(\mathbf{x})\,\|\boldsymbol{\nabla}u\|^2\,\mathrm{d}\mathbf{x} + d_0\int_\Omega \omega(\mathbf{x})\,u^2\,\mathrm{d}\mathbf{x} \tag{3.39}$$

$$\geq \min(1,\,d_0)\,\|u\|_{W_\omega^{\mathrm{KG}}}^2\,. \tag{3.40}$$

On the other hand, inequality (3.38) together with the fact that $\forall a, b \in \mathbb{R}$, $2ab \leq a^2 + b^2$ yields

$$\left|\int_\Omega [\boldsymbol{\nabla}\omega\cdot\boldsymbol{\nabla}u]\,u\,\mathrm{d}\mathbf{x}\right| \leq \frac{2}{R_c}\,\|u\|_{W_\omega^{\mathrm{KG}}}^2\,. \tag{3.41}$$

Consequently, inequalities (3.40–3.41) lead to

$$a_{\mathrm{KG}}^*(u,u) \geq \left(\min(1,\,d_0) - \frac{2}{R_c}\right)\|u\|_{W_\omega^{\mathrm{KG}}}^2\,.$$

□

In particular, one can always choose R_c big enough to make the coercivity constant found in Lemma 3.6 strictly positive. All in all, we end up with the following proposition.

Proposition 3.1. Let $\Omega = \mathbb{R}^2$ and define the weight $\omega\colon \mathbb{R}^2 \to \mathbb{R}_+^*$ according to Eqs. (3.36–3.37). Suppose that there exist two constants d_0, $d_\infty \in \mathbb{R}$ such that $0 < d_0 \leq d(\mathbf{x}) \leq d_\infty$ for all $\mathbf{x} \in \mathbb{R}^2$ and that $f \in L^2(\mathbb{R}^2,\,\omega)$.

Then, according to Lax–Milgram theorem (see Box E), the weak formulation

$$\text{“ Find } u \in W_\omega^{\mathrm{KG}} \text{ such that for all } v \in W_\omega^{\mathrm{KG}},\ a^*_{\mathrm{KG}}(u,v) = l^*_{\mathrm{KG}}(v) \text{”}$$

has a unique solution, which depends continuously on the problem's data.

Remark 3.4. Proposition 3.1 does not generalize easily to other PDEs. Indeed, the coercivity of the weighted bilinear form $a^*(\cdot,\cdot)$ over W_ω^{KG} is lost when considering, for instance, the Poisson equation (3.4) of Example 3.2, for which

$$\forall u,\, v \in W_\omega^{\mathrm{KG}},\ a^*_{\mathrm{P}}(u,v) = \int_{\mathbb{R}^2} \omega(\mathbf{x}) \big[\boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v\big]\, \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^2} \big[\boldsymbol{\nabla}\omega \cdot \boldsymbol{\nabla} u\big] v\, \mathrm{d}\mathbf{x}\,. \tag{3.42}$$

In fact, it is easy to see from Eq. (3.42) that *(i)* $a^*_{\mathrm{P}}(u,u) = 0$ for any constant function u, and *(ii)* that constants belong to W_ω^{KG} since ω defined by Eqs. (3.36–3.37) belongs to $L^1(\mathbb{R}^2)$.[4] Moreover, this procedure results in a non-symmetric bilinear form and *de facto* a non-symmetric stiffness matrix, which is often not desirable in terms of solving linear systems numerically.[5]

To summarize, the explicit weight regularization technique, applied to the linear Klein–Gordon equation (3.3) in $\Omega = \mathbb{R}^2$, results in a well-posed weak formulation. The integrals featured in the definition of the (bi)linear forms a^*_{KG} and l^*_{KG} are free of singular coefficients thanks to the choice of the weight function ω given by Eqs. (3.36–3.37). The next section shows that an adequate finite-dimensional function space can be constructed in the 'usual' way — with piecewise polynomials of the variable $\mathbf{x}$ in Ω_{int} and piecewise polynomials of the variable $\boldsymbol{\xi}$ in $\tilde{\Omega}_{\mathrm{ext}}$.

3.4 The FE framework

In this section, in the same vein as in Sec. 2.1.3, we introduce the finite-dimensional spaces $W^h_{\mathrm{KG}} \subset W^{\mathrm{KG}}_\omega$ and $W^h_{\mathrm{P}} \subset W^{\mathrm{P}}_\omega$. This requires the definitions of meshes. Most notably, the fact that the inverted exterior domain is bounded makes it *meshable* with a finite number of elements, which was the primary motivation for introducing the Kelvin inversion. The present section corresponds to the description of the *ifem* method [242].

3.4.1 Construction of meshes

We define the one-point compactification of $\tilde{\Omega}_{\mathrm{ext}}$ by $\tilde{\Omega}^*_{\mathrm{ext}} \coloneqq \tilde{\Omega}_{\mathrm{ext}} \cup \{\mathbf{0}_{\mathrm{ext}}\}$. We then use two polygonal meshes:

- $\mathcal{T}^h_{\mathrm{int}}$ on Ω_{int}, which is comprised of N_{int} DOFs;
- $\widetilde{\mathcal{T}}^h_{\mathrm{ext}}$ on $\tilde{\Omega}^*_{\mathrm{ext}}$ such that $\mathbf{0}_{\mathrm{ext}}$ belongs to the set of vertices of $\widetilde{\mathcal{T}}^h_{\mathrm{ext}}$ (but does not constitute a true degree of freedom), which is comprised of N_{ext} DOFs.

We further denote by Σ^h and $\tilde{\Sigma}^h$ the surface meshes of $\mathcal{T}^h_{\mathrm{int}}$ and $\widetilde{\mathcal{T}}^h_{\mathrm{ext}}$ respectively, constituted of $(n-1)$-dimensional elements. Each of the two surface meshes are comprised of N_Γ DOFs. The total number of DOFs is denoted N_{tot} and is equal to $N_{\mathrm{int}} + N_{\mathrm{ext}} - N_\Gamma$ (the boundary DOFs should be counted only once).

Definition 3.3. We say that the meshes $\mathcal{T}^h_{\mathrm{int}}$ and $\widetilde{\mathcal{T}}^h_{\mathrm{ext}}$ have the same trace, i.e. $\Sigma^h \equiv \tilde{\Sigma}^h$, if the following two conditions are met:

1. boundary vertices of both meshes have the same coordinates;
2. moreover, the resulting *facet* elements [which are $(n-1)$-dimensional elements] of Σ^h and $\tilde{\Sigma}^h$ have the same connectivity.

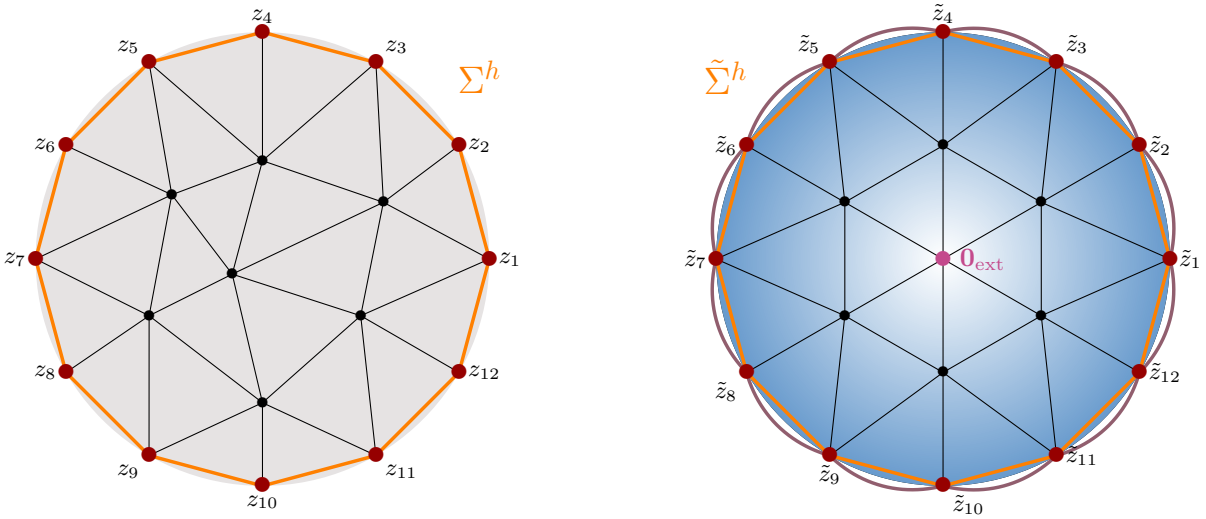
From now on, we demand that $\mathcal{T}^h_{\mathrm{int}}$ and $\widetilde{\mathcal{T}}^h_{\mathrm{ext}}$ have the same trace $\Sigma^h \equiv \tilde{\Sigma}^h$ in the sense of Definition 3.3. An example of such meshes is given in Fig. 3.5.

Remark 3.5. Note that, *stricto sensu*, the condition $\Sigma^h \equiv \tilde{\Sigma}^h$ is not enough to ensure that the method is exactly conforming. Indeed, meshing with polygonal elements means that the boundaries Γ and $\tilde{\Gamma}$ are approximated by $(n-1)$-dimensional polytopes. Unlike $(n-1)$-spheres, polytopes are not invariant under the Kelvin transform (3.25). This is visible on Fig. 3.5 where the elements of Σ^h are actually mapped to the purple curved lines. As a consequence, entries of the stiffness matrix and load vector associated with boundary nodes are spoiled by small errors. We can name at least three ways around this issue:

1. Approximate Γ exactly: this could in theory be achieved by the use of e.g. higher-order curved elements, isoparametric elements or isogeometric analysis (see e.g. Refs. [263, 264]).

[4] Actually, one can show that for any $\alpha \in [0, 1[$ and $C \in \mathbb{R}$, $u_\alpha \colon \mathbf{x} \mapsto C\|\mathbf{x}\|^\alpha$ belongs to W^{KG}_ω and is such that $a^*_{\mathrm{P}}(u_\alpha, u_\alpha) \leq 0$.

[5] For instance, one cannot employ the conjugate gradient method in its standard form to solve non-symmetric linear systems.



Mesh of the interior domain $\mathcal{T}_{\text{int}}^h$ | Mesh of the inversed exterior domain $\mathcal{T}_*^h$

Figure 3.5: Example of meshes $\mathcal{T}_{\text{int}}^h$ and $\widetilde{\mathcal{T}}_{\text{ext}}^h$. The requirements that (i) the two meshes have the same trace $\Sigma^h \equiv \tilde{\Sigma}^h$ and (ii) $\mathbf{0}_{\text{ext}} \in \mathcal{T}_*^h$ are fulfilled. Note that the Kelvin inversion $\mathcal{K}$ does not preserve polygonal simplices, which are actually mapped to curved lines (in purple). As a consequence, the stiffness matrix coefficients involving $(\tilde{z}_i)$ are spoiled by a small error.

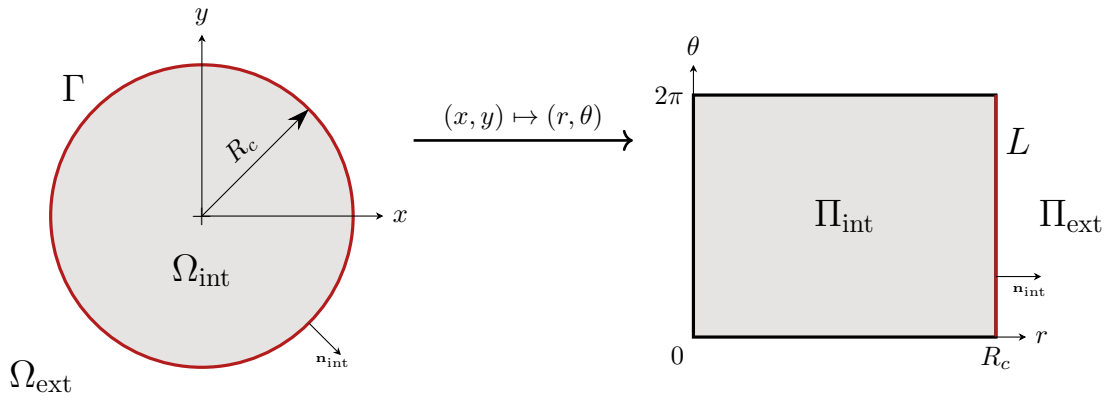


Figure 3.6: Mapping of Ω_{int} to Π_{int} through the change of coordinates from Cartesian coordinates (x, y) to polar coordinates (r, θ). The originally curved boundary Γ is mapped to the straight line L. See Sec. 2.3 for a discussion on how such coordinate transforms can be leveraged in the framework of FEM.

2. Replace the Kelvin inversion by the polygonal inversion introduced in Ref. [242]. This can be achieved at the price of a slightly more complicated numerical implementation.

3. Switch from Cartesian coordinates to a more adapted coordinate system, e.g. spherical coordinates in n-dimensions. This has the effect of mapping Ω_{int} and $\tilde{\Omega}_{\text{ext}}$ into rectangles for $n = 2$ (see Fig. 3.6) or a rectangular cuboid for $n = 3$. The resulting geometry can be approximated exactly with polygonal elements. We use this approach in Refs. [137, 141].

We are now in a position to define the discrete functional spaces W_{KG}^h and W_{P}^h associated with the pair of meshes $(\mathcal{T}_{\text{int}}^h, \widetilde{\mathcal{T}}_{\text{ext}}^h)$.

3.4.2 Discrete spaces

The definition of suitable finite dimensional spaces depends on the regularization technique employed (see Sec. 3.3.3). In both cases, we will make use of the following lemma.

Lemma 3.7. Let K be a polyhedron of dimension n with $\mathbf{0}$ as one of its vertices and $u \in \mathbb{P}_k(K)$ such that $u(\mathbf{0}) = 0$. Then, there exists $C_K > 0$ such that

$$\forall \mathbf{x} \in K, \ |u(\mathbf{x})| \leq C_K \|\mathbf{x}\| \,. \tag{3.43}$$

Proof. The polynomial function u can be written explicitly as

$$u(\mathbf{x}) = \sum_{k=1}^{n} b_k x_k + Q_2(\mathbf{x}) \ , \forall \mathbf{x} \in K \ ,$$

where $b_k \in \mathbb{R}$ are not all-zero and Q_2 is the polynomial gathering all terms of order greater than or equal to two. Note the absence of 0^{th}-degree term owing to the condition $u(\mathbf{0}) = 0$. It follows that

$$u^2(\mathbf{x}) = \left(\sum_{k=1}^{n} b_k x_k \right)^2 + \underbrace{2Q_2(\mathbf{x}) \sum_{k=1}^{d} b_k x_k + Q_2^2(\mathbf{x})}_{:= Q_3(\mathbf{x})} = \sum_{k=1}^{n} b_k^2 x_k^2 + 2 \sum_{i>j} b_i b_j x_i x_j + Q_3(\mathbf{x})$$

$$\leq b_{\max}^2 \|\mathbf{x}\|^2 + \frac{n(n-1)}{2} b_{\max}^2 \|\mathbf{x}\|^2 + Q_3(\mathbf{x}) \ ,$$

with $b_{\max} = \max_{1 \leq k \leq n}(b_k)$. We have used the fact that $\forall i, j \neq i$, $2|x_i x_j| \leq x_i^2 + x_j^2 \leq \|\mathbf{x}\|^2$. Q_3 is a polynomial function gathering all terms of degree at least 3. It can be written as the sum of N monomials

$$Q_3(\mathbf{x}) = \sum_{l=1}^{N} \alpha_l M_l^{(3)}(\mathbf{x}) \ , \forall \mathbf{x} \in K \ ,$$

where $\alpha_l \in \mathbb{R}$ and $M_l^{(3)}$ is a monomial of degree at least 3, with unitary coefficient. Then, for all $l \in \{1, \ldots, N\}$, there exist indices i_l, j_l such that we can factorize $M_l^{(3)}(\mathbf{x}) = x_{i_l} x_{j_l} M_l^{(1)}(\mathbf{x})$, with $M_l^{(1)}$ being a monomial of degree one. K being a compact set,

$$\exists B_K^l > 0 \text{ such that } \forall \mathbf{x} \in K, \ |M_l^{(1)}(\mathbf{x})| \leq B_K^l \ .$$

Setting $B_K = \max_{1 \leq l \leq N}(B_K^l)$ and $\alpha_{\max} = \max_{1 \leq l \leq N}(\alpha_l)$, we have the following inequality:

$$|Q_3(\mathbf{x})| \leq \sum_{l=1}^{N} |\alpha_l M_l^{(3)}(\mathbf{x})| \leq \alpha_{\max} \sum_{l=1}^{N} |x_{i_l} x_{j_l}| |M_l^{(1)}(\mathbf{x})| \leq \alpha_{\max} N B_K \|\mathbf{x}\|^2 \ ,$$

which concludes this proof. □

Definition 3.4. For the sake of clarity, let us define the sub-domain

$$\tilde{\mathcal{T}}_\infty^h \coloneqq \left\{ \tilde{K} \in \widetilde{\mathcal{T}}_{\text{ext}}^h \text{ such that } \mathbf{0}_{\text{ext}} \in \tilde{K} \right\} \setminus \{\mathbf{0}_{\text{ext}}\} \, , \tag{3.44}$$

as well as its image by the Kelvin transform

$$\mathcal{T}_\infty^h \coloneqq \mathcal{K}(\tilde{\mathcal{T}}_\infty^h) = \left\{ \mathcal{K}(\tilde{K} \setminus \{\mathbf{0}_{\text{ext}}\}) \, , \ \tilde{K} \in \widetilde{\mathcal{T}}_{\text{ext}}^h \text{ such that } \mathbf{0}_{\text{ext}} \in \tilde{K} \right\} . \tag{3.45}$$

Discrete counterpart of W_ω^{P} (à la Boulmezaoud)

When discussing the Poisson problem on the whole space in Sec. 3.2.2, we introduced the weighted Sobolev space

$$W_\omega^{\text{P}} = \left\{ u \in L^2\big(\mathbb{R}^2, \omega_0\big) \text{ such that } \|\boldsymbol{\nabla} u\| \in L^2\big(\mathbb{R}^2, \omega_1\big) \right\} , \tag{3.46}$$

where the pair of weights (ω_0, ω_1) is given by Eq. (3.14) in dimension two. For a polynomial degree $k \in \mathbb{N}^*$, we set the discrete spaces

$$W_{\text{P,int}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\bar{\Omega}_{\text{int}}) \text{ such that } \forall K \in \mathcal{T}_{\text{int}}^h, \ u_{|K} \in \mathbb{P}_k(K) \right\} , \tag{3.47a}$$

$$W_{\text{P,ext}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\bar{\Omega}_{\text{ext}}) \text{ such that } \forall \tilde{K} \in \widetilde{\mathcal{T}}_{\text{ext}}^h, \ \widehat{u}_{|\tilde{K}} \in \mathbb{P}_k(\tilde{K}) \text{ and } \widehat{u}(\mathbf{0}_{\text{ext}}) = 0 \right\} , \tag{3.47b}$$

$$W_{\text{P}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\mathbb{R}^2) \text{ such that } u_{|\mathcal{T}_{\text{int}}^h} \in W_{\text{P,int}}^h \text{ and } u_{|\widetilde{\mathcal{T}}_{\text{ext}}^h} \in W_{\text{P,ext}}^h \right\} . \tag{3.47c}$$

Proposition 3.2. Suppose that $\beta > -1$, then W_{P}^h [given by Eq. (3.47c)] is a subspace of W_ω^{P} (defined in Sec. 3.2.2).

Proof. The only non-trivial part of this proposition is associated with the asymptotic behavior of functions belonging to W_{P}^h. Specifically, the critical point is to show that

$$I_1 = \int_{K_F} \omega_0(\mathbf{x}) \left[u(\mathbf{x})\right]^2 \mathrm{d}\mathbf{x} < +\infty \quad \text{and} \quad I_2 = \int_{K_F} \omega_1(\mathbf{x}) \|\boldsymbol{\nabla} u(\mathbf{x})\|^2 \,\mathrm{d}\mathbf{x} < +\infty \,, \quad \forall u \in W_{\mathrm{P}}^h \text{ and } K_F \in \mathcal{T}_\infty^h \,.$$

On the one hand, switching to the $\boldsymbol{\xi}$ coordinates and applying Lemma 3.7 to $\widehat{u}(\boldsymbol{\xi})$ [see Eq. (3.30)] yields

$$\begin{aligned} I_1 &= \int_{\tilde{K}_F} \left[\tilde{u}(\boldsymbol{\xi})\right]^2 \tilde{\omega}_0(\boldsymbol{\xi}) \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^4 \mathrm{d}\boldsymbol{\xi} = \int_{\tilde{K}_F} \left[\widehat{u}(\boldsymbol{\xi})\right]^2 \tilde{\omega}_0(\boldsymbol{\xi}) \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^{2(\beta-2)} \mathrm{d}\boldsymbol{\xi} \\ &\lesssim \int_{\tilde{K}_F} \frac{\|\boldsymbol{\xi}\|^2}{R_c^4 + \|\boldsymbol{\xi}\|^2} \ln\left(\frac{R_c^4 + \|\boldsymbol{\xi}\|^2}{\|\boldsymbol{\xi}\|^2}\right)^{-2} \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^{2(\beta-1)} \mathrm{d}\boldsymbol{\xi} \\ &\lesssim \int_0^{R_c} \frac{r^{2\beta+1}}{\ln(r)^2} \,\mathrm{d}r \,, \end{aligned}$$

where the notation $a \lesssim b$ means that there exists a constant C such that $a \leq Cb$. The latter integral is convergent if and only if $\beta > -1$.

On the other hand, to show that $I_2 < +\infty$, we recall that

$$\|\boldsymbol{\nabla} u\|^2 = \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^4 \|\tilde{\boldsymbol{\nabla}} \tilde{u}\|^2 \quad , \quad \tilde{\boldsymbol{\nabla}} \tilde{u} = \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^\beta \widetilde{\boldsymbol{\nabla}} \widehat{u} + \frac{\beta}{R_c^\beta} \|\boldsymbol{\xi}\|^{\beta-2} \, \widehat{u} \, \boldsymbol{\xi} \quad , \quad \left\|\widetilde{\boldsymbol{\nabla}} \widehat{u}\right\| \in L^\infty\big(\tilde{\Omega}_{\mathrm{ext}}\big) \,.$$

Then, using Cauchy–Schwarz inequality together with Lemma 3.7, we obtain

$$\|\tilde{\boldsymbol{\nabla}} \tilde{u}\|^2 \leq \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^{2\beta} \|\widetilde{\boldsymbol{\nabla}} \widehat{u}\|^2 + \left(\frac{\beta}{R_c^\beta}\right)^2 \|\boldsymbol{\xi}\|^{2(\beta-1)} |\widehat{u}|^2 + 2\frac{|\beta|}{R_c^{2\beta}} \|\boldsymbol{\xi}\|^{2\beta-1} \left\|\widetilde{\boldsymbol{\nabla}} \widehat{u}\right\| |\widehat{u}| \lesssim \|\boldsymbol{\xi}\|^{2\beta}$$

Therefore, we have

$$I_2 = \int_{\tilde{K}_F} \left(\frac{\|\boldsymbol{\xi}\|}{R_c}\right)^4 \|\widetilde{\boldsymbol{\nabla}} \tilde{u}\|^2 \left(\frac{R_c}{\|\boldsymbol{\xi}\|}\right)^4 \mathrm{d}\boldsymbol{\xi} \lesssim \int_0^{R_c} r^{2\beta+1} \,\mathrm{d}r \,.$$

The latter integral is convergent if and only if $\beta > -1$.

Note that this proof could easily be generalized to the case of an arbitrary dimension $n \in \{1, 2, 3\}$ and we would have found the condition $\beta > (n-4)/2$. □

Discrete counterpart of W_ω^{KG} (weight regularization technique)

When discussing the Klein–Gordon problem on the whole space in Sec. 3.3.3, we introduced the weighted Sobolev space

$$W_\omega^{\mathrm{KG}} = \left\{ u \in L^2\big(\mathbb{R}^2, \omega\big) \text{ such that } \|\boldsymbol{\nabla} u\| \in L^2\big(\mathbb{R}^2, \omega\big) \right\}, \tag{3.48}$$

where the weight ω is given by Eqs. (3.36–3.37) in dimension two. For a polynomial degree $k \in \mathbb{N}^*$, we set the discrete spaces

$$W_{\mathrm{KG,int}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\bar{\Omega}_{\mathrm{int}}) \text{ such that } \forall K \in \mathcal{T}_{\mathrm{int}}^h,\, u_{|K} \in \mathbb{P}_k(K) \right\}, \tag{3.49a}$$

$$W_{\mathrm{KG,ext}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\bar{\Omega}_{\mathrm{ext}}) \text{ such that } \forall \tilde{K} \in \widetilde{\mathcal{T}}_{\mathrm{ext}}^h,\, \widehat{u}_{|\tilde{K}} \in \mathbb{P}_k(\tilde{K}) \text{ and } \widehat{u}(\mathbf{0}_{\mathrm{ext}}) = 0 \right\}, \tag{3.49b}$$

$$W_{\mathrm{KG}}^h \coloneqq \left\{ u \in \mathcal{C}^0(\mathbb{R}^2) \text{ such that } u_{|\mathcal{T}_{\mathrm{int}}^h} \in W_{\mathrm{KG,int}}^h \text{ and } u_{|\widetilde{\mathcal{T}}_{\mathrm{ext}}^h} \in W_{\mathrm{KG,ext}}^h \right\}. \tag{3.49c}$$

Proposition 3.3. W_{KG}^h [defined by Eq. (3.49c)] is a subspace of W_ω^{KG} [defined in Sec. 3.3.3.]

Proof. Let $u \in W_{\mathrm{KG}}^h$. First, let us observe that ωu^2 and $\omega \|\boldsymbol{\nabla} u\|^2$ belong to $L^1_{\mathrm{loc}}(\mathbb{R}^2)$ so we only have to prove their integrability over $\mathcal{T}_\infty^h$ defined by Eq. (3.45). Let $\tilde{K}_F \in \tilde{\mathcal{T}}_\infty^h \cup \{\mathbf{0}_{\mathrm{ext}}\}$ and K_F be the corresponding element of

$\mathcal{T}_\infty^h$. Because of the nodal constraint on $\mathbf{0}_{\text{ext}}$, $\tilde{u}(\mathbf{0}) = 0$ and $\tilde{u}_{|\tilde{K}_F} \in \mathbb{P}_k(\tilde{K}_F)$. Using Lemma 3.7, we get

$$\forall \mathbf{x} \in K_F,\ |u(\mathbf{x})| = |\tilde{u}(\boldsymbol{\xi})| \leq C_{K_F} \|\boldsymbol{\xi}\| = \frac{C_{K_F} R_c^2}{\|\mathbf{x}\|},\ \text{with } \boldsymbol{\xi} = \mathcal{K}(\mathbf{x})\,.$$

Consequently,

$$\int_{K_F} \omega(\mathbf{x}) u(\mathbf{x})^2 \,\mathrm{d}\mathbf{x} \leq C_{K_F}^2 R_c^4 \int_{K_F} \frac{|\omega(\mathbf{x})|}{\|\mathbf{x}\|^2} \,\mathrm{d}\mathbf{x} < +\infty\,. \tag{3.50}$$

Similarly, we want an upper bound of the gradient. We have $\boldsymbol{\nabla} u = \mathbf{J}_{\mathcal{K}}(\mathcal{K}^{-1}(\boldsymbol{\xi}))^T \widetilde{\boldsymbol{\nabla}} \tilde{u}$ where the matrix $\mathbf{J}_{\mathcal{K}}(\mathcal{K}^{-1}(\boldsymbol{\xi}))$ is the Jacobian matrix of the Kelvin inversion at point $\mathcal{K}^{-1}(\boldsymbol{\xi})$ and is given in Box I. It is such that

$$\|\mathbf{J}_{\mathcal{K}}(\mathcal{K}^{-1}(\boldsymbol{\xi}))\|^* = \left\{ \rho\Big[\mathbf{J}_{\mathcal{K}}\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big)^T \mathbf{J}_{\mathcal{K}}\big(\mathcal{K}^{-1}(\boldsymbol{\xi})\big) \Big] \right\}^{1/2} \propto \|\boldsymbol{\xi}\|^2$$

where $\|\cdot\|^*$ is the matrix norm induced by the Euclidean norm on $\mathbb{R}^n$ and $\rho(\mathbf{M})$ denotes the spectral radius of a matrix $\mathbf{M}$. Therefore, there exists $C > 0$ such that

$$\forall \mathbf{x} \in K_F,\ \|\boldsymbol{\nabla} u(\mathbf{x})\| = \|\mathbf{A}(\boldsymbol{\xi}) \widetilde{\boldsymbol{\nabla}} \tilde{u}\| \leq \|\mathbf{A}(\boldsymbol{\xi})\|^* \|\widetilde{\boldsymbol{\nabla}} \tilde{u}\| \leq C \|\boldsymbol{\xi}\|^2 \propto \|\mathbf{x}\|^{-2}$$

Indeed, K_F is a compact set and the application : $\boldsymbol{\xi} \in K_F \mapsto \|\widetilde{\boldsymbol{\nabla}} \tilde{u}\|$ is continuous because $\tilde{u} \in \mathbb{P}_k(\tilde{K}_F)$. As a result,

$$\int_{K_F} \omega(\mathbf{x}) \|\boldsymbol{\nabla} u\|^2 \,\mathrm{d}\mathbf{x} \leq C \int_{K_F} \frac{|\omega(\mathbf{x})|}{\|\mathbf{x}\|^4} \,\mathrm{d}\mathbf{x} < +\infty\,. \tag{3.51}$$

Eqs. (3.50–3.51) show that $u \in W_{\text{KG}}^h$. □

3.4.3 Assembling of the stiffness matrix and load vector

To conclude this section on *ifem*, let us equip the finite-dimensional spaces W_{P}^h and W_{KG}^h with their respective bases. Define the basis $(w_i)_{1 \leq i \leq N_{\text{tot}}}$ satisfying

- $w_i \in W^h$ where W^h denotes either W_{KG}^h or W_{P}^h;
- $w_i(M_j) = \delta_{i,j}$ if $M_j \in K$ for some $K \in \mathcal{T}_{\text{int}}^h$;
- for $M_j \in K$ for some $K \in \widetilde{\mathcal{T}}_{\text{ext}}^h$
  - $\tilde{w}_i(M_j) = \delta_{i,j}$ if $W^h = W_{\text{KG}}^h$,
  - $\widehat{w}_i(M_j) = \delta_{i,j}$ if $W^h = W_{\text{P}}^h$.

Then, the computation of the stiffness matrix and load vector is performed as explained in Sec. 2.1.3. Specifically, it requires the numerical evaluation of all individual terms $a(w_j, w_i)$ and $l(w_i)$ for all pairs $(i, j) \in [\![1,\, N_{\text{tot}}]\!]^2$, where $(a,\, l) = (a_{\text{KG}}^*,\, l_{\text{KG}}^*)$ for the Klein–Gordon problem and $(a,\, l) = (a_{\text{P}},\, l_{\text{P}})$ for the Poisson problem. In the interior domain, the basis functions are Lagrange polynomials in the $\mathbf{x}$ coordinate, so that the computation of the latter two terms is achieved as in standard FEM. In the exterior domain however, basis functions w are not Lagrange polynomials, instead

- if $W^h = W_{\text{KG}}^h$, then $\tilde{w}$ is a Lagrange polynomial in the $\boldsymbol{\xi}$ coordinate and so standard FEM techniques can be used to compute integrals involving $\tilde{w}$;
- if $W^h = W_{\text{P}}^h$, then $\widehat{w}$ is a Lagrange polynomial in the $\boldsymbol{\xi}$ coordinate and so standard FEM techniques can be used to compute the integrals involving $\widehat{w}$.

Remark 3.6. Note that the piecewise polynomial function associated with the origin $\mathbf{0}_{\text{ext}}$ is not to be included in the set of basis functions because of the nullity condition. However, for the sole purpose of computing the stiffness matrix and load vector, it is strictly equivalent to include the latter function to the set of basis functions and impose a zero Dirichlet boundary condition at the node $\mathbf{0}_{\text{ext}}$. This is the approach we adopt in order to make the numerical implementation of the method as simple as possible.

3.5 Iterative variant: the alternate inverted finite element method

In this section, we present a novel approach which we call the *alternate inverted finite elements method*, abbreviated as *a-ifem*. It builds on top of two existing and well-studied methods, namely:

- The *inverted finite elements method* (*ifem*), originally proposed by Boulmezaoud in Ref. [242] and put into use in many subsequent articles [243–247]. This approach shares several common features with the so-called *method of auxiliary mapping* described in the work of Oh *et al.* [240, 241]. The *ifem* has already been discussed at length throughout Secs. 3.3 and 3.4.
- Domain decomposition methods, notably developed in the field of high performance computing (HPC), which consist in splitting a large computational domain into smaller, more manageable sub-domains which exchange information at their common boundaries (see e.g. Refs. [200, 201] for reviews on the topic). In particular, the *a-ifem* technique draws on an iterative relaxation procedure introduced by Marini and Quarteroni [265], which belongs to the class of domain decomposition techniques without overlap.

The *a-ifem* technique was first introduced in our work [137], although it was therein referred to as the *ping-pong* method, owing to the iterative nature of the numerical scheme. Because it shares a common ground with the *ifem* technique, several aspects of the method have already been introduced in the previous sections. We thereby focus on the specificities of the *a-ifem* approach here, referring back to relevant previous points when needed.

3.5.1 The iterative procedure

As already mentioned, the iterative algorithm we describe is essentially based on the work of Marini and Quarteroni [265], since the vast majority of results they obtained (in a bounded scenario) are also applicable to our case. The idea consists in splitting the original problem into two sub-problems — one over Ω_{int} and the other one over $\tilde{\Omega}_{\text{ext}}$. Transmission conditions at the interface Γ are taken into account partly in one sub-domain and partly in the other one, without overlap. As a side note, the domain decomposition scheme we employ is discussed in other work, see e.g. Refs. [266–268]. In Ref. [200], it is referred to as the *Dirichlet / Neumann method*.

From the global problem to the split problem

In what follows, we will need the following lemma.

Lemma 3.8. Let U an open set of $\mathbb{R}^n$ and $v \in L^2(U)$. Then, $v \in H^1(U)$ if and only if there exists $C > 0$ such that for all $\varphi \in \mathscr{D}(U)$,

$$\left| \int_U v \frac{\partial \varphi}{\partial x_i} \, \mathrm{d}\mathbf{x} \right| \leq C \, \|\varphi\|_{L^2(U)} \,, \;\; i \in \{1, \dots, n\} \,. \tag{3.52}$$

A good starting point for explaining how *a-ifem* works is to go back to Eq. (3.26), where we used Chasles' relation on integrals to argue that

$$a(u,v) = l(v) \iff a_{\text{int}}(u,v) + a_{\text{ext}}(u,v) = l_{\text{int}}(v) + l_{\text{ext}}(v) \,. \tag{3.53}$$

Ultimately, we want to split this single equation — which we refer to as the *global problem* — into two sub-problems, defined on Ω_{int} and $\tilde{\Omega}_{\text{ext}}$ respectively. Using W to denote either W_ω^{P} (for the Poisson problem 3.2) or W_ω^{KG} (for the weighted Klein–Gordon problem 3.1), and the subscript k to refer to either of the two sub-domains {'int', 'ext'}, we set the additional functional spaces

$$W_k \coloneqq \left\{ v_{|\Omega_k}, \, v \in W \right\} \quad , \quad W_k^0 \coloneqq \left\{ v \in W_k \,/\, \gamma_k v = 0 \right\} \quad \text{and} \quad \Phi \coloneqq \left\{ v_{|\Gamma}, \, v \in W \right\} . \tag{3.54}$$

In these definitions, $\gamma_k \colon W_k \to H^{1/2}(\Gamma)$ is the trace operator. Remark 3.1 and the fact that Γ is bounded make indeed γ_k a well-defined, continuous and surjective operator (see Ref. [269] for a rigorous definition of fractional Sobolev spaces). In particular, we can identify $H^{1/2}(\Gamma) = \gamma_k(W_k) = \Phi$. When v belongs W, we use the notation γ (without subscript) to refer to the trace of v on Γ.[6]

Proposition 3.4. Consider either the linear Klein–Gordon problem 3.1 or the Poisson problem 3.2. Let $\phi \in \Phi$ and k be a subscript used to refer to either the interior domain or the exterior domain. The problem

$$\text{Find } u_k \in W_k \text{ such that } \forall v \in W_k^0 \,, \;\; a_k(u_k, v) = 0 \text{ and } u_k = \phi \text{ on } \Gamma$$

has a unique solution. Here, the bilinear form $a(\cdot, \cdot)$ refers to either $a^*_{\text{KG}}(\cdot, \cdot)$ or $a_{\text{P}}(\cdot, \cdot)$.

[6] One way of defining $\gamma \colon W \to H^{1/2}(\Gamma)$ is to take $\gamma v = \gamma_{\text{int}} v_{|\Omega_{\text{int}}}$, since for any $v \in W$, $v_{|\Omega_{\text{int}}} \in W_{\text{int}}$.

Proof. Cauchy–Schwarz inequality (and Lemma 3.5 in the case of $a^*_{\mathrm{KG,ext}}$) provides the continuity of $a_k(\cdot,\cdot)$ on $W_k \times W_k$. On Ω_{int}, there exists a constant $A > 0$ such that the weight functions ω_0, ω_1 appearing in Eqs. (3.13–3.14), and ω given by Eqs. (3.36–3.37), are bounded between A and 1. Consequently, properties of the space W^0_{int} coincide with those of the classical Sobolev space $H^1_0(\Omega_{\mathrm{int}})$. In particular, one inherits the *usual* Poincaré inequality, which is enough to show that the coercivity of a_{int} over W^0_{int} is preserved. The coercivity of a_{ext} over W_{ext} on the other hand can be proved *(i)* in a similar fashion to the proof of Lemma 3.6 for the case of the weighted Klein–Gordon equation, and *(ii)* by using a Poincaré-type inequality similar to Eq. (3.15) for an exterior domain for the case of the Poisson equation — see e.g. Refs. [252, 256] for the derivation of such Poincaré-like inequalities on exterior unbounded domains of $\mathbb{R}^n$. Since $W^0_{\mathrm{ext}} \subset W_{\mathrm{ext}}$, a_{ext} is also coercive over W^0_{ext}. Lax–Milgram theorem concludes the proof of this proposition. □

Proposition 3.4 allows us to define the following trace extension operator

$$\begin{aligned} R_k \colon \Phi &\to W_k \qquad \text{with} \qquad a_k(R_k\phi, v) = 0 \quad \forall\, v \in W^0_k \\ \phi &\mapsto R_k\phi \qquad\qquad\quad R_k\phi(\mathbf{x}) = \phi(\mathbf{x}) \quad \forall\, \mathbf{x} \in \Gamma \end{aligned} \,, \tag{3.55}$$

which is well-defined according to Proposition 3.4.

Lemma 3.9. The trace extension operator R_k is linear.

Proof. Let ϕ_1, $\phi_2 \in \Phi$ and $\alpha \in \mathbb{R}$. Define $\chi = \alpha\phi_1 + \phi_2$ and $w = \alpha R_k(\phi_1) + R_k(\phi_2)$. On the one hand, $R_k\chi$ is the unique element of W_k such that $\forall v \in W^0_k$, $a_k(R_k\chi, v) = 0$ and $R_k\chi = \chi$ on Γ (by linearity of the trace). On the other hand, $u \in W_k$ happens to satisfy $a_k(u, v) = 0\ \forall v \in W^0_k$ (by bilinearity of a_k) and $u = \chi$ on Γ (by linearity of the trace). Therefore, unicity implies $u = R_k\chi$, i.e. $R_k(\alpha\phi_1 + \phi_2) = \alpha R_k\phi_1 + R_k\phi_2$. □

One can then define the *split problem* as

$$\begin{gathered} \text{Find } u_{\mathrm{int}} \in W_{\mathrm{int}} \text{ and } u_{\mathrm{ext}} \in W_{\mathrm{ext}} \text{ such that} \\ \begin{cases} \forall v \in W^0_{\mathrm{int}}\ , \ a_{\mathrm{int}}(u_{\mathrm{int}}, v) \ = l_{\mathrm{int}}(v) \\ \forall v \in W_{\mathrm{ext}}\ , \ a_{\mathrm{ext}}(u_{\mathrm{ext}}, v) = l_{\mathrm{ext}}(v) - a_{\mathrm{int}}(u_{\mathrm{int}}, R_{\mathrm{int}}\gamma_{\mathrm{ext}}v) + l_{\mathrm{int}}(R_{\mathrm{int}}\gamma_{\mathrm{ext}}v) \\ u_{\mathrm{int}} = u_{\mathrm{ext}} \text{ on } \Gamma \end{cases} \end{gathered} \tag{3.56}$$

Lemma 3.10. $u \in W$ is solution to the global problem (3.53) if and only if $u_{\mathrm{int}} = u_{|\Omega_{\mathrm{int}}}$ and $u_{\mathrm{ext}} = u_{|\Omega_{\mathrm{ext}}}$ are solutions to the split problem (3.56).

Proof. $\Longleftarrow$: Let $u_{\mathrm{int}} \in W_{\mathrm{int}}$, $u_{\mathrm{ext}} \in W_{\mathrm{ext}}$ be solutions to the split problem (3.56). We set

$$\begin{aligned} u \colon \Omega &\to \mathbb{R} \\ \mathbf{x} &\mapsto \begin{cases} u_{\mathrm{int}}(\mathbf{x}) & \text{if } \mathbf{x} \in \bar{\Omega}_{\mathrm{int}} \\ u_{\mathrm{ext}}(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega_{\mathrm{ext}} \end{cases} \end{aligned} .$$

Then u belongs to the weighted Sobolev space W since *(i)* $\|u\|^2_W = \|u_{\mathrm{int}}\|^2_{W_{\mathrm{int}}} + \|u_{\mathrm{ext}}\|^2_{W_{\mathrm{ext}}} < +\infty$, and *(ii)* weak derivatives of u exist on Ω thanks to the continuity condition $u_{\mathrm{int}} = u_{\mathrm{ext}}$ on Γ. Indeed, this latter point can be seen using Lemma 3.8. For any bounded domain $U \subset \mathbb{R}^n$ that encompasses Γ, $\varphi \in \mathscr{D}(U)$ and $i \in \{1, \dots, n\}$, we have

$$\begin{aligned} \left| \int_U u \frac{\partial \varphi}{\partial x_i} \,\mathrm{d}\mathbf{x} \right| &= \left| \int_{U_{\mathrm{int}}} u_{\mathrm{int}} \frac{\partial \varphi}{\partial x_i} \,\mathrm{d}\mathbf{x} + \int_{U_{\mathrm{ext}}} u_{\mathrm{ext}} \frac{\partial \varphi}{\partial x_i} \,\mathrm{d}\mathbf{x} \right| \\ &= \left| - \int_{U_{\mathrm{int}}} \varphi \frac{\partial u_{\mathrm{int}}}{\partial x_i} \,\mathrm{d}\mathbf{x} + \int_\Gamma u_{\mathrm{int}} \varphi \nu^i_{\mathrm{int}} \,\mathrm{d}\gamma - \int_{U_{\mathrm{ext}}} \varphi \frac{\partial u_{\mathrm{ext}}}{\partial x_i} \,\mathrm{d}\mathbf{x} + \int_\Gamma u_{\mathrm{ext}} \varphi \nu^i_{\mathrm{ext}} \,\mathrm{d}\gamma \right| \\ &\leq \left| \int_{U_{\mathrm{int}}} \varphi \frac{\partial u_{\mathrm{int}}}{\partial x_i} \,\mathrm{d}\mathbf{x} \right| + \left| \int_{U_{\mathrm{ext}}} \varphi \frac{\partial u_{\mathrm{ext}}}{\partial x_i} \,\mathrm{d}\mathbf{x} \right| \\ &\leq \left(\left\| \frac{\partial u_{\mathrm{int}}}{\partial x_i} \right\|_{L^2(U_{\mathrm{int}})} + \left\| \frac{\partial u_{\mathrm{ext}}}{\partial x_i} \right\|_{L^2(U_{\mathrm{ext}})} \right) \|\varphi\|_{L^2(U)} \,, \end{aligned}$$

where we have set $U_{\mathrm{int}} = U \cap \Omega_{\mathrm{int}}$ and $U_{\mathrm{ext}} = U \cap \Omega_{\mathrm{ext}}$. There remains to be checked that u is solution to the

global problem (3.53). For $v \in W$, we set

$$\begin{cases} v_{\text{int}} = v - R_{\text{int}}\gamma v & \text{on } \Omega_{\text{int}} \\ v_{\text{ext}} = v_{|\Omega_{\text{ext}}} & \text{on } \Omega_{\text{ext}} \end{cases},$$

so that $v_{\text{int}} \in W^0_{\text{int}}$ and $v_{\text{ext}} \in W_{\text{ext}}$. Finally, one has

$$\begin{aligned} a(u,v) &= a_{\text{int}}(u_{\text{int}}, v_{\text{int}}) + a_{\text{ext}}(u_{\text{ext}}, v_{\text{ext}}) + a_{\text{int}}(u_{\text{int}}, R_{\text{int}}\gamma v) \\ &= l_{\text{int}}(v_{\text{int}}) + l_{\text{ext}}(v_{\text{ext}}) + l_{\text{int}}(R_{\text{int}}\gamma v) \\ &= l(v) \end{aligned}$$

$\Longrightarrow$: Conversely, let $u \in W$ be solution to the global problem. Then, the condition $u_{\text{int}} = u_{\text{ext}}$ on Γ is automatically verified (almost everywhere). Let $v_{\text{int}} \in W^0_{\text{int}}$ and v be its extension by 0 in Ω_{ext}. One can show using Lemma 3.8 that $v \in W$, so that one can write

$$a_{\text{int}}(u_{\text{int}}, v_{\text{int}}) = a(u,v) = l(v) = l_{\text{int}}(v_{\text{int}})\,.$$

Similarly, let $v_{\text{ext}} \in W_{\text{ext}}$ and define v its extension by $R_{\text{int}}\gamma v_{\text{ext}}$ in Ω_{int}. Then $v \in W$ and

$$a_{\text{ext}}(u_{\text{ext}}, v_{\text{ext}}) = a(u,v) - a_{\text{int}}(u_{\text{int}}, R_{\text{int}}\gamma v_{\text{ext}}) = l_{\text{int}}(R_{\text{int}}\gamma v_{\text{ext}}) + l_{\text{ext}}(v_{\text{ext}}) - a_{\text{int}}(u_{\text{int}}, R_{\text{int}}\gamma v_{\text{ext}})\,,$$

which ends the proof. □

Taking a step back, one can see that the split problem (3.56) is nothing but a weak formulation of the set of equations

$$\mathcal{L}u_{\text{int}} = f \quad \text{on } \Omega_{\text{int}} \tag{3.57a}$$

$$\mathcal{L}u_{\text{ext}} = f \quad \text{on } \Omega_{\text{ext}} \tag{3.57b}$$

$$u_{\text{int}} = u_{\text{ext}} \text{ on } \Gamma \tag{3.57c}$$

$$\frac{\partial u_{\text{int}}}{\partial \nu_{\text{int}}} + \frac{\partial u_{\text{ext}}}{\partial \nu_{\text{ext}}} = 0 \quad \text{on } \Gamma\,, \tag{3.57d}$$

where $\mathcal{L}$ denotes the differential operator of the lhs of the PDE [which we assume to be of the generic form (2.3) for this specific discussion], and $\partial/\partial\nu_{\text{int}}$, $\partial/\partial\nu_{\text{ext}}$ denote the co-normal derivatives [see e.g. Eq. (2.5)]. Under suitable regularity conditions, the split problem in its strong form (3.57) is equivalent to the global problem in its strong form $\mathcal{L}u = f$, see e.g. Refs. [270, 271]. The transmission of the normal derivative in Eq. (3.57d) at Γ is encoded in the term $-a_{\text{int}}(u_{\text{int}}, R_{\text{int}}\gamma v_{\text{ext}}) + l_{\text{int}}(R_{\text{int}}\gamma v_{\text{ext}})$ of the weak counterpart Eq. (3.56). Indeed, let $u_{\text{int}} \in W_{\text{int}}$, $u_{\text{ext}} \in W_{\text{ext}}$ be solutions to problem (3.56) and further satisfying the strong split problem (3.57). On the one hand, an integration by parts in the definition of the bilinear form a_{ext} yields

$$\begin{aligned} a_{\text{ext}}(u_{\text{ext}}, v_{\text{ext}}) &= \langle \mathcal{L}u_{\text{ext}}, v_{\text{ext}} \rangle_{\text{ext}} + \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{ext}}) \cdot \mathbf{n}_{\text{ext}}\, v_{\text{ext}}\, \mathrm{d}\gamma \\ &= l_{\text{ext}}(v_{\text{ext}}) + \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{ext}}) \cdot \mathbf{n}_{\text{ext}}\, v_{\text{ext}}\, \mathrm{d}\gamma\,, \quad \forall v_{\text{ext}} \in W_{\text{ext}}\,. \end{aligned}$$

On the other hand, an integration by parts in the split problem (3.56) yields

$$\begin{aligned} a_{\text{ext}}(u_{\text{ext}}, v_{\text{ext}}) &= l_{\text{ext}}(v_{\text{ext}}) - \langle \mathcal{L}u_{\text{int}}, R_{\text{int}}\gamma v_{\text{ext}} \rangle_{\text{int}} - \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{int}}) \cdot \mathbf{n}_{\text{int}}\, R_{\text{int}}\gamma v_{\text{ext}}\, \mathrm{d}\gamma + \langle f, R_{\text{int}}\gamma v_{\text{ext}} \rangle_{\text{int}} \\ &= l_{\text{ext}}(v_{\text{ext}}) - \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{int}}) \cdot \mathbf{n}_{\text{int}}\, v_{\text{ext}}\, \mathrm{d}\gamma\,, \quad \forall v_{\text{ext}} \in W_{\text{ext}}\,. \end{aligned}$$

Equating these two expressions for a_{ext} leads to the transmission condition in the weak form

$$\int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{int}}) \cdot \mathbf{n}_{\text{int}}\, v_{\text{ext}}\, \mathrm{d}\gamma + \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{ext}}) \cdot \mathbf{n}_{\text{ext}}\, v_{\text{ext}}\, \mathrm{d}\gamma = 0\,, \quad \forall v_{\text{ext}} \in W_{\text{ext}}\,.$$

In the case of the Klein–Gordon equation (3.3) or the Poisson equation (3.4), we have $\mathbf{C} = \mathbf{I}_n$ and the latter

expression boils down to

$$\int_\Gamma \frac{\partial u_{\text{int}}}{\partial n_{\text{int}}} v_{\text{ext}} \,\mathrm{d}\gamma + \int_\Gamma \frac{\partial u_{\text{ext}}}{\partial n_{\text{ext}}} v_{\text{ext}} \,\mathrm{d}\gamma = 0 \ , \quad \forall v_{\text{ext}} \in W_{\text{ext}} \, .$$

Remark 3.7. We showed how the transmission of the normal derivative at Γ was encoded in the term $-a_{\text{int}}(u_{\text{int}}, R_{\text{int}}\gamma v_{\text{ext}}) + l_{\text{int}}(R_{\text{int}}\gamma v_{\text{ext}})$ of Eq. (3.56). However, we could also have accounted for it through a Neumann boundary term as

$$a_{\text{ext}}(u_{\text{ext}}, v_{\text{ext}}) = l_{\text{ext}}(v_{\text{ext}}) - \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{int}}) \cdot \mathbf{n}_{\text{int}} v_{\text{ext}} \,\mathrm{d}\gamma \ , \quad \forall v_{\text{ext}} \in W_{\text{ext}} \tag{3.58}$$

instead. While the two weak formulations can be obtained in an equivalent way from the strong split problem (3.57), they lead to distinct problems when turning to finite dimensional (discrete) spaces. In Sec. 3.6, we shall see that the Neumann boundary term approach is the least efficient of the two.

From the split problem to the iterative procedure

The coupled set of equations (3.56) cannot be solved but simultaneously. In order to decouple the two subproblems, we define an iterative scheme. Let $\lambda^0 \in \Phi$ be an initial guess of the solution on the boundary Γ. For $\ell \geq 1$, we construct the sequences of functions $(u_{\text{int}}^\ell)_{\ell\in\mathbb{N}^*} \in W_{\text{int}}^{\mathbb{N}^*}$ and $(u_{\text{ext}}^\ell)_{\ell\in\mathbb{N}^*} \in W_{\text{ext}}^{\mathbb{N}^*}$ which are obtained by iteratively solving the following problems:

$$\begin{cases} a_{\text{int}}(u_{\text{int}}^\ell, v_{\text{int}}) = l_{\text{int}}(v_{\text{int}}) \ , \ \forall v_{\text{int}} \in W_{\text{int}}^0 \\ \gamma_{\text{int}} u_{\text{int}}^\ell = \lambda^{\ell-1} \end{cases} \tag{3.59a}$$

$$a_{\text{ext}}(u_{\text{ext}}^\ell, v_{\text{ext}}) = l_{\text{ext}}(v_{\text{ext}}) - a_{\text{int}}(u_{\text{int}}^\ell, R_{\text{int}}\gamma_{\text{ext}} v_{\text{ext}}) + \langle f, R_{\text{int}}\gamma_{\text{ext}} v_{\text{ext}} \rangle_{\text{int}} \ , \ \forall v_{\text{ext}} \in W_{\text{ext}} \tag{3.59b}$$

$$\lambda^\ell = \theta_\ell \gamma_{\text{ext}} u_{\text{ext}}^\ell + (1-\theta_\ell)\lambda^{\ell-1} \, , \tag{3.59c}$$

where $(\theta_\ell)_{\ell\in\mathbb{N}}$ is a sequence of positive relaxation parameters introduced to ensure the convergence of this iterative scheme. We then study the conditions under which the sequence $(u_{\text{int}}^\ell,\, u_{\text{ext}}^\ell)_{\ell\in\mathbb{N}}$ converges towards the solution $(u_{\text{int}},\, u_{\text{ext}})$ of the split problem (3.56).

Definition 3.5. Using the subscript k to refer to either the interior region or the exterior one, the coercivity of a_k over W_k allows us to define the following norm over W_k

$$\|v\|_k^2 \coloneqq a_k(v, v) \ , \quad \forall v \in W_k \, . \tag{3.60}$$

Definition 3.6. The trace space Φ defined in Eq. (3.54) can be supplemented with the norm

$$|||\phi|||^2 \coloneqq \|R_{\text{int}}\phi\|_{\text{int}}^2 \, , \quad \forall \phi \in \Phi \tag{3.61}$$

which thereby forms a Banach space.

Proof. Let us show that Φ is indeed complete for the norm $|||\cdot|||$. Let $(\phi_\ell)_{\ell\in\mathbb{N}} \in \Phi^{\mathbb{N}}$ be a Cauchy sequence. Using Definition 3.6 and denoting $\alpha > 0$ a coercivity constant of a_{int}, we get for $i, j \in \mathbb{N}$

$$|||\phi_i - \phi_j|||^2 = a_{\text{int}}\big(R_{\text{int}}(\phi_i - \phi_j), R_{\text{int}}(\phi_i - \phi_j)\big) \geq \alpha \|R_{\text{int}}(\phi_i - \phi_j)\|_{W_{\text{int}}}^2 \, .$$

Using the linearity of R_{int} (see Lemma 3.9), we deduce that $(R_{\text{int}}\phi_\ell)_{\ell\in\mathbb{N}}$ is also a Cauchy sequence in W_{int} for the norm $\|\cdot\|_{W_{\text{int}}}$. Yet W_{int}, as a Banach space, is complete, which implies that the sequence $(R_{\text{int}}\phi_\ell)_{\ell\in\mathbb{N}}$ converges to some $u \in W_{\text{int}}$. Denoting $\psi = \gamma_{\text{int}} u \in \Phi$ and $C \geq 0$ a continuity constant of a_{int}, one has

$$\forall \ell \in \mathbb{N} \, , \ |||\phi_\ell - \psi|||^2 = a_{\text{int}}\big(R_{\text{int}}(\phi_\ell - \psi), R_{\text{int}}(\phi_\ell - \psi)\big) \leq C \, \|R_{\text{int}}\phi_\ell - R_{\text{int}}\psi\|_{W_{\text{int}}}^2 \, , \tag{3.62}$$

where we have again used the linearity of R_{int}. To finish this proof, it remains to be shown that $\|R_{\text{int}}\psi - u\|_{W_{\text{int}}} = 0$. This requires the continuity of the operators $R_{\text{int}} \colon \Phi \to W_{\text{int}}$ and $\gamma_{\text{int}} \colon W_{\text{int}} \to \Phi$.

1. First, the equivalence between the norms $\|\cdot\|_{\text{int}}$ [defined by Eq. (3.60)] and $\|\cdot\|_{W_{\text{int}}}$ provides us with a constant $\beta \geq 0$ such that for any $\chi \in \Phi$,

$$\|R_{\text{int}}\chi\|_{W_{\text{int}}}^2 \leq \beta \, a_{\text{int}}(R_{\text{int}}\chi, R_{\text{int}}\chi) \leq \beta \, |||\chi|||^2 \, , \tag{3.63}$$

   i.e. the trace extension operator R_{int} is continuous with respect to the relevant norms.

2. Second, for any $v \in W_{\text{int}}$, the definition of R_{int} in Eq. (3.55) lets us write

$$a_{\text{int}}(R_{\text{int}}\gamma_{\text{int}}v, R_{\text{int}}\gamma_{\text{int}}v - v) = 0 \quad \text{i.e.} \quad a_{\text{int}}(R_{\text{int}}\gamma_{\text{int}}v, R_{\text{int}}\gamma_{\text{int}}v) = a_{\text{int}}(R_{\text{int}}\gamma_{\text{int}}v, v)\,.$$

Thence, we have

$$|||\gamma_{\text{int}}v|||^2 = a_{\text{int}}(R_{\text{int}}\gamma_{\text{int}}v, R_{\text{int}}\gamma_{\text{int}}v) = a_{\text{int}}(R_{\text{int}}\gamma_{\text{int}}v, v) \leq C\,\|R_{\text{int}}\gamma_{\text{int}}v\|_{W_{\text{int}}}\,\|v\|_{W_{\text{int}}} \leq C\beta^{\frac{1}{2}}|||\gamma_{\text{int}}v|||\,\|v\|_{W_{\text{int}}}\,,$$

which implies the continuity of the trace operator γ_{int} with respect to the relevant norms.

All in all, Eq. (3.62) finally implies

$$|||\phi_\ell - \psi|||^2 \leq C\left\|R_{\text{int}}\phi_\ell - R_{\text{int}}\left[\gamma_{\text{int}}\left(\lim_{j\,+\infty} R_{\text{int}}\phi_j\right)\right]\right\|^2_{W_{\text{int}}} \leq C\left\|R_{\text{int}}\phi_\ell - u\right\|^2_{W_{\text{int}}} \underset{\ell\to+\infty}{\longrightarrow} 0\,,$$

so that the sequence $(\phi_\ell)_{\ell\in\mathbb{N}}$ converges in Φ. In particular, the permutation of the limit sign with the composition of R_{int} and γ_{int} in the above expression is legitimized by the continuity of these two operators. □

Theorem 3.1. There exists a positive constant $\theta^* \in]0, 1]$ such that for any sequence of relaxation parameters $\theta_{\min} \leq \theta_\ell < \theta^*$ (where $\theta_{\min} > 0$) and for any initial guess $\lambda^0 \in \Phi$, the solution $(u^\ell_{\text{int}}, u^\ell_{\text{ext}})$ of the iterative scheme (3.59) converges to the solution $(u_{\text{int}}, u_{\text{ext}})$ of the split problem (3.56) in the sense of the norm defined by Eq. (3.60).

Proof. This theorem is proven in Ref. [265] with no additional assumption. Let us nonetheless report a rough sketch of the proof for the sake of completeness. This is done in two stages:

1. First, one shows that the convergence of the sequence $(\gamma_{\text{int}}u^\ell_{\text{int}})_{\ell\in\mathbb{N}^*}$ in Φ for the norm $|||\cdot|||$ implies the convergence of the whole sequence $(u^\ell_{\text{int}}, u^\ell_{\text{ext}})_{\ell\in\mathbb{N}^*}$ towards the solution $(u_{\text{int}}, u_{\text{ext}})$ of the split problem (3.56) for the norms $\|\cdot\|_{W_{\text{int}}}$ and $\|\cdot\|_{W_{\text{ext}}}$ respectively. The proof mainly relies on *(i)* the equivalence between the norm $\|\cdot\|_k$ defined via Eq. (3.60) and the norm $\|\cdot\|_{W_k}$ defined by Eq. (3.9), and *(ii)* on the fact that W_k equipped with the norm $\|\cdot\|_{W_k}$ is complete.

2. Second, one shows that the iteration $\lambda^{\ell-1} \leftarrow \lambda^\ell$ given by Eq. (3.59c) defines a contraction over the space of traces Φ. More precisely, there exists $\lambda^\infty \in \Phi$ such that $|||\lambda^\infty - \lambda^\ell||| \to 0$ as $\ell \to +\infty$. One then concludes with point 1.

□

Remark 3.8. Ref. [200] refers to this iterative scheme as the *Dirichlet / Neumann method*. This name comes from the fact that one iterates between the two sub-domains by imposing the continuity of the solution on $\partial\Omega_{\text{int}}$ with a Dirichlet boundary condition and by imposing the continuity of the normal derivative on $\partial\Omega_{\text{ext}}$ in a weak sense. The scheme can also be understood in terms of Poincaré-Steklov operators [200, 272]: going from Eq. (3.59a) to Eq. (3.59b) involves a Dirichlet to Neumann operator while going from Eq. (3.59b) to Eq. (3.59c) involves a Neumann to Dirichlet operator. As a side note, other domain decomposition techniques could probably be employed, possibly with overlap to enhance the theoretical convergence speed of iterations.

3.5.2 The FE approximation

The FE approximation of the iterative scheme (3.59) shares many points in common with the discretization of the *ifem* method laid out in Sec. 3.4. It requires nonetheless some slight adjustments, which we detail here. We further refer to

$$\text{“Find } u \in W^h \text{ such that } \forall v \in W^h\,,\ a(u, v) = l(v)\text{”} \tag{3.64}$$

as the *global discrete problem*.

Additional discrete spaces

The finite-dimensional spaces of interest have been defined through Eq. (3.49) for the Klein–Gordon problem 3.1 and through Eq. (3.47) for the Poisson problem 3.2. We further define

$$W^{0,h}_{\text{KG,int}} := \left\{u \in W^h_{\text{KG,int}} \text{ such that } u = 0 \text{ on } \Sigma^h\right\}, \tag{3.65}$$

$$W^{0,h}_{\text{P,int}} := \left\{u \in W^h_{\text{P,int}} \text{ such that } u = 0 \text{ on } \Sigma^h\right\}. \tag{3.66}$$

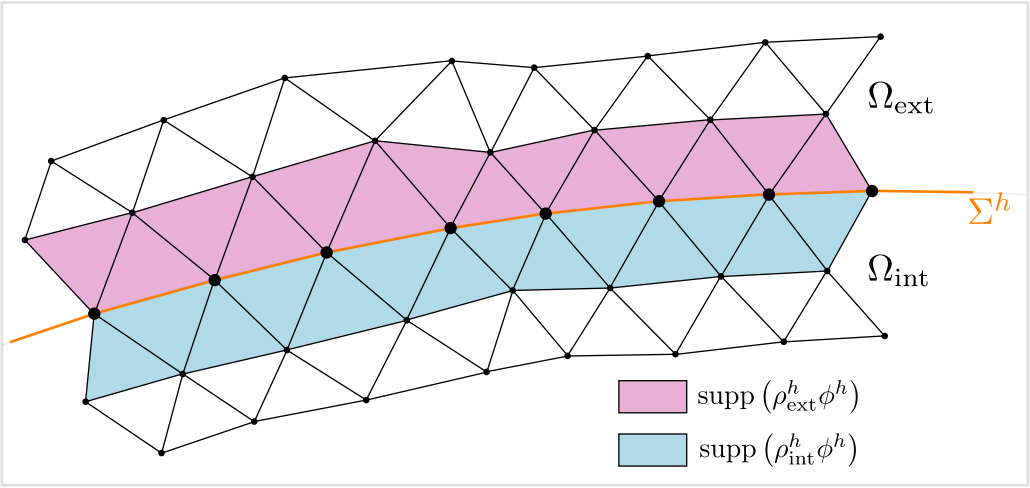


Figure 3.7: Support of extended traces by operator $\rho^h_k\colon \Phi^h \to W^h_k$ [Eq. (3.68)]

Thereafter, we use the generic notation $W^{0,h}_{\text{int}}$ to refer indiscriminately to either $W^{0,h}_{\text{KG,int}}$ or $W^{0,h}_{\text{P,int}}$. The same logic applies to W^h_{ext}. We also need a discrete counterpart of the trace space Φ given by Eq. (3.54), namely

$$\Phi^h \coloneqq \left\{\phi \in \mathcal{C}^0(\Gamma) \text{ such that } \forall I \in \Sigma^h,\ \phi_{|I} \in \mathbb{P}_k(I)\right\}, \tag{3.67}$$

where k designates the polynomial degree here.

From the continuous to the discrete iterative procedure

Following Ref. [265], the trace extension operator R_k appearing in the weak formulation (3.59) is replaced by $\rho^h_k\colon \Phi^h \to W^h_k$, where for $\phi \in \Phi^h$,

$$\left(\rho^h_k\phi\right)_{|\Sigma^h} = \phi \qquad \text{and} \qquad \left(\rho^h_k\phi\right)_{|K} = 0\ \forall K \,/\, K \cap \Sigma^h = \varnothing\,. \tag{3.68}$$

For a clear understanding, we illustrate the support of an extended trace on Fig. 3.7. Therefore, the discrete version of the iterative procedure (3.59) becomes

$$\begin{cases} a_{\text{int}}(u^\ell_{\text{int}}, v_{\text{int}}) = l_{\text{int}}(v_{\text{int}})\ ,\ \forall v_{\text{int}} \in W^{0,h}_{\text{int}} \\ \gamma_{\text{int}} u^\ell_{\text{int}} = \lambda^{\ell-1} \end{cases} \tag{3.69a}$$

$$a_{\text{ext}}(u^\ell_{\text{ext}}, v_{\text{ext}}) = l_{\text{ext}}(v_{\text{ext}}) - a_{\text{int}}(u^\ell_{\text{int}}, \rho^h_{\text{int}}\gamma_{\text{ext}} v_{\text{ext}}) + \langle f, \rho^h_{\text{int}}\gamma_{\text{ext}} v_{\text{ext}}\rangle_{\text{int}}\ ,\ \forall v_{\text{ext}} \in W^h_{\text{ext}} \tag{3.69b}$$

$$\lambda^\ell = \theta_\ell \gamma_{\text{ext}} u^\ell_{\text{ext}} + (1-\theta_\ell)\lambda^{\ell-1}\,. \tag{3.69c}$$

Theorem 3.2. Let $u^h \in W^h$ be the solution of the global problem (3.64) and set $u^h_{\text{int}} = u^h_{|\Omega_{\text{int}}}$, $u^h_{\text{ext}} = u^h_{|\Omega_{\text{ext}}}$. There exists a positive constant $\theta^* \in]0,1]$ such that for any sequence of relaxation parameters $\theta_{\min} \leq \theta_\ell < \theta^*$ (where $\theta_{\min} > 0$) and for any initial guess $\lambda^0 \in \Phi^h$, the solution $(u^\ell_{\text{int}}, u^\ell_{\text{ext}})$ of the discrete iterative scheme (3.69) converges to $(u^h_{\text{int}}, u^h_{\text{ext}})$ in the sense of the norm defined by Eq. (3.60).

Proof. Refer to Ref. [265] for a sketch of the proof. □

Remark 3.9. Several additional points mentioned in Ref. [265] must be highlighted:

- The convergence interval for the iterative scheme (i.e. the range of the relaxation parameters θ_ℓ) does not depend on the mesh discretization parameter h.
- It is possible to compute an *optimal* relaxation parameter at each iteration at low cost, see the algorithm described in Section 5 of Ref. [265].

For this iterative procedure to end, one must supplement the algorithm with a stopping criterion. For instance, one can require the relative change in the solution for two consecutive iterations to be small, i.e.

$$\frac{\|u^\ell_{\text{int}} - u^{\ell-1}_{\text{int}}\|_2}{\|u^{\ell-1}_{\text{int}}\|_2} + \frac{\|u^\ell_{\text{ext}} - u^{\ell-1}_{\text{ext}}\|_2}{\|u^{\ell-1}_{\text{ext}}\|_2} \leq \epsilon\,, \quad \text{for some } \epsilon \in \mathbb{R}^*_+\,, \tag{3.70}$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^n$ for any $n \in \mathbb{N}^*$. Equivalently, one could only restrict this criterion to the trace of the solution on Γ, that is

$$\frac{\|\lambda^\ell - \lambda^{\ell-1}\|_2}{\|\lambda^{\ell-1}\|_2}, \quad \text{for some } \epsilon \in \mathbb{R}_+^* \leq \epsilon\,. \tag{3.71}$$

Note that in the two expressions above, we made an abuse of notation by equating the functions u_{int}, u_{ext} and λ with the vectors of their respective values at the meshes' nodes.

Finally, for the sake of clarity, we provide in Fig. 3.8 a schematic view of the various problems involved in this work and the main logical links between them. In addition, Algorithm 4 summarizes a simplified version of the *a-ifem* technique.

Algorithm 4 the alternate inverted finite element method (simplified)

Inputs, initialization:
  Pick an initial guess for the trace $\lambda^0 \in \Phi^h$
  Choose a relaxation parameter $\theta \in]0, 1]$
  Define a maximum number of iterations $\ell_{\max}$ and $\epsilon > 0$ a parameter for the stopping criterion
  Build the meshes $\mathcal{T}_{\text{int}}^h$ and $\tilde{\mathcal{T}}_{\text{ext}}^h$
  Assemble iteration-independent matrices and load vectors:
    $\mathbf{A}_{\text{int}}$ associated with the bilinear form $a_{\text{int}}(\cdot, \cdot)$
    $\mathbf{L}_{\text{int}}$ associated with the linear form $l_{\text{int}}(\cdot)$
    $\mathbf{A}_{\text{ext}}$ associated with the bilinear form $a_{\text{ext}}(\cdot, \cdot)$
    $\mathbf{L}_{\text{ext}}$ associated with the linear form $l_{\text{ext}}(\cdot)$
  $\ell \leftarrow 1$ and crit $\leftarrow \epsilon + 1$
while $\ell \leq \ell_{\max}$ and crit $\geq \epsilon$ **do**
  Solve the problem $a_{\text{int}}(u, v) = l_{\text{int}}(v)$ for all $v \in W_{\text{int}}$ with $u = \lambda^{\ell-1}$ on Γ, yielding $\mathbf{U}_{\text{int}}^\ell$
  Assemble $\mathbf{B}^\ell$, the vector associated with $-a_{\text{int}}(u_{\text{int}}^\ell, \rho_{\text{int}}^h \gamma_{\text{ext}} v_{\text{ext}}) + \langle f, \rho_{\text{int}}^h \gamma_{\text{ext}} v_{\text{ext}} \rangle_{\text{int}}$
  Solve the linear system $\mathbf{A}_{\text{ext}}\mathbf{U}_{\text{ext}}^\ell = \mathbf{L}_{\text{ext}} + \mathbf{B}^\ell$, yielding $\mathbf{U}_{\text{ext}}^\ell$
  Set $\lambda^\ell = \theta \gamma_{\text{ext}} u_{\text{ext}}^\ell + (1-\theta)\lambda^{\ell-1}$ and crit $= \|\lambda^\ell - \lambda^{\ell-1}\|_2 / \|\lambda^{\ell-1}\|_2$
  $\ell \leftarrow \ell + 1$
end while

3.6 Numerical experiments

This section is devoted to the presentation of several numerical experiments to test both the *ifem* and *a-ifem* methods. They are obtained with a custom Python code using the Finite Element package Sfepy [273]. The objectives of these tests are multiple:

- check that the implementations are correct by implementing several indicators (benchmarks and errors);
- demonstrate the efficiency of the methods to solve problems posed on unbounded domains, especially showing that only few iterations are needed to converge in the case of *a-ifem*;
- compare the efficiency of *ifem vs a-ifem*;
- study empirically what happens if one decides to implement the transmission of the flux from the exterior domain to the interior one using an explicit Neumann boundary term as written in Eq. (3.58).

We make use of acronyms to refer to the various techniques involved:

- *ifem*, 'inverted finite element method', is the technique first introduced in Ref. [242], the only difference being that we make use of the Kelvin inversion rather than the polygonal inversion.
- *a-ifem*, 'alternate inverted finite element method', which is the technique described in Sec. 3.5.
- *a-ifem*$_N$, 'alternate inverted finite element method with Neumann boundary term' [Eq. (3.58)], the alternative implementation of the transmission of the flux across Γ discussed in Remark 3.7.
- *dbc*, 'true Dirichlet boundary condition', refers to the interior problem with exact essential boundary condition imposed on Γ.

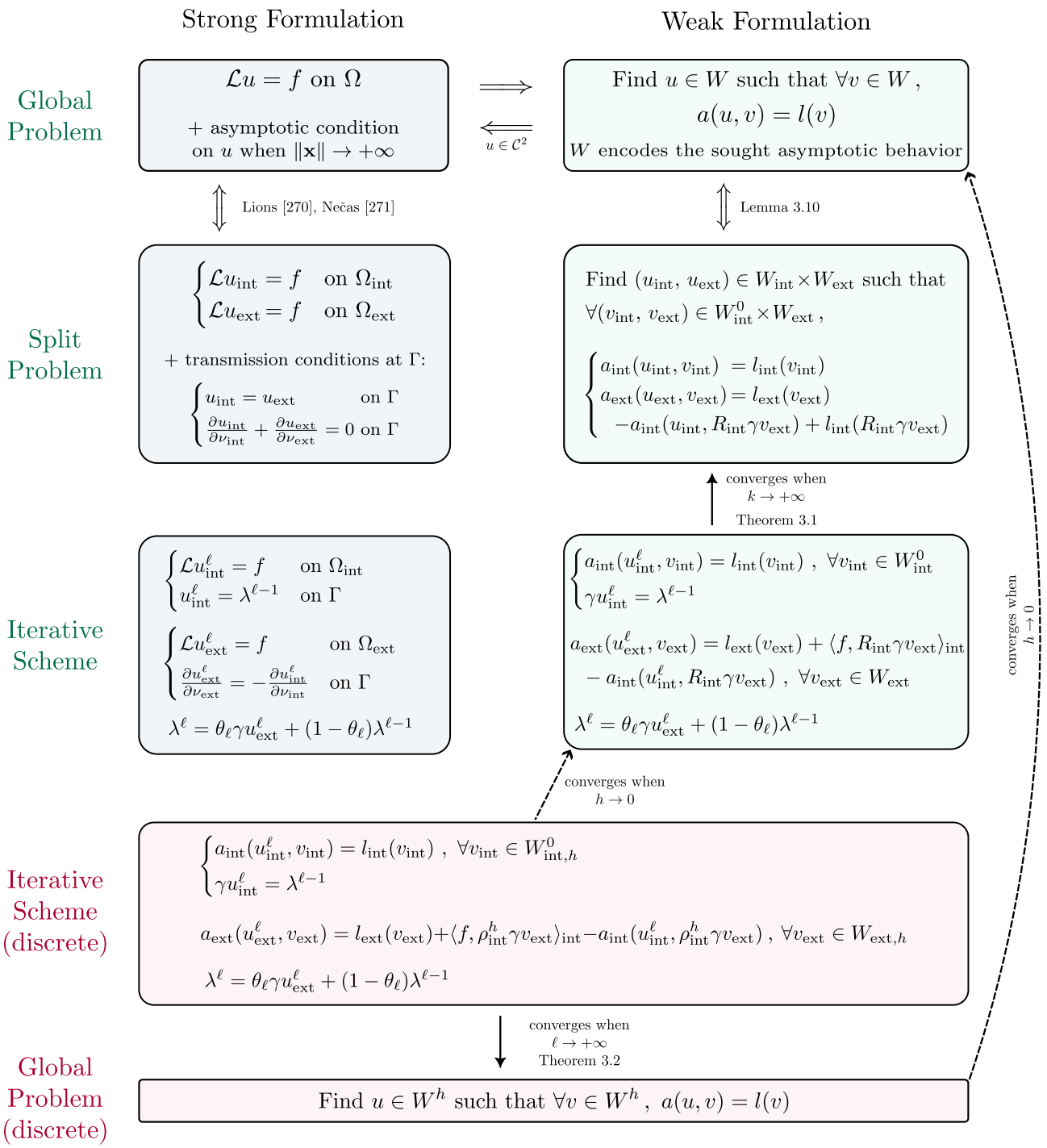


Figure 3.8: Overview of the various problems involved in the description of the *a-ifem* technique (not all logical links are represented).

3.6.1 Notes on the actual implementation

The implementation of these techniques was done using Sfepy [273] as our FEM engine, and GMSH [274] for the generation of meshes.

For *ifem*, the assembling process described in Sec. 3.4.3 results in a single stiffness matrix of size $N_{\text{tot}} \times N_{\text{tot}}$ and a single load vector of size N_{tot}, encompassing DOFs from both the interior domain Ω_{int} and the inverted exterior domain $\tilde{\Omega}_{\text{ext}}$. At the level of the code, this can be achieved relying on Sfepy's implementation of 'Linear Combination Boundary Conditions'. Specifically, the `'match_dofs'` option allows for tying together DOFs from different meshes.

In comparison, *a-ifem* is perhaps less cumbersome to implement on top of an existing FEM code as it does not involve the tying of DOFs from different meshes. The only detail requiring extra care is the assembling of the load vector in the rhs of Eq. (3.69b), which we go through here. Specifically, let us denote by $\mathbf{B}^\ell$ the load vector of size N_{ext} associated with the term $-a_{\text{int}}(u^\ell_{\text{int}}, \rho^h_{\text{int}}\gamma_{\text{ext}}v_{\text{ext}}) + \langle f, \rho^h_{\text{int}}\gamma_{\text{ext}}v_{\text{ext}}\rangle_{\text{int}}$. Then B^ℓ_j is non-zero only if the j^{th} DOF lies on the boundary $\tilde{\Sigma}^h$. For such indices j, denoting w_j the corresponding basis function, one has

$$B^\ell_j = \langle f,\, w_j\rangle_{\text{int}} - \sum_{i=1}^{N_{\text{int}}} U^\ell_{i,\text{int}}\, a_{\text{int}}(w_i, w_j)\,. \tag{3.72}$$

In the rhs of this expression, the first term is nothing but a component of the load vector of the interior problem while the sum corresponds to a component of the matrix-vector product $\mathbf{A}_{\text{int}}\mathbf{U}^\ell_{\text{int}}$, where $(A_{\text{int}})_{ij} = a_{\text{int}}(w_j, w_i)$ is the stiffness matrix associated with the interior problem while $\mathbf{U}^\ell_{\text{int}} \in \mathbb{R}^{N_{\text{int}}}$ is the interior solution vector at iteration ℓ.

3.6.2 Protocol, metrics and validation

Errors

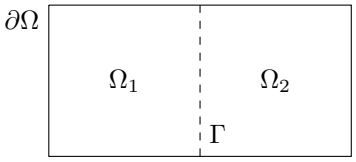
In order to assess the quality of the numerical approximations obtained thereafter, we define several relative errors:

$$\begin{aligned}
e^{\text{int}}_{L^2} &= \frac{\|u-u^h\|_{L^2(\Omega_{\text{int}})}}{\|u\|_{L^2(\Omega_{\text{int}})}} \qquad e^{\text{ext}}_{L^2} = \frac{\|u-u^h\|_{L^2(\Omega_{\text{ext}})}}{\|u\|_{L^2(\Omega_{\text{ext}})}} \qquad e^{\text{tot}}_{L^2} = \frac{\|u-u^h\|_{L^2(\Omega_{\text{int}})} + \|u-u^h\|_{L^2(\Omega_{\text{ext}})}}{\|u\|_{L^2(\Omega_{\text{int}})} + \|u\|_{L^2(\Omega_{\text{ext}})}} \\
e^{\text{int}}_{H^1} &= \frac{\|u-u^h\|_{H^1(\Omega_{\text{int}})}}{\|u\|_{H^1(\Omega_{\text{int}})}} \qquad e^{\text{ext}}_{H^1} = \frac{\|u-u^h\|_{H^1(\Omega_{\text{ext}})}}{\|u\|_{H^1(\Omega_{\text{ext}})}} \qquad e^{\text{tot}}_{H^1} = \frac{\|u-u^h\|_{H^1(\Omega_{\text{int}})} + \|u-u^h\|_{H^1(\Omega_{\text{ext}})}}{\|u\|_{H^1(\Omega_{\text{int}})} + \|u\|_{H^1(\Omega_{\text{ext}})}} \\
e^{\text{int}}_{W} &= \frac{\|u-u^h\|_{W(\Omega_{\text{int}})}}{\|u\|_{W(\Omega_{\text{int}})}} \qquad e^{\text{ext}}_{W} = \frac{\|u-u^h\|_{W(\Omega_{\text{ext}})}}{\|u\|_{W(\Omega_{\text{ext}})}} \qquad e^{\text{tot}}_{W} = \frac{\|u-u^h\|_{W(\Omega_{\text{int}})} + \|u-u^h\|_{W(\Omega_{\text{ext}})}}{\|u\|_{W(\Omega_{\text{int}})} + \|u\|_{W(\Omega_{\text{ext}})}}
\end{aligned} \tag{3.73}$$

where u denotes the exact solution to the problem at stake and u^h the numerical approximation obtained through the finite element approximation.

Checking the implementation of the iterative method

We start by checking the implementation of the iterative algorithm for the simple case where the domain Ω is an open bounded set of $\mathbb{R}^2$. In particular, we construct a rectangular domain that we divide into two squares, which in turn play the role of the two sub-domains. The transmission conditions are imposed on the shared interface Γ. The test problems are a Laplace equation and a Poisson equation with homogeneous Dirichlet boundary conditions on $\partial\Omega$.



$$\begin{cases} \Delta u = 0 & \text{in } \Omega \\ u = 1 & \text{on } \partial\Omega \end{cases} \text{ and } \begin{cases} \Delta u = 1 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

Trivially, the solution to the Laplace problem is given by $u \equiv 1$, while the solution to the Poisson problem can be approximated with standard FEM. In our test, the initial guess λ^0 on Γ is set to zero. The successful numerical implementation yielded accurate solutions with a small number of iterations.

	dbc	*a-ifem*	*a-ifem*$_N$	*ifem*	*dbc*	*a-ifem*	*a-ifem*$_N$	*ifem*	*dbc*	*a-ifem*	*a-ifem*$_N$	*ifem*
int	1.5e-3	1.5e-3	1.6e-3	1.5e-3	1.3e-5	1.6e-5	9.1e-5	1.6e-5	1.0e-7	7.1e-7	3.5e-5	7.1e-7
e_{L^2} ext	—	2.4e-3	1.7e-3	2.4e-3	—	6.2e-5	3.6e-4	6.2e-5	—	4.5e-6	1.1e-4	4.5e-6
tot	—	1.6e-3	1.6e-3	1.6e-3	—	2.5e-5	1.4e-4	2.5e-5	—	1.6e-6	4.8e-5	1.6e-6
int	2.1e-2	2.2e-2	2.2e-2	2.2e-2	8.8e-4	8.9e-4	9.2e-4	8.9e-4	3.5e-5	3.9e-5	5.7e-4	3.9e-5
e_{H^1} ext	—	6.9e-3	2.5e-3	6.9e-3	—	4.1e-4	5.0e-4	4.1e-4	—	5.5e-5	1.5e-4	5.5e-5
tot	—	2.1e-2	2.1e-2	2.1e-2	—	8.7e-4	8.9e-4	8.7e-4	—	4.0e-5	6.9e-4	4.0e-5
int	2.1e-2	2.2e-2	2.2e-2	2.2e-2	8.8e-4	8.9e-4	9.2e-4	8.7e-4	3.5e-5	3.9e-5	5.7e-5	3.9e-5
e_W ext	—	9.3e-3	3.2e-3	9.3e-2	—	5.6e-4	6.3e-4	5.6e-4	—	7.5e-5	1.9e-4	7.5e-5
tot	—	2.1e-2	2.1e-2	2.1e-2	—	8.8e-4	9.1e-4	8.8e-4	—	4.1e-5	6.8e-5	4.1e-5
θ_{opt}		0.5889				0.5880				0.5880		
N_{int}		653				14 160				328 766		
N_{ext}		1009				21 724				512 154		
N_Γ		45				270				1000		

Table 3.1: Compilation of the relative errors [Eq. (3.73)] for the Klein–Gordon problem on $\mathbb{R}^2$, for $k = 2$ and $R_c = 3$. Note that in this special case of weighted Sobolev space, $e_W^{\text{int}} = e_{H^1}^{\text{int}}$.

Creating the meshes

Meshes are created using the GMSH software [274] which is a two- and -three-dimensional finite element mesh generator with a built-in CAD[7] engine. Its Python API[8] enhances flexibility and enables us to automate the meshing of recurring geometries, e.g. disks. As mentioned in Sec. 3.4.1, one must enforce the condition that the interior and exterior meshes have matching surface elements, which is achieved in practice through custom Python routines. While GMSH implements higher-order curved elements which could be used to better approximate the circular or spherical boundary Γ, Sfepy currently cannot handle higher-order curved elements.[9] Nonetheless, other FEM softwares enjoy built-in support of curved elements. See Remark 3.5 for a discussion of this issue.

One must also bear in mind that, in the inversed exterior domain, distances get more and more stretched as we approach the origin. Indeed we already saw that the determinant of the Kelvin transform is $(R_c/\|\boldsymbol{\xi}\|)^{2n}$ — see Box H. As a consequence, we refine the mesh of $\tilde{\Omega}_{\text{ext}}$ around the origin. In that respect, Ref. [242] provides an estimate of the best approximation error, that is how $\|u - \Pi_h u\|_W$ decreases with h, where Π_h is the global interpolation operator.

Protocol

In order to remain consistent with the previous theoretical sections, we assess our iterative method *a-ifem* on the linear Klein–Gordon problem 3.1 and on the Poisson problem 3.2, in two dimensions ($n=2$). The exact form of the equations is selected so that the analytical solution is known. We examine convergence rates — that is how the error decreases when one increases the number of DOFs — for the various relative errors displayed in Eq. (3.73). In order to be able to assess these convergence rates, we define a benchmark solution labeled *dbc* that is obtained by solving the problem on the interior domain only with standard FEM and exact Dirichlet boundary conditions on Σ^h.[10] We perform the same task with the *a-ifem*$_N$ and *ifem* techniques to gain more insight into the efficiency of the various techniques we mentioned.

We also discuss the influence of other parameters such as R_c or θ by performing targeted tests. Finally, with a view to comment on the computational complexity of the *a-ifem* technique, it is crucial to determine how fast the procedure converges. In other words, what is the typical number of iterations required to converge?

3.6.3 First example: linear Klein–Gordon equation

We solve the linear Klein–Gordon equation (3.3) on $\mathbb{R}^2$ with

$$d(\mathbf{x}) = 1 \qquad \text{and} \qquad f(\mathbf{x}) = \frac{5 + \|\mathbf{x}\|^2(\|\mathbf{x}\|^2 - 2)}{(1 + \|\mathbf{x}\|^2)^3}\ , \ \forall \mathbf{x} \in \mathbb{R}^2\,.$$

[7] Computer-Aided Design
[8] Application Programming Interface.
[9] Sfepy implements *Isogeometric Analysis* (IGA) which can deal with virtually any curved geometry, but this is a departure from standard FEM and several functionalities are not available in IGA.
[10] Of course, doing so is only possible here because the analytical solution is known in advance.

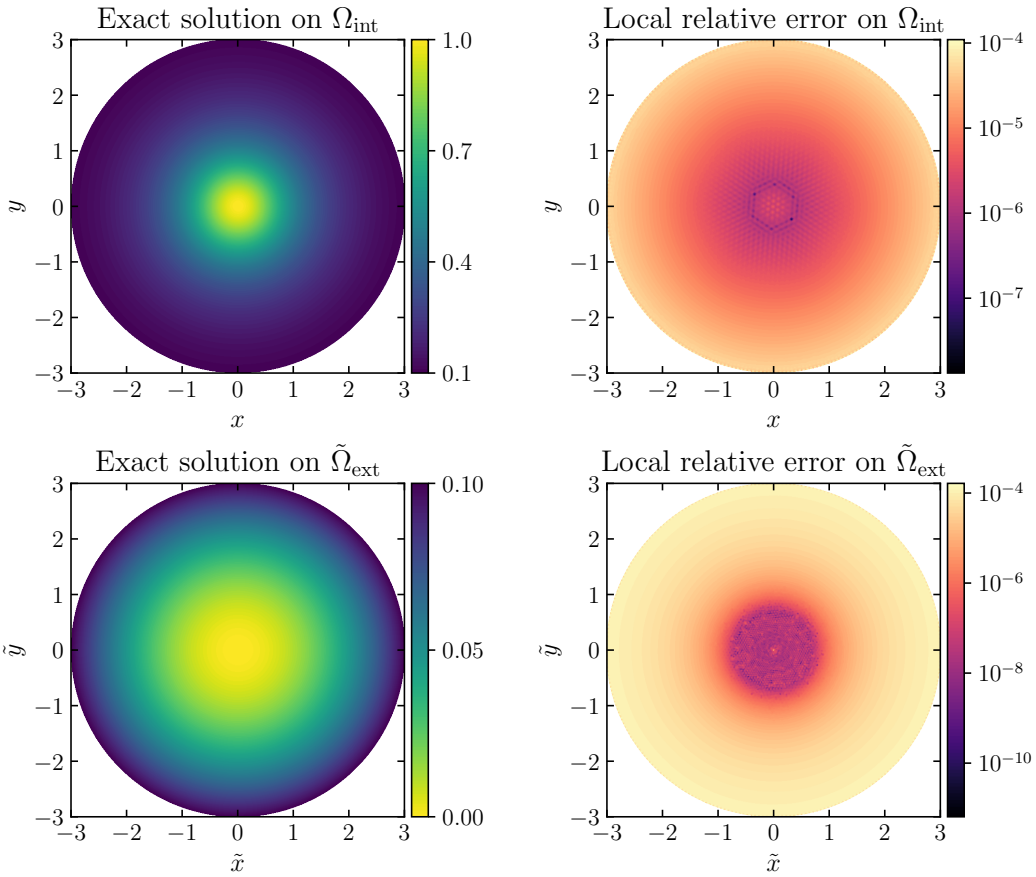


Figure 3.9: Exact solution u to the Klein–Gordon problem (left, linear-scale) and local relative error defined as $|u - u^h|/|u|$ (right, log-scale) on Ω_{int} (first row) and on $\tilde{\Omega}_{\text{ext}}$ (second row). The numerical approximation u^h was obtained with the iterative method *a-ifem* together with parameters $\{R_c = 3,\, k = 2,\, N_{\text{tot}} \simeq 3.6 \times 10^4\}$.

One can check that $f \in L^2(\mathbb{R}^2, \omega)$, where ω is the weight function introduced when discussing the explicit weight regularization technique back in Sec. 3.3.3 [Eqs. (3.36–3.37)], so that theoretical results found there hold and we can implement the explicit weight regularization approach. The solution to that problem is given by $u(\mathbf{x}) = 1/(1 + \|\mathbf{x}\|^2)$, $\forall \mathbf{x} \in \mathbb{R}^2$. With this specific choice for the function d, the coercivity (sufficient) condition derived in Lemma 3.6 becomes $R_c > 2$. In the following, we set $R_c = 3$.

As a first check, we represent in Fig. 3.9 the analytical solution (left) and the local relative error (right) on the interior domain (first row) and the inverted exterior domain (second row). Here the set of computational parameters maintains the relative error below 10^{-4} in most of the numerical domains. The exact solution is particularly well-approximated for $\|\boldsymbol{\xi}\| \leq 1$ in $\tilde{\Omega}_{\text{ext}}$, except at the very origin where the relative error is undefined because $\tilde{u}(\boldsymbol{\xi} = \mathbf{0}) = 0$.

In order to further assess the quality of the various numerical approximation, we report in Table 3.1 the relative errors in the L^2-, H^1- and W-norms for three distinct mesh refinement settings. In addition, Fig. 3.10 shows a graphical version of these results by displaying convergence curves of the various methods implemented. The benchmark, herein referred to as *dbc*, provides an approximate limit to the smallest achievable error with the other three methods implemented in this work. Note that *dbc* is restricted to the interior domain. One take-home message from this table and figure is that *a-ifem* and *ifem* perform equally-well on this test-case while *a-ifem*$_N$ exhibits larger errors for almost all configurations. This latter observation may seem counter-intuitive at first. Indeed, we stated in Remark 3.7 that the strong split problem (3.57) can lead to two equivalent weak formulations at the stage of continuous weighted Sobolev spaces. After discretization however, it is clear that

$$a_{\text{int}}(u_{\text{int}}, \rho^h_{\text{int}}\gamma v_{\text{ext}}) - \langle f, \rho^h_{\text{int}}\gamma v_{\text{ext}}\rangle_{\text{int}} \neq \int_\Gamma (\mathbf{C}\boldsymbol{\nabla} u_{\text{int}}) \cdot \mathbf{n}_{\text{int}} v_{\text{ext}} \, \mathrm{d}\Gamma \,. \tag{3.74}$$

For some unknown reason, this discrepancy is emphasized on the L^2 relative error (in the interior domain), while all three methods seem to perform equally well when we only pay attention to the relative error in norm H^1 (which is the same as the W-norm in the interior domain for this specific case of weighted Sobolev space). Besides, retaining only the results of *a-ifem* and *ifem*, the convergence rates are the same as the ones exhibited by *dbc* (except for the L^2-error with $k = 2$). The offset between *dbc*-curves and the rest is merely due to the fact

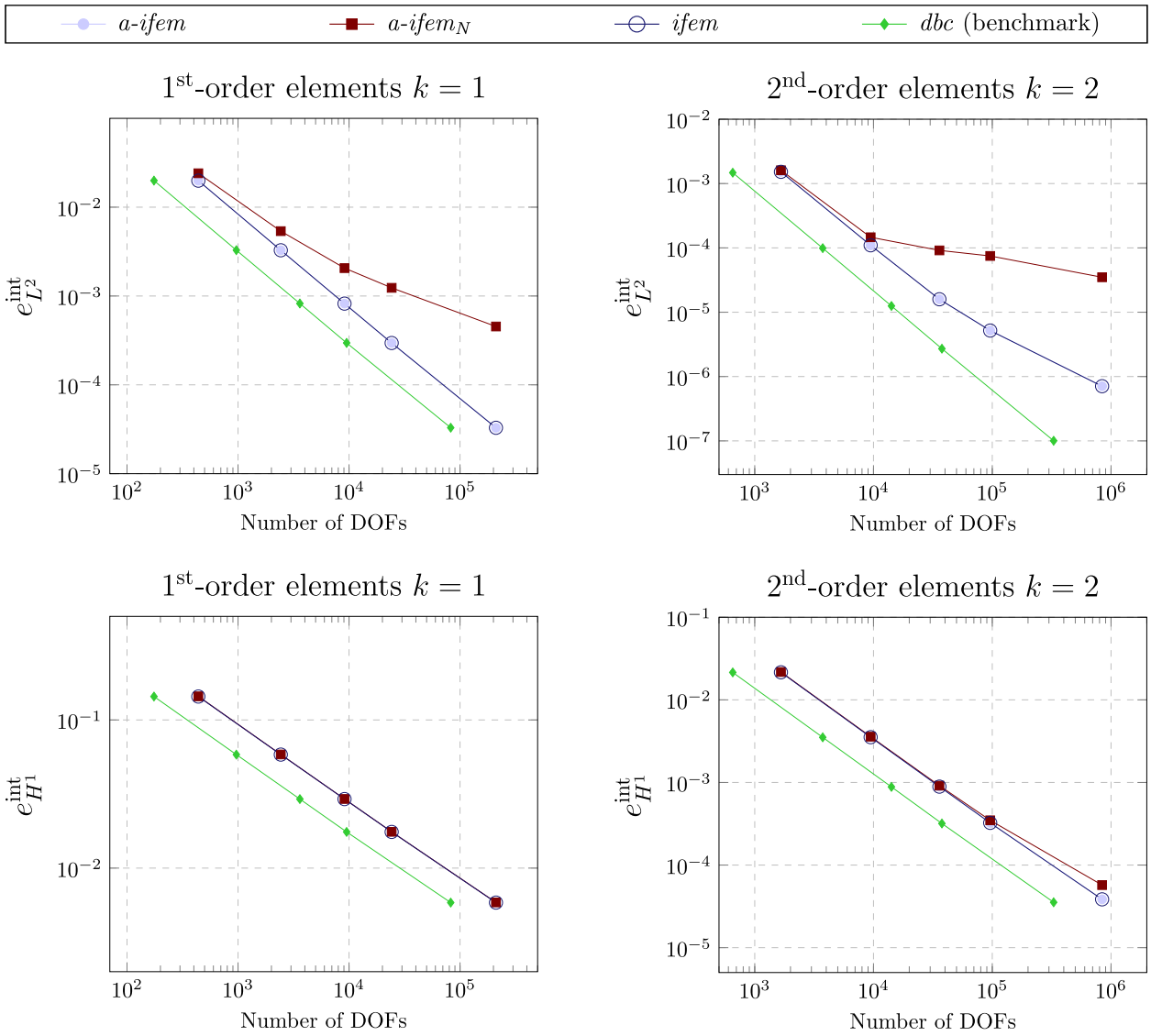


Figure 3.10: Convergence curves (i.e. relative error as a function of the number of degrees of freedom in Ω_{int}) for the unbounded Klein–Gordon problem. The first row (resp. second row) corresponds to the relative error in L^2-norm (resp. H^1-norm). The left-hand column (resp. right-hand column) corresponds to first-order (resp. second-order) triangular Lagrange elements. Note that in this case, $e^{\text{int}}_{H^1} = e^{\text{int}}_{W}$.

		dbc	*a-ifem*	*a-ifem*$_N$	*ifem*	*dbc*	*a-ifem*	*a-ifem*$_N$	*ifem*	*dbc*	*a-ifem*	*a-ifem*$_N$	*ifem*
e_{L^2}	int	7.5e-3	6.5e-2	1.6e-1	6.5e-2	3.5e-5	7.7e-4	9.0e-4	1.9e-3	5.5e-7	1.4e-4	5.6e-4	1.4e-4
	ext	—	2.0	4.8	2.0	—	3.3e-1	1.6e-1	3.3e-1	—	9.3e-2	3.8e-1	9.3e-2
	tot	—	8.0e-1	2.0	8.0e-1	—	1.4e-1	6.4e-2	1.4e-1	—	3.8e-2	1.5e-1	3.8e-2
e_{H^1}	int	6.3e-2	6.9e-2	9.3e-2	6.9e-2	1.7e-3	1.8e-3	1.8e-3	1.9e-3	1.1e-4	1.2e-4	2.7e-4	1.2e-4
	ext	—	1.7	4.2	1.7	—	2.9e-1	1.4e-1	2.9e-1	—	8.2e-2	3.3e-1	8.2e-2
	tot	—	3.8e-1	9.1e-1	3.8e-1	—	6.3e-2	2.9e-2	6.3e-2	—	1.8e-2	7.1e-2	1.8e-2
e_W	int	6.1e-2	6.3e-2	7.3e-2	6.3e-2	1.7e-3	1.7e-3	1.7e-3	1.8e-3	1.1e-4	1.1e-4	1.8e-4	1.1e-4
	ext	—	9.5e-2	2.3e-1	9.5e-2	—	2.8e-3	1.6e-3	2.8e-3	—	2.2e-4	8.5e-4	2.2e-4
	tot	—	6.9e-2	9.6e-2	6.9e-2	—	1.8e-3	1.8e-3	1.9e-3	—	1.3e-4	2.8e-4	1.3e-4
θ_{opt}		0.5041				0.5074				0.5068			
N_{int}		332				9754				147 858			
N_{ext}		472				15 142				229 410			
N_Γ		30				200				800			

Table 3.2: Compilation of the relative errors [Eq. (3.73)] for the Poisson problem on $\mathbb{R}^2$, for $k = 2$ and $R_c = 2$.

that *dbc* employs N_{int} DOFs while the other three methods employ $\sim N_{\text{int}} + N_{\text{ext}}$ DOFs.

3.6.4 Second example: Poisson equation

We solve the Poisson equation given by Eq. (3.4) on $\mathbb{R}^2$ with

$$f(x,y) = -\frac{2}{(1+r^2)^4}\left[a^2(r^4+2x^2-10y^2+1) + b^2(r^4+2y^2-10x^2+1) - (2ab)^2(2r^2-1)\right] \ \forall (x,y) \in \mathbb{R}^2\,,$$

with $r^2 = x^2 + y^2$ and $a, b \in \mathbb{R}$. Unlike the previous example, this is a *true* 2D problem in the sense that the equation is not purely radial (as long as $a \neq b$). One can check that $f \in W^{-1}_{\log}$ and behaves as $\sim \|\mathbf{x}\|^{-4}$ when $\|\mathbf{x}\| \to +\infty$. The solution to that problem is given by

$$u(x,y) = \frac{(bx)^2 + (ay)^2 - (ab)^2}{(1+\|\mathbf{x}\|^2)^2}\,.$$

Thereafter, we set $a = 4$ and $b = 1$.

There remains to fix the value of the free parameter β that appears in the definition of the hat-operator given by Eq. (3.30). With this specific choice of rhs f, one can show that $\tilde{f}(\boldsymbol{\xi})$ is proportional to $\|\boldsymbol{\xi}\|^4$. Consequently, the singularity in $\|\boldsymbol{\xi}\| \to 0$ in Eqs. (3.31–3.32) vanishes for any $\beta \geq 0$. We choose $\beta = 0$ which, as well as greatly simplifying the expression of a_{ext} [Eq. (3.31)], is compliant with the condition $\beta > -1$ — that was obtained in Proposition 3.2 — required to have $W^h_{\text{ext}} \subset W_{\text{ext}}$. Here there is no specific condition to fulfill for R_c and we therefore freely choose $R_c = 2$.

Following the same methodology as in the previous example, we report in Table 3.2 the relative errors in the various norms for three mesh refinement settings and display the corresponding convergence curves in Fig. 3.11. These additional results are in line with the conclusions we drew from the previous example. Namely, *a-ifem* and *ifem* perform equally well — the respective relative errors being equal to at least two significant digits — while *a-ifem*$_N$ delivers slightly inconsistent results depending on the considered metric.

3.6.5 Testing the influence of auxiliary parameters

Here we discuss the influence of auxiliary parameters to address important questions such as the speed at which iterations converge depending on the relaxation parameter θ or the R_c-dependence of the error. Furthermore, we have seen in the above sections 3.6.3 and 3.6.4 that *a-ifem* and *ifem* performed equally well on the two test cases. An interesting question to ask then is: which of the two methods has the lowest computational complexity?

Influence of R_c

The influence of R_c is assessed on the Klein–Gordon example in Sec. 3.6.3 by computing the local relative error along a radial line for several values of R_c. The results of this study are shown in Fig. 3.12 where we varied $R_c \in \{0.1, 0.5, 1, 2, 10, 30\}$. The error on the interior domain Ω_{int} corresponds to the blue part of the curves

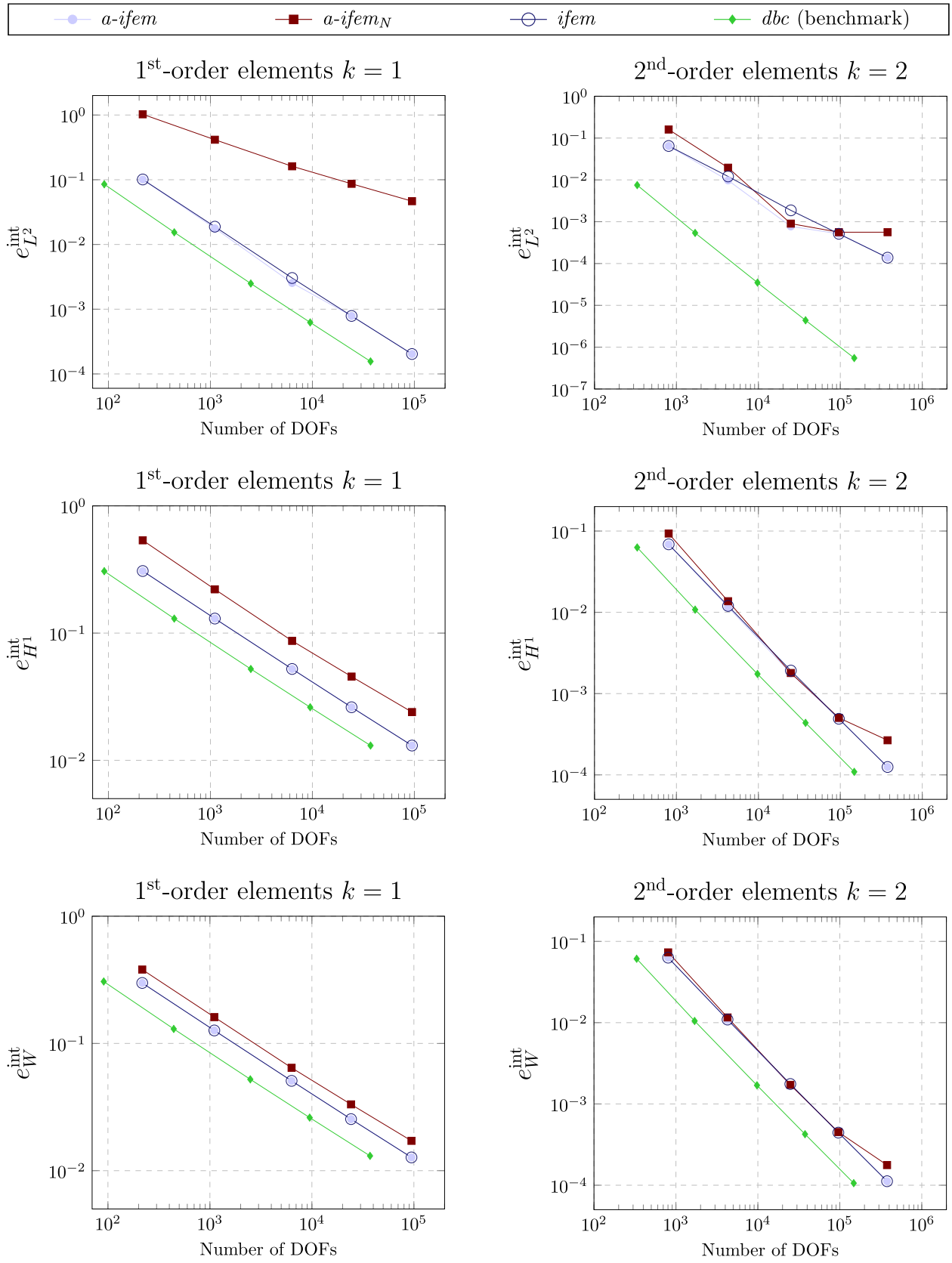


Figure 3.11: Convergence curves (i.e. relative error as a function of the number of degrees of freedom in Ω_{int}) for the unbounded Poisson problem. The first row (resp. second row) corresponds to the relative error in L^2-norm (resp. H^1-norm). The left-hand column (resp. right-hand column) corresponds to first-order (resp. second-order) triangular Lagrange elements.

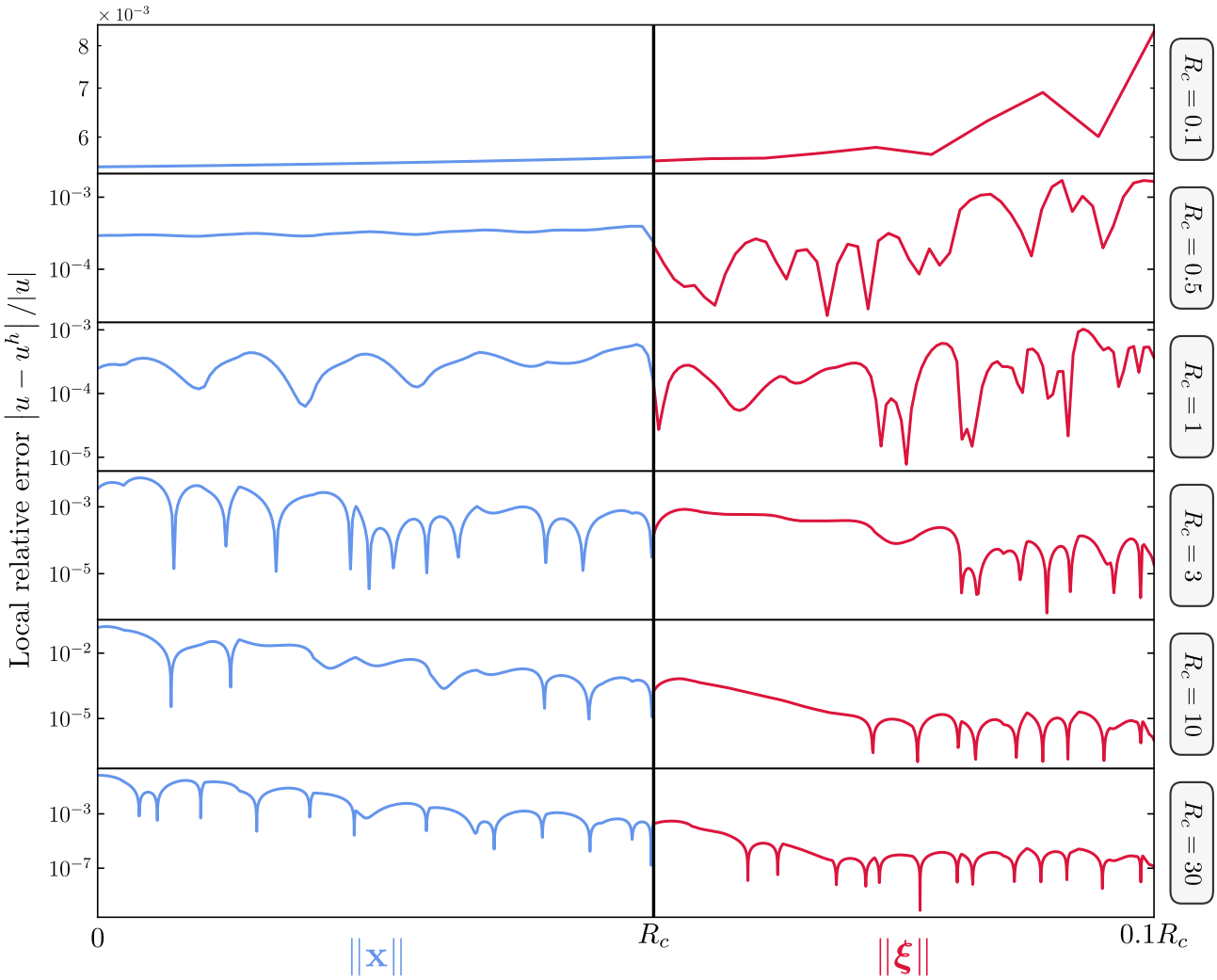


Figure 3.12: Influence of the cutoff radius R_c on the local relative error computed as $|u-u^h|/|u|$ on the Helmholtz problem. Each panel represents the local relative error as a function of $\|\mathbf{x}\|$ in Ω_{int} (blue) and $\|\boldsymbol{\xi}\|$ in $\tilde{\Omega}_{\text{ext}}$ (red), for $R_c \in \{0.1, 0.5, 1, 3, 10, 30\}$. The lower value of $\|\boldsymbol{\xi}\|$ is truncated at $0.1R_c$ due to the fact that the relative error is undefined at the origin because $\tilde{u}(\boldsymbol{\xi} = \mathbf{0}) = 0$. The same mesh is used for all computations, with parameters $\{k = 2,\ N_{\text{tot}} \simeq 1.8 \times 10^3\}$.

whereas the red part corresponds to the error on the inverted exterior domain $\tilde{\Omega}_{\text{ext}}$. Firstly, it is worth noting that taking $R_c \leq 2$ did not make the stiffness matrix singular in this specific case and provides reasonable numerical approximations with respect to the cases $R_c > 2$. It is however difficult to make qualitative comments from these curves. Fig. 3.13 provides more synthetic results in that respect, by showing for both the Helmholtz and Poisson problem how the error e_W^{tot} varies with R_c. In both cases, and regardless of the number of DOFs employed in the computation, it appears that $R_c = 1$ minimizes this error. The particularity of having a global minimum is also brought to light in the *ifem* method in Ref. [242].

Influence of θ

The speed of convergence of the iterations in the context of the *a-ifem* method is a critical point to be examined as this conditions the relevance of the method in practice. Various metrics can be used to determine the number of iterations for the algorithm to converge. In Fig. 3.14, we use the relative error e_W^{int} for that purpose. Specifically, we vary the relaxation parameter $\theta \in \{0.2, 0.4, 0.6, 0.8, 1, \theta_{\text{opt}}\}$ — where θ_{opt} denotes a somewhat optimal relaxation parameter obtained through the algorithm laid out in Ref. [265] — and look for the iteration number beyond which the error does not evolve anymore. For the Helmholtz problem (left), selecting $\theta = \theta_{\text{opt}} \simeq 0.5880$ makes the iterative procedure converge in exactly two iterations. For the Poisson problem (right), the error does not vary significantly beyond the third iteration for the choice $\theta = \theta_{\text{opt}} \simeq 0.5084$. As a remark, we observe that for the Helmholtz problem (left panel of Fig. 3.14), the error at convergence is actually slightly smaller than that of the benchmark, despite the fact that *a-ifem* and *dbc* employ the same mesh for the interior domain.

The speed of convergence of the iterations in the context of the *a-ifem* method is a critical point to be

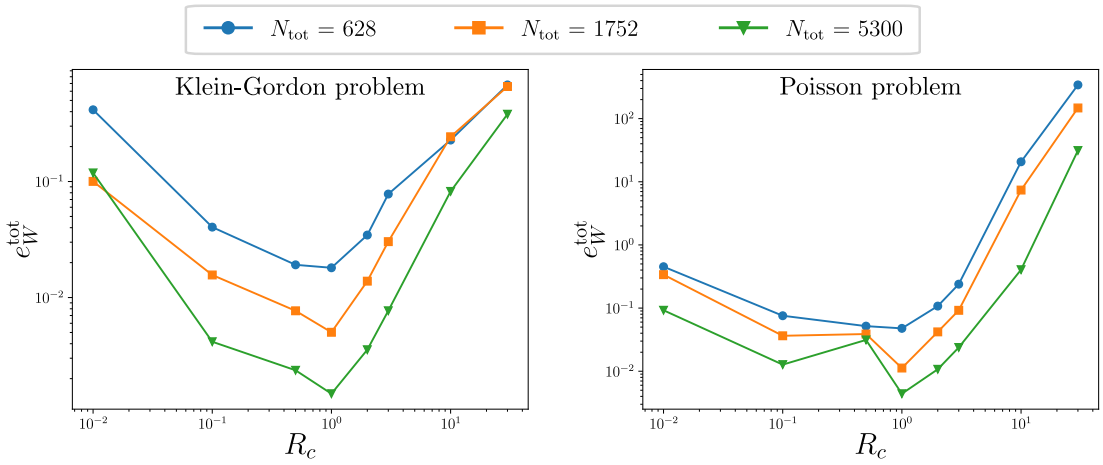


Figure 3.13: Influence of the cutoff radius R_c on e_W^{tot} (the relative error in W-norm in $\Omega_{\text{int}} \cup \Omega_{\text{ext}}$) for both the Helmholtz problem (left) and the Poisson problem (right). The total number of DOFs N_{tot} is varied in $\{628, 1752, 5300\}$.

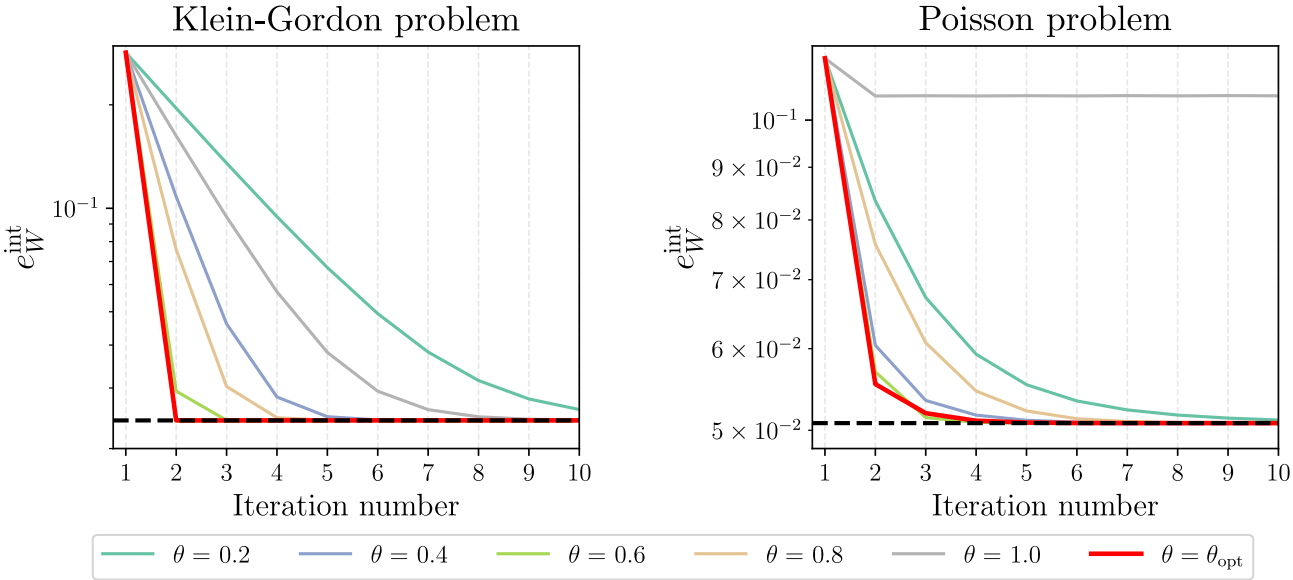


Figure 3.14: Influence of the relaxation parameter θ on the convergence of iterations. For both Klein–Gordon (left) and Poisson (right) problems, we represent the relative error e_W^{int} as a function of the iteration number for several values of $\theta \in \{0.2, 0.4, 0.6, 0.8, 1, \theta_{\text{opt}}\}$. The optimal parameter is given by $\theta_{\text{opt}} \simeq 0.5880$ for the Helmholtz problem and by $\theta_{\text{opt}} \simeq 0.5084$ for the Poisson problem. The black dashed line corresponds to the error obtained with the *dbc* method (benchmark). The numerical approximations were obtained with the iterative method together with parameters $\{R_c = 3,\ k = 1,\ N_{\text{tot}} = 6325\}$.

examined as this conditions the relevance of the method in practice. Various metrics can be used to determine the number of iterations for the algorithm to converge. In Fig. 3.14, we use the relative error e_W^{int} for that purpose. Specifically, we vary the relaxation parameter $\theta \in \{0.2, 0.4, 0.6, 0.8, 1, \theta_{\text{opt}}\}$ — where θ_{opt} denotes a somewhat optimal relaxation parameter obtained through the algorithm laid out in Ref. [265] — and look for the iteration number beyond which the error does not evolve anymore. For the Helmholtz problem (left), selecting $\theta = \theta_{\text{opt}} \simeq 0.5880$ makes the iterative procedure converge in exactly two iterations. For the Poisson problem (right), the error does not vary significantly beyond the third iteration for the choice $\theta = \theta_{\text{opt}} \simeq 0.5084$. As a remark, we observe that for the Helmholtz problem (left panel of Fig. 3.14), the error at convergence is actually slightly smaller than that of the benchmark, despite the fact that *a-ifem* and *dbc* employ the same mesh for the interior domain.

Comment on computational complexity: *ifem vs a-ifem*

The time-complexity of FEM is dominated by the linear system solving stage (source). For a generic linear system of size N, direct methods (e.g. Gauss elimination, LU and Cholesky factorizations, etc.) typically exhibit $O(N^3)$ time-complexity. In practice, the stiffness matrix resulting from a finite-element discretization is sparse and often symmetric. Taking advantage of these properties can greatly speed up the resolution — and this applies to both direct and iterative solvers.[11] Because the calculation of the theoretical complexity is complicated matter in that case, we conducted a mere empirical study on the Poisson problem and found that the time of the solving phase was roughly proportional to N^2.[12] Besides, it should be recalled that not all Galerkin methods lead to sparse linear systems. Indeed, using global or non-local basis functions — as is the case for spectral methods [275] — typically produces dense matrices.

In the following demonstration, let us remain general by assuming a $O(N^\alpha)$ complexity, $\alpha > 0$. On the one hand, the linear system to be solved in *ifem* is of size $(N_{\text{int}} + N_{\text{ext}} - N_\Gamma)$, since in this case the boundary nodes should be counted only once. On the other hand, completing a single iteration of the *a-ifem* method requires the resolution of two linear systems: the one associated with the interior domain is of size $N_{\text{int}} - N_\Gamma$ (because Dirichlet boundary conditions are applied on the boundary Γ) whereas the one associated with the inversed exterior domain is of size N_{ext}. Let O_{ifem} and $O_{a\text{-}ifem}$ the time-complexities associated with the two methods, one has

$$O_{ifem} = C_{\text{ls}}(N_{\text{int}} + N_{\text{ext}} - N_\Gamma)^\alpha \quad ; \quad O_{a\text{-}ifem} = n_{\text{iter}} C_{\text{ls}} \left[(N_{\text{int}} - N_\Gamma)^\alpha + N_{\text{ext}}^\alpha\right] ,$$

where n_{iter} is the total number of iterations undertaken to reach convergence and C_{ls} is a solver-dependent constant. For *a-ifem* to be be competitive against *ifem*, one must have

$$O_{a\text{-}ifem} \leq O_{ifem} \iff n_{\text{iter}} \leq \frac{(N_{\text{int}} + N_{\text{ext}} - N_\Gamma)^\alpha}{(N_{\text{int}} - N_\Gamma)^\alpha + N_{\text{ext}}^\alpha} .$$

This inequality on n_{iter} can be simplified in the case where $N_{\text{int}} \simeq N_{\text{ext}} = N$ and N is large enough to neglect the term N_Γ; it becomes $n_{\text{iter}} \leq 2^{\alpha-1}$. That number is expected to be somewhere between 2 and 4 depending on the properties of the linear system to be solved — which in turn determines the value of α. This has to be put into perspective with the results obtained in Fig. 3.14, where we observed that convergence is indeed reached in less than four iterations for both the Helmholtz and Poisson problems. In other words, at least for these two specific test cases, *a-ifem* exhibits comparable time-complexity to *ifem*. Let us further recall that, as observed in Secs. 3.6.3 and 3.6.4, both techniques are equivalent in terms of numerical error.

Chapter summary

In this chapter, we adapted the finite element framework to the case of unbounded domains. Guided by both theoretical and numerical considerations, we delved into techniques based on compactification transforms, more specifically on the Kelvin inversion. In that perspective, our main contribution was to propose a novel technique — the *alternate inverted finite element method* (*a-ifem*) — which builds on top of the inverted finite element method and a specific domain decomposition scheme. We established a firm mathematical ground for *a-ifem* and empirically proved its relevance on test cases in two dimensions.
This excursion into applied mathematics directly serves the objective we set ourselves, that is to be able to study scalar-tensor gravity on unbounded domains. The next chapter is dedicated to the application of the numerical techniques discussed so far to the physical situations of interest.

[11] For relatively small FEM problems, typically less than one million DOFs, direct solvers are generally preferred.
[12] This empirical study was performed with both a direct solver (`spsolve` from `scipy.sparse.linalg`) and an iterative solver (conjugate gradient, `cg` from `scipy.sparse.linalg`).

Chapter 4

Modeling gravity in scalar-tensor theories of gravity with *femtoscope*

Outline of the current chapter

In Chapter 1, we established a list of desired specifications for a numerical code dedicated to the investigation of scalar-tensor models with screening mechanisms. The two subsequent chapters 2 and 3 elucidated how such specifications can be fullfiled within the finite element framework. The present chapter deals with the actual implementation of a PYTHON code — called *femtoscope* — developed specifically for this PhD work. In particular, we provide insights into the program's architecture, its features, and present the various physical models that are used in the studies that were conducted with it. Then, *femtoscope* is showcased on the two examples of interest, namely the Poisson equation governing the Newtonian potential, and the nonlinear Klein–Gordon equation driving the dynamics of the chameleon scalar field.

This chapter is obviously not intended to be a full user's manual, but rather to shed light on the tool's main features and limitations. The code is available on GitHub at https://github.com/onera/femtoscope.

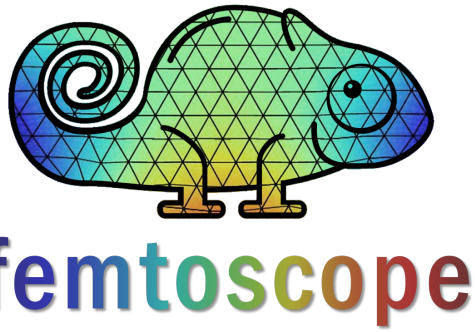


Figure 4.1: Logo of the *femtoscope* software. The original purpose of this numerical tool was to study the chameleon scalar field in the context of scalar-tensor theories of gravity with screening mechanisms, see Sec. 1.2.

Specification	FDM / multigrid [180, 186, 187, 194]	FEM codes [189–193]	*femtoscope* [137]
Asymptotic boundary condition	✗	✗	✓
Complex geometries	rectangular grid	✓	✓
Spatial dimensions	1D, 2D, 3D	1D, 2D, 3D	1D, 2D, 3D
Coordinate system	Cartesian only	Cartesian, cylindrical	Cartesian, cylindrical, spherical
Time-dependence	✗	✗	✗[a]
Multi-scale simulations	✓	possible through h-adaptivity	possible through h-adaptivity

Table 4.1: *femtoscope vs* existing numerical codes in terms of specifications (see also Table 1.4).

[a]See Secs. 2.1.4 and 2.2.4 for a possible implementation of time-dependent problems on top of *femtoscope*.

4.1 Overview of *femtoscope*

This first section is devoted to the presentation of the numerical PYTHON code *femtoscope*. After recalling the main reasons behind its development, we review the various physical problems that it can currently handle and discuss the implementation itself; i.e. the overall program's architecture. We also briefly introduce density models that are used in subsequent studies of chameleon gravity in the Earth environment.

As a side note, the choice for the program's name is threefold: *(i)* echo the MICROSCOPE space mission, *(ii)* contain 'FEM' which is the commonly adopted acronym for 'Finite Element Method', and *(iii)* contain the Danish prefix 'femto' $\rightarrow 10^{-15}$ to convey the ideas of accuracy and precision. Fig. 4.1 is the logo of *femtoscope*.

4.1.1 Motivations

The need for a new numerical tool was already expressed in Sec. 1.4. Of course, several numerical codes predate the beginning of this PhD thesis. In this respect, Sec. 1.4.1 is an attempt to take stock of all existing tools targeting the resolution of the scalar field equations of interest. The list of desired specifications we drew up, nevertheless, brought out the limitations of available options. In order to put things into perspective, Table 4.1 is an update of Table 1.4 where we have added a column for *femtoscope*. Most notably, *femtoscope* stands out from other codes because of its implementation of asymptotic boundary conditions, following the techniques thoroughly described in Chapt. 3. Setting this (crucial) feature aside, there is no denying that there are significant overlaps between *femtoscope* and the SELCIE code [190], the latter falling into the category of 'FEM codes' in Table 4.1. While the treatment of asymptotic boundary conditions could have been implemented on top of SELCIE, the development on *femtoscope* was already on track by the time SELCIE was publicly released. The two codes are compared in Sec. 4.2.2.

There are numerous advantages granted by versatile numerical tools. While analytical approximations are valuable when it comes to getting an idea of how the scalar field behaves in a given setup, some questions cannot be answered but through the use of numerical simulations. For instance, getting a little bit ahead of ourselves, we demonstrate in Chapt. 5 how *femtoscope* allows us to tackle issues related to the screening of spacecraft and the influence of the Earth's atmosphere in the context of chameleon gravity. The use of FEM, which can accommodate for virtually any distribution of matter, empowers us to study complex setups, where going from one geometry to another is merely a matter of a few lines of code (this point is perhaps best illustrated by Ref. [276]). In contrast, analytical prescriptions are generally derived on a case-by-case basis and restricted to simple geometries.

Given the above, *femtoscope* is a powerful tool for exploring scalar-tensor models in setups and regimes that are not accessible by any other means. Beyond easing model exploration, the interpretation of experimental data (collected e.g. from tests of gravity) within a given scalar-tensor model can greatly benefit from such a numerical tool when it comes to inferring accurate model constraints.

4.1.2 Physical problems and nondimensionalization of equations

Let us now turn to the physical problems that can be addressed using *femtoscope*. From a mathematical viewpoint, the code was designed to handle virtually any semi-linear elliptic PDE problem posed on $\Omega \subseteq \mathbb{R}^3$ (bounded or not). This scope encompasses, in particular:

- the Poisson equation $\Delta u = f(\mathbf{x})$, which governs the Newtonian potential Φ sourced by a given matter distribution as well as the electric potential sourced by a given charge distribution;[1]
- the linear Klein–Gordon equation of the form $\Delta u = m^2 u + f(\mathbf{x})$, obeyed by the Yukawa potential (see Sec. 1.1.3 for a derivation of this equation);
- the semi-linear Klein–Gordon equation of the form $-\Delta u = f(\mathbf{x}, u)$ where f is nonlinear in its second variable, which includes the chameleon and symmetron field equations in the static regime.

These four physical problems (namely Poisson, Yukawa, chameleon and symmetron) are *hard coded* into *femtoscope* due to their frequent use. This arguably may sound like a quite restrictive framework to work with. In reality, Newton's method, as presented in Sec. 2.2.1, not only applies to semi-linear PDEs (e.g. chameleon field equation), but also to quasi-linear and fully-nonlinear PDEs which arise in more complex screening mechanisms, see Table 1.3. Beside some technical details not worth mentioning here, the implementation of a new nonlinear PDE problem in *femtoscope* merely boils down to entering the corresponding Newton-linearized weak form in the code. Once translated in this general framework, the PDE problem at stake is automatically being handled as any other nonlinear problem in *femtoscope*. Section 4.1.3 provides further details about the program's architecture.

Regardless of the PDE at stake and whatever the physics it describes, it is always good practice in numerics to work with a *dimensionless* version of that equation. In fact, nondimensionalization has the effect of normalizing physical quantities, simplifying comparisons across different scenarios, and enhancing numerical stability by removing physical constants. Additionally, depending on how this process is performed, the resulting dimensionless equations may exhibit fewer free parameters to tune compared to their dimensional counterparts, effectively reducing the parameter space. This is known as *parameter degeneracy* — several combinations of parameters give rise to the same solution, up to a scaling factor. The detection of such degeneracies is formalized by the so-called Buckingham π theorem [277, 278], although this dimensional analysis tool is not used in the following cases.

In what follows, we go over each of the four aforementioned PDE problems and derive the dimensionless form that is implemented in *femtoscope*. In particular, we denote by L_0 and ρ_0 characteristic length and density scales respectively. Dimensionless quantities are written with a tilde (not to be confused with the tilde notation introduced for the Kelvin inversion in Sec. 3.3.2).

Newtonian potential — Poisson equation

The Poisson equation (1.4) governing the Newtonian potential Φ is straightforward to process as the theory does not have any free parameter. Denoting Φ_0 a characteristic quantity with units $\mathrm{m}^2/\mathrm{s}^2$, a dimensionless version of the Poisson equation is $\Phi_0 \tilde{\Delta}\tilde{\Phi}/(4\pi G\rho_0 L_0^2) = \tilde{\rho}$, where we have set

$$\tilde{\mathbf{x}} = \mathbf{x}/L_0 \,, \qquad \tilde{\Phi}(\tilde{\mathbf{x}}) = \Phi(\mathbf{x})/\Phi_0 \,, \qquad \tilde{\rho}(\tilde{\mathbf{x}}) = \rho(\mathbf{x})/\rho_0 \,.$$

Denoting α_{P} the dimensionless constant $4\pi G\rho_0 L_0^2/\Phi_0$, we get $\tilde{\Delta}\tilde{\Phi} = \alpha_{\mathrm{P}}\tilde{\rho}$. Therefore, the implementation of the Poisson equation in *femtoscope* reads

$$\Delta u = \alpha_{\mathrm{P}}\rho(\mathbf{x}) \,, \tag{4.1}$$

where the numerical value of the dimensionless constant α_{P} depends on the specific choice of the characteristic scales (L_0, ρ_0, Φ_0). We have dropped the tilde notation in Eq. (4.1).

Note that the form (4.1) is, as desired, 'physics-agnostic' in the sense that one could well be dealing with non-gravitational physics, e.g. electrostatics.[1] Additionally, nothing prevents us from first fixing (L_0, ρ_0) and then choosing $\Phi_0 = 4\pi G\rho_0 L_0^2$ which results in $\alpha_{\mathrm{P}} \equiv 1$. While this is correct, there is no particular gain in doing so,[2] which is why *femtoscope* implements Eq. (4.1) as is, with a user-defined parameter α_{P}.

Yukawa potential — linear Klein–Gordon equation

In Eq. (1.109), we denoted the Yukawa potential by V_Y, which encompasses both the Newtonian gravitational potential through the term GM/r, and the non-Newtonian gravity contribution through the term $\alpha\exp(-r/\lambda)/r$. Thence, this 'total potential' can be decomposed as $V_Y = \Phi + U$, where Φ is the Newtonian potential satisfying $\Delta\Phi = 4\pi G\rho$, while U satisfies a linear Klein–Gordon equation $\Delta U = U/\lambda^2 + 4\pi\alpha G\rho$. The latter PDE is the one we are interested in solving here. Given an undetermined constant U_0 with units $\mathrm{m}^2/\mathrm{s}^2$, the linear Klein–Gordon equation can be expressed as

$$\frac{U_0}{L_0^2}\tilde{\Delta}\tilde{U} = \frac{U_0}{\lambda^2}\tilde{U} + 4\pi\alpha G\rho_0\tilde{\rho} \,.$$

[1] The application to electrostatics is deliberately mentioned here as *femtoscope*, in an early version prior to this PhD work, was used to study the electrostatic stiffness of the MICROSCOPE's accelerometers.

[2] In particular, the choice of scales that leads to $\alpha_{\mathrm{P}} = 1$ does not further reduce the dimension of the parameter space, since the theory has no free parameter whatsoever!

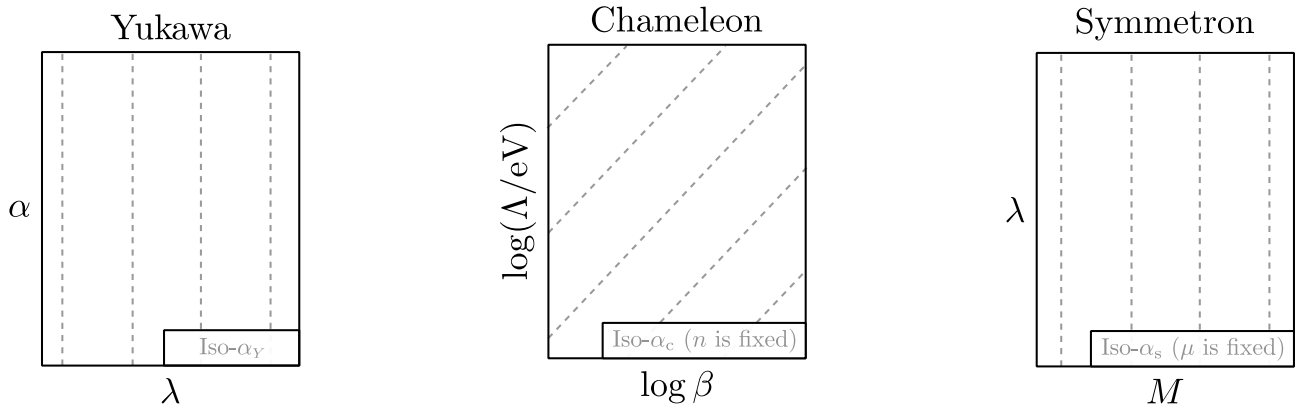


Figure 4.2: Parameter spaces of the Yukawa model, chameleon model (at fixed n), symmetron model (at fixed μ). For each of these three models, the gray dashed lines correspond to iso values of the dimensionless parameters α_Y, α_c and α_s, respectively.

Upon setting $U_0(\alpha, \lambda) = 4\pi\alpha G\rho_0\lambda^2$, this equation simplifies to $(\lambda/L_0)^2\tilde{\Delta}\tilde{U} = \tilde{U} + \tilde{\rho}$. Therefore, the implementation of this linear PDE in *femtoscope* reads

$$\alpha_Y \Delta u = u + \rho \,, \tag{4.2}$$

where $\alpha_Y = (\lambda/L_0)^2$ is the only dimensionless free parameter of the resulting dimensionless problem.

This example showcases the reduction of the parameter space of the theory. Indeed, we just went from two parameters (α, λ) in the original Yukawa model, to a single dimensionless parameter α_Y (which happens to be independent of α) appearing in Eq. (4.2). This proves extremely useful when it comes to exploring large regions of the (two-dimensional) parameter space: given a solution $u(\alpha_Y)$ to Eq. (4.2) for some fixed α_Y (which fixes λ), we have access to the dimensionful U for any α through the rescaling $U = U_0(\alpha, \lambda)u(\alpha_Y) = 4\pi\alpha G\rho_0\lambda^2 u(\alpha_Y)$ — see the left panel of Fig. 4.2. It is tempting to further use λ as our typical length scale L_0, so that $\alpha_Y \equiv 1$ and we are left with no free parameters at all. However, having the dimensionless spatial coordinate $\tilde{\mathbf{x}}$ depending on a free parameter of the model at stake is not desirable. It is rather preferable to work with Eq. (4.2) and a model-independent length scale L_0 when exploring the (α, λ)-plane, the main reason being the same mesh (and underlying numerical domain) can be used for all FEM computations in this endeavor.

Chameleon field — nonlinear Klein–Gordon equation

The chameleon field equation (1.117) features three model parameters (β, Λ, n) and is expressed in natural units (see Appendix A). Setting $\phi_0 = [nM_{\rm Pl}\Lambda^{n+4}/(\beta\rho_0)]^{1/(n+1)}$ and proceeding as in the two previous examples, we arrive at

$$\alpha_{\rm c}\tilde{\Delta}\tilde{\phi} = \tilde{\rho}(\tilde{\mathbf{x}}) - \tilde{\phi}^{-(n+1)} \,, \qquad \text{with} \qquad \alpha_{\rm c} = \left(\frac{\Lambda M_{\rm Pl}}{\beta\rho_0 L_0^2}\right)\left(\frac{nM_{\rm Pl}\Lambda^3}{\beta\rho_0}\right)^{\frac{1}{n+1}} . \tag{4.3}$$

Therefore, the implementation of this nonlinear Klein–Gordon equation in *femtoscope* reads

$$\alpha_{\rm c}\Delta u = \rho - u^{-(n+1)} \,. \tag{4.4}$$

Once again, this particular nondimensionalization reduces the number of free parameters from three (β, Λ, n) to two dimensionless parameters ($\alpha_{\rm c}$, n). For a fixed exponent n, exploring the (β, Λ)-plane can be done by solving Eq. (4.4) for various $\alpha_{\rm c}$ and adequately rescaling the resulting dimensionless solutions — see the center panel of Fig. 4.2.

Symmetron field — nonlinear Klein–Gordon equation

The symmetron model was briefly discussed in Sec. 1.2.1. With conformal factor Ω and potential V given by Eq. (1.110), the equation of motion for the scalar field is

$$\Delta\phi = \left(\frac{\rho}{M^2} - \mu^2\right)\phi + \lambda\phi^3 \,, \tag{4.5}$$

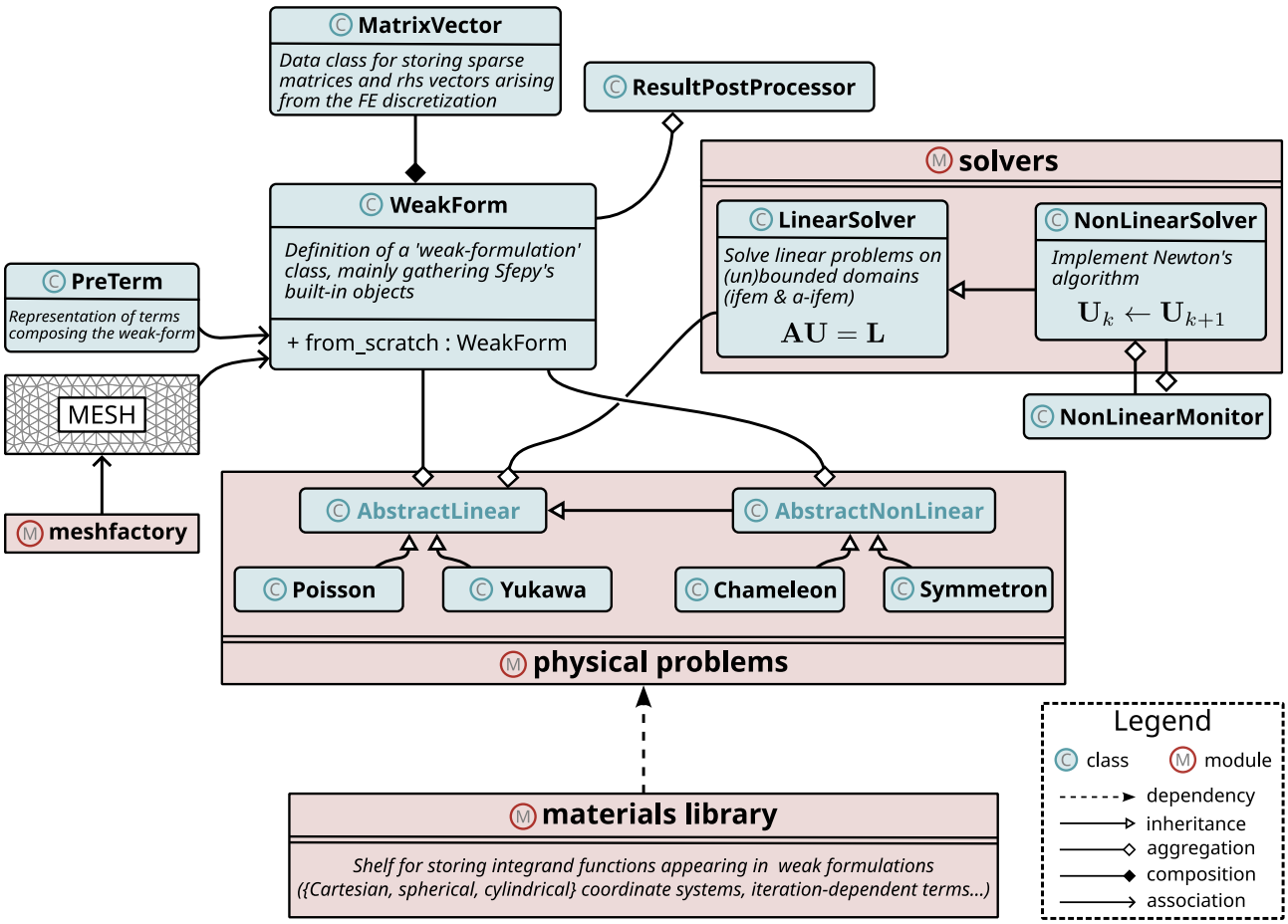


Figure 4.3: Simplified UML diagram of *femtoscope*.

where we have further made the assumption $\phi^2 \ll M^2$. Denoting μ_0 an undetermined mass scale and $\beta_\mathrm{s} = \mu/\mu_0$, this equation can be rewritten in terms of the dimensionless quantities

$$\frac{1}{\lambda\phi_0^2 L_0^2}\tilde{\Delta}\tilde{\phi} = \left(\frac{\rho_0}{\lambda\phi_0^2 M^2}\tilde{\rho} - \frac{\mu_0^2}{\lambda\phi_0^2}\beta_\mathrm{s}^2\right)\tilde{\phi} + \tilde{\phi}^3 \,.$$

Now, a relevant choice for the undetermined mass scale is $\mu_0 = \sqrt{\rho_0}/M$, so that the dimensionless equation boils down to $(M/L_0)^2\tilde{\Delta}\tilde{\phi}/\rho_0 = (\tilde{\rho} - \beta_\mathrm{s}^2)\tilde{\phi} + \tilde{\phi}^3$. Therefore, the implementation of the symmetron field equation in *femtoscope* reads

$$\alpha_\mathrm{s}\Delta u = (\rho - \beta_\mathrm{s}^2)u + u^3 \,, \qquad \text{with} \qquad \alpha_\mathrm{s} = \frac{M^2}{\rho_0 L_0^2} \,. \tag{4.6}$$

Again, this nondimensionalization has reduced the number of model parameters from three (λ, M, μ) to only two dimensionless parameters (α_s, β_s) — see the right panel of Fig. 4.2.

A summary of this nondimensionalization procedure is provided in Fig. 4.2. We have represented, from left to right, the two-dimensional parameter space of the Yukawa model (α, λ), the chameleon model (Λ, β) at fixed n, and the symmetron model (λ, M) at fixed μ. The dashed gray lines correspond to iso values of α_Y, α_c and α_s, respectively. In plain language, solutions to the dimensionless problems (4.2, 4.4, 4.6) allow one to explore the parameter spaces along these iso curves. The existence and uniqueness of solutions to the nonlinear Klein–Gordon equations (4.4, 4.6) is examined in Appendix C.

4.1.3 Program architecture

This part is a short glimpse into the inner workings of *femtoscope*. The program is coded in an object-oriented fashion — in that respect Fig. 4.3 provides its simplified UML[3] diagram, to show how the main classes and modules interact with one another. The tree structure of the software is further set out in Fig. 4.4.

[3]Unified Modeling Language.

```
~root
├── doc           # documentation of the code
├── data          # gathers all data files generated or used by femtoscope
│   ├── mesh            # mesh files (.vtk extension)
│   │   └── geo               # CAD files used by GMSH (.geo extension)
│   ├── model           # density models (PREM & US76, Solar density)
│   ├── result
│   └── tmp
├── femtoscope    # source code of femtoscope
│   ├── core            # weak form representation & solvers
│   ├── display         # visualize FEM results
│   ├── inout           # mesh generation, creation of .vtk files, post-processing
│   ├── misc            # analytical expression, unit conversions, constants...
│   ├── physics         # physical problems (Poisson, Yukawa, chameleon, symmetron)
│   └── tests           # unit & integration tests
├── images
├── script
├── tutorials     # hands-on Jupyter notebooks
├── femtoscope.yml
└── README.md
```

Figure 4.4: Simplified *femtoscope*'s tree structure.

Program workflow

Nowadays, open-source finite element codes are legion. We identified the PYTHON package Sfepy [273] as a flexible open-source FEM library[4] that meets our requirements. As the chosen FEM engine, Sfepy is the genuine corner stone of *femtoscope*: the assembling of stiffness matrices and load vectors (described in Sec. 2.1.3), the solving linear systems and the tying of DOFs from different meshes, among other critical operations, are performed through Sfepy's internal routines which can be used as black boxes. In that respect, *femtoscope* builds on top of this robust FEM engine to implement our desired features, such as a custom Newton solver to deal with nonlinear problems, or the handling of problems posed on the whole space (see Chapt. 3). The actual implementation of these features requires a significant number of lines of code $\sim O(10^4)$, which justifies calling *femtoscope* a program in its own right rather than just a collection of scripts following the nominal use Sfepy.

For solving PDE problems from the list given in Sec. 4.1.2, *femtoscope* must be provided with a mesh representing the system of interest as well as a density map $\rho(\mathbf{x})$. From there, PDE problems are solved according to the decision tree depicted in Fig. 4.5. Essentially, the philosophy is that the solving of virtually any PDE problem can be reduced to the solving of a finite sequence of simpler PDE problems — linear and elliptic — for which the basics of FEM laid out in Chapt. 2 apply. The steps of the *femtoscope*'s algorithm are (in this order):

1. If the problem is time-dependent, one can apply the techniques laid out in Secs. 2.1.4 and 2.2.4, which results in a sequence of problems independent of the time variable. Such time stepping schemes (using FDM) are not yet implemented in *femtoscope*.

2. Stationary problems fall into two categories: linear ones, and nonlinear ones. Nonlinear problems, especially semi-linear PDEs, are handled through the use of Newton iterations, where one merely solves a sequence of linearized problems until convergence is reached — see Sec. 2.2. Line search is available (see Sec. 2.2.3).

3. Finally, the last hurdle is when, after going through steps 1 and 2, the linear stationary problem is posed on an unbounded domain. There, one can leverage the various techniques exposed in Chapt. 3; let us mention *ifem* and *a-ifem*, which have been the two most used techniques throughout this PhD work.

We lay emphasis on the fact that each difficulty (time-dependence, nonlinearity, unboundedness of Ω) are thereby decoupled from each other and treated in a specific order which makes sense implementation-wise. As for the post-processing part, saving and inspecting the numerical solution are possible through dedicated routines.

[4]This code is being actively developed and maintained on GitHub: https://github.com/sfepy/sfepy. Last visited July 16$^{\text{th}}$, 2024.

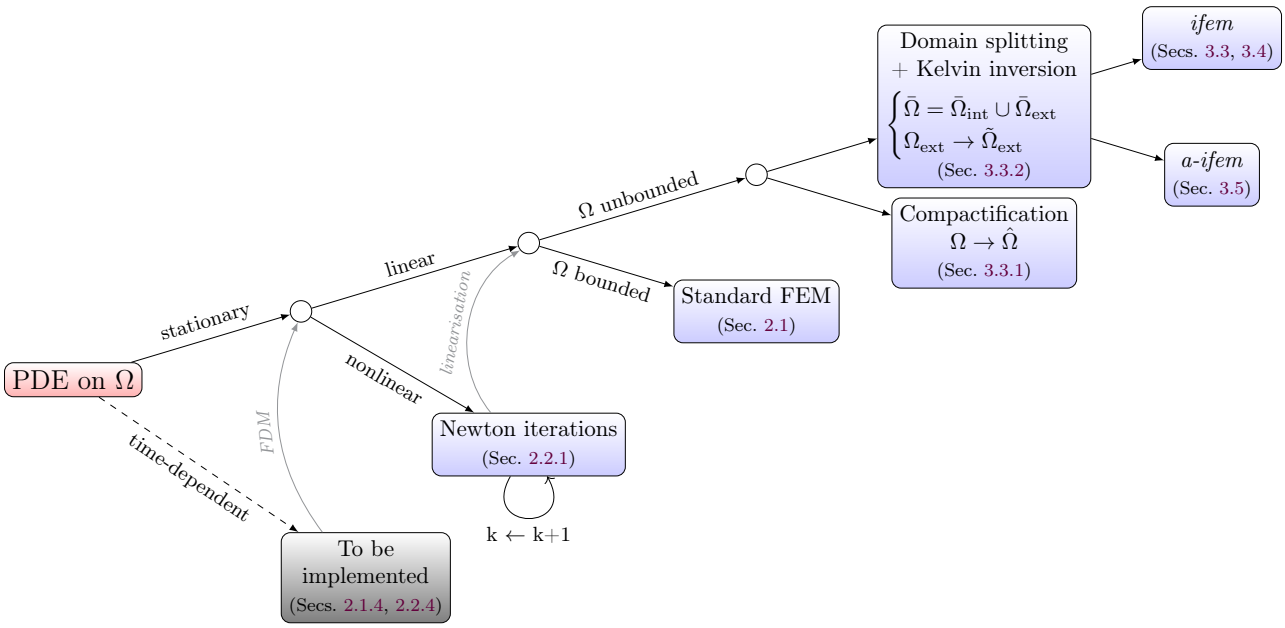


Figure 4.5: Overview of *femtoscope* decision tree depending on the nature of the PDE problem to be solved.

Mesh generation

Meshes are created using the GMSH software [274], which has already been introduced in Sec. 3.6.2. All meshes are created either through the PYTHON API (for simple geometries) or by means of `.geo` scripts (for more complex ones), and are saved using the legacy Visualization Toolkit format (VTK). Special care is in order when constructing meshes for use with *ifem* or *a-ifem*. In fact, as pointed out in Sec. 3.4.1, the mesh of the interior domain and the mesh of the inverted exterior domain must have the same trace. In 2D, this is easily achieved by imposing the coordinates of each individual line element on the boundary. In 3D, things get more complicated as controlling the position of triangles' vertices on the surface boundary is not enough to satisfy Definition 3.3. Indeed, one further has to provide a connectivity table of the surface mesh, since there is not a single way of drawing triangles from a given collection of points in 2D.

Dimensional reduction

The techniques introduced in Sec. 2.3 for taking advantage of the continuous symmetries exhibited by some problems are all implemented in *femtoscope*. In practice, this involves writing the weak form at stake in the adequate coordinates system, i.e. for which the symmetry allows one to drop at least one coordinate. In this perspective, Cartesian, spherical and cylindrical coordinate system are implemented in *femtoscope*. This process, called dimensional reduction, should be leveraged whenever possible as it greatly reduces the computational cost of FEM computations. Indeed, the representation of a symmetrical geometry in 3D ineluctably requires more DOFs (more finite elements) than in lower dimensions. Fig. 4.6 illustrates this very point on the example of a cylinder, where one greatly benefits from using cylindrical coordinates in which the geometry can be described independently of the azimuthal angle φ.

4.1.4 Implementation of physical models

For studies related to fifth force effects in the Earth environment, we need to equip ourselves with realistic density models of both the Earth's interior and its surrounding atmosphere.

Earth density model

The density inside the Earth is modeled using the so-called Preliminary Reference Earth Model (PREM) [279]. This radial model specifically integrates data from seismic waves recorded globally, laboratory measurements of material properties, and theoretical calculations to provide a detailed and widely accepted representation of Earth's density distribution from the crust to the core. The left panel of Fig. 4.7 shows the density in $\mathrm{kg\,m^{-3}}$ as a function the distance from the Earth's center r in Earth's radius. The discontinuities correspond to the interface between layers with different physical properties. The data can be downloaded from `http://ds.iris.edu/spud/earthmodel/9991844` (last visited: July 16$^{\text{th}}$, 2024).

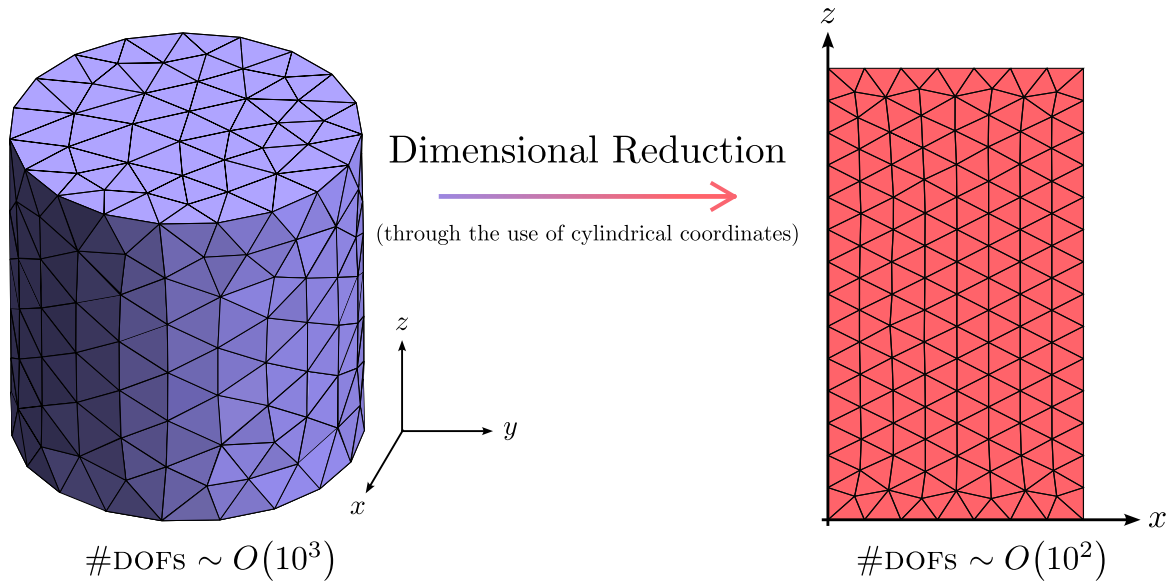


Figure 4.6: Illustration of the mesh size reduction process when dealing with axisymmetric setups (a cylinder here).

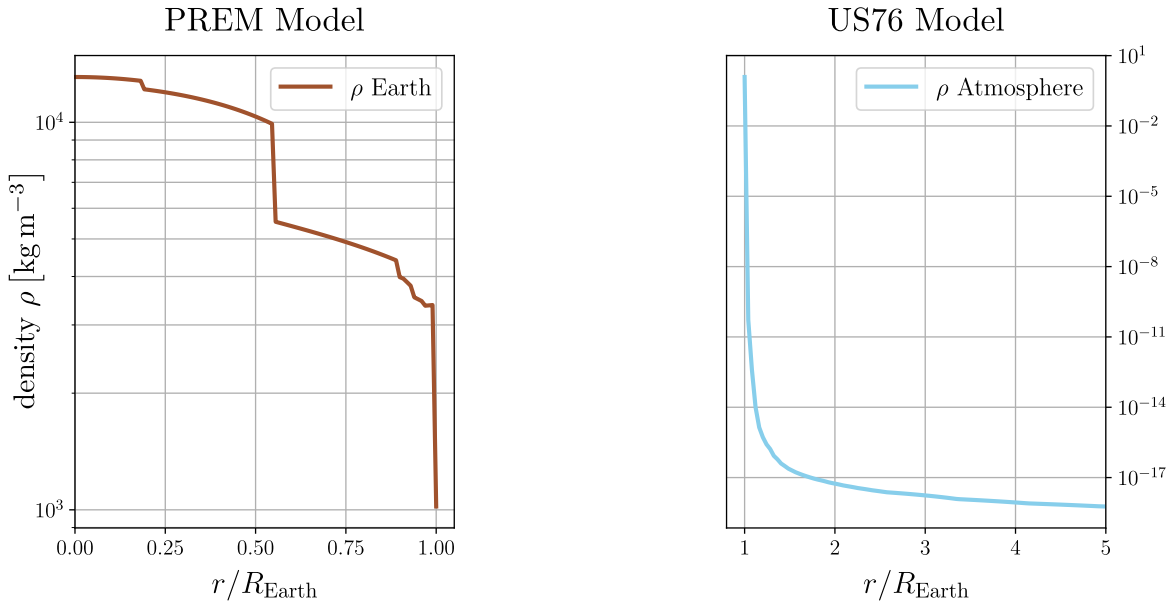


Figure 4.7: Density of matter inside the Earth according to the PREM model (left) and in the atmosphere according to the US76 model (right).

Atmospheric density models

We also use a static radial model for the Earth atmosphere — the US76 model [280] — which describes the evolution of various physical fields with respect to altitude, including density. The latter is represented on the right panel of Fig. 4.7. The data can be downloaded from http://www.braeunig.us/space/atmos.htm (especially for the density between 1000 km and 36000 km altitude; last visited: July 16$^{\text{th}}$, 2024).

4.1.5 Miscellaneous functionalities and possible improvements

Some additional functionalities of *femtoscope* are worth mentioning, starting with analytical formulas. In the context of Newtonian gravity, we provide the exact analytical expression for the potential created by homogeneous oblate spheroids (including the simpler case of the perfect sphere) — see Sec. 4.2.1. For more complex matter distributions, it can be computed semi-analytically via the integral representation

$$\Phi(\mathbf{x}) = -G \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x}')}{\|\mathbf{x} - \mathbf{x}'\|} \, \mathrm{d}\mathbf{x}' \,, \tag{4.7}$$

which follows from using the Green function associated with the Laplacian. The three-dimensional integral of Eq. (4.7) can be evaluated numerically, for instance, with Scipy's `tplquad` routine [281]. This semi-analytical method, however, should be used only sparingly — e.g. for validation purposes — because it yields the potential at a single point in space and the numerical evaluation of a triple integral is not particularly cheap. As for the chameleon model, we implement an analytical approximation for the case of the perfect sphere immersed in a lower density background, which is reported in Ref. [137]. Note that the latter slightly differs from Eqs. (1.121–1.123).

There are of course missing features from *femtoscope*, some of which have already been mentioned by now, most notably the absence of a time-dependent solver — which is necessary for studying systems that violate the quasi-static approximation. Speaking of which, we stress that the existing features of the program are as decoupled as possible from each other as conveyed by Fig. 4.5, which would definitely facilitate the smooth implementation of time-dependent problems on top of the current version of the code.

4.2 Validation of the code

Broadly speaking, code validation is the process of checking that the code is correct. In our case, we want to ensure that *femtoscope* correctly solves the model equations. This statement is quite qualitative, in practice such a validation takes different forms:

- First of all, it seems natural to ask that the PDE problem at stake is well-posed, particularly that solutions do indeed exist, otherwise there is absolutely no meaning in trying to solve such a PDE problem numerically. This concern, which mostly falls under the fields of PDE analysis and applied mathematics, is taken seriously in this PhD work: Chapt. 2 and 3 deal with the well-posedness of weak formulations of elliptic problems on bounded and unbounded domains respectively, while Appendix C focuses on the nonlinear PDEs governing the chameleon and symmetron scalar fields.

- *femtoscope* hinges on Sfepy for FEM computations [273]. Despite not being a commercial software, Sfepy is thoroughly tested and we do not question its integrity. However, *femtoscope*'s added features are prone to errors in their implementation — things can go wrong at many different stages — and preventing the apparition of bugs is all the more challenging as the code grows larger and larger. It is common practice in software engineering to create so-called *unit tests* in order to test and validate the expected behavior of individual building blocks that make up the program. In this perspective *femtoscope* is supplemented, although not exhaustively, with a number of unit tests implemented using the `pytest` framework (see Fig. 4.4).

- In the world of scientific computing, the computational engine of FEM is being used far outside what is theoretically understood, especially in the realm of nonlinear problems. In this regard, error estimates and simple problems with known solutions are vital to verification. A priori error estimates were discussed in Sec. 2.1.3 and in Box G, whereas a posteriori error estimates employ the FEM solution itself, e.g. to plot convergence curves (see Figs. 3.10 and 3.11). They are key to match FEM metaparameters (mesh size, order of the Lagrange elements...) with a given error level. The production of such convergence curves is not possible but with the knowledge of analytical solutions (or other types of benchmarks) to simple problems.

- Comparison of numerical codes can also help to identify problems related to the implementation. In the case of the chameleon, *femtoscope* was tested against SELCIE, revealing good agreement between the two codes on bounded domains.

The present section puts these various points into effect with a view to validating the implementation of both the Poisson and chameleon problems.

4.2.1 Poisson equation

Solving the Poisson equation on a bounded domain with Dirichlet boundary conditions imposed on the domain's boundary is one of the simplest and most well-known applications of FEM. Solving that same PDE on the whole space is way less common. Here, we test this important feature of *femtoscope* by solving the Poisson equation governing the gravitational potential of a flat ellipsoid of revolution on $\mathbb{R}^3$, using the various techniques discussed in Chapt. 3. Tests of this kind involve the interplay between many objects from different PYTHON classes (see Fig. 4.3) and therefore fall into the category of *integration tests*. Integration tests, unlike unit tests which focus on testing individual components or modules in isolation, aim to verify interactions between these components to ensure that they work together correctly as a cohesive system. In the following, we consider the gravitational potential of a Maclaurin spheroid (oblate spheroid) as a first integration test.

An oblate spheroid is a volume bounded by a surface defined by the equation

$$x^2 + y^2 + \frac{z^2}{1-e^2} = a^2 \iff \frac{x^2+y^2}{a^2} + \frac{z^2}{c^2} = 1\,, \tag{4.8}$$

where a is the semi-major axis and $e \in [0,\,1]$ is the ellipticity which is related to the semi-minor axis by $c = a\sqrt{1-e^2}$. The gravitational potential created by such a homogeneous body can be computed analytically — it was found by Maclaurin in the interior of the ellipsoid [282, 283], and solutions for the whole space can be found e.g. in Refs. [284, 285]. In particular, we use

$$\Phi(x,y,z) = \pi G\rho\sqrt{1-e^2}\Big[(x^2+y^2)A_1 - a^2A_2 + z^2A_3\Big] \tag{4.9}$$

with

$$A_1 = \frac{\arcsin(e) - e\sqrt{1-e^2}}{e^3}\,, \quad A_2 = 2\,\frac{\arcsin(e)}{e}\,, \quad A_3 = 2\,\frac{e - \sqrt{1-e^2}\arcsin(e)}{e^3\sqrt{1-e^2}}$$

for the interior of the ellipsoid [283], and

$$\Phi(\alpha,\,\beta) = -\frac{GM}{f}\left[\arctan\left(\frac{1}{\sinh\alpha}\right) + q_2(\sinh\alpha)P_2(\cos\beta)\right] \tag{4.10}$$

for the exterior, where the curvilinear coordinates $(\alpha,\,\beta)$, the length parameter f, the Legendre polynomial P_2 and the function q_2 are given by Eqs. (7, 8, 17, 23) of Ref. [285] respectively. Note that we have taken z as the symmetry axis, while the semi-axes a and b are aligned with the x-axis and y-axis respectively. This analytical solution [Eqs. (4.9–4.10)] serves as a benchmark for testing the implementation of the Poisson equation on unbounded domain in several coordinate systems — spherical, cylindrical (for which dimensional reduction applies, see Sec. 2.3.2) and Cartesian (in 3D).

This test case is used to produce convergence curves, giving a sense of how the 'error' decreases as the mesh underlying the FEM computation gets finer, for the various techniques discussed throughout Chapt. 3. This is more or less the same exercise as what we did with Figs. 3.10 and 3.11, except here $\Omega = \mathbb{R}^3$ and we do not use any of the errors defined by Eq. (3.73) but a mean pointwise relative error instead. By way of illustration, Fig. 4.8 displays such convergence curves obtained with *femtoscope* for spherical coordinates. The methods showcased here — *compactification* of the whole space, *a-ifem*$_N$ and *ifem* — are all described in Chapt. 3. The green curve labeled '*dbc*' is obtained via standard FEM by setting the exact Dirichlet boundary condition on the boundary of the interior domain. It constitutes a benchmark insofar it provides the convergence rate of standard FEM, for which there exist several well-known a priori error estimates, some of which are reported in Sec. 2.1.3 and in Box G. In particular, we lay emphasis on the fact that, as was observed in Figs. 3.10 and 3.11, the convergence rate of *ifem* (and *compactification*) is similar to that of the benchmark. The slight offset of the former with respect to the latter is merely due to the fact that the mesh of the inverted exterior domain $\widetilde{\mathcal{T}}^h_{\text{ext}}$ inescapably introduces additional DOFs. Last but not least, while the mean pointwise relative error may appear as an unconventional metric to choose,[5] it is particularly relevant when it comes to selecting a suitable mesh size given a desired level of precision.

[5]Most works in the literature prefer to compute the relative errors in L^2- or H^1-norm as they can be readily compared against the a priori error estimates.

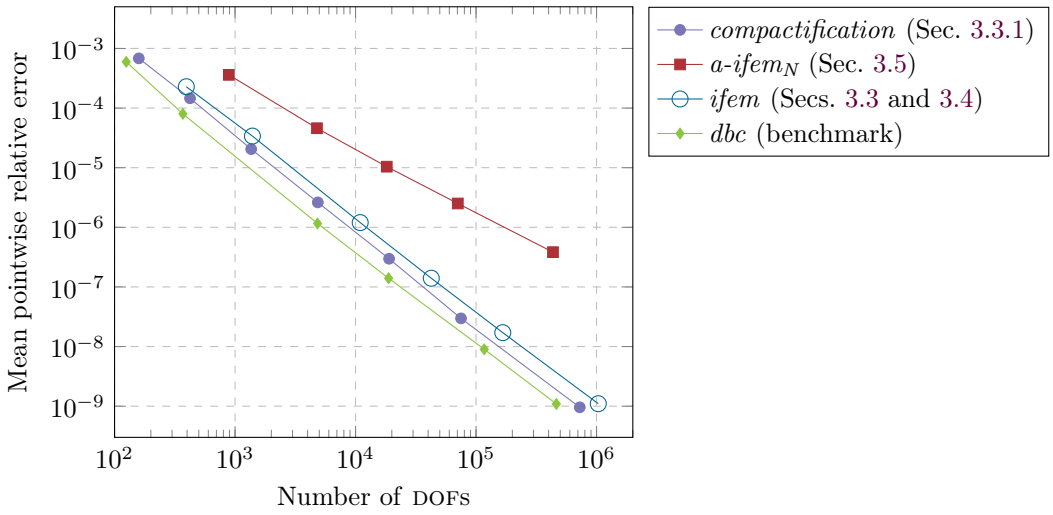


Figure 4.8: Convergence curves produced using various techniques implemented in *femtoscope* to solve problems posed on unbounded domains, namely *(i)* the *compactification* of the whole domain, *(ii)* the *a-ifem*$_N$ method, and *(iii)* the *ifem* method — see Chapt. 3 and in particular the beginning of Sec. 3.6 for the acronyms. The test problem is a Poisson equation governing the gravitational potential of a flat ellipsoid of revolution, which is solved using spherical coordinates as described in Sec. 2.3.2. The 'mean pointwise relative error' (y-axis) is computed by randomly sampling 10^4 points in the interior domain over which the FEM solution is compared against the analytical one. The *dbc* curve (green) serves as another benchmark and is obtained by applying the exact Dirichlet boundary condition on the boundary of the interior domain. The FE approximation order is set to two.

4.2.2 Klein–Gordon equation

Solving Klein–Gordon equations of the form (4.4, 4.6) on unbounded domains is one of the genuine *raisons d'être* of *femtoscope*. This type of PDE problems is the most complex to handle as one has to take care of the nonlinearity on top of *ifem* or *a-ifem*. At the level of the code however, these two issues hardly interact as depicted by the decision tree in Fig. 4.5. In particular, the custom Newton solver (class `NonLinearSolver` in Fig. 4.3) was first tested and validated for nonlinear problems posed on bounded domains. Here, we focus on the chameleon model [Eq. (4.4)]. The numerical solutions we obtain with *femtoscope* are compared against the analytical approximation (1.121–1.123) for the homogeneous perfect sphere and against SELCIE.

Comparison against the analytical approximation for the perfect sphere

Our first test case is the homogeneous sphere with radius $\tilde{R}_\mathrm{b} = 1$. Computations conducted with *femtoscope* use the *ifem* method together with $R_c = 5$. As depicted by Fig. 4.9, this process results in two solution vectors:

(a) $\tilde{\phi}_\mathrm{int}$ on the interior domain Ω_int plotted against the radial coordinate $\tilde{r}$;

(b) $\tilde{\phi}_\mathrm{ext}$ on the inverted exterior domain $\tilde{\Omega}_\mathrm{ext}$ plotted against the inverted coordinate $\tilde{\eta} = R_c^2/\tilde{r}$.

Panel (c) illustrates how $(\tilde{\phi}_\mathrm{int}, \tilde{\phi}_\mathrm{ext})$ can be put back together to form the solution in the real space for an arbitrarily large radius $\tilde{r}$ (25 in this case), continuity being ensured by *ifem*.

Fig. 4.10 shows various chameleon field profiles obtained with *femtoscope* (solid lines) for several values of the dimensionless parameter α_c [Eq. (4.3)], which we simply denote by α from now on. It is computed with the characteristic scales $L_0 = R_\mathrm{Earth} = 6371\,\mathrm{km}$ and $\rho_0 = 1\,\mathrm{kg\,m^{-3}}$, while the model parameter n is set to one. The top panel of this figure brings to the fore the importance of the FEM implementation on unbounded domains:

1. in the deeply screened regime ($\alpha \lesssim 10^{-1}$), the field quickly reaches the value that minimizes the effective potential in vacuum, so that truncating the numerical domain at R_c and setting the Dirichlet boundary condition $\tilde{\phi}(R_c) = \tilde{\phi}(\tilde{r} \to +\infty)$ would not result in too large an error;

2. for $\alpha \gtrsim 1$ however, the scalar field grows more slowly towards its asymptotic value so that it would not have any physical sense to impose such a Dirichlet boundary condition at R_c in this scenario.

Still in Fig. 4.10, we represent for each α the analytical approximation in dashed lines in the top panel, while the bottom panel depicts the relative difference of the latter with respect to the numerical solution. While there is an overall agreement between the five pairs of profiles, the relative difference systematically rises above the

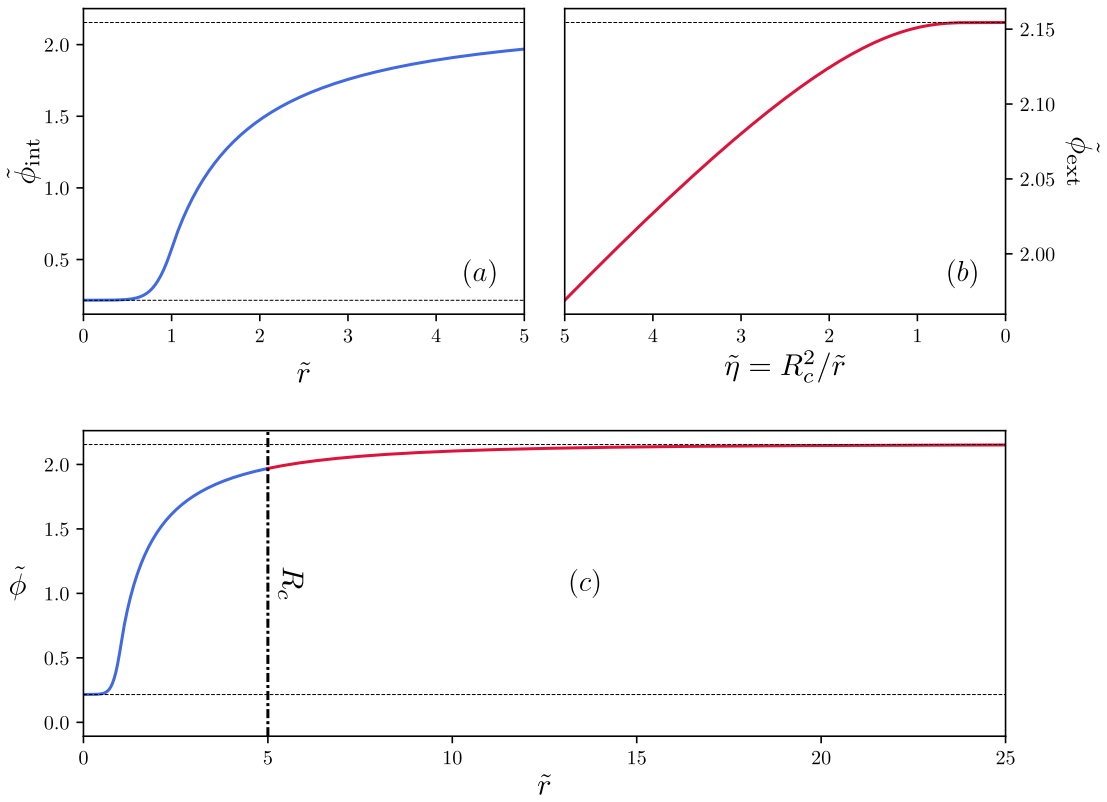


Figure 4.9: Numerical solution (dimensionless) to the radial chameleon Klein–Gordon equation on $\mathbb{R}_+$ obtained through *ifem* or *a-ifem* (see Chapt. 3). Panel (a) represents the solution on the interior domain $\Omega_{\rm int} = [0,\, R_c] \ni \tilde{r}$ (dimensionless). Panel (b) represents the solution on the inverted exterior domain $\tilde{\Omega}_{\rm ext} = [0,\, R_c] \ni \tilde{\eta}$ (dimensionless). Panel (c) shows how the solution can be reconstructed in the real space for an arbitrarily large $\tilde{r}$. The dashed lines are set at $\tilde{\phi}_{\rm min} = \tilde{\phi}(\tilde{r} = 0)$ and at $\tilde{\phi}_\infty = \tilde{\phi}(\tilde{r} \to +\infty)$. Note that there is no way to guess *a priori* the value of the field $\tilde{\phi}$ at R_c. Imposing the Dirichlet boundary condition $\tilde{\phi}(R_c) = \tilde{\phi}(\tilde{r} \to +\infty)$ would have resulted in a manifest error.

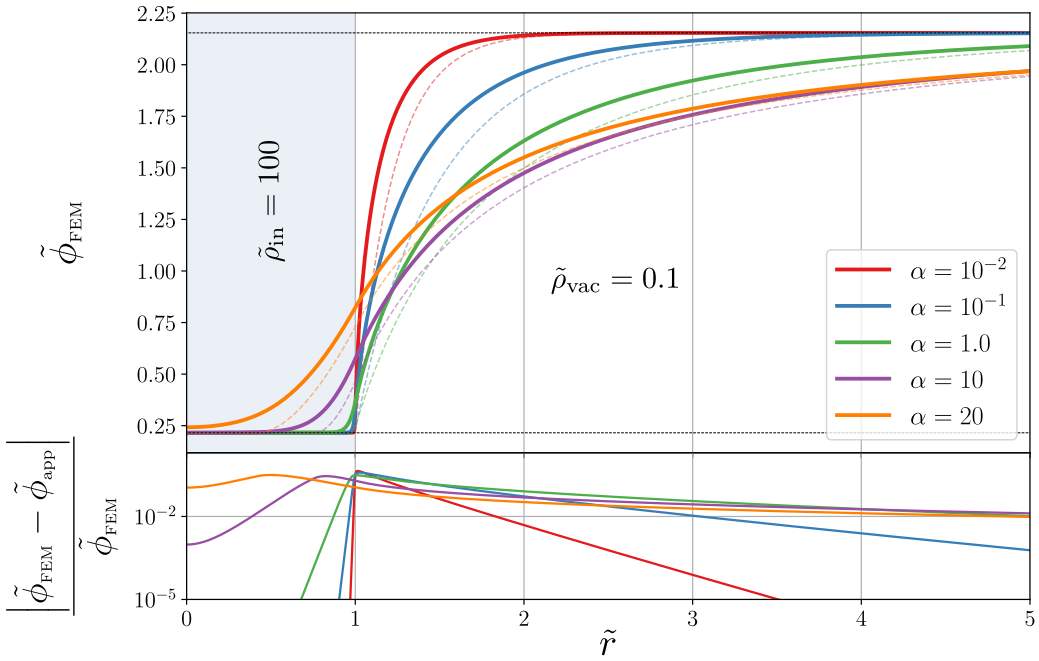


Figure 4.10: Radial profiles of the FEM chameleon field sourced by an homogeneous sphere of radius 1 and density $\tilde{\rho}_{\rm in} = 100$ immersed in a medium of lower density $\tilde{\rho}_{\rm vac} = 0.1$, for several values of α defined through Eqs. (4.3–4.4). On the top panel, the solid lines correspond to *femtoscope*'s outputs $\tilde{\phi}_{\rm FEM}$ while the dashed ones are the associated analytical approximations $\tilde{\phi}_{\rm app}$ given by Eqs (1.121–1.123). This shows the transition between the screened regime ($\alpha \in \{10^{-2},\, 10^{-1}\}$) and the unscreened regime ($\alpha \geq 20$). The bottom panel depicts the relative difference between the numerical and analytical approximations.

	Residual analytical approximation	Residual numerical solution
$\alpha = 10^{-2}$	2.5×10^{-2}	9.7×10^{-8}
$\alpha = 10^{-1}$	4.0×10^{-2}	3.6×10^{-8}
$\alpha = 1$	6.0×10^{-2}	4.2×10^{-6}
$\alpha = 10$	8.1×10^{-2}	6.7×10^{-5}
$\alpha = 20$	1.1×10^{-1}	1.2×10^{-4}

Table 4.2: Euclidean norm of the residual vector (2.78) associated with the chameleon profiles displayed in Fig. 4.10 for both the analytical approximation and *femtoscope*'s solution.

one-percent level. The analytical approximation is to blame for this observed discrepancy. Indeed, as emphasized in Refs. [136, 286], Eqs. (4.3–4.4) are only valid in certain regions of the chameleon parameter space and should not serve as a trustworthy benchmark. In this respect, Table 4.2 clearly shows that the residual of the analytical approximation is larger by many orders of magnitude than that of the numerical solution (after convergence). Besides, one may have noticed that there seems to be a relation between α and the size of the residual: especially, the greater α, the bigger the residual. This relation cannot be ascribed to a poor convergence of the Newton algorithm as the relative change of the numerical approximation between the last two iterations (in 2-norm) is consistently below 10^{-14} for all α. A better explanation is linked to the fact that the residual, as defined in this article, is an *absolute* quantity and not a *relative* one (see Appendix B from Ref. [137] for a more thorough investigation of this question).

Comparison against SELCIE

The comparison against the analytical approximation (Fig. 4.10) together with the inspection of the residual (Table 4.2) were first steps to help build confidence in *femtoscope*'s reliability. In the same spirit, we now conduct a short comparison between SELCIE and *femtoscope*.[6]

Boundary value problems in SELCIE To the best of our knowledge, SELCIE [190] is the only alternative publicly available code that can be used to compute the chameleon field for arbitrary density distributions. Despite sharing many similarities with SELCIE, *femtoscope* was developed in an independent way to achieve close aims. It is therefore all the more important to check that the outputs of the two codes coincide as no *exact* analytical solution is available. With this aim in mind, we selected a set of physical parameters and computed the chameleon field for a solid sphere with the two softwares. Unlike *femtoscope*, SELCIE cannot deal with asymptotic boundary conditions and must therefore truncate the numerical domain at some finite distance from the origin — see Table 4.1. The artificial border thereby created is left free of any Dirichlet boundary condition, hence the following natural boundary condition applies

$$\tilde{\boldsymbol{\nabla}}\tilde{\phi} \cdot \mathbf{n} = 0\,. \tag{4.11}$$

In addition to not being physically relevant in all situations — there is no reason for the field to have an everywhere-null flux across the boundary —, such an homogeneous Neumann boundary condition should raise concerns regarding the well-posedness of this PDE problem. Indeed, we saw in Sec. 2.1 that Dirichlet boundary conditions were (literally) *essential* to making the Poisson problem well-posed. Here, coerciveness is preserved without having to rely on the Poincaré inequality (see Box F) because of the very structure of the Newton-linearized version of the chameleon equation, see Eqs (2.74–2.75). This situation is purely fortuitous, and does not hold e.g. for Eq. (4.4) with $n = -4$.

Furthermore, SELCIE uses the same Newton's algorithm as *femtoscope* for dealing with the nonlinearity of the field equation (see Sec. 2.2.1). By default,[7] the field is initialized with a constant value $\tilde{\phi}_{\min}$, which is computed from the maximum density $\tilde{\rho}_{\max}$ within the numerical domain as

$$\tilde{\phi}_{\min} = \tilde{\rho}_{\max}^{-\frac{1}{n+1}}\,. \tag{4.12}$$

Comparison protocol Fig. 4.11 aims at comparing SELCIE and *femtoscope* on two different simulations performed in cylindrical coordinates with dimensional reduction. Because *femtoscope* is not limited to bounded domains, we solved the Klein–Gordon equation by way of two techniques:

1. Applying the Dirichlet boundary condition $\tilde{\phi} = \tilde{\phi}_{\text{vac}}$ at the artificial border (blue crosses).

[6] This comparison is based on the SELCIE version from September 16$^{\text{th}}$, 2022.

[7] The initial field profile ('initial guess') can also be user supplied since version 1.4.0.

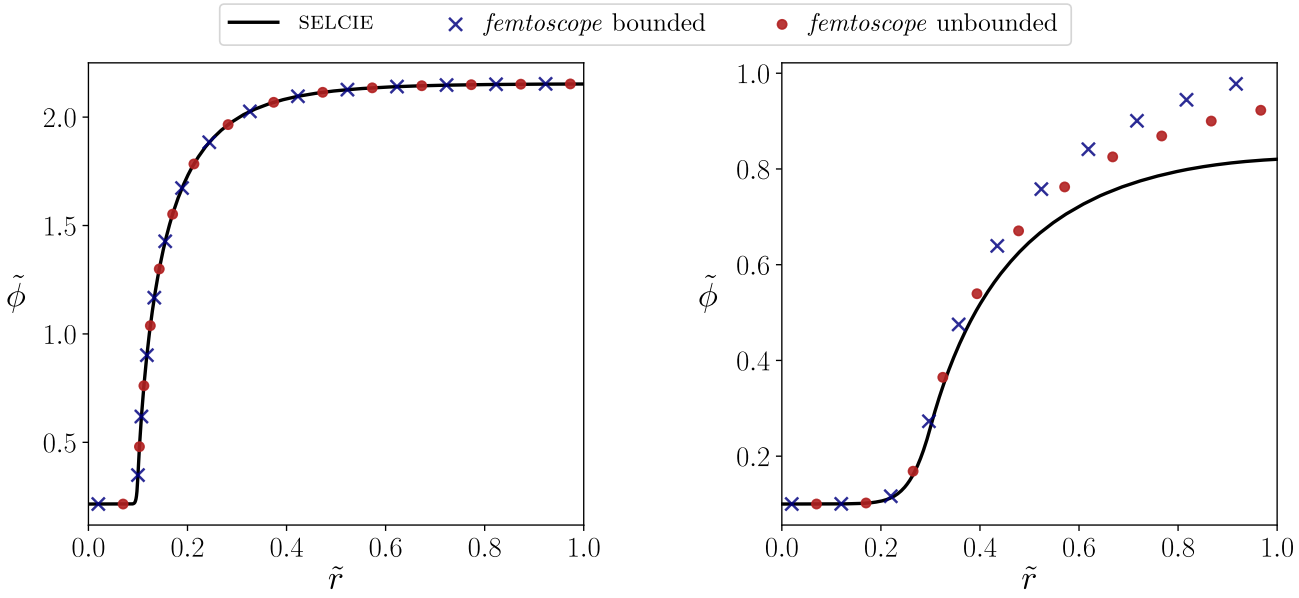


Figure 4.11: Comparison between SELCIE (black solid line) and *femtoscope* (*i*) with Dirichlet boundary condition at the truncation radius $R_c = 1$ (blue crosses), and (*ii*) by means of the *ifem* technique (red dots). The parameters used for producing this figure are $\{n = 2, \alpha = 5 \times 10^{-3}, R_c = 1, \tilde{R}_\mathrm{b} = 0.1, \tilde{\rho}_\mathrm{in} = 100, \tilde{\rho}_\mathrm{vac} = 0.1\}$ for the left panel, and $\{n = 1, \alpha = 1, R_c = 1, \tilde{R}_\mathrm{b} = 0.3, \tilde{\rho}_\mathrm{in} = 100, \tilde{\rho}_\mathrm{vac} = 1\}$ for the right panel.

2. Using the *ifem* method to enforce the correct asymptotic behavior of the field at infinity (red dots). We recall once more that this is the most general approach as no particular assumptions have to be made regarding the physical parameters of the simulation.

Left panel of Fig. 4.11 In order for SELCIE to produce a reasonable numerical approximation, we chose a set of physical parameters such that:

- The ball is screened. In this manner, the field is correctly initialized deep inside the ball via Eq. (4.12).
- The field value at the truncation boundary is close to the value that minimizes the effective potential outside the ball, denoted $\tilde{\phi}_\mathrm{vac}$. As a result, the field's gradient is expected to be small near the boundary and Eq. (4.11) makes physical sense.

The three outputs — 'SELCIE', '*femtoscope* bounded' and '*femtoscope* unbounded' — almost perfectly overlap. Indeed, the relative difference between any two of the three numerical approximations is bounded below 0.3 %. There are several potential causes to explain this sub-percentage difference:

- We did not use the same meshes for SELCIE and *femtoscope*. Yet, the quality of the FEM solution is known to be intimately interrelated with that of the associated mesh.
- For all three outputs, the field is initialized and constrained differently (see the discussion above).
- SELCIE and *femtoscope* do not use the same linear solver on these specific simulations.
- The FE approximation order is set to one in SELCIE (by default) whereas we used third-order polynomials for the computations performed with *femtoscope*.

Right panel of Fig. 4.11 This plot aims at showing the limits of the domain truncation approach. Here, '*femtoscope* unbounded' (red dots) should be regarded as the benchmark as it is the only simulation that correctly implements the asymptotic boundary condition. We can see that, as we move away from the ball ($\tilde{r} \geq 0.3$), the three outputs start diverging:

- 'SELCIE' (black solid line) implements condition (4.11) which is why we observe $[\mathrm{d}\tilde{\phi}/\mathrm{d}r](\tilde{r} = 1) = 0$. This is wrong because the field should keep increasing to $\tilde{\phi}_\mathrm{vac}$ at infinity and results in a significant deviation from the benchmark (relative difference up to 11 % at $\tilde{r} = 1$).
- '*femtoscope* bounded' (blue crosses) implements the Dirichlet boundary condition $\tilde{\phi}(\tilde{r} = 1) = \tilde{\phi}_\mathrm{vac}$. This results in a 6 % relative difference with respect to the benchmark at $\tilde{r} = 1$.

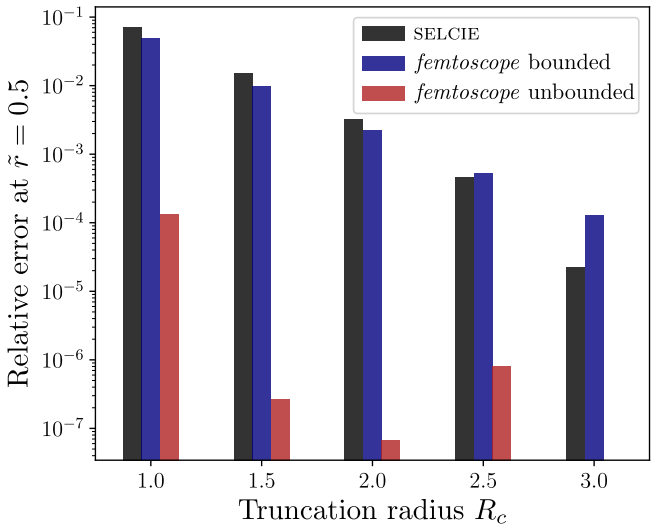


Figure 4.12: Influence of the truncation radius R_c on the relative error (not in percentage) for 'SELCIE', '*femtoscope* bounded' and '*femtoscope* unbounded'. The simulations were performed with the set of parameters $\{n = 1, \alpha = 1, R_{\rm b} = 0.3, \rho_{\rm in} = 100, \rho_{\rm vac} = 1\}$ (the same as the right panel of Fig. 4.11). Here, the chosen benchmark is { '*femtoscope* unbounded', $R_c = 3.0$}.

Influence of the truncation radius R_c Finally, we complement this comparative study by addressing the question of the influence of the size of the truncated domain on accuracy. To do so, we start from $R_c = 3$ and retrieve the numerical value $\tilde{\phi}(\tilde{r} = 0.5)$ using *ifem*: this is our benchmark. Then we decrease the truncation radius down to 1 in steps of 0.5 and compute the relative error at $\tilde{r} = 0.5$ for all three outputs. The results of this experiment are shown in Fig. 4.12. The takeaway here is that approaches based on truncation ('SELCIE' and '*femtoscope* bounded') become increasingly inaccurate as R_c decreases. Moreover, for an arbitrary set of parameters (β, Λ, n), the truncation radius ensuring an acceptable level of error cannot be known in advance. One thus has to be very cautious when using codes relying on truncation, and must have enough physical insights into how to choose the truncation radius. As for '*femtoscope* unbounded', the dependence between the error and R_c is much less pronounced, except for $R_{\rm cut} = 1$ where the relative error goes beyond 10^{-4}. This brief investigation, although it is merely based on a single example, further illustrates why properly dealing with boundary conditions is of key importance.

4.3 Examples of usage

In the above, we have endeavored to show that *femtoscope* can be used to obtain correct results on the PDE problems that must be solved in order to study screened scalar-tensor models of gravity. In this section, the use of *femtoscope* is extended to physical cases that could not be investigated by analytical means. This includes the study of chameleon gravity around the Earth with the realistic density models discussed in Sec. 4.1.4 and the mutual attraction of two spherical bodies. As a side note, let us mention that *femtoscope* is also currently being used in order to revisit Ref. [171] (constraining short range Yukawa deviation from Newtonian gravity using MICROSCOPE's technical sessions aimed at estimating the electrostatic stiffness of the sensors).

4.3.1 Chameleon gravity around the Earth — radial model

We now consider a more realistic treatment of chameleon gravity in the Earth vicinity. We look for quantitative values of the fifth force acting on test particles as predicted by the chameleon model in Earth orbit. When relevant, the altitude is chosen to be that of the GRACE-FO satellites,[8] i.e. around 500 km [287]. In order to study the effect of the chameleonic force on test particles at such altitudes, the latter needs to be quantified. To put things into perspective, the amplitude of the fifth force can be compared to other known physical effects taking place in orbit around the Earth. Especially, it is meaningful to compare it against the relativistic correction to Newtonian gravity $\delta a_{\rm GR}$, and to Newtonian gravity itself. At first order, this relativistic correction reads

$$\delta a_{\rm GR} = \frac{3}{r^3}\left(\frac{\mu_{\rm Earth}}{c}\right)^2 \simeq 10^{-9} a_{\rm Newton} \tag{4.13}$$

[8] https://gracefo.jpl.nasa.gov/. Last visited: July 13th, 2024.

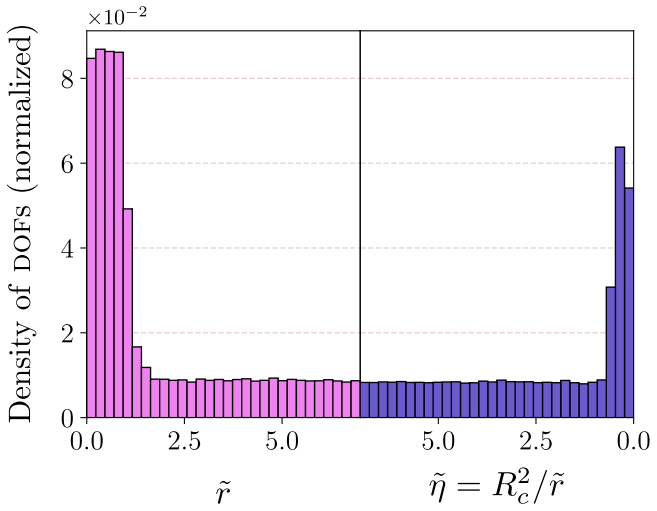


Figure 4.13: Distribution of DOFs in the mesh of the 'realistic' radial Earth model. In the interior domain (left side), the mesh is particularly refined around the various density jumps inside the Earth and at the transition between the crust and the atmosphere (see Fig. 4.7). In the inverted exterior domain (right side), the mesh is refined around the characteristic slope break at $\tilde{\eta} = \tilde{m}_{\rm vac} R_c^2/3$ which can be seen on Fig. 4.9 (b). Again, the tilde notation here is used to denote dimensionless quantities.

for a circular orbit [288], where $\mu_{\rm Earth}$ is the Earth's standard gravitational parameter and $a_{\rm Newton} = \mu_{\rm Earth}/r^2$. This is already about nine orders of magnitude smaller than Newtonian attraction for typical satellite altitudes (from low Earth orbits to the geostationary one).

Computation of the chameleon field in a realistic Earth environment

The first step is to implement a realistic model of the density inside and around the Earth as in Fig. 4.7. The use of purely radial models allows us to conduct numerical simulations in 1D, much cheaper than their 2D or 3D counterparts.[9] The density decreases from $1.3 \times 10^4\,\mathrm{kg/m^3}$ at the center of the Earth to barely $4.0 \times 10^{-19}\,\mathrm{kg/m^3}$ beyond the geostationary altitude, which represents a variation over nearly twenty-three orders of magnitude. Moreover, it is subject to a three-order-of-magnitude jump at the interface between the Earth and the atmosphere. Density being the source of the field, the mesh employed in numerical simulations has to be very fine around such rapid variations (see Fig. 4.13), and we set the relaxation parameter to $\omega = 0.5$ (experimentally determined to ensure convergence). The truncation radius is set at $7R_{\rm Earth}$ because the density is assumed constant beyond this altitude. To check the relevance of such models, we computed the Newtonian potential potential with *femtoscope* and found the conventional value of gravitational acceleration on Earth g, of about $9.8\,\mathrm{m\,s^{-2}}$.

In Fig. 4.14, we represent profiles of the chameleon field and its gradient for different values of the α parameter. The computed dimensionless field is further normalized in such a way that it tends to 1 at infinity, while the gradient is mapped onto $[0, 1]$ — which allows for a better side-by-side comparison of the profiles. The α-values are chosen so as to span over both the so-called screened regime ($\alpha \in \{10^{-8}, 1.5 \times 10^{-6}\}$) and unscreened regime ($\alpha \in \{3.5 \times 10^{-6}, 10^{-5}\}$). As can be seen on the inset, in the screened regime, the field is subject to jumps occurring at density jumps within the Earth, before stalling when the density crosses some threshold. This is the region where the corresponding gradient curve peaks to its highest value, before decreasing as r^{-2} in the upper atmosphere and beyond. In the unscreened regime, the field does not reach the value that minimizes the effective potential at the center of the Earth. One point worth mentioning is that, in the latter case, the field profiles are all identical up to an affine transformation. As a consequence, the associated normalized gradients almost perfectly overlap. A physical interpretation of this phenomenon is that in the unscreened regime, the field is sourced by the entire mass of the Earth and there is no thin-shell effect. The overall shape of the gradient is reminiscent of the Earth Newtonian gravity, which makes sense considering that the gravitational potential is not subject to any screening mechanism. Finally, let us denote by $\alpha_{\rm screened} \simeq 2.6 \times 10^{-6}$ the value at which the transition between the two regimes occurs.[10]

[9] *De facto*, the Earth flattening at the poles cannot be taken into account despite being one of the major perturbing accelerations [288]. Ref. [143] shows that ellipsoidal departures from spherical symmetry results in an enhancement of the chameleonic force.

[10] For $n = 2$, one would have $\alpha_{\rm screened} \simeq 3.1 \times 10^{-3}$.

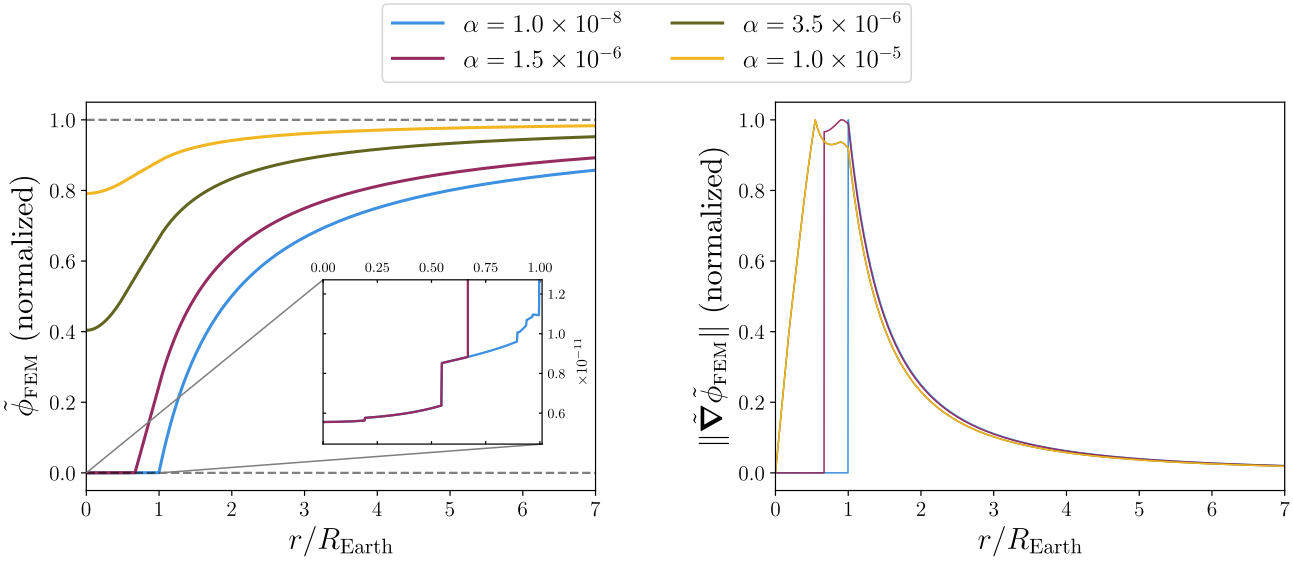


Figure 4.14: Normalized chameleon radial profile (left) and gradient (right) for the realistic 1D model of the Earth with densities depicted in Fig. 4.7. The chosen values for α are such that the Earth is either screened or unscreened, where the transition between the two regimes occurs at $\alpha \sim 2 \times 10^{-6}$ roughly.

Fifth force in orbit

Let us now proceed to a more quantitative analysis of the chameleon field effects by computing the fifth force supposedly applied on satellite in orbit — under the assumption that the latter is unscreened.[11] This force is computed according to Eq. (1.98). The mapping $(\beta, \Lambda) \mapsto \alpha$ not being injective (see e.g. Fig. 4.2, center panel), it is relevant to study the shape of the iso-fifth-forces in the (β, Λ)-plane. Using the analytical approximation (1.121–1.123) is a good starting point to get a sketch of such contour lines. Because this approximation can only handle constant density profiles inside and outside the sphere, we separately average the PREM and US76 models depicted in Fig. 4.7 and keep the two mean values. The result of this process is shown in the left panel of Fig. 4.15, where we can clearly see the demarcation between the two regimes across the line $\alpha_{\rm screened} \simeq 2 \times 10^{-6}$. The Earth is screened (respectively unscreened) below (respectively above) this line. Note that we obtain the same characteristic iso-force contours as in Fig. 7 from Ref. [143] [plotted in the $(\log \Lambda, -\log \beta)$-plane].

It is striking to note that in the unscreened regime, the fifth force almost no longer depends on the energy scale Λ. This is particularly visible on the analytical approximation. From Eq. (1.123), one has

$$\tilde{\phi}'(\tilde{r}) = K(1 + \tilde{m}_{\rm vac}\tilde{r})\,\tilde{r}^{-2} \exp\Big[-\tilde{m}_{\rm vac}\big(\tilde{r} - \tilde{R}_{\rm b}\big)\Big] .$$

However, the vacuum density used in this study is so small ($\rho_{\rm vac} = 4.04 \times 10^{-19}$ kg/m^3) that, at a satellite's altitude, $m_{\rm vac} r \ll 1$ and thus

$$\tilde{\phi}'(\tilde{r}) \sim K/\tilde{r}^2, \text{ with } K \sim \tilde{\rho}_{\rm in}/3\alpha \implies \tilde{\phi}' \propto \Lambda^{(n+4)/(n+1)} .$$

The dimensionful version of the force is recovered by multiplying the dimensionless gradient by the factor $\beta\phi_0/(\mathrm{M_{Pl}} L_0) \propto \Lambda^{(n+1)/(n+4)}$. Consequently, the result of this multiplication does not depend on Λ — QED.

The insights gained in the above helps us to comment on the results obtained with *femtoscope* and the realistic density model. The right panel of Fig. 4.15 is the numerical counterpart of its left panel, where we have represented the curves of equation $a_\phi = 10^k \delta a_{\rm GR}$ for $-2 \leq k \leq 1$. We obtain the same characteristic iso-fifth-forces (L-shaped), whose equations roughly reads

$$\begin{cases} \beta \sim \text{const.} & \text{for } \alpha \geq \alpha_{\rm screened} \\ \Lambda \sim \kappa \beta^{-1/5} & \text{for } \alpha < \alpha_{\rm screened} \end{cases} \tag{4.14}$$

for some positive constant κ. The power $-1/5$ can be recovered from the analytical approximation which gives $-n/(n+4)$ in the general case.

This kind of plot has to be put into perspective with the current existing constraints on the chameleon model, see e.g. Fig. 4 from Ref. [152]. Cases for which the Earth is unscreened (i.e. $\alpha > \alpha_{\rm screened}$) are excluded unless β

[11] We shall take a closer look at this specific assumption in Chapt. 5.

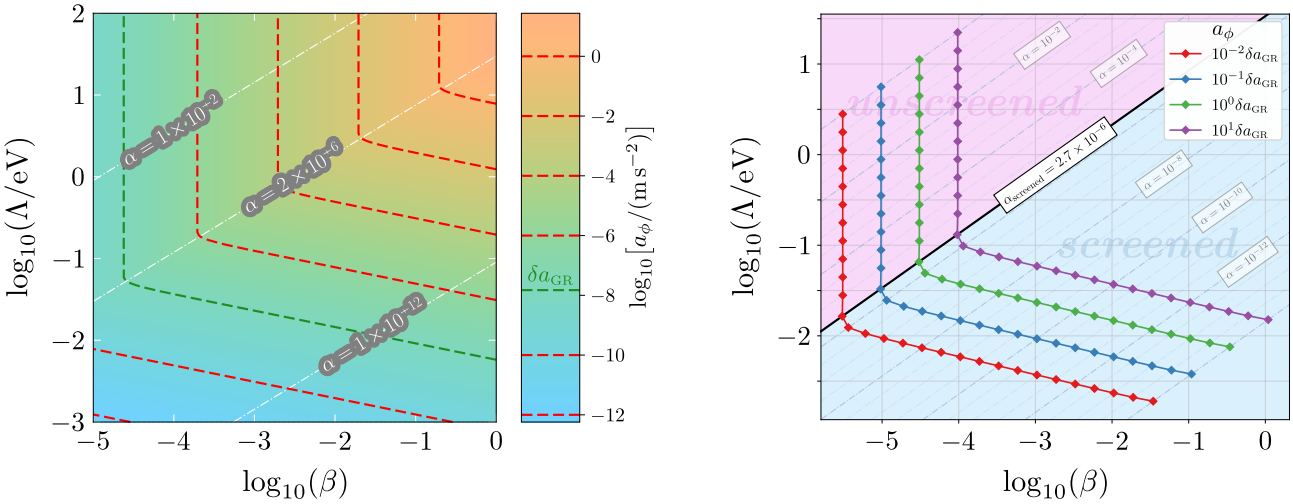


Figure 4.15: Amplitude of the fifth force in orbit (altitude = 500 km) in the (β, Λ)-plane, $n = 1$. The left panel is derived from the analytical approximation (1.121–1.123). Iso-a_ϕ values are depicted by red dashed lines, except for the contour $a_\phi = \delta a_{\rm GR}$ which is represented in green [Eq. (4.13)]. Likewise, the right panel shows fifth force iso-lines of the form $a_\phi = 10^k \delta a_{\rm GR}$, with k ranging from -2 to 1, obtained with *femtoscope*. Gray lines in the background correspond to the iso-values of the α parameter used in the numerical computations. The screened regime (blue shade) and the unscreened regime (purple shade) are unmistakably separated on both sides of $\alpha_{\rm screen} \simeq 2.7 \times 10^{-6}$.

is very small [157]. In the screened regime, it appears that the chameleon acceleration is an increasing function of β, at fixed energy scale Λ and fixed altitude. This is true only up to a certain threshold on β, above which the field is so coupled to matter that its dynamics become frozen, i.e. $\phi \propto \rho^{-1/(n+1)}$. This effect is discussed at more length in Chapt. 5.

Finally, it is useful to see how the chameleonic force compares to our current description of gravity in space. To that extent, we reproduced in Fig. 4.16 the traditional representation of satellite perturbations as a function of the altitude (see e.g. Fig. 3.1 from Ref. [288]). It features the Newtonian gravity $a_{\rm Newton} = \mu_{\rm Earth}/r^2$ and its relativistic correction at first order given by Eq. (4.13) as well as fifth force profiles. Yet, the two pairs (β, Λ) which result in an unscreened Earth are already ruled-out by experiments — see e.g. Fig. 2 from Ref. [289]. Below the threshold $\alpha_{\rm screened}$, the freezing of the field inside the Earth means that the exterior field profile is sourced only by the mass outside the thin-shell radius. This puts into question the validity of a purely radial density model. Indeed, the shell sourcing the field might be so thin that it is no longer possible to make the assumption that the Earth is spherically-symmetric. In which case, it is reasonable to expect the fifth force to be dependent on the local landform. FEM would then be necessary to capture the aspherical shape of the topography. Again, this issue is investigated in Chapt. 5.

Finally, a commentary has to be made with respect to the use of realistic physical quantities. Specifically, we noticed that numerical issues can arise when density varies widely within the simulation domain. Part of the chameleon parameter space associated with an unscreened Earth ended up inaccessible to our numerical tool as the relative variation of the field $(\phi_{\rm max} - \phi_{\rm min})/\phi_{\rm max}$ would be of order $\sim 10^{-14}$, close to machine epsilon in double-precision floating-point format.

4.3.2 Fifth force between two spheres

So far, we have showcased *femtoscope* on test cases where gravity is sourced by a single body. The field profiles thereby obtained, through their gradient, allow us to compute the geodesics of the Jordan-frame metric which test particles follow — this is the *standard approach* that is used in most works relying on analytical approximation. For an extended test body, this approximation is justified as long as it is unscreened. Now take the Moon — the test body — in free fall around the Earth — the source body. In Newtonian gravity, the linearity of the Poisson equation allows us to decompose the total potential as $\Phi_{\rm tot} = \Phi_{\rm Earth} + \Phi_{\rm Moon}$. This basic *superposition principle* does not apply a priori to the chameleon field (or any other scalar field from scalar-tensor models with screening) because its equation of motion is plagued with a nonlinearity. Yet, there are various interesting physical situations involving two spherical bodies: not only the {Earth, Moon} system but essentially any two (isolated) planetary bodies in the Solar system, a binary neutron star inspiral, or test masses in a Cavendish-like laboratory experiment. All these relevant scenarios are hardly accessible by analytical means only, although some approximations exist in the literature [72, 142, 149, 290, 291].

This part is not intended to be a complete study of the two-body problem in chameleon gravity, but rather a

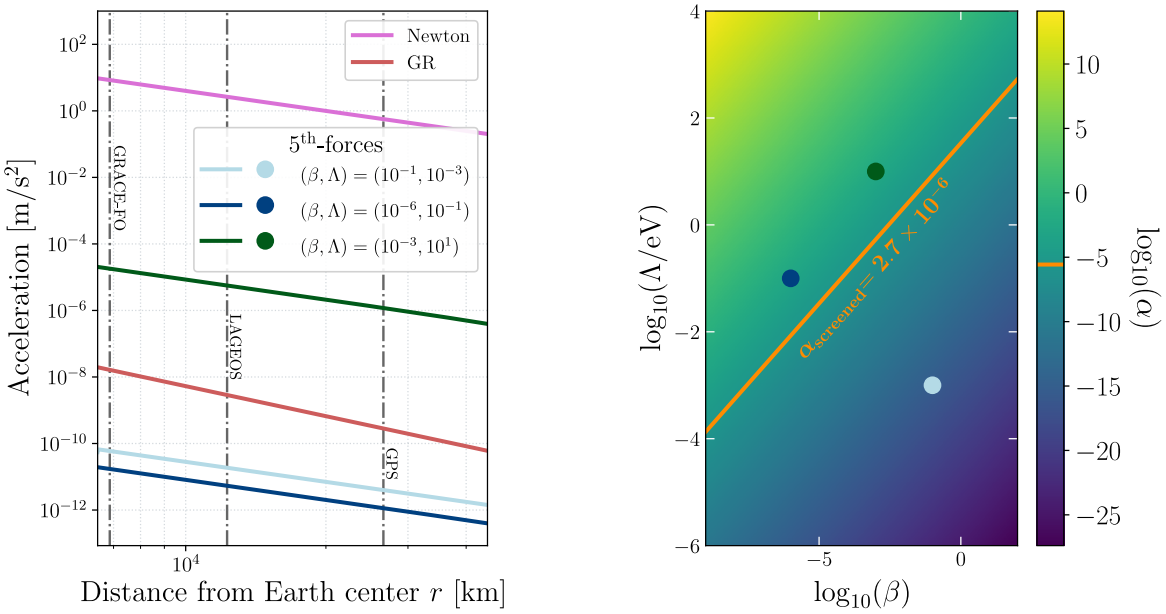


Figure 4.16: The chameleonic force as a perturbing acceleration for satellites. Left panel: orders of magnitude of hypothetical fifth forces alongside known forces [Newtonian gravity and its first order relativistic correction (4.13)] as a function of r. Right panel: the (β, Λ) values used to compute such fifth forces.

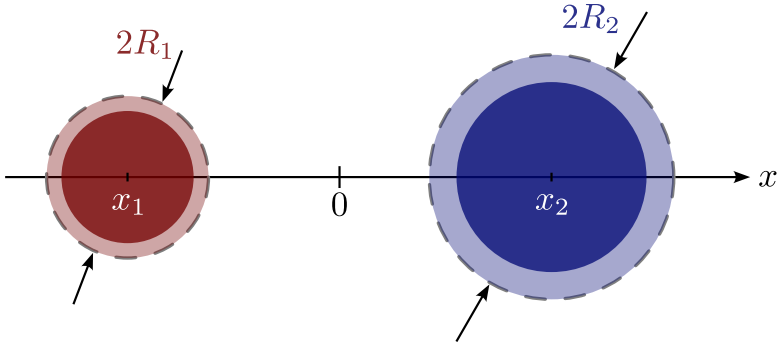


Figure 4.17: The two-body problem in chameleon gravity. The x-axis passes through the center of the two spherical bodies of radius R_1 and R_2. The dashed lines indicate the boundary of each body and the use of two shades represent the screening of the bodies (as in Fig. 1.4 where we illustrate the chameleon field profile for screened and unscreened spheres).

demonstration of the possibilities offered by FEM-based numerical tools.

Example of computation

The notation of the problem we tackle here are reported in Fig. 4.17. The x-axis, passing through the center of each body, is such that the problem is invariant by rotation about it. Consequently, all FEM computations can be performed in 2D thanks to the dimensional reduction technique (see Sec. 2.3). In particular, we make use of the cylindrical coordinate system. Moreover, the asymptotic boundary condition $\tilde{\phi} \to \tilde{\phi}_{\text{vac}}$ as $\|\tilde{\mathbf{x}}\| \to +\infty$ is handled via the *ifem* technique.

Fig. 4.18 shows the scalar field profile together with the magnitude of its gradient for a test case with parameters:

- $\tilde{R}_1 = 1.0$, $\tilde{x}_1 = -3/2$, $\tilde{\rho}_1 = 10^2$ [sphere 1];
- $\tilde{R}_2 = 0.8$, $\tilde{x}_2 = +3/2$, $\tilde{\rho}_2 = 5 \times 10^2$ [sphere 2];
- $\tilde{\rho}_{\text{vac}} = 10^{-3}$, $\alpha = 0.1$, $n = 1$ [other parameters].

The left panels of the figure show the dimensionless scalar field profile. The top panel corresponds to the full 2D map. There, the horizontal faint dashed line is the $\tilde{x}$-axis, along which the solution is plotted on the bottom panel, where the shaded areas represent the space occupied by the two spheres. In this situation, the spheres are *screened* and exhibit a very thin shell. The right panels of the figure show the 2-norm of the gradient of

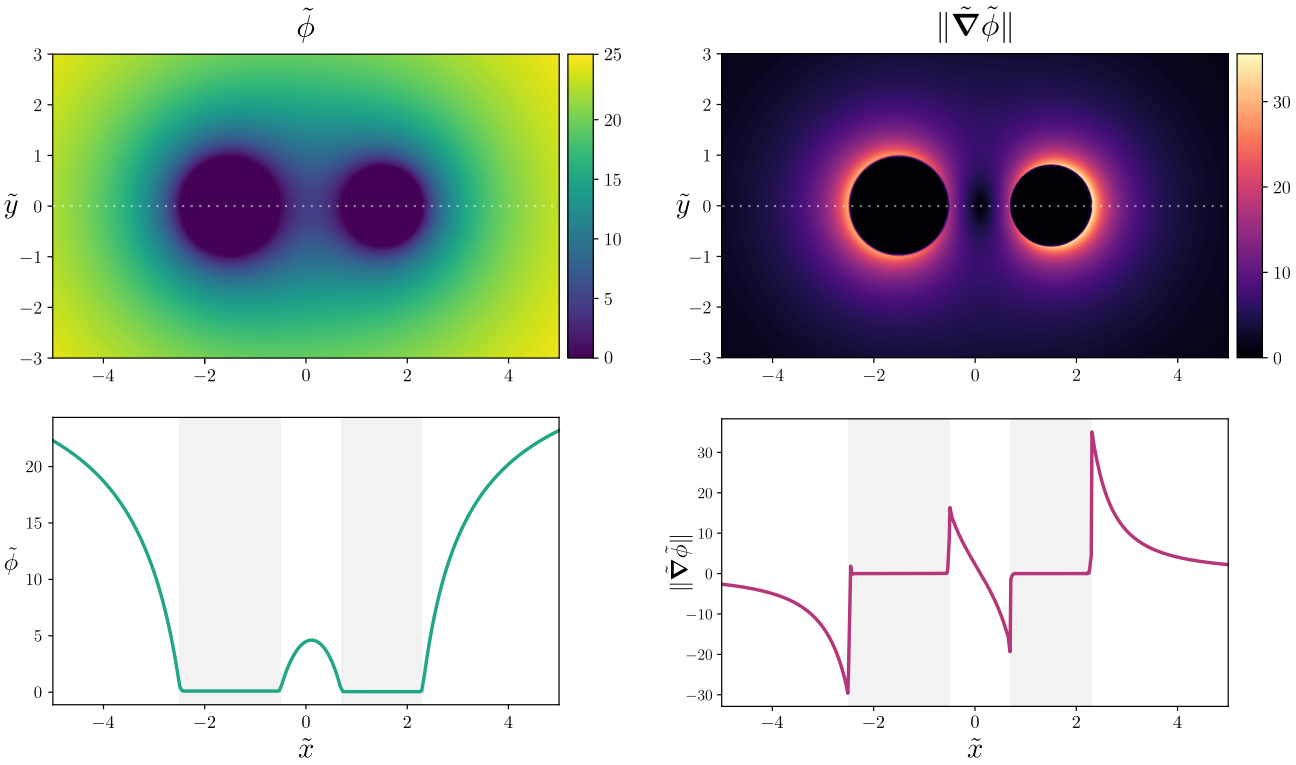


Figure 4.18: Example of numerical solution to the two-body problem for the chameleon field. The two panels on the left illustrate the value taken by the dimensionless scalar field $\tilde{\phi}$ in the $(\tilde{x}\tilde{y})$-plane (top panel) and along the $\tilde{x}$-axis (bottom panel). Likewise, the two panels on the right depict the magnitude of the gradient of the scalar field $\|\tilde{\boldsymbol{\nabla}}\tilde{\phi}\|$. Parameters: $\{\tilde{R}_1 = 1,\, \tilde{x}_1 = -3/2,\, \tilde{\rho}_1 = 10^2,\, \tilde{R}_2 = 0.8,\, \tilde{x}_2 = 3/2,\, \tilde{\rho}_2 = 5\times 10^2,\, \tilde{\rho}_{\rm vac} = 10^{-3},\, n = 1,\, \alpha = 0.1\}$.

the scalar field — let us recall that the fifth force amplitude is directly proportional to this quantity in the chameleon model. Again, the top panel is the full 2D map while the bottom panel specifically represents $\partial_{\tilde{x}}\tilde{\phi}$ along the $\tilde{x}$-axis (one has $\partial_{\tilde{y}}\tilde{\phi} \equiv 0$ along this particular line due to cylindrical symmetry).

Shifting of equilibrium points

The interactions mediated by the Newtonian potential on the one hand, and by scalar fields on the other hand, are both *attractive*. Because of this simple fact, it may be difficult to tell them apart in the context of e.g. laboratory experiments. It is therefore relevant to look for situations which somehow disentangle Newtonian gravity from fifth forces. This quest for experiments allowing to make fifth force effects stand out from other known physical effects is not new. A fairly innovative concept using a charged particle in an electromagnetic field was proposed in Ref. [292]. Likewise, Ref. [104] shows that space geodesy experiments could detect the signature of a Yukawa fifth force by measuring the Earth's J_2 coefficient at two different altitudes. Ref. [293] claims that non-spherical test bodies immersed in a background field will experience a net torque caused by the scalar field, an effect which has no counterpart in Newtonian gravity. Ref. [157] describes a torsion pendulum experiment for which the existence of a putative chameleon fifth force should create small torques while not being sensitive to the effects of massless fields. Here, we imagine an idea for an experiment using two spheres where a clear line can be drawn between a chameleonic fifth force and classical gravity.

Suppose that the two spheres are of equal mass $M_1 = M_2$, but with different radii $R_1 \neq R_2$. This condition can be expressed mathematically as

$$M_1 = M_2 \iff \left(\frac{R_1}{R_2}\right)^3 = \frac{\rho_2}{\rho_1}\,, \tag{4.15}$$

and we denote by M this common mass. From the perspective of an uncharged test particle, the two spheres look identical in the sense that the gravitational potential they individually source on the outside cannot be distinguished from that of a point mass. The Newtonian acceleration undergone by such a test particle is simply

$$\mathbf{a}_{N,\rm tot} = \mathbf{a}_{N,1} + \mathbf{a}_{N,2} = -GM\left(r_1^{-2}\mathbf{e}_1 + r_2^{-2}\mathbf{e}_2\right), \tag{4.16}$$

with $\mathbf{r}_i$ the vector joining the center of the sphere i to the test particle, $r_i = \|\mathbf{r}_i\|$ and $\mathbf{e}_i = \mathbf{r}_i/r_i$, $i = 1, 2$.

Based on the sketch provided in Fig. 4.17, the plane perpendicular to the x-axis at the origin is such that

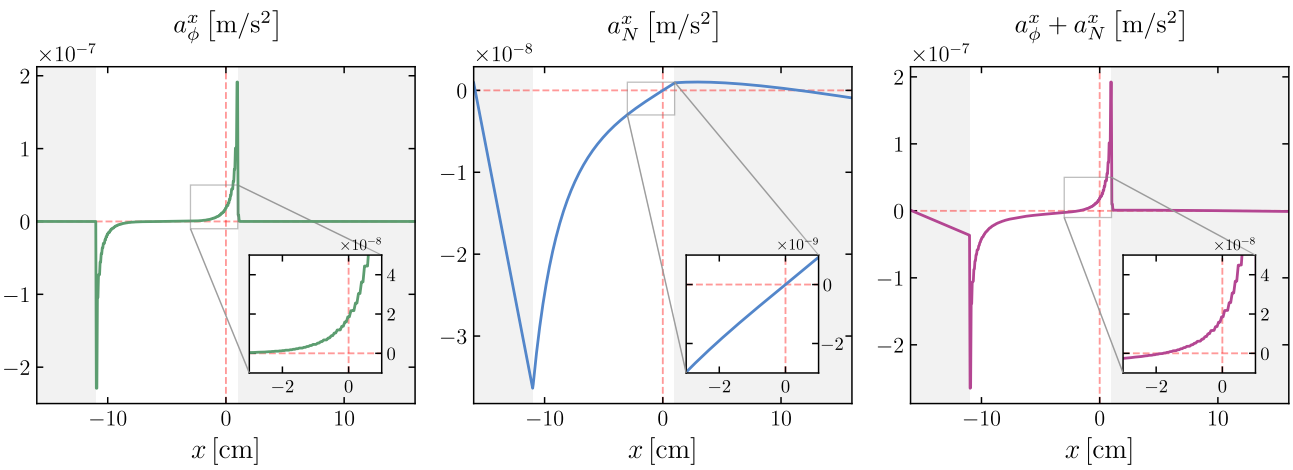


Figure 4.19: Gravitational acceleration along the x-axis (see Fig. 4.17) for the two-body problem with physical parameters given by Eq. (4.3.2). The pale shaded areas correspond to the space occupied by the spheres. Left panel: *chameleonic acceleration* a_ϕ^x. Center panel: *Newtonian acceleration* a_N^x. Right panel: *total acceleration* $a_\phi^x + a_N^x$. The small (unphysical) oscillations visible on the inset plot of the left and right panels is due to the finite element discretization and could be mitigated through the use of a more refined mesh and/or higher order Lagrange elements.

$\|\mathbf{a}_{N,1}\| = \|\mathbf{a}_{N,2}\|$. At the origin, we have $\mathbf{a}_{N,\rm tot} = \mathbf{0}$: the point where the Newtonian attraction is canceled coincides with the geometric center of the setup. This symmetry property does not generalize to the chameleon when at least one of the two bodies is partially screened. This mainly follows from the qualitative discussion we had in Sec. 1.2.2 when looking for possible violations of the WEP in chameleon gravity. Indeed, the two spheres are susceptible to develop different thin shell parameters ($\Delta R/R$) [Eq. (1.122)] and the center of chameleon forces might well be displaced from the origin point. Consequently, the point of equilibrium in the modified gravity setting must not coincide with $x = 0$. The question then boils down to getting an estimate of this displacement scale.

Unlike the example given in Fig. 4.18, this study requires the definition of physical quantities, with realistic numerical values. In this respect, we select the following set of parameters

$$\begin{gathered}\rho_1 = 2700\,\mathrm{kg\,m^{-3}}\,, \qquad \rho_2 = 100\,\mathrm{kg\,m^{-3}}\,, \qquad \rho_{\rm vac} = 0.1\,\mathrm{kg\,m^{-3}}\,,\\ x_1 = -16\,\mathrm{cm}\,, \qquad x_2 = +16\,\mathrm{cm}\,, \qquad R_1 = 5\,\mathrm{cm}\,, \qquad R_2 = 15\,\mathrm{cm}\,,\\ \beta = 10^2, \qquad \Lambda = \Lambda_{\rm DE} \simeq 2.4\,\mathrm{meV}\,, \qquad n = 1\,,\end{gathered} \tag{4.17}$$

which are representative of experimental systems that fit in the laboratory. The sphere 1 is roughly the density of aluminum while the sphere 2 could be made e.g. of polyurethane foam. The Newtonian part of gravity is derived analytically (taking advantage of the superposition principle) whereas the fifth force is accessed via *femtoscope*.

The results of this case study are reported in Fig. 4.19. The left panel represents the chameleon acceleration along the x-axis, $a_\phi^x \propto -\partial_x\phi$. The net zero acceleration in the spheres is characteristic of the screened regime. Most notably, the center of chameleon forces is *not* at the geometric center $x = 0$ but lies somewhere a few centimeters to the left of the origin. The center panel corresponds to the Newtonian acceleration a_N^x, which is derived analytically. The condition of mass equality [Eq. (4.3.2)] is consistent with the fact that the acceleration exactly cancels at $x = 0$. This means that test particles with $x < 0$ are attracted toward the left whereas those with $x > 0$ are attracted to the right. Note that the Newtonian acceleration is of the same order of magnitude as the chameleon acceleration, if not smaller (see Sec. 2 of Refs. [276] for additional insights into this observation; note that we are not talking about the gravitational attraction between the two spheres here). Finally, the right panel represents the total gravitational acceleration, which is merely the sum of the Newtonian and chameleon contributions. As can be seen on the corresponding inset, the pristine symmetry exhibited by the Newtonian acceleration is broken: the geometric center no longer coincides with the point where the gravitational pull of the spheres exactly balances. This shift of the $L1$ Lagrange point of the system is just under 2 cm, which is quite significant given the characteristic centimetric length scale of the setup. At the origin however, the total acceleration is no longer zero but rather $\sim 20\,\mathrm{nm/s^2}$.

Now that the principle has been established and that we have a conclusive order of magnitude, the difficult part is the actual design of an experiment that takes advantage of it. Measuring small effects (a few $\mathrm{nm/s^2}$ essentially) is made easier if one can benefit from long integration times in the measurement. Yet, long integration times may turn out to be out of reach, because the point of equilibrium between the various gravitational forces at stake is an *unstable* one. Moreover, the gravitational acceleration of the Earth has been completely ignored

in this picture. Further study of these issues is left for future work. Finally, it is to be noted that this idea of creating an asymmetry in the chameleon field profile is explored in Ref. [294] for parallel plates.

Chapter summary

In this chapter, we have introduced *femtoscope* as a novel numerical tool dedicated to the study of scalar-tensor models of gravity with screening mechanisms. It implements the various FEM-related techniques that were discussed at length in the previous two chapters, which makes it sufficiently versatile for our needs. Indeed, the program can handle virtually any semi-linear elliptic PDE problem posed on bounded or unbounded regions of space, provided that the corresponding Newton-linearized weak formulation is supplied by the user. Some specific cases of interest are already implemented — the Poisson, Yukawa, chameleon and symmetron problems — and are ready for use.

We then showcased *femtoscope* on the chameleon model specifically, whose nonlinear Klein–Gordon equation of motion restricts the use of analytical techniques to the simplest, most symmetrical cases. In that respect, the use of FEM with non-uniform meshes frees us from this limitation and opens the way to the realistic study of complex setups, hitherto inaccessible. In particular, *femtoscope* complements the recent SELCIE code by further being able to deal with asymptotic boundary conditions, making it the only publicly available code with such features.

Although the possibilities thereby offered by *femtoscope* sparked many ideas, we could not pursue them all in this thesis and had to make choices. The next chapter is a follow up to the preliminary study of chameleon gravity in the Earth environment. Specifically, we have underlined the fact that modeling the Earth as a sphere is no longer realistic in the screened regime where the chameleonic force is sourced by the outer layers. It is interesting to compute the imprint of the local relief on the chameleon field in Earth orbits and to investigate whether or not we could discriminate between the fifth-force signature and known effects with the current technology embedded on navigation and potential science satellites.

Chapter 5

Fifth force effects in Earth orbit

Outline of the current chapter

This chapter follows on from the study of the chameleon model in the Earth environment begun in Sec. 4.3.1. Its ultimate goal is to quantitatively assess the testability of fifth force effects in space. To this end, we use *femtoscope* in order to model this special case of modified gravity, solving for both the Newtonian potential and the chameleon field. In particular, numerical simulations allow us to go beyond the simplifying assumptions and modeling traditionally found in the literature. Building on these FEM computations, we study the dynamics of satellites in orbit around the Earth with and without the putative chameleonic force, which roughly amounts to comparing geodesics of the Einstein-frame metric *vs* those of the Jordan-frame metric, respectively. Given the level of precision achieved by recent space geodesy missions, we look whether it is possible to discriminate between the two in the presence of model uncertainties.

Unlike for other chapters, we reproduce here our work [141], published in *Physical Review D*.

5.1 Introduction and Summary

One of the goal of this PhD work is to determine whether space-based experiments are well-suited for testing gravity. This question being way too broad to be covered in its entirety here, we narrow it down to the specific case of scalar-tensor theories of gravity with screening mechanisms, focusing on the prototypical example of the chameleon model. In this perspective, Sec. 4.3.1 is a first step towards a realistic modeling of modified gravity in the Earth vicinity. There, we implement a radial density model based on PREM (Earth interior) and US76 (atmosphere) which, upon using the *femtoscope* code, yields solutions for the Newtonian potential and chameleon scalar field. Regarding the latter, this preliminary study makes clear the fact that viable regions of its parameter space (n, β, Λ) all map to a *screened* Earth. In such configurations, the fifth force is sourced only by the outermost layers of the Earth, which seriously calls into question the use of radial models: locally, the Earth's landform is very irregular and does not look much like the surface of a sphere. Furthermore, there is no such thing as the *thin shell effect* in Newtonian gravity and so the chameleon field should in principle leave a distinctive imprint on the total gravitational field, however small it is.[1] The main goal of the present article is to quantify this 'gravitational imprint' and assess whether the signature of a chameleon fifth force can be extracted from space-based gravitational measurements.

Studying the impact of the Earth's slight deviation from spherical symmetry on chameleon gravity is a very challenging task from a numerical perspective. On the one hand, the Earth's rotation and self-gravity are responsible for its equatorial bulge and flattening at the poles, giving it the overall shape of an oblate ellipsoid [Eq. (4.8)]. This J_2 effect is not taken into account in this article. On the other hand, going to smaller scales reveals a very complex landform, featuring mountains, ridges, craters, etc. Modeling this whole variety of topographies would result in a very complex 3D model, expensive to run FEM computations on, and yielding results difficult to interpret physically. In the light of this remark, we implement a highly simplified model where the Earth is represented by a sphere with a mountain on top of it. The overall shape created in this way exhibits

[1] By that, we mean that the total gravitational acceleration $\mathbf{a}_N + \mathbf{a}_\phi$ is not merely $(1+\varepsilon_\phi)\,\mathbf{a}_N$, for some $\epsilon_\phi > 0$.

azimuthal symmetry, which makes it possible to run relatively cheap 2D computations (see Sec. 2.3) and eases the interpretation of results.

Nonetheless, capturing the gravitational impact of this single source of asphericity with *femtoscope* is nontrivial because the mountain, given its size and mass, only represents a tiny deviation from the dominating Earth's monopole. Spherical coordinates are a natural choice for performing FEM computations for this specific problem. The handling of asymptotic boundary conditions is all the more crucial here as setting homogeneous Dirichlet boundary conditions at some finite distance from the Earth would impose spherical symmetry as we approach such an artificial boundary. This is particularly undesirable in this study where we look for small deviations from spherical symmetry. In this respect, we employ the *ifem* technique[2] to impose the correct asymptotic behavior on the unknowns (see Chapt. 3). A great deal of work is being done to ensure that *femtoscope*'s outputs are actually good enough approximations to the true solutions. This is a critical part of this article as all the subsequent physical discussions are based on these numerical solutions.

From there, we derive the multipole expansion of the chameleon field ϕ in the atmosphere-free case and compare it against that of the Newtonian potential Φ, at a fixed altitude. Upon normalization, this side-by-side comparison shows that the spherical harmonic coefficients of Φ and ϕ share the same distribution with respect to the degree ℓ when the Earth is unscreened. In the screened scenario however, we observe, as was anticipated, the emergence of a distinctive signature in the coefficients' distribution. In terms of acceleration, the chameleon acceleration $\mathbf{a}_\phi$ is a bit more orthoradially-directed than the Newtonian acceleration $\mathbf{a}_N$. Their norm ratio $\|\mathbf{a}_\phi\|/\|\mathbf{a}_N\|$ remains small though, bounded from above by $\sim 10^{-6}$ at the equivalent of LEO altitudes.

With the knowledge of the characteristics of this sought 'chameleonic signature' at hands, the next logical question is: *can we detect it ?* Again, this is a difficult question to answer quantitatively given all the actual physics involved in a realistic setup. Sticking to our simplified {sphere + mountain} model, we first tackle the issue related to the influence of the atmosphere on the fifth force. To that end, we implement three different atmospheric models — *tenuous*, *Earth-like*, *dense* — and reproduce the same analysis as above. Fixing (n, Λ) and gradually increasing the coupling parameter β underlines the existence of four regimes: (i) for low values of β, the atmosphere is *transparent* to the fifth force, (ii) above a certain threshold, it acts as an attenuator, effectively reducing the chameleon acceleration, (iii) for even stronger couplings, any non-radial dependence of the scalar field vanishes so that the mountain is plainly *invisible*, and (iv) the atmosphere itself eventually becomes screened.

On another note, in practice, the measurement of the Earth gravity from space involves satellites. The fifth force such extended objects undergo crucially depends on their thin shell parameter, and the point mass approximation is only valid as long as they are not screened. A whole part of this article is thus dedicated to the study of the backreaction of a spacecraft on the scalar field. For the first time, we go beyond the various qualitative screening criteria found in the literature by computing the full solution of the {Earth + satellite} system. We show that the transition from the unscreened to the screened regime occurs over a very narrow band in the chameleon parameter space. In the latter regime specifically, the resulting fifth force acting on the satellite is suppressed extremely efficiently.

We then consider a 'best-case scenario' with no atmosphere and follow the trajectory of a satellite — treated as a point mass — orbiting the {sphere + mountain} system, with and without the putative chameleonic force field. The orbit propagation code we use features a *projection technique* to numerically ensure the conservation of energy, see Appendix E. In particular, we compute the resulting *anomaly*[3] on several observables, most notably the distance variations between two satellites following each other as in the GRACE-FO setup. The anomaly levels we find for this idealized model are technically well within the detection range of current onboard and ground-based space technology, which may come as a surprise at first. Yet, taking into account model uncertainties allows for degeneracies which greatly lowers this hope. Specifically, it is possible to make Newtonian gravity mimic the fifth force — at a given altitude — by slightly tweaking the density model of the {sphere + mountain} system, while reasonably staying within the error bars.

Finally, we study the possibility of breaking this degeneracy by performing such a space geodesy experiment at two (or more) different altitudes. Suppose that the chameleon field actually exists. Then, the density model of the Earth inferred from two distinct altitudes, under the assumption of Newtonian gravity, would be inconsistent with each other. In the final part of the article, we endeavor to quantify such a *tension*. Given the orders of magnitude involved and the optimistic model underlying them, our take-home message is that space geodesy is not likely to result in competitive constraints on the chameleon model in the near future.

5.2 Article

[2] In the article, this technique is referred to as 'virtual connection of d.o.f.'.

[3] The term 'anomaly' is used to refer to the difference for a given observable between the {Newtonian gravity} case and the {Newtonian gravity + fifth force} case.

What to expect from scalar-tensor space geodesy

Hugo Lévy,[1,2,*] Joël Bergé,[1] and Jean-Philippe Uzan[2]

[1] *DPHY, ONERA, Université Paris Saclay F-92322 Châtillon - France*

[2] *Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris, 98 bis bd Arago, 75014 Paris, France*

Scalar-tensor theories with screening mechanisms come with non-linearities that make it difficult to study setups of complex geometry without resorting to numerical simulations. In this article, we use the *femtoscope* code that we introduced in a previous work in order to compute the fifth force arising in the chameleon model in the Earth orbit. We go beyond published works by introducing a departure from spherical symmetry — embodied by a mountain on an otherwise spherical Earth — as well as by implementing several atmospheric models, and quantify their combined effect on the chameleon field. Building on the numerical results thus obtained, we address the question of the detectability of a putative chameleon fifth force by means of space geodesy techniques and, for the first time, quantitatively assess the back-reaction created by the screening of a satellite itself. We find that although the fifth force has a supposedly measurable effect on the dynamics of an orbiting spacecraft, the imprecise knowledge of the mass distribution inside the Earth greatly curtails the constraining power of such space missions. Finally, we show how this degeneracy can be lifted when several measurements are performed at different altitudes.

I. Introduction

Scalar fields appear in most of the extensions beyond the standard models. Theories involving extra dimensions, from Kaluza-Klein theories up to string theories in the low energy limit, predict the existence of a light spin-0 particle. Scalar fields are also key ingredients in cosmology phenomenology, in particular for the dark sector and inflation. Coupling the scalar field to matter[1] automatically gives rise to a so-called *fifth force*, resulting in deviations from general relativity (GR) in gravitational phenomena. Evading the Solar system tests of GR and laboratory experiments [79] comes at the price of introducing non-linearities in the model which enable *screening mechanisms* (e.g. Damour-Polyakov [13], chameleon [40, 41], K-mouflage [5, 11], or Vainshtein [56, 73]).

Although screening mechanisms are precisely designed to recover GR — and thus in the weak field regime, Newtonian gravity — at Solar system scales, they leave nonetheless a small imprint which we can attempt to measure. Tests can be performed in a very wide range of length scales, from laboratory experiments [14, 36, 72], to spacecraft in orbit around the Earth [24, 25, 71] or traveling through the Solar system [8], planetary motion [26, 80, 81], and to astrophysical tests [37, 74, 78] (see Refs. [12, 18] and references therein for a more comprehensive review). Here, we are interested in the category of space-based experiments, which have long been expected to provide new constraints in the case of the chameleon model [41]. Several space missions were successfully launched in the past decades: MICROSCOPE [6, 71] for testing the weak equivalence principle (see Refs. [61, 62] for how constraints on the chameleon model could be derived from those data), Gravity Probe A and B [28], LAGEOS 1 and 2, LARES 1 and 2 [25].

Beside these space missions specifically tailored for fundamental physics, artificial satellites have also given rise to space geodesy. Initially, space geodesy primarily focused on measuring the Earth's shape and size, but technological advancements have propelled it into a realm of unprecedented accuracy and multifaceted applications. Cutting edge instruments onboard satellites allow for the implementation of complementary geodetic techniques such as laser and Doppler ranging, Global Navigation Satellite Systems, gravimetry (e.g. GOCE, CHAMP, GRACE-FO satellite missions), etc. The determination of the Earth's figure (mass distribution) constitutes an inverse problem: given the data $d_{\rm obs}$ collected by the various satellite missions and a model describing the laws of gravitation $\mathcal{M}$ with forward map $F_{\mathcal{M}}$, the goal is to determine the model parameters p such that the residual $d_{\rm obs} - F_{\mathcal{M}}(p)$ is minimized (in some specific sense, e.g. least-squares or probabilistic approaches). In space geodesy, this inverse problem is solved with the central assumption that the governing equation is Newton's law of gravity (and p would represent the distribution of mass) [52].

The goal of the present article is to assess the pertinence of orbitography techniques to test screened scalar-tensor theories, illustrated with the chameleon model, and to characterize the *best site* in the Solar system to perform such tests. This is a follow-up to our previous article Ref. [47] where we laid the foundations in terms of numerical simulations. There, we saw that the unconstrained region of the chameleon parameter space (see Fig. 3 of Ref. [82]) corresponds to a situation where the Earth is screened, i.e. where the chameleonic force is sourced only by its outer layers. This mere observation suggests that the local landform — specifically any local deviation from spherical symmetry — can leave a significant imprint on the chameleon profile. Consequently, if the chameleon's effects differ sufficiently from Newto-

* hugo.levy@onera.fr

[1] From a quantum mechanical perspective, the introduction of a scalar field in the gravity sector *always* generate interactions between this scalar and matter fields [12].

nian gravity, it should leave a distinctive signature on the Earth's gravity.

Mountains and craters are typical examples of asphericities that can be sensed through space geodesy. Relative to the size of a planet, a mountain represents a spiky feature. Several works bring to light the parallel between chameleon (and symmetron) gravity in the screened regime and electrostatics: the behavior of the scalar field is roughly the same as the behavior of the electrostatic potential for a perfect conductor[2] [38, 57, 64]. Taking the analogy a step further, Ref. [38] mentions the "lightning rod effect" in electromagnetism, exhibited by needle-like conductors around which the electric field ($\propto$ gradient of the potential) is enhanced. In the case of the chameleon, the counterpart of the electric field would be the fifth force ($\propto$ gradient of the scalar field) — making the mountain an interesting case study. Nevertheless as Ref. [57] underlines, while this analogy provides valuable qualitative insights, numerical computations remain essential to establish a quantitative connection with real-world observations and experimental data. In that respect, we aim to address the long-standing question of how much an atmosphere *smooths out* the mountain's contribution to the fifth force in space. More generally, existing work accounting for the atmosphere [34, 39–41, 53, 76] are, in our opinion, not extensive enough: the models are not realistic (one layer of constant density) and conclusions are drawn on qualitative arguments that can be misleading (see e.g. the introduction of Ref. [43]). We shall also pay attention to the influence of a spacecraft on the background field, and evaluate how this perturbation impacts the overall fifth force that it experiences.

The article is organized as follows. In Sec. II, we briefly recall the main equations describing both Newtonian and chameleon gravity, and give precise meaning to physical models outlined above, namely the modeling of the mountainous planet together with its atmosphere. In this setup, the total gravitational potential is computed numerically using *femtoscope*, a code that was specifically designed to solve these equations with asymptotic boundary conditions [47]. It allows for the computation of both the Newtonian potential and the chameleon field in space. The numerical results are presented and discussed in Sec. III. We explore a vast region of the chameleon parameter space and ascertain the influence of an atmosphere in several scenarios, making this a quite comprehensive study compared to what has been done in previous work. Finally, Sec. IV takes us back to space geodesy as we compare the dynamics of a spacecraft with and without a fifth force acting on it as it orbits the mountainous planet. We address the issue of being able to discriminate between the two in the presence of model uncertainties, and further suggest ways to break this source of degeneracy. These analyses pave the way to the design of orbitography experiments in the Solar system and their subtle interpretation. We conclude in Sec. V.

[2] Indeed, it can be shown that the equation of motion of the chameleon field in the quasi-static Newtonian limit with thin-shell can be well-approximated by the electrostatic potential equation. Then, same differential equations lead to same solutions.

II. Model & Numerical techniques

A. General equations

1. Newtonian gravity

It is well known that, in the weak-field regime and when the sources are moving very slowly compared to the speed of light, GR reduces to Newtonian gravity which is described by the *Newtonian potential* Φ with dimension $[L^2 \cdot T^{-2}]$. For a static configuration, we define it as

$$\Phi(\mathbf{x}) = -G \int_{\mathbb{R}^3} \frac{\rho(\mathbf{x}')}{\|\mathbf{x} - \mathbf{x}'\|} \, \mathrm{d}^3 x' \,, \tag{1}$$

where G is the Newtonian gravitational constant and $\rho = \rho(\mathbf{x})$ is the matter density function which depends on position $\mathbf{x}$. Assuming that the *weak equivalence principle* holds perfectly (Ref. [71] shows that it holds at less than 10^{-15}) and from a classical mechanics perspective, the gravitational acceleration undergone by a point-like particle is simply $\mathbf{a}_\Phi = -\boldsymbol{\nabla}\Phi$. Eq. (1) provides a straightforward way of computing the Newtonian potential by evaluating some three-dimensional integral (see e.g. Ref. [31]). However, it may be more convenient from a numerical standpoint to solve the following Poisson's equation

$$\Delta\Phi = 4\pi G\rho \,, \tag{2}$$

implied by the definition of Φ. Indeed, on the one hand one has to evaluate the integral appearing in Eq. (1) for each point $\mathbf{x}$ where the Newtonian potential is sought, whereas on the other hand solving the partial differential equation (PDE) (2) provides an approximation of Φ over the whole numerical domain.

Assuming that the mass density vanishes as one moves away from the source of gravity, the gravitational acceleration $\mathbf{a}_\Phi = -\boldsymbol{\nabla}\Phi$ is expected to decay to zero at infinity. The essential boundary condition is therefore defined at infinity and a very common choice for the constant of integration is

$$\Phi(\mathbf{x}) \underset{\|\mathbf{x}\| \to +\infty}{\longrightarrow} 0 \,. \tag{3}$$

2. Chameleon gravity

In the Newtonian limit, the chameleon field ϕ is governed by a nonlinear Klein-Gordon equation which takes

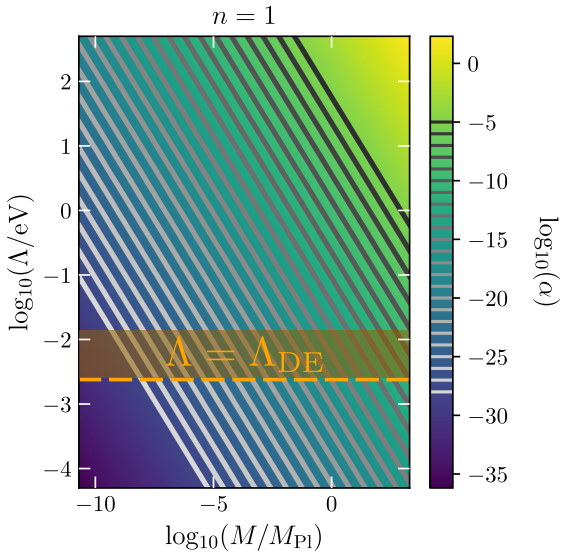


FIG. 1. Mapping from the chameleon parameter space in the plane $n = 1$ to the dimensionless parameter α appearing in Eq. (7), where $M = M_{\rm Pl}/\beta$. The gray lines represent the isovalues of the α parameter covered in this study, ranging from 10^{-5} to 10^{-28}. The orange horizontal dashed line corresponds to $\Lambda = \Lambda_{\rm DE} = 2.4 \times 10^{-3}\,\mathrm{eV}$, the dark energy scale.

the form

$$\Delta\phi = \frac{\mathrm{d}V_{\rm eff}}{\mathrm{d}\phi} = \frac{\beta}{\mathrm{M_{Pl}}}\rho - \frac{n\Lambda^{n+4}}{\phi^{n+1}}\,, \tag{4}$$

where $\mathrm{M_{Pl}} \equiv 1/\sqrt{8\pi G}$ is the reduced Planck mass and $V_{\rm eff}$ is the so-called *effective potential* of the scalar field. The model further has three parameters — β a positive dimensionless constant which encodes the coupling of the scalar field to matter, Λ a mass scale and n a natural number. The 3-acceleration experienced by a point-like particle induced by its coupling to the chameleon field is proportional to the gradient of the scalar field and takes the form

$$\mathbf{a}_\phi = -\frac{\beta}{\mathrm{M_{Pl}}}\boldsymbol{\nabla}\phi\,. \tag{5}$$

If we assume that the density uniformly decays to some vacuum density $\rho_{\rm vac}$ far away from the source, then the chameleonic acceleration is expected to decay to zero at infinity, just as in the Newtonian gravity case discussed above. Equating the r.h.s. of Eq. (4) to zero and solving for ϕ yields the following asymptotic boundary condition:

$$\phi(\mathbf{x}) \underset{\|\mathbf{x}\|\to+\infty}{\longrightarrow} \left(\mathrm{M_{Pl}}\frac{n\Lambda^{n+4}}{\beta\rho_{\rm vac}}\right)^{\frac{1}{n+1}} \equiv \phi_{\rm vac}\,. \tag{6}$$

In Ref. [47], we introduced *femtoscope* — a PYTHON numerical tool based on the finite element method which enables us to solve Eq. (4) on spatially unbounded domains. We perform the same nondimensionalization as in Refs. [16, 47] by introducing (i) ρ_0 a characteristic density of the problem, (ii) $\phi_0 \equiv (n\mathrm{M_{Pl}}\Lambda^{n+4}/\beta\rho_0)^{1/(n+1)}$ the expectation value of the chameleon field in an ambient medium of density ρ_0 and (iii) L_0 a characteristic length scale of the system under study. Denoting the new dimensionless quantities with a tilde, trivial algebra leads to

$$\begin{aligned} &\alpha\tilde{\Delta}\tilde{\phi} = \tilde{\rho} - \tilde{\phi}^{-(n+1)}\,, \\ &\text{with } \alpha \equiv \left(\frac{\mathrm{M_{Pl}}\Lambda}{\beta L_0^2\rho_0}\right)\left(\frac{n\mathrm{M_{Pl}}\Lambda^3}{\beta\rho_0}\right)^{1/(n+1)}\,. \end{aligned} \tag{7}$$

The mapping $(\beta, \Lambda) \mapsto \alpha$ for $n = 1$ is illustrated in Fig. 1. Note that Eq. (7) now only depends on two parameters, α and n, instead of the three initial ones, which allows for a more efficient numerical exploration of the chameleon parameter space[3]. The chameleonic acceleration (5) then scales as

$$\mathbf{a}_\phi \propto \Lambda^{\frac{n+4}{n+1}}\beta^{\frac{n}{n+1}}\tilde{\boldsymbol{\nabla}}\tilde{\phi}\,. \tag{8}$$

We denote a_0 the multiplicative constant appearing in front of the dimensionless gradient, which reads

$$\begin{aligned} a_0\,[\mathrm{m/s^2}] = (\Lambda\,[\mathrm{eV}] \times \mathfrak{e}\,[\mathrm{J/eV}])^{\frac{n+4}{n+1}} \\ \frac{\beta^{\frac{n}{n+1}}}{\mathrm{M_{Pl}}L_0}\left[\frac{n\mathrm{M_{Pl}}}{\rho_0(\hbar c)^3}\right]^{\frac{1}{n+1}}\,. \end{aligned} \tag{9}$$

In Eq. (9), physical quantities are expressed in SI units unless specified using square brackets and $\mathfrak{e} \sim 1.6022 \times 10^{-19}\,\mathrm{J/eV}$ is the conversion factor from electron-volts to joules. As a rule of thumb, the smaller α, the more screened the setup. All physical results issued in this article are evaluated with $L_0 = R_\oplus = 6371\,\mathrm{km}$ (the Earth radius) and $\rho_0 = 1\,\mathrm{kg/m^3}$.

The Newtonian potential and the chameleon field do not have the same physical dimension. In order to be able to compare these two quantities, we define a new field

$$\Psi = \frac{\beta}{M_{\rm Pl}}\phi \tag{10}$$

which can be expressed in $\mathrm{m^2/s^2}$. We refer to Ψ as the *chameleon potential* since it plays the same role as Φ. The total gravitational acceleration undergone by a point-like particle will simply be $-\boldsymbol{\nabla}(\Phi + \Psi)$. Furthermore, the term 'fifth-force' will be used loosely throughout this article. Most occurrences of it should be taken as a synonym for 'chameleon acceleration', i.e. a quantity homogeneous to an acceleration and not a force per say. Finally, we will often refer to the 'screened regime' or to the 'thin-shell of a body' in this article. These notions

[3] Naturally, the mapping $(\beta, \Lambda, n) \mapsto (\alpha, n)$ described above is not bijective.

can be given precise meanings now that we have introduced the main notations. A macroscopic body is said to be *screened* when the chameleon field reaches the value that minimizes its effective potential $V_{\rm eff}$ deep inside the body. In that case, the field remains essentially frozen in that body except in a (usually) thin surface layer, which is referred to as the thin-shell.

B. Physical models

1. Mountains

At first order and seen from afar, planetary-mass objects have a rounded, ellipsoidal shape due to their self-gravity and rotation. It is only when we take a closer look at such bodies in the Solar System that smaller, more complex features become visible: mountains, ridges, craters, volcanoes, etc. This rich variety of topographies results in perturbations (with respect to the spherically symmetric case) in the gravitational field which, in the case of the Earth, can be measured by geodetic satellites. With a view to understand how $5^{\rm th}$-forces affect Newtonian gravity in the vicinity of these topographical features, it is desirable to first work with a simple toy-model. We thus consider a spherical body together with a single, axisymmetric mountain on top of it as depicted in Fig. 2. It is mainly described by two dimensionless parameters:

- h_m, the height of the mountain divided by the radius $R_{\rm body}$ of the spherical body (which is unitary on Fig. 2);
- θ_m, the mountain's half-angle.

Note that these two parameters are deliberately exaggerated on Fig. 2 for better visualization, and are clearly not representative of any *realistic* mountain in the Solar system — see Table I. All numerical computations presented in this article were performed with $h_m = 10^{-2}$ and $\theta_m = 10^{-2}$ rad comparatively. The resulting setup is itself axisymmetric which means FEM computations can be performed in two dimensions rather than three, greatly reducing computational complexity.

For the model to be complete, we further need to specify the density function $\rho(\mathbf{x})$ inside and outside the body. For the sake of simplicity, we assign a constant density to the body $\rho_{\rm body}$. The body may or may not be surrounded by an atmosphere. In either case, the density outside the body depends solely on the radial distance from the center r and always goes down to a constant vacuum value $\rho_{\rm vac}$. For all FEM computations, we set

$$\tilde{\rho}_{\rm body} = \frac{\rho_{\rm body}}{\rho_0} = 10^3 \text{ and } \tilde{\rho}_{\rm vac} = \frac{\rho_{\rm vac}}{\rho_0} = 10^{-15}\,.$$

Additionally, we will work most of the time with the dimensionless variable $\tilde{r} = r/L_0$, and set $L_0 = R_{\rm body}$. The

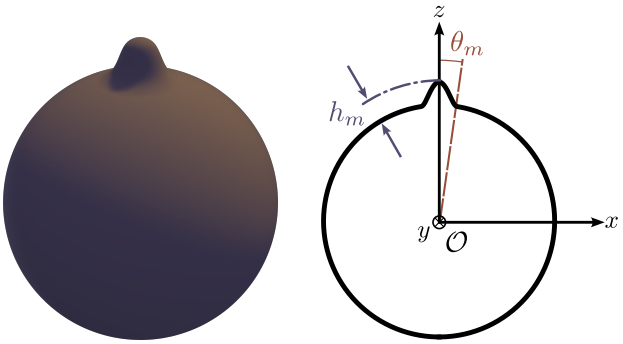


FIG. 2. Mountain visualization and notations. The Cartesian frame $(\mathcal{O}, x, y, z)$ is centered at the geometric center of the sphere devoid of mountain. The actual mountain profile used in numerical computations is drawn using B-splines in polar coordinates so as to form a smooth manifold.

various fields involved in this study (chameleon potential, Newtonian potential, together with their gradient) will be probed at fixed discrete values of $\tilde{r}$ for the sake of consistency. We made the choice to show results for $\tilde{r} \in \{1.059, 1.111, 1.314, 4.645, 6.617\}$, which for the case of the Earth corresponds roughly to peculiar orbits: the International Space Station, MICROSCOPE, a Medium Earth orbit, Galileo and geostationary satellites, respectively.

2. Atmospheres

Some Solar system bodies are surrounded by an atmospheric layer — a gas envelop held in place by the gravity of the body. This slight over-density with respect to the case with no-atmosphere is expected to have an influence on the chameleon field profile and, therefore, on the $5^{\rm th}$-force in space [40, 41, 53]. However, works that take account of the atmosphere often model it as an additional shell of matter with constant density satisfying $\rho_{\rm body} > \rho_{\rm atm} > \rho_{\rm vac}$, or at best as a constant piecewise function [34, 39, 40, 53, 76]. It is actually difficult to be more precise than this using analytical techniques only. Here, we go a step further by taking advantage of *femtoscope* to analyze the chameleon field profile in more realistic atmospheric setups. To avoid confusion, the requirement that the atmosphere must have a thin-shell stipulated in Ref. [40] only holds in the case of non-universal coupling, wherein unacceptably large violations of the weak equivalence principle would be observed in ground based experiments. Here, we work on the assumption of a universal coupling (characterized by a single dimensionless constant β) and so there is no particular reason for imposing this condition[4].

[4] The fact remains that, even in the case of a universal coupling, deviations from the inverse square law can be suppressed by the

5

TABLE I. List of some peculiar mountains in the Solar system[a].

Site	Body density [kg/m^3]	Atmosphere density [kg/m^3]	Atmosphere thickness [km]	height (base to peak) [km]	height (base to peak) h_m	θ_m [rad]
Earth Mount Everest	2.6×10^3 (Earth crust)	1.2 (sea level)	~ 100	4.6	7.2×10^{-4}	$\sim 10^{-3}$
Earth Mauna Kea	2.6×10^3 (Earth crust)	1.2 (sea level)	~ 100	10.2	1.6×10^{-3}	$\sim 10^{-2}$
Mars Mons Olympus	2582 (Mars crust)	2×10^{-2} (max.)	~ 10	21.9	6.5×10^{-3}	$\sim 9 \times 10^{-2}$
Moon Mons Huygens	2550 (Moon crust)	no atmosphere		5.5	3.2×10^{-3}	$\sim 6 \times 10^{-2}$
Io Boösaule Montes	3500 (mean density)	$< 10^{-6}$	—	18.2	10^{-2}	$\sim 1.5 \times 10^{-2}$
Vesta Rheasilvia central peak	2800 (crust estimate)	no atmosphere		25	10^{-2}	~ 0.4

[a] Mainly based on https://en.wikipedia.org/wiki/List_of_tallest_mountains_in_the_Solar_System, last visited: August 22$^{\text{th}}$, 2023

Three atmospheric density profiles are considered in this study: *Earth-like*, *Tenuous* and *Dense*. The Earth-like model is built from the 1976 version of the U.S. Standard Atmosphere model [1], commonly known as the US76 model[5]. It provides an estimate of the Earth atmospheric density ρ_{US} as a continuous function of the altitude, up to $R_{\text{atm}} \sim 36 \times 10^3$ km. Because we want the minimum dimensionless density in the numerical domain to be exactly $\tilde{\rho}_{\text{vac}} = 10^{-15}$, we apply the following transformation on the original data:

$$\log \tilde{\rho}_{\text{Earth-like}} = \log \tilde{\rho}_{\text{US}} + k\Big[\log \tilde{\rho}_{\text{US}} - \log(\min \tilde{\rho}_{\text{US}})\Big]$$

$$\text{with } k = \frac{\log(\tilde{\rho}_{\text{US}}/\tilde{\rho}_{\text{vac}})}{\log(\max \tilde{\rho}_{\text{US}}/\min \tilde{\rho}_{\text{US}})}$$

for $r < R_{\text{atm}}$, which is nothing but an affine transformation on the logarithmic densities. Beyond R_{atm}, we set $\tilde{\rho}_{\text{Earth-like}} = \tilde{\rho}_{\text{vac}}$. The other two models — Tenuous and Dense — are purely empirical in the sense that they are not based on actual atmospheric data. Both are constructed via the expression

$$\log \tilde{\rho}(r) = \begin{cases} A \exp\left[\dfrac{(r - R_{\text{atm}})^2}{\sigma^2}\right] + B & \text{if } r < R_{\text{atm}} \\ \log \tilde{\rho}_{\text{vac}} & \text{otherwise} \end{cases},$$

where the parameters (A, B, σ) are adjusted by hand to obtain either a very tenuous, thin atmosphere or a very dense, thick one instead. The resulting density profiles are depicted in Fig. 3.

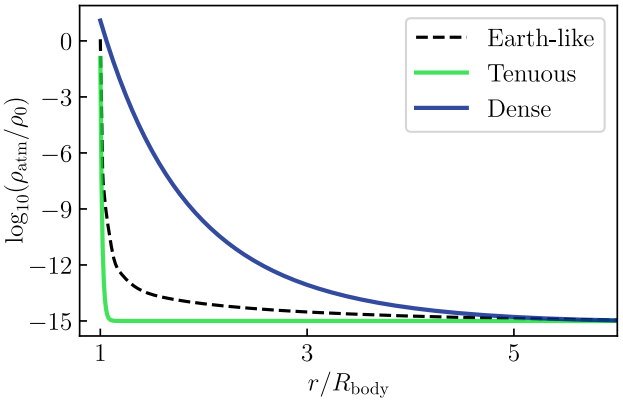


FIG. 3. Atmospheric profiles investigated in this study.

C. Decomposition of scalar fields into spherical harmonics

In geophysics and physical geodesy, the Earth gravitational potential is conveniently modeled as a spherical harmonics expansion [59]. Any well-behaved function $f : \mathbb{R}^3 \to \mathbb{R}$ may be decomposed as

$$f(r, \mathbf{n}) = \sum_{l=0}^{+\infty} \sum_{m=-l}^{+l} f_{lm}(r) Y_{lm}(\mathbf{n}) \,, \tag{11}$$

where r, $\mathbf{n} = (\theta, \varphi)$ refer to spherical coordinates, Y_{lm} is the *real* spherical harmonic function of *degree* l and *order* m (see Ref. [77] for its definition), and f_{lm} are the spherical harmonic coefficients that only depend on the radial coordinate — they are referred to as the *bare coefficients* in this article. There are several normalization conventions for an unequivocal definition of spherical harmonic functions. In this study, we stick to the *orthonormalized*

atmosphere.

[5] Data downloaded from http://www.braeunig.us/space/atmos.htm, (especially for the density between 1000 km - 36000 km altitude). Last visited: June 1$^{\text{st}}$, 2022.

TABLE II. Notations for the spherical harmonic coefficients.

	Bare coefficients	Rescaled coefficients
Newtonian potential	$\Phi_{lm}(r)$	y^N_{lm}
Chameleon potential	$\Psi_{lm}(r)$	$y^C_{lm}(r)$

convention for which

$$\int_{\mathcal{S}^2} Y_{lm}(\mathbf{n}) Y_{l'm'}(\mathbf{n})\, \mathrm{d}^2\Omega = \delta_{ll'}\, \delta_{mm'}\,, \tag{12}$$

where $\mathcal{S}^2$ is the unit 2-sphere, $\mathrm{d}\Omega$ is the differential surface $\sin(\theta)\mathrm{d}\theta\mathrm{d}\varphi$ and δ_{ij} is the Kronecker delta function. The notations used to refer to the spherical harmonic coefficients of the Newtonian potential Φ and the chameleon potential Ψ are gathered in Table II.

1. *Rescaled coefficients*

The bare spherical harmonic coefficients of the Newtonian potential Φ_{lm} further exhibit a scaling property. Let us denote by $\mu_{\rm body} \equiv GM_{\rm body}$ the standard gravitational parameter of the central body of mean radius $R_{\rm body}$ and mass $M_{\rm body}$. Then, the rescaled coefficients

$$y^N_{lm} = \frac{r}{\mu_{\rm body}} \left(\frac{r}{R_{\rm body}}\right)^l \Phi_{lm}(r) \tag{13}$$

can be shown to be independent of r [7],[6] owing to the specific form of the Newtonian potential (1). Such rescaled coefficients are thus *universal* to the body under consideration. Similarly to Eq. (13), we denote by $y^C_{lm}(r)$ the rescaled coefficients of the chameleon potential which, for their part, have no particular reason to be independent of the radial distance. In that sense, Ref. [7] shows the explicit dependence of such coefficients with respect to r in the case of a Yukawa interaction.

This relation can also serve as a means of checking the numerical results obtained for the Newtonian potential. This test is performed in Appendix B.

2. *Recovery of the coefficients*

We use the software `SHTools` [77] to compute the spherical harmonic coefficients of the scalar fields of interest. The Python package `pyshtools` comes with the routine `SHGrid.expand` which calculates the coefficients by means of some numerical quadrature[7]. The only detail worth mentioning is the fact that this routine outputs separate variables for the *cosine* C_{lm} and *sine* S_{lm} coefficients (sometimes referred to as the *Stokes coefficients*). The conversion from (C_{lm}, S_{lm}) to bare coefficients is outlined in Appendix A — Eq. (A5).

D. Numerical techniques

1. *Using femtoscope to solve linear and nonlinear PDEs with asymptotic boundary conditions*

As mentioned earlier, *femtoscope* is a ready-to-use Python program which plays a central role in this study as it enables us to compute both the Newtonian potential and the chameleon field by solving Eqs. (2) and (4) respectively. It is based on the finite element method — building on top of the open-source package *Sfepy* [23] — and further implements techniques to deal with nonlinearities and asymptotic boundary conditions (3, 6).

The proper treatment of these asymptotic boundary conditions is of noticeable importance in this study. Indeed, it is tempting to simply truncate the numerical domain at a fixed radius and apply a homogeneous Dirichlet boundary condition on the artificial border resulting from that process. This procedure has several flaws:

1. For the error that arise therefrom to be small, the domain must be sufficiently large, which translates to higher computational cost.

2. Selecting the size of that domain is a *blind experiment* in the sense that the dependence of the error on the truncation radius is not easily accessible without additional tricks.

3. It wantonly imposes spherical symmetry on the solution as we approach the artificial boundary. This is particularly undesirable in this study where we are interested in the small deviations from spherical symmetry introduced by the presence of the mountain.

This latter point is illustrated on Fig. 4 where it can be seen that, as we approach the artificial boundary, the truncation method (labeled 'FEM bounded', dash-dotted pink line) exhibits a poor approximation.

Instead, we employ a technique based on the splitting of the numerical domain Ω into two subdomains $\Omega_{\rm int}$ and $\Omega_{\rm ext}$ such that $\bar{\Omega} = \bar{\Omega}_{\rm int} \cup \bar{\Omega}_{\rm ext}$. $\Omega_{\rm int}$ is the bounded, interior domain, while $\Omega_{\rm ext}$ is the unbounded, exterior domain. An *inversion* transform is then applied to $\Omega_{\rm ext}$, resulting in a bounded domain $\tilde{\Omega}_{\rm ext}$ (called the *inversed exterior domain*) which can be meshed on a

[6] The numerical values of $\mu_{\rm body}$ and $R_{\rm body}$ could in theory be chosen arbitrarily. However the numerical values of the rescaled coefficients are tied to this choice.

[7] In this study, we use a $\mathsf{N} \times 2\mathsf{N}$ *Driscoll and Healy* sampled grid.

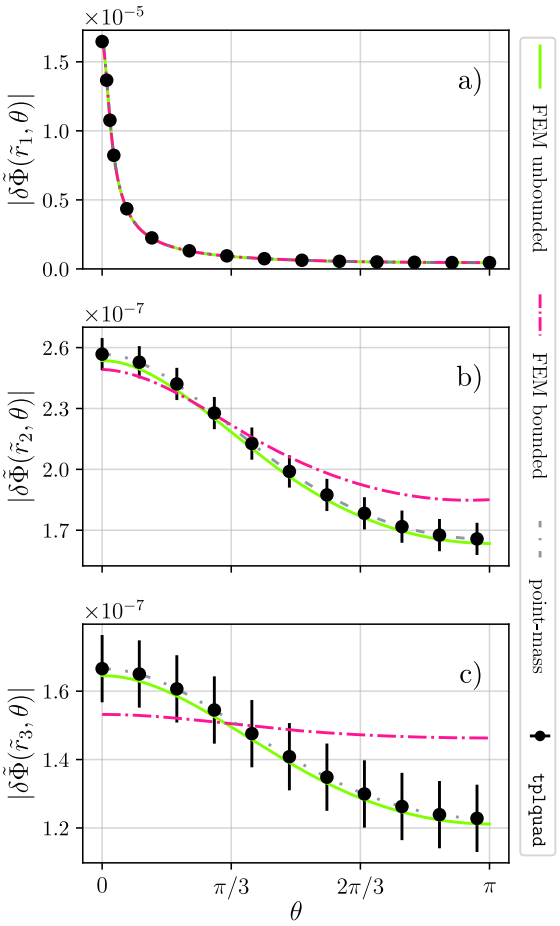


FIG. 4. Orthoradial profiles of the dimensionless Newtonian potential $\delta\tilde{\Phi}$ sourced by the mountain at three different altitudes, corresponding to $\tilde{r}_1 = 1.059$, $\tilde{r}_2 = 4.645$, and $\tilde{r}_3 = 6.617$ (top, middle and bottom panels respectively). The black dots together with their error bar represent the benchmark solution, obtained through the computation of the integral Eq. (1) with `scipy`'s `tplquad` routine. The pink dash-dotted line is obtained by solving Poisson's equation (2) with an homogeneous Dirichlet boundary condition applied at $\tilde{r} = \tilde{R}_c = 7$ while the green solid line is the solution provided by *femtoscope* with asymptotic boundary condition. Finally, the gray dash-dotted line is an analytical approximation where the mountain is replaced by a point-mass, whose location and mass were fitted to provide a good match with respect to the benchmark: $m_{\rm mountain}/M_{\rm body} = 2.23 \times 10^{-7}$ and $z/R_{\rm body} = 2.22 \times 10^{-3}$.

computer. There are many possible numerical implementations based on this method, see e.g. Refs. [10, 55, 58]. In this study, we make use of the so-called *virtual connection of DOFs* described in our previous work [47].

2. *Numerical challenges and verification*

There are several inconspicuous challenges associated with the numerical computation of the field profiles in the setup described in Sec. II B. To start with, let us stress the fact that we are looking for small deviations from spherical symmetry, owing to the presence of a very localized over-density at the pole that we here call a mountain. Quantitatively speaking, a back-of-the-envelope calculation shows that — at a fixed altitude h — the relative variation of the Newtonian potential $\Phi(R_{\rm body}+h,\theta)$ along the latitudes with respect to its mean value at this altitude is no larger than a few 10^{-6}. The higher we go, the smaller this ratio, which means our numerical approximations have to be correct up to at least seven significant digits to be deemed *good*. This mere order-of-magnitude calculation raises an additional concern: how do we actually check that the numerical approximations we obtain are compliant with the required levels of precision?

The Poisson's equation (2) governing the Newtonian potential being linear, it is possible to apply the superposition principle, where the total field is simply the mountain's contribution on top of a spherically symmetric background: $\Phi_{\rm tot}(r,\theta) = \delta\Phi(r,\theta) + \Phi_0(r)$. Turning to the chameleon field, the nonlinearity in the r.h.s. of the Klein-Gordon equation (4) prevents us from following the same path. Even if one were to decompose the chameleon field as $\phi_{\rm tot}(r,\theta) = \delta\phi(r,\theta) + \phi_0(r)$, the term $(\phi_0+\delta\phi)^{-(n+1)}$ becomes linearizable only under the assumption that $\delta\phi \ll \phi_0$ *everywhere*. Unfortunately, this assumption has no reason to hold in all scenarios, owing to the very nature of the screening mechanism. Indeed, it is far from being valid in the case where the mountain itself becomes screened, which turns out to be the most interesting case given the current constraints on the chameleon field [82]. For lack of a better workaround, we abandoned perturbation-based techniques and put our efforts into solving for the full field. It is therefore necessary to compare the FEM approximation obtained with *femtoscope* against some benchmark. Failing to have an analytical solution for the Newtonian potential of a mountain, we can still resort to the numerical integration of Eq. (1). In this respect, we use `scipy`'s `tplquad` routine [75] to evaluate the integral with an estimated relative error of a few 10^{-9}. This semi-analytical approach constitutes our benchmark and is depicted by the black dots together with their error bar in Fig. 4. Note that while it takes only a few seconds to evaluate the potential at a single point with this method, it is not conceivable to construct a full map of the field in this way. Rather, this semi-analytical computation should be employed sparingly to assess the error of the FEM computations.

In contrast, the chameleon field does not enjoy a similar integral representation which in turns means that we cannot easily define a benchmark profile. Nonetheless, we came up with the following strategies:

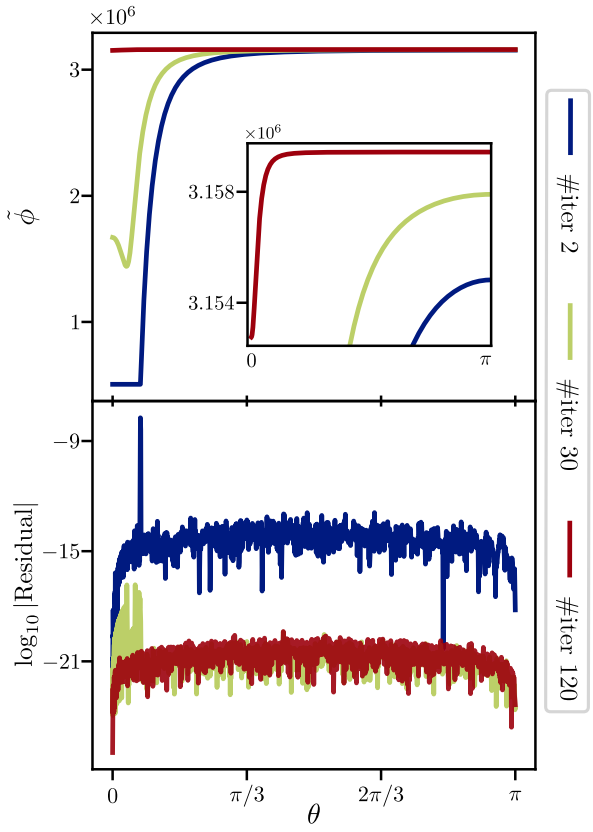


FIG. 5. Evolution of the dimensionless chameleon profile $\tilde{\phi}$ (top) and the associated residual (bottom) after different numbers of iterations of Newton's method (2, 30 and 120). These two quantities are displayed as a function of θ at fixed altitude $\tilde{r} = 1.059$. The residual becomes stationary and thus no longer decreases after a sufficient number of iterations has been reached.

– Select the set of FEM-related parameters (number and distribution of DOFs, order of the base functions, etc.) so that the FEM-approximation of the Newtonian potential matches the benchmark and use those parameters for the FEM computation of the chameleon field. The light green curves on Fig. 4 correspond to such FEM-approximations (using the aforementioned 'virtual connection of DOFs' method) and show that it is indeed possible to reach a high level of accuracy as they stay within the error bars of the benchmarks.

– It is also good practice to refer to pre-established *FEM convergence curves*, which are simple charts relating the error to the number of DOFs — see e.g. Fig. 1 of Ref. [46]. We can then construct our meshes in an enlightened way, ensuring they are fine enough to meet the stated accuracy.

– Evaluate the *strong residual*, which can be done by inputting the FEM approximation obtained for the chameleon field into its equation of motion (7) — schematically: Residual $= \alpha\tilde{\Delta}\tilde{\phi} - \tilde{\rho} + \tilde{\phi}^{-(n+1)}$. The closer the quantity is to zero, the better the numerical approximation. In order to make this criterion more quantitative, we can monitor (i) the strong residual's decrease throughout the Newton's iterations (see Fig. 5) and especially how small the final residual is compared to the initial one, and (ii) its size relative to the size of each term in it: the final residual should be at least a few orders of magnitude smaller than the dominant terms. This criterion is assessed on all 2D numerical computations of the chameleon field discussed in this article. As an example, Fig. 17 in Appendix C demonstrates that this criterion is indeed met on three distinct numerical solutions, at three altitudes.

Yet, formulating criteria based on the strong residual alone is not entirely satisfactory as it is an absolute quantity. Consequently, there is a priori no simple connection between the relative error committed on the approximation and the strong residual, since the latter quantity is dependent on the PDE's parameters (value of α[8], density model, etc.). Computing a reduction factor, that is by how much the strong residual has decreased over the Newton's iterations, is not sufficient either as it depends on how well the initial guess has been chosen (see discussion in the next paragraph). One idea to break this deadlock is to compare our numerical approximations with the chameleon radial profile around a ball. Indeed, the spherically symmetric case is much more under control as we have analytical approximations at our disposal (see e.g. Refs. [40, 63]) and the Klein-Gordon equation boils down to a one-dimensional ordinary differential equation (ODE) which can be solved numerically with a much higher density of DOFs and higher-order finite elements. In terms of residual, the numerical solutions are actually better than their analytical counterparts (see e.g. Tab. II from Ref. [47]), which is why we propose to use 1D numerical solutions as a benchmark for the spherically symmetric case. Because the addition of a mountain on top of the spherical planet is not expected to have a huge impact on the field's strength outside it, we can check that the evolution of the field along the outgoing radial direction follows that of the benchmark. We provide a quantitative way of assessing that statement in Appendix C, which is applied for all the numerical solutions discussed in this article. Finally, the orthoradial variations of the field at fixed altitudes seems more difficult to verify. As a rough check, we can set $h_m = 0$ and verify that this leads to $\partial_\theta\phi \equiv 0$. In practice, we do not expect this equality to hold exactly so we rather make sure that the amplitude $\max_\theta \phi(r,\theta) - \min_\theta \phi(r,\theta)$

[8] In particular, we observed in Ref. [47] that the 2-norm of the strong residual was increasing with α, all other things being equal.

is much smaller in the case $h_m = 0$ compared to the case $h_m = 0.01$. Doing this sanity check on a handful of cases (doing it on all cases would have been too costly) consistently shows that the two quantities differ by at least two orders of magnitude, so that we can confidently state that the orthoradial profiles showed later on do not originate from numerical noise.

Ultimately, the most critical point in this FEM computation is the convergence of the Newton's iterations. Whether or not the method converges depends on a lot of factors. Unfortunately, there are no miracle techniques to address convergence issues but rather *recipes* and good practices which we concisely report here. Perhaps the most important one is to start from a *good* initial guess, i.e. an initial approximation that is as close as possible to the true solution. In most cases, we use a pre-computed 1D radial profile of the field to this end. Other common practices are to refine the meshes where the field is expected to vary quickly (large gradient) — that is near the transition between the inside and the outside of the body, and near the area representing spatial infinity in the inversed exterior domain $\tilde{\Omega}_{\text{ext}}$ — or to tweak the relaxation parameter [45]. Additionally in the particular case of the chameleon field entering the so-called screened regime, we can get rid of the region of the mesh $r < R_{\text{screened}}$ where the field is screened (i.e. constant) and apply a Dirichlet boundary condition at $r = R_{\text{screened}}$. When all the above failed, we resorted to so-called *ramping* [29, 30] or *numerical continuation* methods [3, 45]. For example in some cases, we would gradually vary the α parameter from Eq. (7) from a value where the solution is known to the desired value which is problematic convergence-wise, using the solution at each intermediate step as an initial guess of the next one. In spite of all these additional tricks, some combinations of $\{\alpha$, atmosphere model$\}$ resisted all our attempts and were thus discarded from this study. As a closing remark, let us emphasize the fact that we made use of many widely spread techniques in the literature for nonlinear FEM problems (see e.g. Ref. [45] chapter 4), both for implementation and verification purposes. While we are unable to quantify the relative error made on each solution obtained in this study, we grant them a sufficiently high level of confidence that the orders of magnitude discussed hereafter are correct, leaving the physical conclusions unchanged.

In total, we ran FEM computations for four different density profiles outside the main body — the constant vacuum value $\tilde{\rho}_{\text{vac}} = 10^{-15}$ as well as the three atmospheric models depicted in Fig. 3 — and for $\log_{10}\alpha \in \{-5, \dots, -28\}$. This amounts to nearly a hundred problems to solve on meshes with roughly 10^6 $\mathbb{P}_2$-triangles. The computations were performed on an ONERA's computing platform equipped with Broadwell and Cascade Lake nodes.

III. Modified gravity around and above a mountain

In this section, we present and analyze simulations results. We start off with the atmosphere-free case before discussing the influence of each atmospheric models. Due to the parameters' degeneracy mentioned in Sec. II A 2, we decided to fix $\Lambda = \Lambda_{\text{DE}}$ for all numerical results and figures presented in the following. One can refer to Fig. 1 to get a better grasp of the $(\beta, \Lambda, n) \mapsto (\alpha, n)$ mapping.

A. Simulation of an atmosphere-free planet

1. *Gravitational potential profiles and spherical harmonics decomposition*

The total gravitational potential is the sum of the Newtonian potential Φ and the chameleon potential Ψ as defined in Sec. II A. These are the direct results of FEM computations, i.e. *femtoscope*'s outputs. As this raw data can sometimes be noisy, we had recourse to smoothing splines notably for post-process operations requiring the evaluation of the fields outside mesh data points like the computation of spherical harmonic coefficients using `SHTools` [77]. Note that the azimuthal symmetry of our setup imposes that the only nonzero coefficients are the ones for which the order m is equal to zero.

In Fig. 6, we represent the potential profiles as a function of the colatitude θ — the radial coordinate being fixed at $\tilde{r} = 1.059$ — (left column) and their associated spherical harmonic coefficients for degrees $l \in \{1, \cdots, 100\}$ (right column). The top row corresponds to the specific case of the Newtonian potential while the following four rows correspond to chameleon potentials for $\log_{10}\alpha \in \{-4, -6, -15, -25\}$ respectively. The Newtonian potential Φ is by far the largest contribution to the total potential: roughly $-10^7\,\text{m}^2/\text{s}^2$ compared to $0.2\,\text{m}^2/\text{s}^2$ for the chameleon potential in the $\alpha = 10^{-25}$ case (last row of the same figure). The variation of the potential with respect to θ around this mean value has the same kind of shape where both ends of the curves have a slope that goes down to zero for symmetry reasons. Note that the potential is always smaller at $\theta = 0$ than at $\theta = \pi$. This is because the mass excess that the mountain represents is located at $\theta = 0$, forming a deeper potential well.

While all four potential profiles share this apparently common trend, the spherical harmonic coefficients displayed on the right column reveal important differences and two types of spectrum emerge. On the one hand, the Newtonian potential and the chameleon potential for $\alpha = 10^{-4}$ have a similar, monotonically decreasing spectrum. This is due to the fact that here, the chameleon field is unscreened which means that all the mass of the main body contributes to the field just like in the Newtonian case. On the other hand, as soon as $\alpha < 10^{-5}$, the chameleon field enters the screened regime, changing the shape of the spectra. We recall that the smaller α,

the more screened the setup. These spectra all have a maximum for $l > 1$. This distinctive feature of screened chameleon potentials, which could not be seen by eye on the left-hand-side curves, is nevertheless small in front of the Newtonian potential's coefficients Φ_{l0}.

As α decreases, the chameleon potential mean value increases. Indeed, we have $\Psi = K\tilde{\phi}$ where $K \propto \alpha^{-(n+1)/(n+2)}$ at fixed Λ. As a result, the spherical harmonic coefficients also get amplified as α decreases, leading to a more disturbed gravitational potential.

Fig. 18 in Appendix D further shows how the spectra evolve as the altitude is increased for the Newtonian potential and the chameleon potential ($\alpha = 10^{-25}$).

2. *Newtonian gravity and fifth-forces*

Once the gravitational potential is known, the actual gravitational acceleration is easily derived by computing its gradient. It is convenient to decompose the acceleration vector $\mathbf{a}$ onto the unit vectors $(\mathbf{e}_r, \mathbf{e}_\theta)$ (there is no component of the acceleration on $\mathbf{e}_\varphi$ due to rotational invariance) such that

$$\mathbf{a} = a_r \mathbf{e}_r + a_\theta \mathbf{e}_\theta \, .$$

In practice, the dimensionless gradient is computed numerically and then multiplied by the relevant coefficient a_0 with units m/s^2 — see Eq. (9). Fig. 7 gives an overview of both Newtonian acceleration (top panel) and fifth-forces for $\log_{10}\alpha \in \{-15, -27, -28\}$. Specifically, we represent the component a_r (purple curve) and a_θ (crimson curve) as a function of θ while the altitude is held fixed at $\tilde{r} = 1.059$.

An important point to discuss here is the limit $\alpha \to 0$. On the one hand, we have seen that for chameleon gravity, a_0 is proportional to $\alpha^{-(n+1)/(n+2)}$ at fixed Λ, and consequently

$$a_0 \underset{\alpha \to 0^+}{\longrightarrow} +\infty \text{ with constraint } \Lambda = \Lambda_{\rm DE} \, .$$

This gives the impression that one can make the fifth-force as large as desired simply by taking an ever-decreasing value of α. On the other hand, we know that in the limit $\alpha = 0$, the chameleon field profile is trivially given by $\tilde{\phi} = \tilde{\rho}^{-1/(n+1)}$ (take $\alpha = 0$ in Eq. 7). Yet for altitudes higher than the mountain's height, our models are such that $\partial_\theta \tilde{\rho} \equiv 0$ so that $\tilde{a}_\theta$ is expected to vanish for sufficiently small values of α. In front of this *apparent* paradox, we raise two points:

1. Taking the limit $\alpha \to 0$ at fixed Λ coerces $\beta \to +\infty$. Yet, a glimpse at the chameleon constraints plot (see e.g. Fig 3 from Ref. [18]) reveals that chameleon models with $\beta > 10^{14}$ are ruled-out by precision atomic tests. In our case, this corresponds to $\alpha < 10^{-37}$, which is out of the range of α values covered in this study.

	$\arg\max_\alpha$	value
$\tilde{a}_r$	10^{-24}	1.47×10^8
a_r	10^{-25}	1.46×10^{-7} [m/s^2]
$\tilde{a}_\theta$	10^{-25}	1.09×10^6
a_θ	10^{-25}	1.40×10^{-9} [m/s^2]
$\|\tilde{\mathbf{a}}\|$	10^{-24}	4.58×10^9
$\|\mathbf{a}\|$	10^{-25}	4.50×10^{-6} [m/s^2]

TABLE III. Assessment of the maximum fifth-force at $\tilde{r} = 1.059$

2. Another argument that does not involve referring to current chameleon constraints can be made on the basis of Figs. 7 and 8. On Fig. 8, we decompose a_θ into the product $a_0(\alpha)$ times $\tilde{a}_\theta = \partial_\theta \tilde{\phi}/\tilde{r}$. The two terms of this product both depend on α: while a_0 is simply a power law of α, $\tilde{a}_\theta$ clearly exhibits the phenomenon aforementioned, namely that the dimensionless gradient — after reaching a peak for $\alpha = 10^{-25}$ — vanishes for $\alpha < 10^{-28}$. When multiplied together, these two quantities result in a_θ which is scattered in log-scale on the bottom panel of this figure. We recognize the power law behavior $a_\theta \propto \alpha^{-n/(n+2)}$ in the range $[10^{-10}, 10^{-21}]$ where $\tilde{a}_\theta$ is roughly constant, followed by a sharp decline due to the vanishing of $\tilde{a}_\theta$. This explains why on Fig. 7, the transition from $\alpha = 10^{-27}$ to $\alpha = 10^{-28}$ completely destroys $\boldsymbol{\nabla}\Psi$. There only remains numerical noise, whose amplitude has no genuine physical meaning.

We also conducted the same analysis as performed in Fig. 8 for the radial component of the acceleration vector as well as for its norm (still for $\tilde{r} = 1.059$). The results are reported in Table III. In terms of dimensionless quantities, the orthoradial component $\tilde{a}_\theta$ is maximum for $\alpha = 10^{-25}$ while the radial component and the norm of the gradient are both maximized for $\alpha = 10^{-25}$. When these quantities are expressed in units homogeneous to an acceleration (m/s^2), $\alpha = 10^{-25}$ is again the argument that maximizes them all. This corresponds to the traditional parameters $(\beta, \Lambda, n) \sim (10^6, \Lambda_{\rm DE}, 1)$

Besides, we can discuss the direction of the acceleration vector by computing the ratio a_θ/a_r for both Newtonian and chameleon gravity. We conclude that the Newtonian part of the total acceleration vector is very radial, with $\max_\theta (a_\theta/a_r) \leq 3 \times 10^{-5}$, whereas the chameleon acceleration has a more significant orthoradial component since $\max_\theta (a_\theta/a_r) \leq 10^{-2}$ for $\log_{10}\alpha = -25$. The physical interpretation for this discrepancy is that in the screened regime, the chameleon acceleration is sourced only by a thin outer layer of the planet which is commonly referred to as the *thin-shell* [40].

Finally, one may be surprised by the fact that maximum fifth forces are obtained for values of the parameters (β, Λ, n) which belong to the thin-shell regime; as

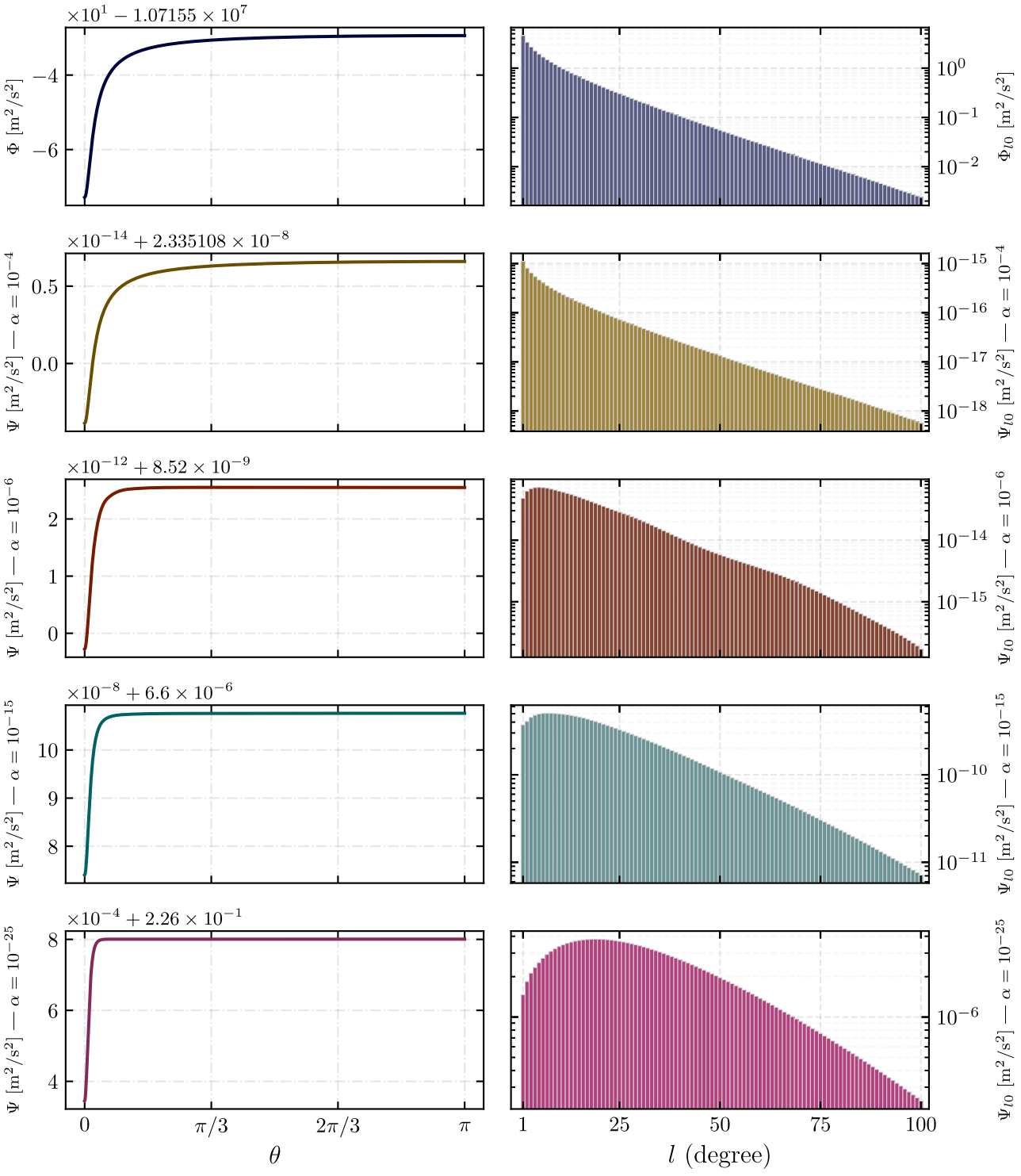


FIG. 6. Newtonian and chameleon potential profiles (left column) together with their spherical harmonic coefficients up to degree 100 (right column) computed at $\tilde{r} = 1.059$. The top row corresponds to the Newtonian case while the four remaining rows are associated with chameleon potentials with $\log_{10}\alpha \in \{-4, -6, -15, -25\}$ from top to bottom. The monopole (Φ_{00}, Ψ_{00}) is deliberately excluded from the bar graphs because it is only dependent on the field's mean value. All quantities are expressed in $\mathrm{m}^2/\mathrm{s}^2$.

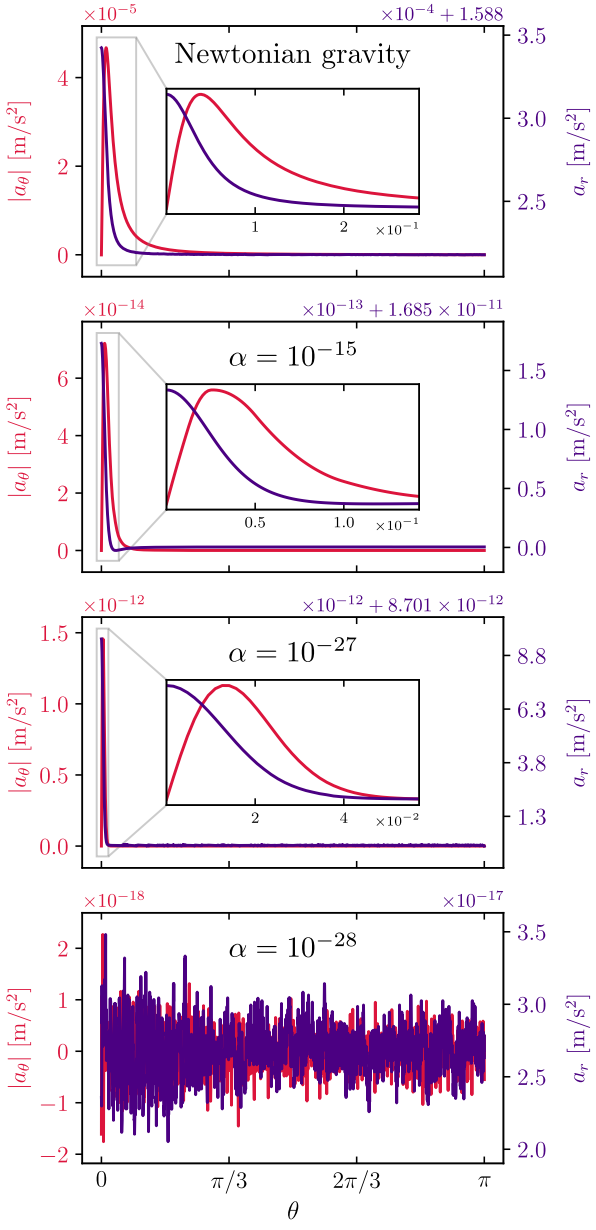


FIG. 7. Gravitational field $\mathbf{a} = a_r \mathbf{e}_r + a_\theta \mathbf{e}_\theta$ in Newtonian gravity (top) and in the chameleon model for the set of parameters $\{\alpha \in \{10^{-15}, 10^{-27}, 10^{-28}\}, n = 1, \Lambda = \Lambda_{\rm DE}\}$ at $\tilde{r} = 1.059$. The orthoradial acceleration a_θ is depicted by the red curve (left axis) while the radial acceleration a_r is depicted by the purple curve (right axis).

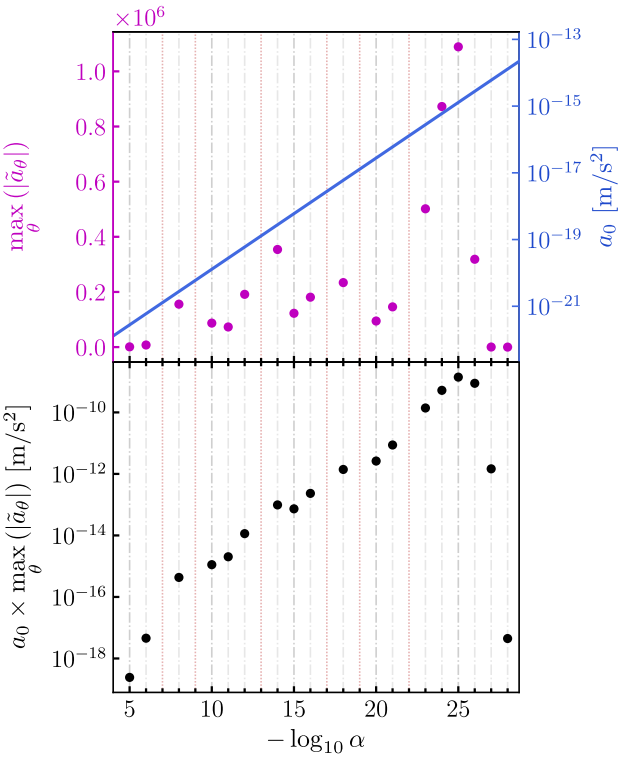


FIG. 8. Study of the chameleon orthoradial acceleration $a_\theta = a_0 \times \tilde{a}_\theta$ with respect to α at fixed Λ and fixed altitude $\tilde{r} = 1.059$. The top panel features each term separately, $\tilde{a}_\theta$ in magenta dots (dimensionless) and a_0 as the blue curve (in m/s^2). The bottom panel is simply the product of these two terms a_θ (in m/s^2). Finally, the red vertical dotted lines correspond to values of α for which the FEM computation was deemed unsatisfactory (failure of Newton's method to converge or unacceptably large residuals).

this appears to contradict the usual rule of thumb that fifth forces should be suppressed in this regime. There is actually no contradiction, provided we clearly define the context. Indeed, the total fifth force acting on a given *macroscopic* body can be computed via the integration of the gradient of the field on its whole volume — as done later in Sec. IV A 2 for instance. It is true that, if the macroscopic object at stake has a thin-shell, the integral of the gradient of the field vanishes everywhere but in that thin-shell, greatly reducing the overall fifth force experienced by that body. Here however, the situation is radically different: we are interested in the fifth force *experienced by* a point-mass (which by essence, cannot possess a thin-shell) *sourced by* a mountainous planet. In this framework, we merely observe that increasing the value of the coupling constant β in a given range, while keeping n and Λ fixed, results in greater fifth forces. Incidentally, increasing β while keeping n and Λ fixed means decreasing α (see Fig. 1) and results in a more screened body. This phenomenon was already observed in Figs. 14 and 15 of our previous work [47] and in Fig. 7 of Ref. [17] for instance. This can also be understood in the framework of the analytical approximation of the chameleon fifth force for spherical objects. Taking Eq. (2.64) from Ref. [63] reads

$$a_\phi = 3\beta^2 \frac{\Delta R}{R} \frac{G M_{\rm ball}}{r^2} (1 + m_\phi) \mathrm{e}^{-m_\phi (r - R_{\rm ball})} .$$

In this expression, $\Delta R/R \propto \beta^{-1}$ and $m_\phi \propto \beta^{\frac{n+2}{2(n+1)}}$ so that, at fixed $r > R_{\rm ball}$, the function : $\beta \mapsto a_\phi$ is increasing on the interval $]0, \beta^*[$ and decreasing on $]\beta^*, +\infty[$, for a certain parameter $\beta^* > 0$. We also see that $a_\phi \to 0$ when $\beta \to \infty$, because of the exponential term. This is exactly the phenomenology that we observe on numerical simulations: beyond a certain value of $\beta \sim 10^8$ (i.e. below a certain value of α, which is around 10^{-28}), the fifth force vanishes — see Figs. 7 and 8.

B. Adding an atmosphere

Here we study the influence of adding an atmosphere to the density model. To the best of our knowledge, only a handful of studies deal with the influence of the atmosphere [34, 39–41, 76]. In this section, we address simple questions: how is the fifth-force mitigated by the presence of an atmosphere? Can the mountain still be somehow *seen* in the field profile? How does all of this depend on the atmospheric model?

Part of the answer can be unveiled by first studying a simplified version of the setup. More precisely, we got rid of any orthoradial dependence in the density distribution function — which amounts to taking the mountain out of our model to end up with a purely radial setup. This simplification allows us to perform computationally inexpensive 1D FEM simulations with *femtoscope* and still get valuable insight into how the chameleon field behaves in the presence of an atmosphere. We ran computations for all atmospheric models outlined in Sec. II B 2 and for all values of $\alpha \in \{10^{-5}, \cdots, 10^{-29}\}$. Part of this simulation campaign has been compiled into Fig. 9, where the vast range of α values explored has been boiled down to only four distinct values for the sake of clarity and conciseness. On each sub-panel, the grey dash-dotted line is associated with the *fully screened* profile obtained in the limit $\alpha = 0$ which is given by $\tilde{\phi}(\tilde{r}) = \tilde{\rho}(\tilde{r})^{-1/(n+1)}$. Contrary to the previous atmosphere-free case, where the radial density would have been a mere Heaviside step function, the atmospheric density function smoothly interpolates between $\tilde{\rho}_{\rm atm}(\tilde{r} = 1)$ to 10^{-15} such that the asymptotic profile's gradient $\partial_r \tilde{\phi}(\alpha \to 0)$ is not identically zero.

It is only when we put these 1D chameleon profiles into perspective with the full 2D simulation's results that a clear understanding of the influence of the atmosphere emerges. Starting from $\alpha = 10^{-5}$ and gradually decreasing the value of this parameter, we witness the succession of several regimes:

1. For the larger values of α, the planet is not fully screened, i.e. there is still a thin-shell. This can be seen on the first column of Fig. 9 ($\alpha = 10^{-7}$) where the chameleon *kicks in* (i.e. departs from limit profile $\tilde{\phi}(\alpha \to 0)$) before $\tilde{r} = 1$ (which corresponds to the transition between the planet and the atmosphere). This regime is particularly visible on sub-panels a) to d). The impact of the atmosphere on the fifth-force at higher altitudes is then minor — see Tab. IV thereafter where we compare the amplitude of the fifth-force with and without atmosphere at $\tilde{r} \in \{1.059, 1.314\}$.

2. At some point when decreasing α, the lowest part of the atmosphere becomes screened itself. This is especially visible on sub-panels e) to h). We provide a zoomed-in view of this very region in order to be able to compare the fraction of the atmosphere that is screened against the relative size of the mountain $\tilde{h}_m = 0.01$. As soon as the screened area overflows the mountain, i.e. everything below $\tilde{r} = 1.01$ is screened, the imprint of the mountain of the chameleon field is definitely lost at higher altitudes. In other words, the orthoradial acceleration vanishes, giving way to numerical noise. This is why some entries of Table IV are set to N/A. When it comes to radial component of the fifth-force, it is hardly modified compared to the scenario without atmosphere.

3. For even smaller values of α, the screening eventually reaches the probed region at high altitude. This is particularly clear in sub-panels i) to l), where the chameleon field profile is getting closer to the limit profile (grey dash-dotted line). Here, the orthoradial acceleration remains drowned in the numerical noise while the radial acceleration is fully dictated by the density profile. This is why in some cases, a_r can even become larger with an atmosphere than without (see entries of Table IV greater than unity).

Once we know that these three regimes exist regardless of the specific form of the atmospheric profile (as long as density decreases with altitude), we can start to be more quantitative by

- specifying where the transition between each regime occurs for the atmospheric models at stake;
- computing the attenuation factor on the fifth-force for the different density models.

These quantitative results are reported in Table IV where each entry is a pair (μ_r, μ_θ) defined as

$$\begin{aligned} \mu_r &= \max_\theta a_r^{\text{with-atm}}(\tilde{r}, \theta) / \max_\theta a_r^{\text{no-atm}}(\tilde{r}, \theta) \\ \mu_\theta &= \max_\theta a_\theta^{\text{with-atm}}(\tilde{r}, \theta) / \max_\theta a_\theta^{\text{no-atm}}(\tilde{r}, \theta) \end{aligned} \tag{14}$$

at a specific radial coordinate $\tilde{r}$. We refer to these coefficients as the attenuation factors, which are of course dependent on the atmospheric model as well the altitude at which they are computed.

The take home message from this study of atmospheric models is that the presence of an atmosphere, as tenuous as it may be, prevents access to the biggest fifth-force attainable without atmosphere. Indeed, we saw earlier

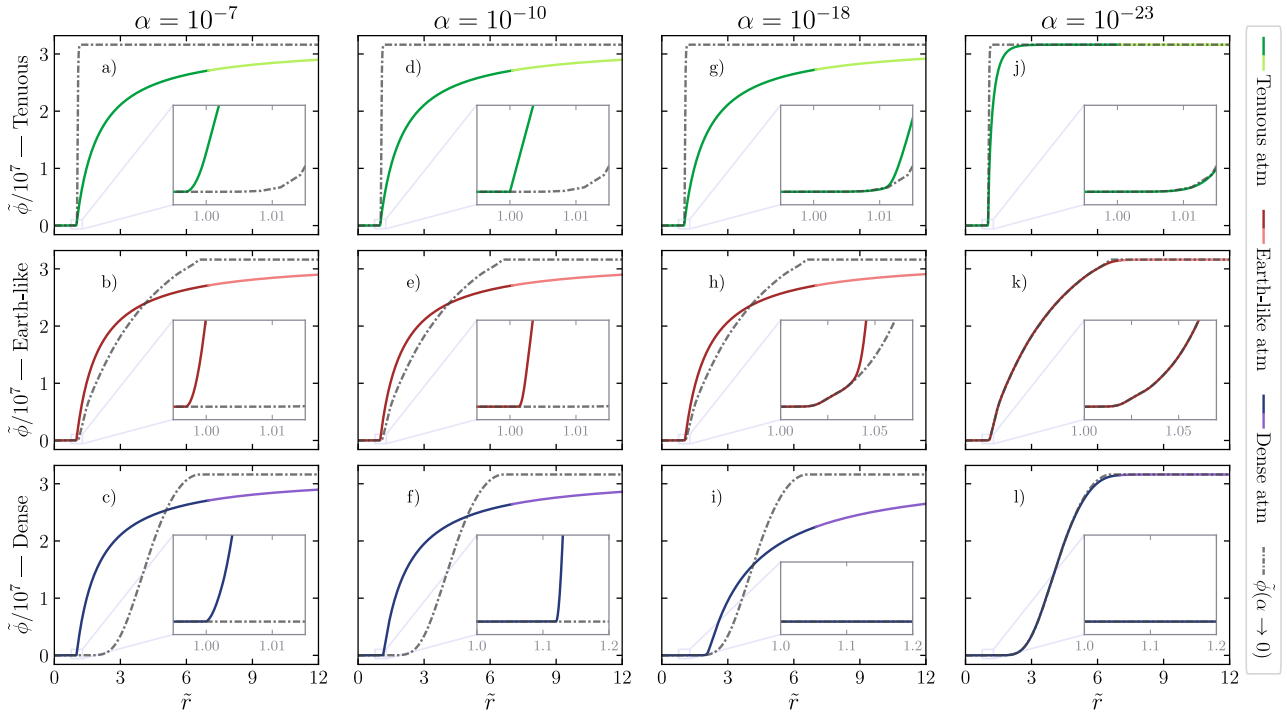


FIG. 9. Radial profiles of the chameleon field for $\log_{10}\alpha \in \{-7, -10, -18, -23\}$ (columns) and for all three atmospheric models defined in Sec. II B 2, namely *Tenuous*, *Earth-like* and *Dense* (rows). On each sub-panel, the grey dash-dotted line corresponds to *fully screened* profile, that is obtained in the limit $\alpha = 0$ and is given by $\tilde{\phi}(\alpha \to 0) = \tilde{\rho}^{-1/(n+1)}$. The radial chameleon profile is depicted by the bi-color solid line, where the transition from the darker color to the lighter one occurs at the chosen interface ($\tilde{r} = 7$) between the interior domain and the kelvin-inversed exterior domain (see Ref. [47] for more details).

	$\tilde{r} = 1.059$				$\tilde{r} = 1.314$			
	$\alpha = 10^{-6}$	$\alpha = 10^{-10}$	$\alpha = 10^{-15}$	$\alpha = 10^{-20}$	$\alpha = 10^{-6}$	$\alpha = 10^{-10}$	$\alpha = 10^{-15}$	$\alpha = 10^{-20}$
Tenuous	(1.00 – 1.00)	(1.00 – 0.89)	(1.00 – 0.15)	(1.03 – N/A)	(1.00 – 0.99)	(1.00 – 0.72)	(1.01 – 0.11)	(1.02 – N/A)
Earth-like	(1.00 – 1.00)	(1.00 – 0.76)	(1.01 – N/A)	(0.07 – N/A)	(1.00 – 0.99)	(1.00 – 0.61)	(1.02 – N/A)	(1.03 – N/A)
Dense	(1.00 – 0.99)	(7×10^{-7} – N/A)	(7×10^{-7} – N/A)	(6×10^{-7} – N/A)	(1.00 – 0.98)	(1.12 – N/A)	(6×10^{-5} – N/A)	(6×10^{-5} – N/A)

TABLE IV. Attenuation coefficients of the radial and orthoradial component of the chameleon acceleration with an atmosphere compared to the atmosphere-free case. The first number of each pair corresponds to the radial part and is computed as $\max_\theta a_r^{\text{with-atm}}(\tilde{r},\theta)/\max_\theta a_r^{\text{no-atm}}(\tilde{r},\theta)$. Similarly, the second figure of each pair is the orthoradial attenuation factor and is defined by $\max_\theta a_\theta^{\text{with-atm}}(\tilde{r},\theta)/\max_\theta a_\theta^{\text{no-atm}}(\tilde{r},\theta)$. These attenuation factors are computed for $\tilde{r} \in \{1.059, 1.314\}$.

that, around $\tilde{r} = 1.059$, the fifth-force was reaching its maximum value for α in the order of 10^{-25}. Yet in all three atmospheric models under study, the screening of the atmosphere at low altitude occurs for much bigger values of α, putting a lower threshold on the maximum accessible fifth-force. When put into perspective with current bounds on $n = 1$ chameleon theory, these results show that the largest part of the unconstrained region maps to a screened atmosphere in the LEO altitude range. All other things remaining equal, the radial component of the fifth force can be recovered by going higher up in altitude, where the atmospheric density is lower.

IV. Influence on spacecraft trajectory

In this section, we shift our focus to how geodesics get modified in the presence of a putative chameleon fifth-force with respect to the purely Newtonian case. We want to ascertain the effects of the fifth force in a rather quantitative way: is the deviation from Newtonian dynamics large enough to be detected by current satellite technology? Is it possible to discriminate the presence of a fifth force from the imperfect knowledge of the model at stake or small perturbations of the initial conditions? When does a satellite in Low Earth Orbit (LEO) become screened? Besides, we will refrain from commenting too

much on secular drifts that can arise between modified gravity and Newtonian gravity. That is because in any realistic scenario — where many additional forces of different nature come into play —, it would be merely impossible to discriminate the fifth force from such forces. We thus keep our analysis *local*, by focusing our attention on the dynamics at the passage over the mountain.

A. Screening of the fifth force by the spacecraft

1. *Existing criteria*

We stress that modeling a spacecraft by a material point (in the framework of chameleon gravity) roughly amounts to making the hypothesis that it does not possess a thin-shell. Ref. [40] derives an analytical criterion for a typical satellite in low Earth orbit not to have a thin-shell (see Eq. 80 of this reference). Applying this criterion with the density values employed in our study (except for that of the satellite itself which is set at $8 \times 10^3\,\mathrm{kg/m^3}$) leads straightforwardly to the requirement that $\beta \lesssim 2\times10^2 \iff \alpha \gtrsim 3\times10^{-20}$ (for $\Lambda = \Lambda_{\rm DE}$ and $n = 1$). The following orbit propagation results being performed with $(n = 1, \Lambda = \Lambda_{\rm DE}, \alpha = 10^{-25})$, the satellite would be partially screened according to this criterion and the chameleon effects would thus be smaller than presented.

Ref. [61] derives another criterion based on numerical simulations claiming that the satellite will be fully screened when the thickness of its walls is larger than $100\lambda_{\rm c,wall}$, where $\lambda_{\rm c,wall}$ refers to the Compton wavelength in the wall. However, this criterion must be taken with a grain of salt as it was derived for a density contrast $\rho_{\rm vac}/\rho_{\rm wall} = 10^{-3}$, far from a realistic setup. Still, applying this second criterion for a wall of thickness 10 cm leads to the fact that the satellite will not have a thin-shell if $\beta \lesssim 2\times10^{-2} \iff \alpha \gtrsim 4\times10^{-14}$.

2. *Computation of the field with femtoscope*

Although quite qualitative, these two criteria provide us with a comprehension of how the parameters in our model affect the screening of the satellite. For instance, increasing the overall density of the satellite (other things being equal) results in more screening. Ideally, one would compute the scalar field profile sourced by the Earth and the spacecraft all at once — which would avoid having to rely on such criteria and provide a definitive answer. The problem then becomes a numerical one, because the simulation should accommodate a thousand-kilometer-size object (a planet) together with a meter-size object (a satellite). We create a mesh using the Gmsh software [33] that captures both scales (whose ratio is equal or less than 10^{-6}) thanks to *h-adaptivity* — a technique that adjusts the mesh resolution by refining or coarsening elements to focus computational resources where they are most needed. The setup is as follows: we place a cylindrical object centered at coordinates $(\tilde{x}_{\rm Sat}, \tilde{z}_{\rm Sat}) = (0, 1.1)$ whose axis is aligned with the z-axis (Fig. 2). The diameter and height of the cylinder are set equal to $L_{\rm Sat}$ and we denote by $\rho_{\rm Sat}$ its density. In order to get an order of magnitude of a satellite mean density, we take the example of a CubeSat[9] whose density is around $\sim 10^3\,\mathrm{kg/m^3}$. We then compute the chameleon field map in the (x, z)-plane for various combinations of $\rho_{\rm Sat}$, $L_{\rm Sat}$ and α. The global acceleration undergone by the cylindrical satellite $\mathbf{a}^{\rm tot}_{\rm cham}$ is obtained by integrating the gradient of the scalar field over its whole volume. Under the assumption that the satellite is made of a material of constant density, one gets

$$\mathbf{a}^{\rm tot}_{\rm cham} = -\frac{1}{V}\int_V \boldsymbol{\nabla}\Psi\,\mathrm{d}V = -\frac{\beta}{V M_{\rm Pl}}\int_V \boldsymbol{\nabla}\phi\,\mathrm{d}V\,, \quad (15)$$

where $V = \pi L^3_{\rm Sat}/4$ is the volume of the cylinder. Now because the setup admits $\mathcal{O}_{xz}$ and $\mathcal{O}_{yz}$ as planes of symmetry, $a_z^{\rm tot} = \mathbf{a}_{\rm cham}\cdot\mathbf{e}_z$ is the only non-zero component of the acceleration vector. Setting $x_{\max} = L_{\rm Sat}/2$, $z_\pm = z_{\rm Sat} \pm L_{\rm Sat}/2$, the calculation thus simplifies to

$$\begin{aligned} a_z^{\rm tot} &= -\frac{1}{V}\int_0^{2\pi}\int_0^{x_{\max}}\int_{z_-}^{z_+} x\partial_z\Psi\,\mathrm{d}z\,\mathrm{d}x\,\mathrm{d}\theta \\ &= \left(\frac{2}{L_{\rm Sat}}\right)^3 \int_0^{x_{\max}} x\left[\Psi(x,z_-) - \Psi(x,z_+)\right]\mathrm{d}x\,. \end{aligned} \quad (16)$$

The resulting 1D integral can easily be computed using any numerical integration routine.

Fig. 10 shows the chameleon potential Ψ (top row) together with the elementary acceleration $a_z = -\partial_z\Psi$ (bottom row) along the axis of the cylinder that passes through the Earth. On panel a), we recognize the customary chameleon field profile of a screened ball, perturbed nearby $z = 1.1\,R_\oplus$ by the presence of the satellite. When we zoom-in, we see the potential well imputed to the satellite in panel b). This localized variation of the chameleon field results in a large gradient in absolute value (bigger than anywhere else in the numerical domain). However big the field's gradient may be, looking at panel d) with naked eyes could lead us to believe that it is an odd function with respect to $z = z_{\rm Sat}$. If that turned out to be the case, then performing the integration (15) would result in a net zero acceleration and the satellite's trajectory would coincide with GR geodesic (in the absence of any non-gravitational perturbation).

We tackle this issue by computing $a_z^{\rm tot}$ using Eq. (16) for several physical parameters ($\rho_{\rm Sat}$, $L_{\rm Sat}$) and several chameleon parameters α. From there, the whole question is to determine how the total chameleon acceleration undergone by the satellite compares against that of a

[9] CubeSats have a form factor of 10 cm cubes and have a mass of no more than 2 kg.

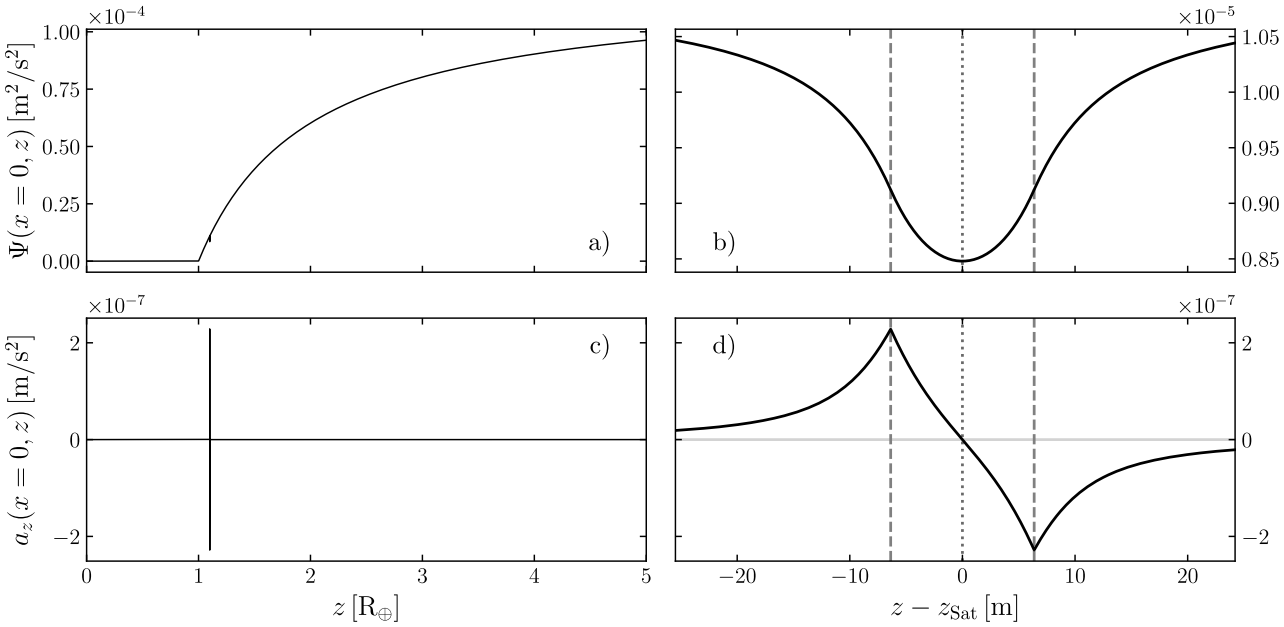


FIG. 10. Chameleon potential Ψ (top row) and acceleration a_z (bottom row) along the z-axis. The panels b) and d) are a zoomed version around $z_{\rm Sat} = 1.1\,R_\oplus$ of panels a) and c) respectively. The parameters used to produce this figure are: $\rho_\oplus = 10^3\,{\rm kg/m^3}$, $\rho_{\rm vac} = 10^{-15}\,{\rm kg/m^3}$, $\rho_{\rm Sat} = 10^3\,{\rm kg/m^3}$, $L_{\rm Sat} = 2\times10^{-6}\,R_\oplus \sim 12.7\,{\rm m}$, $z_{\rm Sat} = 1.1\,R_\oplus \sim 7\times10^3\,{\rm km}$, $\alpha = 10^{-15}$, $n = 1$, $\beta = 0.24$, $\Lambda = \Lambda_{\rm DE}$. On panels b) and d), the dotted line is centered at $z = z_{\rm Sat}$ while the dashed lines represent the extent of the satellite.

	Case 1 — benchmark ($\rho_{\rm Sat} = 10^3\,{\rm kg/m^3},\ L_{\rm Sat} = 2\times10^{-6}\,R_\oplus$)			*Case 2 — denser* ($\rho_{\rm Sat} = 10^4\,{\rm kg/m^3},\ L_{\rm Sat} = 2\times10^{-6}\,R_\oplus$)			*Case 3 — smaller* ($\rho_{\rm Sat} = 10^3\,{\rm kg/m^3},\ L_{\rm Sat} = 5\times10^{-7}\,R_\oplus$)		
α	10^{-14}	10^{-15}	10^{-16}	10^{-14}	10^{-15}	10^{-16}	10^{-14}	10^{-15}	10^{-16}
$\lvert a_z^{\rm tot}\rvert$	7.25×10^{-12}	1.56×10^{-11}	~ 0	7.25×10^{-12}	~ 0	~ 0	7.25×10^{-12}	1.56×10^{-11}	3.36×10^{-11}
$\lvert a_z\rvert$	7.25×10^{-12}	1.56×10^{-11}	3.37×10^{-11}	7.25×10^{-12}	1.56×10^{-11}	3.37×10^{-11}	7.25×10^{-12}	1.56×10^{-11}	3.37×10^{-11}

TABLE V. Total chameleon acceleration undergone by a satellite (extended object) $\lvert a_z^{\rm tot}\rvert$ compared to that of a point-like particle $\lvert a_z\rvert$. The accelerations are expressed in ${\rm m/s^2}$. Each of the three cases corresponds to three different satellites: *Case 1* is a benchmark, *Case 2* represents a 10 times denser satellite, *Case 3* represents a 4 times smaller satellite. As long as the satellite is not screened, $\lvert a_z^{\rm tot}\rvert \simeq \lvert a_z\rvert$. When the satellite is screened (which occurs at a different α depending on the satellite's characteristics), $\lvert a_z^{\rm tot}\rvert$ drops down to nearly zero. Note that Fig. 10 corresponds to *Case 1* with $\alpha = 10^{-15}$.

point-like particle not affecting the background field. The results set out in Table V provide some answers. We consider three cases which correspond to three satellites with distinct characteristics, namely different length scale and density. For each case, we vary $\alpha \in \{10^{-14}, 10^{-15}, 10^{-16}\}$ and compute the total chameleon acceleration undergone by the satellite (extended object) $\lvert a_z^{\rm tot}\rvert$ and that of a point-like particle $\lvert a_z\rvert$. Surprisingly, the outcome of this experiment is binary:

– When the satellite is unscreened — that is when the scalar field does not reach the value that minimizes the effective potential inside the cylinder $\phi_{\rm Sat}$ — we find that the total chameleon acceleration it undergoes is equal to that of a test particle placed at $z_{\rm Sat}$. This is a remarkable fact, which we did not anticipate by simply looking at Eq. (16) and we thus provide an attempt to explain this phenomenon in Appendix E. In other words, the satellite *feels* the fifth force sourced by the Earth as if it did not perturb the field at all. Consequently, it behaves as a point-like particle and will follow the geodesics of the Jordan frame metric $\tilde{g}_{\mu\nu} = \exp\left(2\beta\phi/M_{\rm Pl}\right) g_{\mu\nu}$, where $g_{\mu\nu}$ refers to the Einstein frame metric.

– When the satellite is screened, the integral of the field's gradient over the volume occupied by the satellite vanishes almost completely — essentially because there, the gradient is null. The satellite only *feels* the Newtonian part of the gravitational force and thus follows the geodesics of the Einstein frame metric $g_{\mu\nu}$.

Of course, there actually exists an intermediate case where the satellite would only be *partially* screened, i.e. where the field would indeed reach $\phi_{\rm Sat}$ deep inside the cylinder while still having some space to vary in its outermost regions. In this specific case, the ratio $|a_z^{\rm tot}|\,/\,|a_z|$ lies somewhere between 0 and 1. However, the results reported in Table V suggest that the transition from the unscreened case and the fully screened case does not cover a wide region of the chameleon parameter space. Indeed, taking the *Case 1* as an example, the transition occurs between $\alpha = 10^{-15}$ and $\alpha = 10^{-16}$ — refer to Fig. 1 to get a better idea of the narrowness of this region in the chameleon parameter space.

3. Discussion

We can check that the reported results are in accordance with the qualitative predictions made by the first two criteria discussed earlier. They both predict that increasing the density and/or the length of the satellite should make it more likely to be screened. This is in agreement with our findings: (i) going from *Case 1* to *Case 2* shows the effect of an increase by one order of magnitude of the satellite's density, (ii) going from *Case 3* to *Case 1* illustrates the effect of increasing the satellite's overall size. A follow-up question is whether it is possible to find a distribution of mass inside the satellite $\rho_{\rm Sat}(x,z)$ such that $|a_z^{\rm tot}| > |a_z|$. The simple tests we performed so far — for instance, setting different densities for the upper and lower halves of the satellite — all resulted in $|a_z^{\rm tot}| \leq |a_z|$. The question remains open. Additionally, dealing with an extended object means that new rotational degrees of liberty can enter the scene and it would be interesting to look at similar optimization process in order to find the maximum torque (note that Refs. [38, 57] mention this effect and highlight the fact that it can stand out from Newtonian gravity).

Although the satellite model implemented in this section is very simple, this study shows that it is possible for a realistic satellite not to be screened in parts of the chameleon parameter space. This has implications for space-based tests of gravity. For instance, in chameleon models where the scalar field does not couple universally to all matter fields, violations of the weak equivalence principle are not necessarily suppressed by the satellite walls or the experimental setup (as opposed to what was claimed in Ref. [62]). Another example (which holds for a universal coupling constant β) is that of an accelerometer with a screened test mass onboard an unscreened satellite: the accelerometer would measure a force akin to a bias.

B. Orbital dynamics of an artificial satellite

Sec. IV A made it clear that in practice, a satellite orbiting some planetary body in the framework of chameleon gravity cannot be treated as a point-like particle in the entire parameter space. We have highlighted that there is a narrow transition zone beyond which the satellite becomes fully screened and the net fifth force acting on it vanishes almost completely. In what follows however, we make the assumption that we can treat the satellite as a point-like particle. This is justified by at least two reasons:

1. This is a valid approximation in parts of the parameter space (see Table V).
2. One can always, at least at the thought experiment stage, make the satellite smaller or less dense so that is it not subject to screening.

That being said, in all the orbit propagation results presented in the following, we choose the chameleon parameters that produce the strongest fifth force at $\tilde{r} = 1.059$ (which represents an altitude of approximately 376 km) in the absence of atmosphere: ($n = 1$, $\Lambda = \Lambda_{\rm DE}$, $\beta = 1.1 \times 10^6$). Notice that this point of the parameter space is already constrained by atom interferometry, see e.g. Ref. [19].

Suppose a point-like particle is placed in a gravitational potential U. The equations of motion in spherical coordinates are

$$\begin{aligned} \ddot{r} - r\left(\dot{\theta}^2 + \dot{\varphi}^2 \sin^2\theta\right) &= -\partial_r U \\ r\left(\ddot{\theta} + 2\frac{\dot{r}}{r}\dot{\theta} - \dot{\varphi}\sin\theta\cos\theta\right) &= -\frac{1}{r}\partial_\theta U \\ r\sin\theta\left(\ddot{\varphi} + 2\frac{\cos\theta}{\sin\theta}\dot{\theta}\dot{\varphi} + 2\frac{\dot{r}}{r}\dot{\varphi}\right) &= -\frac{1}{r\sin\theta}\partial_\varphi U \end{aligned} \quad , \tag{17}$$

where dots refer to time derivatives. The massic energy is given by

$$\mathcal{E} = \frac{1}{2}\left(\dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\dot{\varphi}^2\right) + U \tag{18}$$

and it is conserved along the trajectory, i.e. $\dot{\mathcal{E}} \equiv 0$. Our setup being axisymmetric, we can get rid of the φ-dependence. Then, note that Eq. (17) implies that the angular momentum $L \equiv r^2\dot{\theta}$ satisfies

$$\dot{L} = -\partial_\theta U\,. \tag{19}$$

The problem at stake is a perturbed Kepler problem (the mountain and fifth force contributions are small compared to the central force), whose total gravitational potential U can therefore be decomposed into

$$U = -\mu/r + \delta U\,,$$

where $\mu \equiv GM_{\rm body}$ is the standard gravitational parameter of the main body (note that μ does not encompass the mass contained in the mountain itself). The perturbation δU is the sum of the Newtonian potential of the mountain $\delta\Phi$ and the chameleon potential Ψ of the whole system. With this in mind, it is also more appropriate to

decompose the motion into a Keplerian part — that we assume to be circular — and a perturbed part, reading

$$\begin{aligned} r &= a + \delta r & \theta &= \theta_0 + \omega t + \delta\theta \\ \dot{r} &= \dot{\delta r} & \dot{\theta} &= \omega + \dot{\delta\theta} \\ \ddot{r} &= \ddot{\delta r} & \ddot{\theta} &= \ddot{\delta\theta} \\ L &= L_0 + \delta L & \dot{L} &= \dot{\delta L} \end{aligned} \quad . \tag{20}$$

In the above, a is the radius of the circular orbit and ω is the Keplerian pulsation, satisfying $\omega^2 = \mu/a^3$. L_0 is the initial angular momentum with $L_0^2 = \mu a$ and θ_0 is the initial co-latitude for a Keplerian motion. This lets us rewrite the equations of motion (17) as

$$\ddot{\delta r} = L^2/r^3 - \partial_r U\,, \quad \dot{\delta\theta} = L/r^2 - \omega\,, \quad \dot{\delta L} = -\partial_\theta U\,; \tag{21}$$

while the energy conservation reads

$$(\dot{\delta r})^2 + (L/r)^2 + 2U - 2\mathcal{E} = 0\,. \tag{22}$$

Note that at the 0^{th}-order, Eq. (22) boils down to the usual energy conservation in a circular Keplerian orbit

$$(a\dot{\theta})^2 - \frac{2\mu}{a} - 2\mathcal{E} = 0\,.$$

C. Numerical integration with energy conservation

The state vector that we wish to propagate over time is $\mathbf{X} = (\delta r, \dot{\delta r}, \delta\theta, \delta L) \in \mathbb{R}^4$. It is governed by the ordinary differential equation (ODE) $\dot{\mathbf{X}} = F(t, \mathbf{X})$, where $F\colon \mathbb{R}\times\mathbb{R}^4 \to \mathbb{R}^4$ is given by Eq. (21). Note that while energy conservation (22) is derivable from the ODE itself, there is no *a priori* reason for it to hold on the numerical approximation. For one thing, the energy might fluctuate on short time scales depending on the numerical integrator employed, leading to an increase or decrease over longer time scales. Additionally, the r.h.s. of ODE (21) is obtained through FEM computation and is hence *noisy*, meaning that even so-called *energy-preserving integrators* would exhibit the energy-drift phenomenon.

Appending the energy conservation (22) to the ODE defines an over-determined differential-algebraic system of equations (DAE). One convenient way to preserve first integrals such as energy conservation when numerically integrating dynamical systems is to resort to *projection techniques*. The idea behind this class of techniques is to slightly perturb the state after each solver's step so that the energy remains constant. This technique is described in Ref. [35]. As for the implementation, one can easily modify any existing general purpose ODE solver to perform this projection. We provide a minimally modified version of `scipy`'s Runge-Kutta solvers that was used for the numerical integration as supplementary material.

The simulations presented below are performed at $\tilde{r} = 1.059$, which corresponds to an altitude of roughly 376 km. The Newtonian potential and its gradient are evaluated using the point-mass approximation introduced in Sec. II D 2 as it is hardly distinguishable from the semi-analytical solution. As for the chameleon field, we have the freedom to select an operating point in the parameter space. We choose $(n = 1, \Lambda = \Lambda_{\text{DE}}, \alpha = 10^{-25}, \beta = 1.1 \times 10^6)$ which we have identified as the point that concurrently results in the strongest fifth force and the greatest field's strength (in the atmosphere-free scenario, see Sec. III A). The experiment performed in Sec. IV A indicates that any medium to large size satellite would presumably be screened in this case. Consequently, the point-like approximation we adopt can be understood as a *best case scenario*, i.e. an upper bound on the maximum fifth force. Indeed, escaping the screened regime comes at the prize of restricting the allowed range of parameter α to $\alpha > \alpha_{\text{screened}}$, which limits the maximum fifth force — see Sec. III A. In terms of initial conditions, the point-like particle is set in a Keplerian motion so that the initial state vector reads $\mathbf{X}(t=0) = \mathbf{0} \in \mathbb{R}^4$, with $\theta_0 = \pi$.

D. Results and discussion

Here we present and discuss the orbit propagation results. For the sake of clarity and concision, we denote by $\mathbf{X}^{\text{New}}$ and $\mathbf{X}^{\text{Cham}}$ the state vectors in the purely Newtonian case and in the modified gravity (i.e. the sum of chameleon and Newtonian gravity) respectively.

1. *Results of the simulations*

The evolution with respect to time of the main quantities of interest are presented in Fig. 11. The time spans 10 hours which encompasses roughly three full orbits. The first row of this figure shows how, in a purely Newtonian setting, the presence of the mountain breaks the Keplerian, circular motion. Some elements, such as δL are very correlated to the passage of the point-mass above the mountain (denoted by the vertical light-gray dashed lines on each panel). The angular momentum is indeed roughly constant along the trajectory, except nearby $\theta \sim 0$ where it peaks very sharply (L_0 being negative, this corresponds to an increase in the absolute value of L). On the other hand, some other elements are *irreversibly* imprinted by the mountain after the first passage above it, see e.g. δr, $\dot{\delta r}$ or $\delta\theta$, leaving traces on the longer term. The physical intuition for this is that, although the gravity field is symmetric with respect to $\theta = 0$, the dynamics is not. Indeed, in the $(\theta > 0)$-plane, the system acquires non-zero velocity $\dot{\delta\theta}$ which leads $\delta\theta$ to slowly drift away from zero initial state. Right after the passage of the mountain — that is when θ becomes negative — the mountain's gravity acts as a restoring force, which has the immediate effect of slowing down $\delta\theta$. But it is already too late: in the meantime, the altitude has been disturbed ($\delta r \neq 0$) and θ continues on its run (at

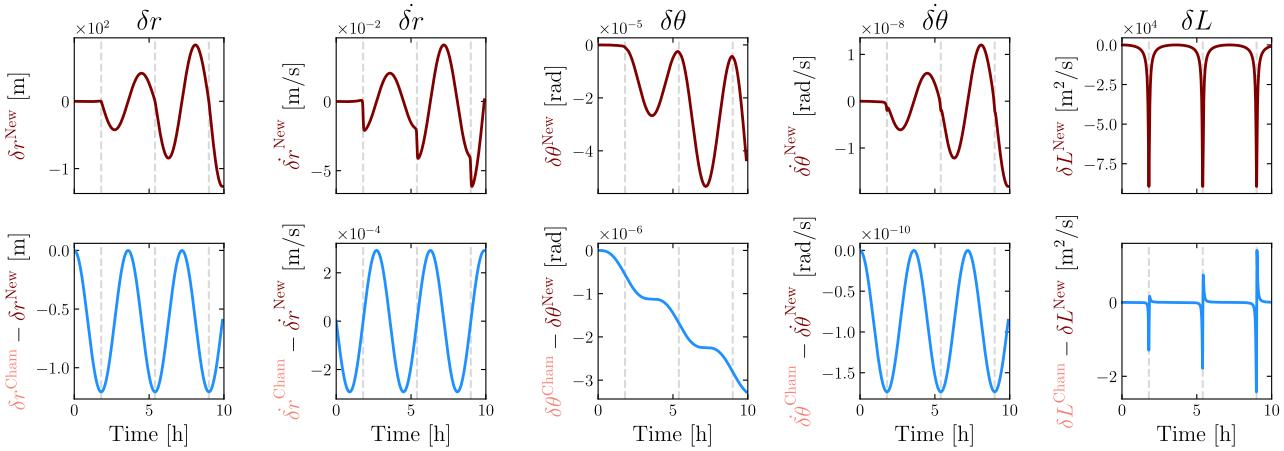


FIG. 11. Orbit propagation over three Keplerian periods. The first row shows the evolution of the state vector $\mathbf{X}^{\rm New} = (\delta r^{\rm New}, \dot{\delta r}^{\rm New}, \delta\theta^{\rm New}, \delta L^{\rm New})$ and $\dot{\delta\theta}^{\rm New}$ with respect to time, where the dynamics is purely Newtonian. The second row lays emphasis on the orbital dynamics in modified gravity by showing $\mathbf{X}^{\rm Cham} - \mathbf{X}^{\rm New}$ and $\dot{\delta\theta}^{\rm Cham} - \dot{\delta\theta}^{\rm New}$. The vertical light-gray dashed lines correspond to the instants at which the point-like particle passes over the mountain, at $\theta = 0$, in the purely Newtonian case.

an increased $\omega + \dot{\delta\theta}$ pace), so that the restoring force at $-\theta_* < 0$ is not equal to the force that disturbed the Keplerian motion at θ_*. Once the symmetry is broken, the orbit can no longer be circular — it has a non-zero osculating eccentricity — which is why $(\delta r, \dot{\delta r}, \delta\theta, \dot{\delta\theta})$ exhibit an oscillatory behavior at approximately the Keplerian frequency. We dedicate Appendix F to prove this point in a more rigorous way.

In the second row of Fig. 11, we illustrate what we call the *anomaly* $\mathbf{X}^{\rm Cham} - \mathbf{X}^{\rm New}$, that is simply the difference between the geodesic in modified gravity and in Newtonian gravity — for the same set of initial conditions. Surprisingly, apart from $\delta\theta^{\rm Cham} - \delta\theta^{\rm New}$ which undergoes a steady decline, the other elements of the anomaly seem to be periodic and remain around zero. In Fig. 12, we show the slow drift of the distance anomaly between the two trajectories, that is $\|\mathbf{r}^{\rm Cham} - \mathbf{r}^{\rm New}\|$. This steady increase of the distance has a mean slope of $\sim 2\,$m/h, but the rate of increase is maximized at each passage of the point-mass above the mountain where it exceed $4\,$m/h.

All these orders of magnitude relating to the anomaly should be put into perspective with the current level of precision with which we are able to determine a satellite's position and other orbital elements. This process goes under the name '*Precise Orbit Determination*' (POD) and involves analyzing various observational data, often obtained from ground-based tracking stations or satellite-based instruments — see e.g. Refs. [49, 67, 68] for the implementation of these techniques and the reachable orders of magnitude in terms of precision. One of the main space geodetic techniques is *Satellite Laser Ranging* (SLR) which measures the time it takes for a laser beam to travel from the ground station to a retro-reflector on the satellite and back again, providing unambiguous range measurements to millimeter precision [4, 60]. This

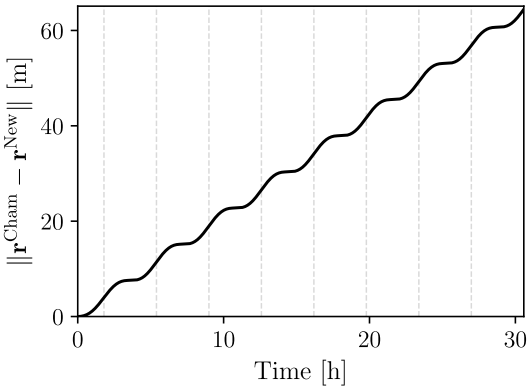


FIG. 12. Distance anomaly as a function of time. The various passages above the mountain, depicted by the vertical light-gray dashed lines, correspond to the most rapid increase in this distance.

technique is also placed at the service of fundamental physics; in this respect let us mention the recent launch of the LARES 2 satellite [25] to test GR.

Therefore, with no uncertainty on the model and initial conditions, the anomaly caused by the fifth force is of the order of a meter (see the leftmost column of Fig. 11), which is around three orders of magnitude larger than the best attainable precision. At this stage of the discussion, it would seem easy to detect the fifth force.

2. The GRACE-FO scenario

The GRACE-FO[10] mission, currently in operation, aims at monitoring the Earth's gravitational field. It uses a pair of satellites flying on the same orbital path, approximately 220 km apart. As they orbit the Earth, the spacecraft are affected by the uneven gravity field caused by the uneven distribution of mass inside the planet — e.g. the presence of a mountain, which produces a slightly stronger gravitational pull. As a result, the distance between the two satellites varies continuously over time. This distance variation is measured down to the micron level thanks to a microwave ranging system[11] [44]. Ultimately, the changes in the distance between the satellites are used to monitor the time variations of the Earth gravity field due to mass changes (ice melting, droughts, floods, etc.).

Given the extreme level of precision GRACE-FO is able to reach in terms of ranging, we investigate whether or not fifth force effects would end up being in its sensitivity range. To do so, we simulate a pair of satellites following each other by duplicating the trajectory and shifting it in time by a few minutes, mimicking the real mission configuration. We can then reconstruct the change in inter-spacecraft distance with respect to time $d(t)$. An example of such a curve is given in Fig. 13 (red solid line or salmon dashed line, the two being indistinguishable by eye), where roughly three orbits have been completed. The passage above the mountain can easily be spotted on the curve by the little *spikes* they spawn. They can be understood fairly intuitively: approaching the mountain's latitude, the leading satellite starts feeling a slightly stronger gravity relative to the trailing one and is pulled slightly ahead, increasing the distance between the two satellites. When the first satellite has eventually passed on the other side of the mountain (that is $\theta < 0$), it is slowed down while the trailing satellite is accelerating, resulting in a decrease in the inter-satellite distance overall. Long after the occurrence of this short-term event, this distance continues to vary in sinusoidal fashion. This is due to the fact that, as discussed above and brought to light in Fig. 11, the orbit is no longer circular after the passage of the mountain and therefore the velocity varies along an orbit.

The difference between the modified gravity case $d_{\rm Cham}$ and the Newtonian case $d_{\rm New}$ can hardly be seen on those curves. It is depicted by the solid blue line on Fig. 13 (again called the 'anomaly'). Choosing the same set of initial conditions for both models ensures that the anomaly is null at time $t = 0$. The anomaly exhibits three maxima — at a level of a few centimeters — corresponding to the passage of the pair of satellites above the mountain.

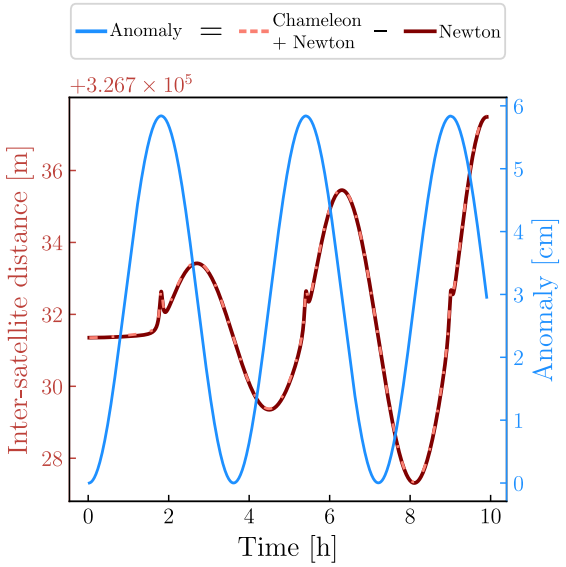


FIG. 13. Left y-axis: inter-spacecraft distance with respect to time in Newtonian gravity (solid red line) and in modified gravity (dashed salmon line). Right y-axis: the blue curve corresponds to the anomaly, that is the difference between the two models. The initial time-delay between the two satellites is set to 100 seconds.

In view of these results, we may believe that chameleon-like fifth forces should be detectable with our current space technology as the anomaly is $\sim 10^4$ times larger than the sensitivity threshold of GRACE-FO. That would be true under the (unrealistic) assumptions that:

1. the initial conditions are *perfectly* known, that is there is no uncertainty in our initial state $\mathbf{X}_0$ prior to the propagation;

2. the density model of the main body (the Earth) is *perfectly* known.

Neither of the two hypotheses can be fulfilled in practice. In the two forthcoming sections, we tackle these points and strongly mitigate our previous statement in regard to fifth force detectability in space.

3. Perturbation of initial conditions

Here, we investigate whether a slight modification of the initial state vector $\mathbf{X}_0 \leftarrow \mathbf{X}_0 + \delta\mathbf{X}_0$ could account for the anomaly that we unveiled in Fig. 13. For that purpose, we employ a Nelder-Mead optimizer where our objective function is

$$\begin{aligned} g\colon \mathbb{R}^4 &\to \mathbb{R}_+ \\ \delta\mathbf{X}_0 &\mapsto \|d_{\rm Cham}(\mathbf{t}, \mathbf{X}_0) - d_{\rm New}(\mathbf{t}, \mathbf{X}_0 + \delta\mathbf{X}_0)\|_2 \,, \end{aligned} \tag{23}$$

[10] Gravity Recovery and Climate Experiment Follow-On.
[11] GRACE-FO also employs laser-ranging interferometry for a more accurate inter-satellite ranging which can improve the separation distance measurement by a factor of more than 20 relative to the GRACE mission [2].

where $\|\cdot\|_2$ is the two-norm in $\mathbb{R}^4$ and $\mathbf{t} = [t_0, t_1, \ldots, t_N]$ is the discrete time vector with $t_N \sim 10\,\mathrm{h}$ and $N = 10^4$. We find an optimum at

$$\delta\mathbf{X}_0^{\mathrm{opt}} = \begin{bmatrix} -6.04 \times 10^{-1} & \mathrm{m} \\ +6.16 \times 10^{-7} & \mathrm{m/s} \\ -7.49 \times 10^{-7} & \mathrm{rad} \\ +1.98 \times 10^{3} & \mathrm{m^2/s} \end{bmatrix}, \tag{24}$$

with a residual smaller than 5 mm[12]. This is an extremely good fit given the characteristic length of the problem (several hundred kilometers). The conclusion to be drawn from this is clear: on this specific inter-satellite distance tracking example, an extra chameleonic acceleration cannot be distinguished from a small perturbation of the initial state vector. A brief analysis indicates that the parameter that has the biggest *weight* in Eq. (24) is δr_0. Now the question is how this small perturbation compares to the precision with which we have access to the initial state. As discussed previously in Sec. IV D 1, it turns out that the initial radial distance r_0 could in principle be determined with at most centimetric precision which is smaller than the 60 cm perturbation found in Eq. (24). Although this does not constitute a rigorous proof, this brief study tends to indicate that the 'unknown initial state' hypothesis can be ruled out.

4. *Perturbation of the mass distribution*

Nonetheless, the knowledge of initial conditions is not the only potential source of degeneracy. Indeed, the mass distribution inside the main body — the very source of gravity — is perhaps the most important degree of freedom to have knowledge of. In that perspective, can the fifth force effects on a satellite be interpreted in the framework of Newtonian gravity as a slightly altered density model? In order to answer that question, we continue in the same spirit as in Sec. IV D 3 by constructing an optimization problem. We saw earlier on, notably in Fig. 4, that the Newtonian potential of the mountain could very well be approximated by a point-mass. We can thus try and perturb the density model — and consequently the Newtonian potential — by adding a point-mass somewhere along the z-axis (see Fig. 2), as we do not wish to break the azimuthal symmetry. This simple model has only two parameters:

- m_* the mass of the point-mass;
- z_* the z coordinate of the point-mass.

[12] The last entry of vector $\delta\mathbf{X}_0^{\mathrm{opt}}$ in Eq. 24 corresponds to the perturbation in the initial angular momentum and may attract attention due to the fact it is orders of magnitude bigger than the other entries. To provide a benchmark, the unperturbed initial angular momentum is $L_0 \simeq 2.2 \times 10^{10}\,\mathrm{m^2/s}$.

	$\tilde{r} = 1.059$	$\tilde{r} = 1.111$
$\tilde{m}_*^{\tilde{r}}$	6.6×10^{-6}	1.8×10^{-6}
$\tilde{z}_*^{\tilde{r}}$	1.03	1.06
$f(m_*^{\tilde{r}}, z_*^{\tilde{r}})/f(0,1)$	3.1×10^{-2}	5.6×10^{-2}
$\|\partial_\theta(\Phi_* - \Psi)\|_{L^2}/\|\partial_\theta(\Phi_* + \Psi)\|_{L^2}$	0.14	0.22
$\|\partial_r(\Phi_* - \Psi)\|_{L^2}/\|\partial_r(\Phi_* + \Psi)\|_{L^2}$	0.07	0.08

TABLE VI. Best fit parameters of the approximation of the chameleon acceleration by a point-mass in Newtonian gravity.

The goal is then to find the pair (m_*, z_*) for which Newtonian gravity best mimics the modified gravity case. Precisely, our objective function is

$$\begin{aligned} f &: \mathbb{R}^2 \to \mathbb{R}_+ \\ (m_*, z_*) &\mapsto \int_0^\pi (\partial_\theta\Phi_* - \partial_\theta\Psi)^2 \,\mathrm{d}\theta \\ &\quad + \int_0^\pi (\partial_r\Phi_* - \partial_r\Psi)^2 \,\mathrm{d}\theta \end{aligned} \tag{25}$$

where Φ_* is the Newtonian potential created by the extra point-mass and the integral is carried out at fixed $\tilde{r}$. We denote by $(m_*^{\tilde{r}}, z_*^{\tilde{r}})$ the pair that minimizes the function f at radius $\tilde{r}$. Using $\|\cdot\|_{L^2}$ to denote the L^2-norm over the space of square-integrable function on $[0, \pi]$, one can rewrite

$$f(m_*, z_*) = \|\partial_\theta(\Phi_* - \Psi)\|_{L^2}^2 + \|\partial_r(\Phi_* - \Psi)\|_{L^2}^2 \,.$$

Basically, we aim at approximating both the radial and orthoradial parts of the chameleon acceleration at the same time. This optimization problem being low-dimensional, we can dispense with a sophisticated optimization algorithm and do a full exploration of the parameter space instead (see Fig. 14). Note that our point-mass model *cannot* reproduce the chameleon monopole (which is, in other words, a constant radial acceleration offset). Therefore, we removed it by hand before proceeding to the optimization phase. This offset is tiny: $\sim 1.4 \times 10^{-7}\,\mathrm{m/s^2}$ which corresponds to relative change of the mean density of the main body of only $\sim 9 \times 10^{-8}$. In comparison, let us mention that the Earth mass is known with a relative uncertainty of 10^{-4}.

The results for $\tilde{r} \in \{1.059, 1.111\}$ are reported in Table VI, where $\tilde{m}_*^{\tilde{r}} = m_*^{\tilde{r}}/M_{\mathrm{mountain}}$ and $\tilde{z}_*^{\tilde{r}} = z_*^{\tilde{r}}/R_{\mathrm{body}}$.

To further assert the quality of the fit in quantitative terms, we compute the ratio $f(m_*, z_*)/f(0,1)$ (where $\{m_* = 0, z_* = 0\}$ corresponds to flat profiles) as well as relative *errors* in L^2-norm. Several comments must be made:

- With only a simplistic model (i.e. a single point-mass has been added to the pre-existing model, contributing to the global Newtonian potential), we manage to approximate the fifth-force profile at a

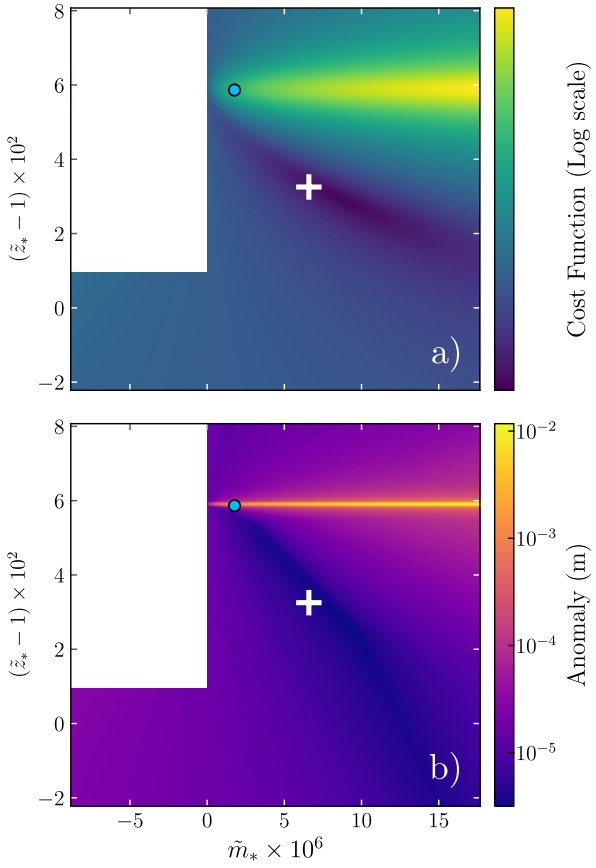


FIG. 14. a) Contour plot of the objective function (in log scale) in the (m_*, z_*)-plane[a]. b) Contour plot of the anomaly in the (m_*, z_*)-plane. The white cross and blue circle are located at the objective function's minimum when $\tilde{r} = 1.059$ and $\tilde{r} = 1.111$ respectively. The area left in plain white is not physically accessible as it corresponds to a 'negative extra mass' in vacuum. Warning: looking at the anomaly (bottom panel), one might expect the special case $m_* = 0$ to reduce to the case displayed on Fig. 13 and exhibit an anomaly of a few centimeters. The difference lies in the fact that here, the chameleon radial acceleration offset is artificially reproduced by slightly increasing the main body's mass.

[a] Note that negative mass (which would correspond to an *extrusion* for $\tilde{z}_* < 1 + h_m$) cannot represent the chameleon acceleration as well as positive mass.

given altitude with remarkable accuracy (see the various figures in Table VI).

– This approximation is good enough to *almost* reproduce the dynamics of the satellite's orbit over the mountain. In fact, we can repeat for instance the same exercise as we did in Fig. 13 and compute the so-called *anomaly*, i.e. the difference between the "modified gravity without extra mass" case and the "Newtonian gravity with extra mass" case. We find it to be no greater than 15 μm. This is more than a thousand times smaller than the anomaly computed in Fig. 13. This invites us to moderate the statements made earlier, since we are approaching here the precision limits of the GRACE-FO's LRI system.

– The objection could be made that the characteristics of the point-mass associated with the best fit do not correspond to any physical reality. Indeed, taking the second column of Table VI with entries $(\tilde{m}_* = 6.6 \times 10^{-6}, \tilde{z}_* = 1.03)$ suggests that there would be a $\sim 2 \times 10^{12}$ kg mass at an altitude of 186 km (or equivalently, 123 km above the mountain's top) — which is obviously absurd! Nevertheless, it can be seen on the top panel of Fig. 14 — which represents the cost function (25) in the (m_*, z_*)-plane — that lowering a bit z_* from the optimum (depicted by the white cross) while maintaining m_* constant has only a slight effect on the cost function. For this reason, the dynamics is not much affected by a shift of z_* towards the planet. As a matter of fact, setting $z_* = 1$ (i.e. bringing the extra mass at the planet's surface) leads to an anomaly bounded below 40 μm. The bottom panel of that same figure is intended to illustrate this phenomenon, and the strong correlation between the cost function and the anomaly is visible to the naked eye.

We can even place this extra mass at the same location as the point-mass Newtonian approximation of the mountain itself (see caption of Fig. 4) without any major change in the dynamics. This can therefore be interpreted as slightly increasing the mountain's density, by roughly 10^{-3} %. Such a slight deviation could equivalently be attributed to the fact that the gravitational constant G is only known with some certainty with four significant digits [48, 69].

These orders of magnitude on the density must be put into perspective with our current knowledge of the Earth inner density, with all the attendant uncertainties. Despite advancements in geophysical techniques, our knowledge of mass distribution is still imperfect, for simple reasons:

1. The planet's interior is out of reach. As a matter of fact, the deepest human-made hole ever dug is *only* 12.3 km deep (less than 0.2 % of the Earth radius).

2. The density is not uniform (even at fixed depth), which means extrapolation is not a valid procedure unless strong assumptions are made about the Earth's composition and structure.

3. As it happens, we also have to rely on indirect measurements, ranging from gravitational anomalies and magnetic anomalies [21, 32, 83] to seismic analysis [20, 27, 50]. All these techniques are in turn limited in both resolution and accuracy.

On this latter point, we stress that one should be careful when trying to put constraints on a given modified grav-

ity model, using a model of the Earth that comes from gravitational measurements in the first place. Indeed, the inversion of a gravity map into, say, a density map is model dependent (and unless contraindicated, would have been performed in a Newtonian framework). See Ref. [7], where this topic is discussed at length. In this regard, let us mention a recent work [42] that proposes to use the preliminary reference Earth model (PREM) [27], which is a radial seismic model, to constrain some alternative theories to GR.

5. *Breaking the degeneracy*

We have seen with two simple examples that drawing a distinction between a fifth force and model uncertainties (of different natures) is no easy task. These uncertainties spearhead degeneracies, which we partially address here.

In Sec. IV D 3, we provided the relevant orders of magnitude of the perturbation of a satellite initial state vector necessary to alone mimic a fifth force influence. The perturbation on the initial altitude was then put in comparison against the available level of precision for LEO satellites. It turned out POD techniques are good enough to the relatively large perturbation found in Eq. (24). Yet one must bear in mind that this was done on a very specific test-case and the conclusion may not generalize to others.

In Sec. IV D 4, we looked at how to distinguish a chameleon acceleration on top of Newtonian dynamics from a slight change of the density model in a purely Newtonian framework. We showed that it was possible to imperceptibly tweak the mass distribution in the mountain and in the planet to reproduce the chameleon acceleration profile *at a given altitude* (see Table VI). This naturally raises the question of whether such a fit works at different altitudes, or rather *how well*. Elements of response can be found in Table VI and Fig. 14. The first rows of Table VI bring out the fact that inferring the mountain's characteristics at two orbital radii leads to two clashing physical realities: strikingly, the extra mass inferred at $\tilde{r} = 1.059$ is almost 4 times greater than it appears at $\tilde{r} = 1.111$. In order to sharpen the analysis, we represent in Fig. 14 by a white cross and a blue circle the cost function's minimum at $\tilde{r} = 1.059$ and $\tilde{r} = 1.111$ respectively, while the contour plots are performed for $\tilde{r} = 1.059$. Following the notations introduced above, $(m_*^{1.059}, z_*^{1.059})$ and $(m_*^{1.111}, z_*^{1.111})$ are coordinates of the white cross and the blue circle respectively. Similarly, $f_{1.059}$ and $f_{1.111}$ refer to the cost functions at the two altitudes. At $(m_*^{1.111}, z_*^{1.111})$, we see that the cost function $f_{1.059}$ is much above its minimum and, in turn, corresponds to a large anomaly. Quantitatively speaking, we have

$$\frac{f_{1.059}(m_*^{1.111}, z_*^{1.111})}{f_{1.059}(m_*^{1.059}, z_*^{1.059})} \simeq 7.8 \times 10^2 \,,$$

which is a big ratio and reflects that $(m_*^{1.111}, z_*^{1.111})$ does not produce a good fit of the chameleon acceleration profile at $\tilde{r} = 1.059$. On the other hand, changing our perspective to $f_{1.111}$, we have

$$\frac{f_{1.111}(m_*^{1.059}, z_*^{1.059})}{f_{1.111}(m_*^{1.111}, z_*^{1.111})} \simeq 13 \,,$$

meaning that $(m_*^{1.059}, z_*^{1.059})$ is better *tolerated* by $f_{1.111}$ and $f_{1.059}$ than $(m_*^{1.111}, z_*^{1.111})$.

Fig. 15 provide more visual insights into these tensions. As in panel b) of Fig. 14, we computed the anomaly as a function of the pair (m_*, z_*), in a scenario where the two GRACE-FO-like satellites orbit at $\tilde{r} = 1.059$ (associated with blue colors) and in another scenario where they orbit at $\tilde{r} = 1.111$ (associated with orange colors). On panel a), we display two contours corresponding to anomaly thresholds of 3×10^{-6} m and 7×10^{-6} m, for both altitudes. The less the blue contours overlap with the orange ones, the greater the tension. Panels b) and c) complement the figure by representing the anomaly along the dotted lines visible on panel a) which pass through the minimal anomaly for each altitude. Ideally, the performances showcased by the GRACE-FO laser-link technology would allow for an exclusion of any (m_*, z_*) pair mapping to an anomaly greater than a micrometer, revealing the incompatibility between the two density models.

In conclusion of this section, the use of different altitudes in the analysis is a first step toward breaking the degeneracy. This idea was already put forward in Ref. [7] where the authors study the impact of a Yukawa potential on the spherical harmonic coefficients of the Earth. Precisely, the rescaled coefficients y_{lm} (introduced in Sec. II C 1) become dependent on the altitude meaning for instance that measurements of the J_2 zonal term at GOCE and GRACE altitudes could provide a test of the model[13]. On the whole, the difficulty lies in being able to find a set of several physical measures that would be in tension one with another when adding a fifth force to the play:

- The greater the tension, the tighter the potential constraints on the modified gravity model.
- The more measurements we have, the greater the likelihood of ending-up with a significant tension.

We tried to bring out such a tension with the GRACE-FO setup deployed at two different altitudes (see Table VI and Fig. 14). Nevertheless, this does not look practical in actual experiments. Even though we manage to create a small tension in the anomaly, one must bear in mind that our fitting model consisting of a single point-mass remains overly simplistic. More complex (and viable) density models may relieve this tension very well. In

[13] Note that this would also be true for the chameleon model as y_{lm}^C depends on the radial coordinate r.

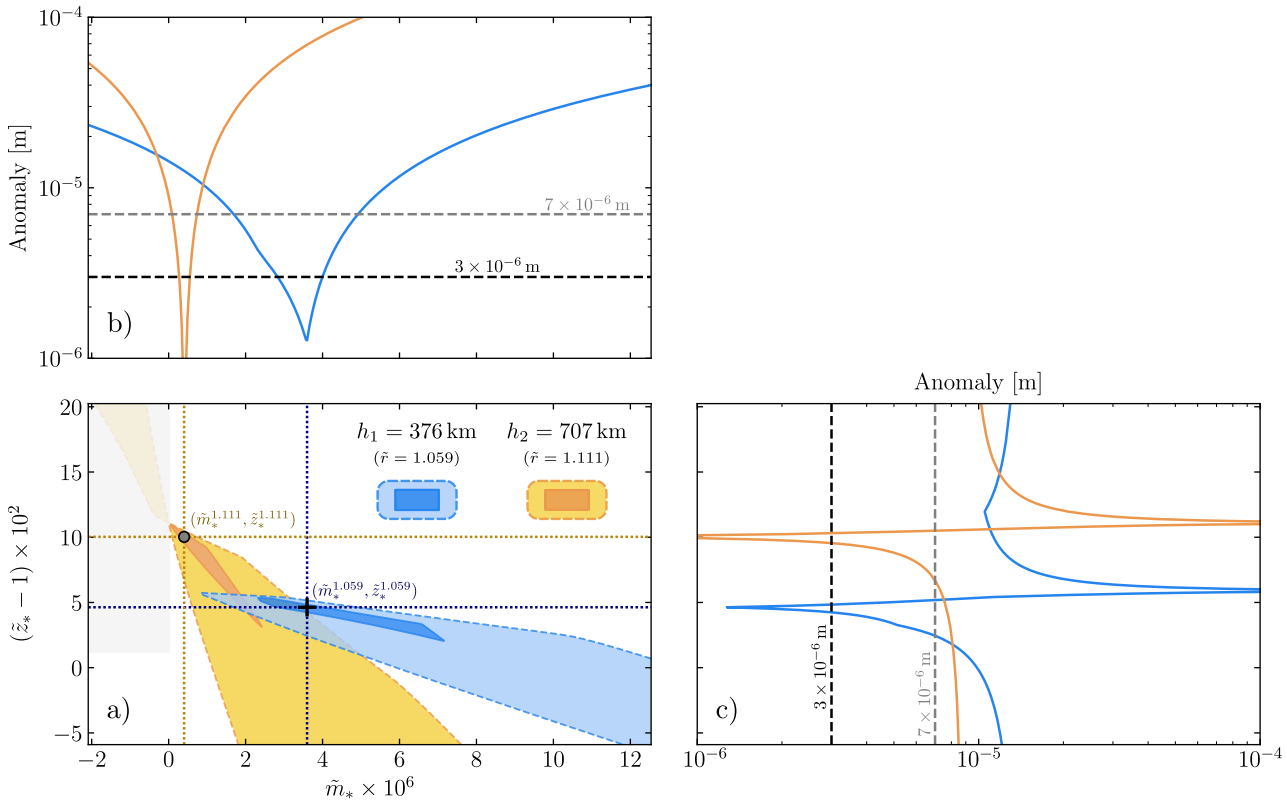


FIG. 15. Tensions in the inferred mountain's characteristics. The blue elements refer to the altitude $h_1 = 376\,\mathrm{km}$ ($\tilde{r} = 1.059$) and the orange elements refer to the altitude $h_2 = 707\,\mathrm{km}$ ($\tilde{r} = 1.111$). The anomaly observed in the inter-satellite distance at h_1 (resp. h_2) is best explained in the framework of Newtonian gravity by an extra point-mass with characteristics $(m_*^{1.059}, z_*^{1.059})$ (resp. $(m_*^{1.111}, z_*^{1.111})$) depicted by the black cross (resp. gray circle) in panel a)[a], which minimizes the anomaly down to $1.3 \times 10^{-6}\,\mathrm{m}$ (resp. $2.6 \times 10^{-7}\,\mathrm{m}$). In panel a), the darker contours map to an anomaly below $3 \times 10^{-6}\,\mathrm{m}$ while the lighter ones map to an anomaly below $7 \times 10^{-6}\,\mathrm{m}$. The gray shaded area is not physically accessible as it corresponds to a 'negative extra mass' in vacuum. Panel b) (resp. c)) represent the anomaly along $\tilde{m}_*$ at $\tilde{z}_*^{\tilde{r}}$ (resp. $\tilde{z}_*$ at $\tilde{m}_*^{\tilde{r}}$). The anomaly peaks (maxima) visible on panel c) correspond to the horizontal feature seen previously in Fig. 14-b) around $\tilde{z}_* = 1.06$.

[a] Note that the pairs $(m_*^{\tilde{r}}, z_*^{\tilde{r}})$ are different from the ones reported in Table VI for two reasons: (i) here, we minimize the anomaly in the inter-satellite distance between the "modified gravity without extra mass" case and the "Newtonian gravity with extra mass" case, which is different from minimizing the objective function f given by Eq. (25); and (ii) we allowed ourselves to modify the initial inter-satellite distance for the two considered altitudes in order to better showcase the tension.

short, the ultimate goal of constraining the chameleon model with space-based geodesy is impeded by:

- the uncertainty on the source of gravity, that is the density model;
- the uncertainty on the measurements themselves;
- all the other forces acting on a satellite, ranging from atmospheric drag and solar radiation pressure, to third-body perturbation, which also come with error bars in our models. Note that these perturbing forces are also a nuisance for geodesy, hence the use of accelerometers on board satellites [22, 70].

In Appendix G, we further look at orbital periods at two different altitudes and compute their difference, in a Newtonian framework and in a modified gravity framework.

V. Discussion and conclusion

Effect of a mountain

This article investigated the testability of chameleon gravity by space geodesy experiments, with a focus on the influence of the local landform and the atmosphere. The motivations were twofold. First, viable regions of the chameleon parameter space all map to a screened Earth, that is only a thin-shell contributes to the fifth force. Therefore, it seemed important to study departures from spherical symmetry, hereby embodied by a mountain. Second, while published works sometimes account for the atmosphere in their study, the models implemented are simplistic (often one layer of constant density) and the determination of whether it has a thin-shell is based on

rather qualitative arguments. Addressing such questions is not possible by means of analytical techniques alone due to the complexity of the physical models we wished to study, and to the nonlinear nature of the chameleon equation of motion. We thus resorted to numerical simulations — performed with the code *femtoscope* — to conduct this work.

We obtained the chameleon contribution to the total gravitational potential of a mountainous planet, scanning through an extended region of the parameter space. As already pointed out in Ref. [47], the unscreened regime shares similarities with pure Newtonian gravity in that, in both cases, the fields are sourced by the entire mass of the main body. Consequently, the chameleon potential in the unscreened regime is roughly the same as the Newtonian potential up to an affine transform (and the same goes for the accelerations). As we enter the screened regime however, the multipole expansion of the chameleon field starts to depart from that of the Newtonian potential, revealing a distinct signature. In terms of acceleration, the chameleon acceleration vector is a bit more directed towards the mountain compared to the Newtonian acceleration. Their norm ratio remains small though, bounded from above by $\sim 10^{-6}$ at the equivalent of LEO altitudes in the atmosphere-free case[14].

Effect of an atmosphere

Based on our study of three distinct atmospheric density profiles, we found that the addition of an extra layer of air surrounding the main body can mitigate the effect of the fifth force. We showed that there exists a threshold on the value of the parameter α. Above this threshold, the atmosphere acts as an attenuator, effectively reducing the chameleon acceleration by a certain amount compared to the case without atmosphere. Below this cutoff, the effect of the atmosphere is more drastic: any non-radial dependence of the scalar field vanishes — the mountain is plainly *invisible*. For even smaller values of α, the atmosphere itself becomes screened, and the chameleon field is thereupon fully determined by the atmospheric density profile. As we saw, it is even possible in this case to enhance the radial fifth force at given altitude with respect to the atmosphere-free case. This study represents a step forward with respect to previous work discussing the influence of the atmosphere. Moreover, this clearly gives the edge to bodies devoid of atmosphere when it comes to select a Solar System site for testing this screened scalar-tensor model.

Space geodesy experiments

Our knowledge of the geopotential comes to a large extent from spaceborne geodesy. From this standpoint, we thus investigated whether constraints could be put on modified gravity models using satellites in orbit. For that purpose, we performed orbit propagations, with and without the putative fifth force, and studied the resulting anomaly[15] on several observables (such as the variations of the distance between two satellites following each other as in the GRACE and GRACE-FO setups). While the anomalies we find are technically well within the detection range of current on-board and ground-based space-technology, we showed that uncertainties in the model for the distribution of matter are large enough to allow for degeneracies.

We laid emphasis on the fact that one way to distinguish a chameleon acceleration from a slight change of the Earth density model in a purely Newtonian framework is to rely on experiments performed at (at least) two different altitudes. Indeed, in the regime where the Earth (or any other planetary body) is screened, the chameleon acceleration does not decrease as r^{-2} like the Newtonian acceleration (this is particularly stressed in Refs. [7, 66]). If the chameleon field *actually exists*, then inferring density models under the assumption of Newtonian gravity at several altitudes should result in tensions between those models. Of course, these tensions should be accounted for in a probabilistic way, which is beyond the scope of this article. Conversely, if it were not for all the other perturbing forces that greatly complexify the model, this method could be used to put constraints on the chameleon model, and more generally on massive scalar-tensor theories.

Back-reaction of a satellite on the scalar field

We also took into account the back-reaction of an object as small as a satellite in orbit on the scalar field. For the first time, we went beyond the various approximations found in the literature and computed the *full solution* of the {Earth + Satellite} system using *femtoscope*. This involves taking advantage of the h-adaptivity technique granted by FEM. We could then compute the overall fifth force acting on an object with a simple geometry and characteristics close to that of a real spacecraft (length-scale and density). Surprisingly, as long as

[14] It is insightful to compare this ratio with the ratio of the Solar radiation pressure over the Newtonian acceleration, which is around 10^{-8} [51]. Despite being so tiny, the Solar radiation pressure perturbing acceleration, when integrated over many orbits, is enough to cause significant drifts of orbital elements [54]. What makes the chameleon acceleration difficult to distinguish from the Newtonian acceleration is the fact that they are both sourced by the same body.

[15] The term 'anomaly' is used to refer to the difference for a given observable between the {Newtonian gravity} case and the {Newtonian gravity + fifth force} case.

the satellite is not screened and despite the background field being disturbed, the global chameleon acceleration undergone by the satellite is the same as the one a point-particle (not disturbing the background field) would experience. We provide mathematical insights into why this is the case in Appendix E. In the screened regime however, the net fifth force vanishes to zero. The transition between those two regimes occurs over a narrow band in the chameleon parameter space.

Outlook

While we focused on fifth force searches, other venues exist to test scalar-tensor theories. For instance, in any such theory involving a conformal coupling of the scalar field ϕ to matter fields in the Einstein frame, the gravitational redshift effect has a ϕ-dependence (see e.g. Ref. [40] for the chameleon's contribution to this effect). We will take a deeper look at this effect in an upcoming article. Most importantly, we will tackle the question of what is actually *measurable*, and with which precision. As a complement to fifth force searches where we look for dynamical effects whose amplitude depends inherently of the field's gradient, gravitational redshift (or equivalently, gravitational time-dilation) can be measured in a static configuration and is sensitive to the field's strength. Whether clocks are put into orbit (as envisaged in the ACES mission [65]) or left on Earth, they have become so precise[16] that their constraining power (i.e. the possibility to use this technology to rule out modified gravity models) has to be quantified. The bound given in Ref. [40] has to be revisited, given two decades elapsed since the writing of this article.

Acknowledgments

HL thanks Pablo Richard for fruitful discussions about numerical computations. We acknowledge the financial support of CNES through the APR program ("MICROSCOPE" project).

A. Conversion of cosine and sine coefficients to bare coefficients

Let $f\colon \mathcal{S}^2 \to \mathbb{R}$ be a real-valued function on the unit sphere and $L \in \mathbb{N}^*$ a maximum spherical harmonic degree. The truncated spherical harmonic expansion, which defines an approximation $f_{\rm trunc} \simeq f$, may be written as

$$f_{\rm trunc}(\mathbf{n}) = \sum_{m=0}^{L}\sum_{l=m}^{L}\Big[C_{lm}\bar{P}_{lm}(\cos\theta)\cos(m\varphi) + S_{lm}\bar{P}_{lm}(\cos\theta)\sin(m\varphi)\Big] \tag{A1}$$

with $\mathbf{n} = (\theta, \varphi)$ and $\bar{P}_{lm}$ the normalized associated Legendre functions[17] which relate to their unnormalized counterparts P_{lm} via

$$\bar{P}_{lm}(x) = \sqrt{\frac{(2-\delta_{m0})(2l+1)}{4\pi}\frac{(l-m)!}{(l+m)!}}P_{lm}(x)\,. \tag{A2}$$

We want to convert the cosine and sine coefficients (C_{lm}, S_{lm}) into the bare coefficients f_{lm} that appear in the usual expansion

$$f_{\rm trunc}(\mathbf{n}) = \sum_{l=0}^{L}\sum_{m=-l}^{+l} f_{lm}Y_{lm}(\mathbf{n})\,. \tag{A3}$$

In order to express (C_{lm}, S_{lm}) as a function of f_{lm}, we start from Eq. (A3) and interchange the order of summations over l and m to obtain

[16] High-precision clocks, such as optical lattice clocks, currently achieve astonishing levels of accuracy with a fractional frequency uncertainty of approximately 10^{-19} to 10^{-20} [9, 15].

[17] In this respect, the definition of *normalized associated Legendre functions* is consistent with the definition of orthonormalized spherical harmonic functions. See Table 1 from Ref. [77].

27

$$
\begin{aligned}
f_{\rm trunc}(\mathbf{n}) &= \sum_{l=0}^{L}\sum_{m=-l}^{+l} f_{lm} Y_{lm}(\mathbf{n}) \\
&= \sum_{l=0}^{L}\left[\sum_{m=0}^{+l} f_{lm}\bar{P}_{lm}(\cos\theta)\cos(m\varphi) + \sum_{m=-l}^{-l} f_{lm}\bar{P}_{l|m|}(\cos\theta)\sin(|m|\varphi)\right] \\
&= \sum_{l=0}^{L}\left[\sum_{m=0}^{+l} f_{lm}\bar{P}_{lm}(\cos\theta)\cos(m\varphi) + \sum_{m'=1}^{+l} f_{l,-m'}\bar{P}_{lm'}(\cos\theta)\sin(m'\varphi)\right] \\
&= \sum_{l=0}^{L}\left[\sum_{m=0}^{+l} C_{lm}\bar{P}_{lm}(\cos\theta)\cos(m\varphi) + \sum_{m=0}^{+l} S_{lm}\bar{P}_{lm}(\cos\theta)\sin(m\varphi)\right] \\
&= \sum_{m=0}^{L}\sum_{l=m}^{L}\left[C_{lm}\bar{P}_{lm}(\cos\theta)\cos(m\varphi) + S_{lm}\bar{P}_{lm}(\cos\theta)\sin(m\varphi)\right] .
\end{aligned}
\tag{A4}
$$

The above computation is consistent if we set

$$
\begin{aligned}
C_{lm} &= f_{lm} \quad \text{if } m \geq 0\,, \\
S_{lm} &= \begin{cases} 0 & \text{if } m = 0 \\ f_{l,-m} & \text{if } m \geq 1 \end{cases} .
\end{aligned}
\tag{A5}
$$

B. Verification of the scaling relation for the spherical harmonic coefficients of the Newtonian potential

The Newtonian potential defined by Eq. (1) is special in that its bare spherical harmonic coefficients $\Phi_{lm}(r)$ can be rescaled according to Eq. (13) which yields altitude-independent coefficients. We denote these rescaled coefficients y^N_{lm}. This peculiar property can be used as an additional means of test ascertaining the quality of our numerical approximations. Indeed, from our numerical $\Phi(r,\theta)$ maps of the Newtonian potential, we can compute the rescaled coefficients y^N_{l0} at several altitudes and check whether or not they actually depend on the altitude. Fig. 16 shows the result of this process for $\tilde{r} \in \{1.059, 1.111, 1.314\}$ and $l \in \{1, \cdots, 10\}$. At first sight, the scaling relation seems consistent with the numerical data at low degree. It is however more difficult to verify it at higher altitude and for higher degrees as the rescaling process involves multiplying the bare coefficients by $\tilde{r}^{l+1}$ which quickly blows up to infinity. The bare coefficients being themselves plagued with numerical errors — they are derived from spherical harmonic decomposition algorithm on top of FEM computations — we clearly do not expect this relation to perfectly hold in this regime.

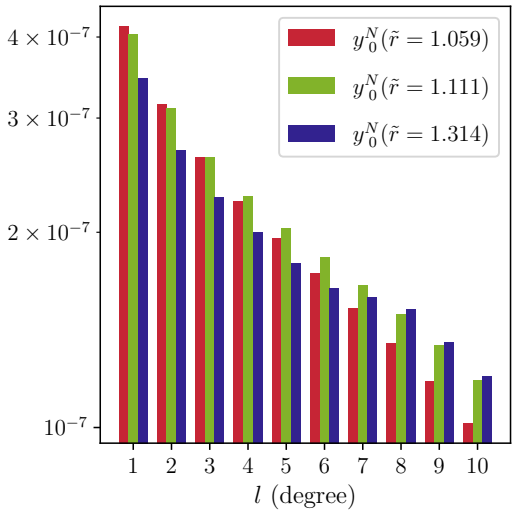


FIG. 16. Verification of the scaling relation between bare spherical harmonic coefficients $\Phi_{l0}(\tilde{r})$ and dimensionless coefficients $y^N_{l0}(\tilde{r})$ obtained numerically. The rescaled coefficients should in principle be independent of the altitude at which they are computed.

C. Additional checks on 2D numerical computations

In this Appendix, we present two additional checks that were performed on all FEM computations of the chameleon field done in this article. We mainly elaborate on the ideas introduced in Sec. II D 2.

1. Check of the radial evolution of the chameleon field

The {Earth + mountain} system constitutes a small departure from spherical symmetry. Therefore, the behavior of the chameleon field in the outgoing radial direction should be close to that of the {Earth} system, which in turn is spherically symmetrical, and so purely radial. From a numerical viewpoint, such radial profiles are much easier to obtain than a less symmetrical case. Indeed in the former case, the Klein-Gordon equation (7) boils down to a simple ODE

$$\alpha \frac{\mathrm{d}}{\mathrm{d}\tilde{r}} \left(\tilde{r}^2 \frac{\mathrm{d}\tilde{\phi}}{\mathrm{d}\tilde{r}} \right) = \tilde{r}^2 \tilde{\rho} - \tilde{r}^2 \tilde{\phi}^{-(n+1)} \,, \tag{C1}$$

where numerical resources (density of DOFs, order of the finite elements) can be increased without blowing up the time complexity of the algorithm. As a result, we can obtain benchmark solutions at relatively low cost, for all the cases discussed in this study treated as purely radial (i.e. without mountain, all other physical parameters being equal). We denote $\tilde{\phi}_{1\mathrm{D}}(\tilde{r})$ such benchmarks, and $\tilde{\phi}_{2\mathrm{D}}(\tilde{r},\theta)$ the 2D field profiles presented throughout the article. We then implement the following metric:

$$\text{for } \tilde{r} > 1 + h_m, \; \Gamma_{\tilde{r}} = \frac{\min\limits_{\theta\in[0,\pi]} \left| \tilde{\phi}_{1\mathrm{D}}(\tilde{r}) - \tilde{\phi}_{2\mathrm{D}}(\tilde{r},\theta) \right|}{\left| \tilde{\phi}_{1\mathrm{D}}(\tilde{r}) \right|} \,. \tag{C2}$$

Note that this metric has the advantage of being relative, as opposed to the absolute criteria discussed in Sec. II D 2.

Tab. VII includes $\Gamma_{\tilde{r}}$ for radial coordinates $\tilde{r} \in \{1.059, 1.111, 1.314, 4.645, 6.617\}$ and for all (α, atmospheric scenario) pairs considered in this work. Although there is no physical motivation for having $\Gamma_{\tilde{r}} \equiv 0$ systematically, the fact that it remains below one part in a thousand in the vast majority of cases reflects the good agreement between the radial benchmark and $\tilde{\phi}_{2\mathrm{D}}$. By way of comparison, applying the same metric on the Newtonian potential yields $\Gamma_{\tilde{r}} \sim 10^{-7}$. Evaluating this metric at different altitudes is also a way to make sure that none of the 2D solutions behaves unexpectedly in the radial direction.

2. Check of the strong residual amplitude with respect to each term

As mentioned in Sec. II D 2, the strong residual alone does not provide much insight into how good the numerical approximation is at the end of Newton's iterations. However, it is meaningful to compare locally the size of the residual against the size of each term in it, namely

$$\left| \alpha \tilde{\Delta} \tilde{\phi} \right| \quad , \quad |\tilde{\rho}| \quad , \quad \left| \tilde{\phi}^{-(n+1)} \right| \,. \tag{C3}$$

A numerical approximation deemed acceptable must be such that the residual should be at least a few orders of magnitude smaller than the dominant term in (C3).

This criterion was assessed for five specific values of $\tilde{r} \in \{1.059, 1.111, 1.314, 4.645, 6.617\}$ on all numerical approximations discussed in this study. In Fig. 17, we show several examples of scatter plots that allowed us to do this monitoring. Each sub-panel corresponds to a given altitude and a given (α, atmospheric scenario) pair — both randomly chosen —, and depicts the absolute value of the residual (black dots) *vs* terms appearing in (C3) (pastel-colored squares) as functions of θ. We see that the residual remains well-below the dominant term in absolute values.

D. Spherical harmonic coefficients at different altitudes

For the sake of comprehensiveness, we provide histograms of the spherical harmonic coefficients of both the Newtonian potential Φ and the chameleon potential Ψ (up to degree $l = 200$) at three altitudes in Fig. 18 (see Sec. III A 1 in the main text). The specific shapes of both potential decompositions hold at all three altitudes, although they get squashed toward lower degrees the higher we go. The oscillations that we observe at high degrees for $\tilde{r} \in \{1.111, 1.314\}$ are not deemed physical but can be rather attributed to numerical noise.

E. Further insights into the fifth force experienced by an unscreened satellite

In Sec. IV A, we computed the total chameleon acceleration undergone by a satellite in orbit. We found that, as long as the satellite was not screened, the resulting force (computed numerically by integrating the gradient of the scalar field over the whole volume occupied by the satellite) was equal to that acting on an equal mass point-like particle. In other words, the back-reaction of the satellite on the scalar field, in the unscreened regime, is such that there is no *self-force* perturbing the Jordan frame geodesics. In this Appendix, we provide an explanation for this phenomenon observed through numerical simulations, based on several approximations that can be justified in the {Earth + Satellite} system.

1. The case of Newtonian gravity

Given two massive bodies labeled by the subscripts $i \in \{1, 2\}$, the total gravitational force acting on the second body is

$$\mathbf{F}_2 = -\int_{V_2} \boldsymbol{\nabla} \Phi_N(\mathbf{x}) \, \mathrm{d}m(\mathbf{x}) \,.$$

	α	$\tilde{r} = 1.059$	$\tilde{r} = 1.111$	$\tilde{r} = 1.314$	$\tilde{r} = 4.645$	$\tilde{r} = 6.617$
No-Atmosphere	10^{-4}	0	0	0	8.7×10^{-7}	8.0×10^{-7}
	10^{-5}	0	0	0	7.9×10^{-7}	7.4×10^{-7}
	10^{-6}	2.6×10^{-3}	1.4×10^{-3}	5.0×10^{-4}	4.2×10^{-5}	2.7×10^{-5}
	10^{-10}	0	0	0	2.5×10^{-6}	1.7×10^{-6}
	10^{-11}	6.9×10^{-6}	1.7×10^{-6}	1.7×10^{-6}	1.7×10^{-6}	1.7×10^{-6}
	10^{-12}	6.9×10^{-6}	6.9×10^{-6}	6.9×10^{-6}	6.8×10^{-6}	6.9×10^{-6}
	10^{-14}	3.9×10^{-5}	3.9×10^{-5}	3.9×10^{-5}	3.8×10^{-5}	3.8×10^{-5}
	10^{-15}	1.5×10^{-5}	1.4×10^{-5}	1.4×10^{-5}	1.3×10^{-5}	1.3×10^{-5}
	10^{-16}	1.3×10^{-5}	7.5×10^{-6}	3.3×10^{-6}	9.7×10^{-7}	9.0×10^{-7}
	10^{-18}	1.7×10^{-4}	1.9×10^{-4}	2.1×10^{-4}	2.0×10^{-4}	2.0×10^{-4}
	10^{-20}	0	0	1.1×10^{-4}	9.0×10^{-6}	5.1×10^{-6}
	10^{-21}	0	0	5.1×10^{-5}	2.9×10^{-6}	1.3×10^{-6}
	10^{-23}	0	0	1.7×10^{-5}	5.7×10^{-9}	4.5×10^{-9}
	10^{-24}	0	0	1.1×10^{-6}	4.8×10^{-10}	4.3×10^{-10}
	10^{-25}	0	0	0	4.5×10^{-11}	4.9×10^{-11}
	10^{-26}	0	0	1.1×10^{-12}	4.7×10^{-12}	4.7×10^{-12}
	10^{-27}	0	0	0	4.0×10^{-13}	5.8×10^{-13}
	10^{-28}	5.7×10^{-14}	0	0	2.8×10^{-14}	4.3×10^{-14}
Tenuous	10^{-6}	4.5×10^{-6}	4.3×10^{-6}	4.1×10^{-6}	4.0×10^{-6}	3.6×10^{-6}
	10^{-10}	4.7×10^{-6}	4.7×10^{-6}	4.7×10^{-6}	4.8×10^{-6}	4.4×10^{-6}
	10^{-11}	0	0	0	6.5×10^{-6}	6.2×10^{-6}
	10^{-12}	1.7×10^{-3}	9.4×10^{-4}	3.4×10^{-4}	4.0×10^{-5}	3.0×10^{-5}
	10^{-14}	1.0×10^{-3}	5.7×10^{-4}	2.4×10^{-4}	6.0×10^{-5}	5.4×10^{-5}
	10^{-15}	2.4×10^{-4}	1.5×10^{-4}	7.4×10^{-5}	2.6×10^{-5}	2.3×10^{-5}
	10^{-17}	1.9×10^{-5}	1.3×10^{-5}	8.4×10^{-6}	5.3×10^{-6}	4.7×10^{-6}
	10^{-20}	1.9×10^{-6}	2.0×10^{-6}	2.1×10^{-6}	2.5×10^{-6}	2.3×10^{-6}
Earth-like	10^{-5}	7.5×10^{-6}	6.9×10^{-6}	5.7×10^{-6}	4.3×10^{-6}	3.9×10^{-6}
	10^{-6}	4.5×10^{-6}	4.3×10^{-6}	4.1×10^{-6}	4.0×10^{-6}	3.6×10^{-6}
	10^{-8}	2.6×10^{-3}	1.4×10^{-3}	5.0×10^{-4}	4.7×10^{-5}	3.1×10^{-5}
	10^{-10}	3.8×10^{-7}	2.8×10^{-6}	4.3×10^{-6}	4.7×10^{-6}	4.4×10^{-6}
	10^{-12}	1.1×10^{-2}	6.2×10^{-3}	2.2×10^{-3}	2.2×10^{-4}	1.5×10^{-5}
	10^{-14}	2.5×10^{-3}	1.4×10^{-3}	5.5×10^{-4}	8.8×10^{-5}	7.2×10^{-5}
	10^{-15}	4.9×10^{-5}	3.6×10^{-5}	2.5×10^{-5}	1.9×10^{-5}	1.9×10^{-5}
	10^{-17}	1.1×10^{-5}	8.3×10^{-6}	6.7×10^{-6}	5.7×10^{-6}	5.3×10^{-6}
	10^{-20}	0	9.7×10^{-7}	2.1×10^{-6}	2.4×10^{-6}	2.2×10^{-6}
	10^{-23}	0	7.7×10^{-5}	1.7×10^{-5}	7.3×10^{-4}	1.4×10^{-3}
Dense	10^{-6}	4.3×10^{-6}	4.3×10^{-6}	4.1×10^{-6}	4.0×10^{-6}	3.6×10^{-6}
	10^{-8}	1.9×10^{-5}	1.1×10^{-5}	5.9×10^{-6}	4.1×10^{-6}	3.7×10^{-6}
	10^{-10}	0	0	4.8×10^{-2}	3.2×10^{-3}	2.1×10^{-3}
	10^{-15}	0	0	0	2.8×10^{-5}	2.7×10^{-5}
	10^{-18}	0	0	0	3.3×10^{-6}	2.9×10^{-6}
	10^{-20}	0	0	0	1.7×10^{-6}	1.6×10^{-6}
	10^{-21}	0	0	0	5.3×10^{-7}	6.3×10^{-7}
	10^{-25}	0	0	0	0	0

TABLE VII. Metric $\Gamma_{\tilde{r}}$ for $\tilde{r} \in \{1.059, 1.111, 1.314, 4.645, 6.617\}$ and for all (α, atmospheric scenario) pairs considered in this work.

30

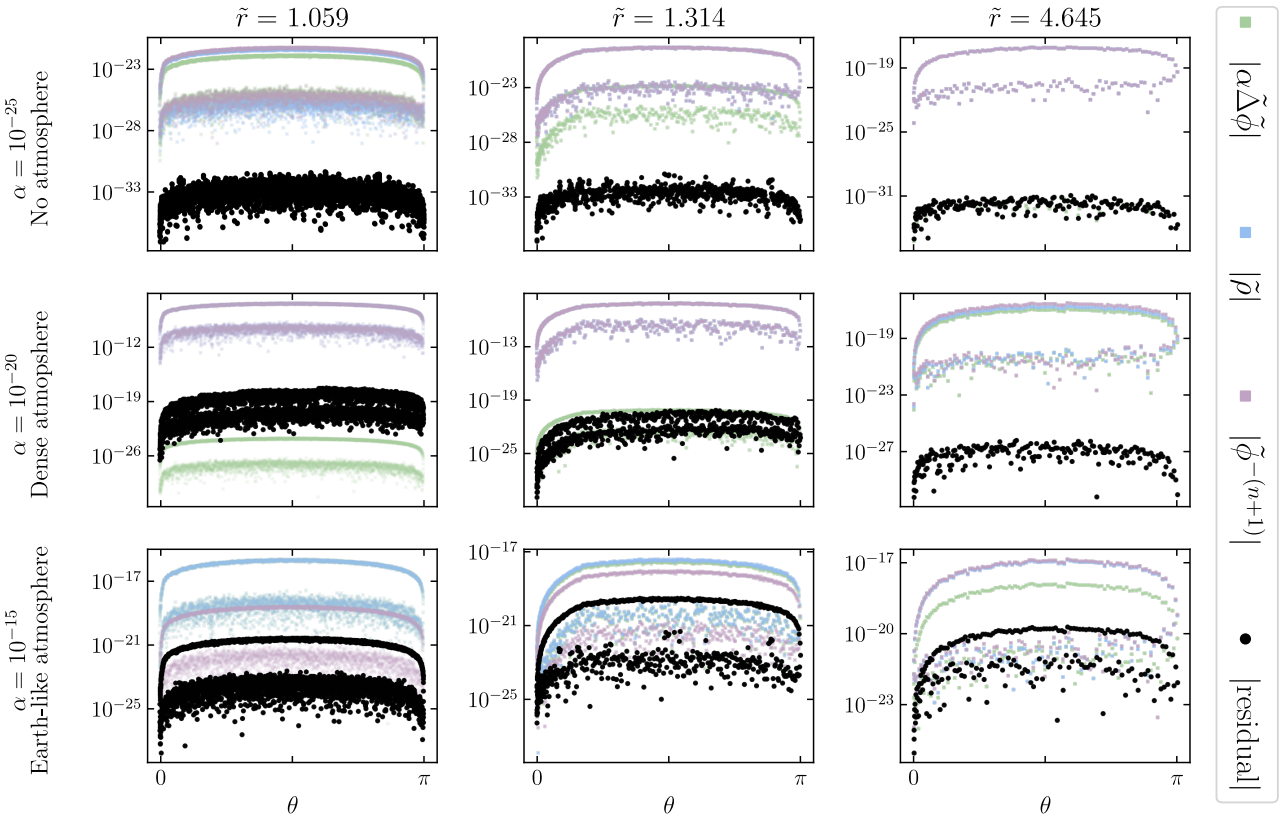


FIG. 17. Representation of the strong residual (black circles) and the various terms of the dimensionless Klein-Gordon equation (7) (pastel-colored squares) in absolute values as a function of $\theta \in [0, \pi]$. Each column corresponds to a given radial coordinate $\tilde{r} \in \{1.059, 1.314, 4.645\}$ whereas each row corresponds to a given pair (α, atmospheric scenario). In all cases, the absolute value of the strong residual remains at least several orders of magnitude below the dominant term of the Klein-Gordon equation, which is in line with the criterion set out in Sec. II D 2. The splitting of the curves associated with each term is due to the fact that we use second-order finite elements.

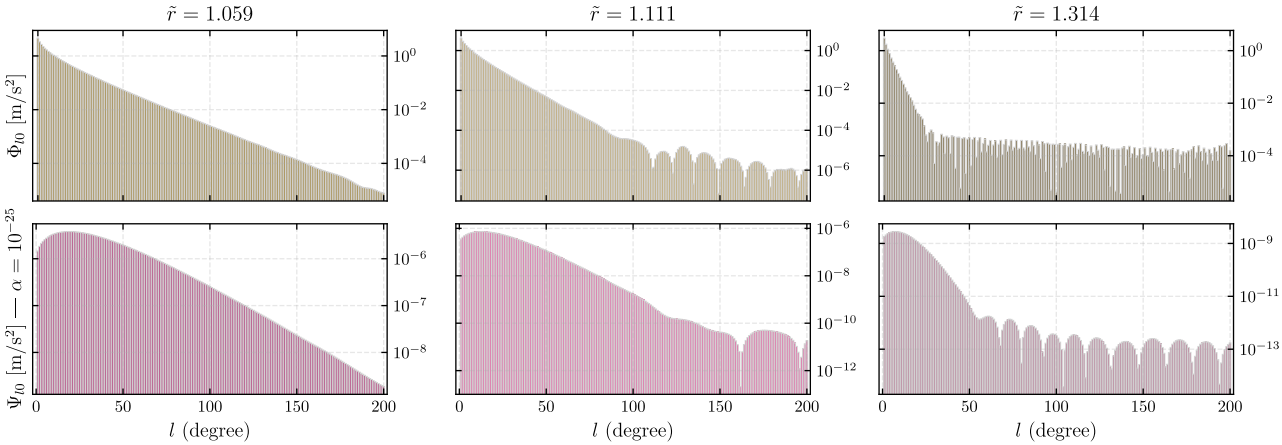


FIG. 18. Spherical harmonic coefficients of the Newtonian potential (top row) and of the chameleon potential for $\alpha = 10^{-25}$ (bottom row). The spectra are computed at three different altitudes, namely $\tilde{r} \in \{1.059, 1.111, 1.314\}$.

In the above expression, V_2 is the volume occupied by the body 2 and Φ_N is the total Newtonian potential created by the two bodies. Thanks to the linearity of the Poisson equation governing the Newtonian potential, one can apply the superposition principle $\Phi_N = \Phi_1 + \Phi_2$, where Φ_i is the potential sourced by the body i alone. The $\boldsymbol{\nabla}$ operator and the integral being linear, we get

$$\mathbf{F}_2 = -\int_{V_2} \boldsymbol{\nabla}\Phi_1(\mathbf{x})\,\mathrm{d}m(\mathbf{x}) - \int_{V_2} \boldsymbol{\nabla}\Phi_2(\mathbf{x})\,\mathrm{d}m(\mathbf{x})\,.$$

Physically speaking, the first integral represents the force exerted by 1 on 2 while the second integral is the force exerted by 2 on 2, which must be zero according to Newton's third law. This can be mathematically proven fairly easily given that

$$\Phi_2(\mathbf{x}) = -G\int_{V_2} \frac{\mathrm{d}m(\mathbf{x}')}{\|\mathbf{x}-\mathbf{x}'\|}$$

$$\text{and } \boldsymbol{\nabla}\left(\|\mathbf{x}-\mathbf{x}'\|^{-1}\right) = -\frac{\mathbf{x}-\mathbf{x}'}{\|\mathbf{x}-\mathbf{x}'\|^3}\,.$$

We thus get

$$\begin{aligned}\int_{V_2} \boldsymbol{\nabla}\Phi_2(\mathbf{x})\,\mathrm{d}m(\mathbf{x}) &= G\int_{V_2}\left(\int_{V_2}\frac{\mathbf{x}-\mathbf{x}'}{\|\mathbf{x}-\mathbf{x}'\|^3}\,\mathrm{d}m(\mathbf{x}')\right)\,\mathrm{d}m(\mathbf{x})\\ &= \frac{G}{2}\int_{V_2}\int_{V_2}\frac{\mathbf{x}-\mathbf{x}'}{\|\mathbf{x}-\mathbf{x}'\|^3}\,\mathrm{d}m(\mathbf{x}')\,\mathrm{d}m(\mathbf{x})\\ &\quad - \frac{G}{2}\int_{V_2}\int_{V_2}\frac{\mathbf{x}'-\mathbf{x}}{\|\mathbf{x}'-\mathbf{x}\|^3}\,\mathrm{d}m(\mathbf{x})\,\mathrm{d}m(\mathbf{x}')\\ &= 0\end{aligned}$$

In conclusion, despite disturbing the overall Newtonian potential, the body 2 experiences the force sourced by the body 1 only. Furthermore, if the body 2 is small enough that $\boldsymbol{\nabla}\Phi_1$ is approximately constant over V_2, we recover the point-mass approximation, i.e. $\mathbf{F}_2 \simeq -m_2\boldsymbol{\nabla}\Phi_1(\mathbf{x}_2)$.

2. The case of the chameleon field in the unscreened regime

The above demonstration relies mainly on the superposition principle, which is lost in the case of the chameleon field because of the nonlinear nature of the Klein-Gordon equation governing the scalar field Eq. (4). Nonetheless, let $\phi_{\rm tot} \coloneqq \phi_\oplus + \delta\phi$ be the chameleon field of the {Earth + Satellite} system, where $\phi_\oplus$ is the background field of the Earth alone. Working with the dimensionless version of the Klein-Gordon equation (7), we have by definition

$$\begin{cases}\alpha\Delta\phi_{\rm tot} &= (\rho_\oplus + \rho_{\rm Sat} + \rho_{\rm vac})\,(\mathbf{x}) - \phi_{\rm tot}^{-(n+1)}\\ \alpha\Delta\phi_\oplus &= (\rho_\oplus + \rho_{\rm vac})\,(\mathbf{x}) - \phi_\oplus^{-(n+1)}\end{cases}.$$

In the unscreened case, $\delta\phi$ can indeed represent a small perturbation with respect to the background field $\phi_\oplus$ — see e.g. the case illustrated in Fig. 10. Then, the nonlinear term can be approximated as

$$\phi_{\rm tot}^{-(n+1)} \simeq \phi_\oplus^{-(n+1)} - (n+1)\phi_\oplus^{-(n+2)}\delta\phi\,,$$

so that

$$\alpha\Delta\delta\phi \simeq \rho_{\rm Sat}(\mathbf{x}) + (n+1)\phi_\oplus^{-(n+1)}\frac{\delta\phi}{\phi_\oplus}\,. \tag{E1}$$

The r.h.s. of Eq. (E1) can be further simplified if we assume that, at the satellite's altitude, $\phi_\oplus$ is close to its asymptotic value in vacuum, that is $\phi_\oplus(\mathbf{x}_{\rm Sat}) \sim \rho^{-1/(n+1)}$. Then we have, depending on whether $\mathbf{x} \in V_{\rm Sat}$,

- Inside the satellite: $\rho_{\rm Sat}(\mathbf{x}) \neq 0$ and so

$$\frac{\phi_\oplus^{-(n+2)}\delta\phi}{\rho_{\rm Sat}(\mathbf{x})} \sim \frac{\delta\phi}{\phi_\oplus}\frac{\rho_{\rm vac}}{\rho_{\rm Sat}} \ll 1\,.$$

  Consequently, Eq. (E1) can be legitimately approximated by a Poisson equation inside the satellite.

- Outside the satellite: $\rho_{\rm Sat}(\mathbf{x}) = 0$. We still have $\delta\phi/\phi_\oplus^{n+2} \ll 1$ and $\delta\phi \to 0$ as one moves away from the satellite (while $\phi_\oplus^{-(n+2)}$ remains bounded) so that we essentially recover a Laplace equation.

In brief, we showed that, under some assumptions, $\delta\phi$ obeys a Poisson equation inside the satellite, and a Laplace equation outside the satellite. The Newtonian potential sourced by the satellite (denoted by Φ_2 in the previous discussion) is governed by the same partial differential equation. Yet, *same equations have the same solutions*, which means that $\delta\phi$ has a role similar to Φ_2. Therefore, following the demonstration made in the case of the Newtonian potential above, we get

$$\begin{aligned}\mathbf{F}_{\rm Sat}^{5^{\rm th}} &= -\frac{\beta}{M_{\rm Pl}}\int_{V_{\rm Sat}}\boldsymbol{\nabla}\phi_{\rm tot}\,\mathrm{d}m(\mathbf{x})\\ &\simeq -\frac{\beta}{M_{\rm Pl}}\int_{V_{\rm Sat}}\boldsymbol{\nabla}\phi_\oplus\,\mathrm{d}m(\mathbf{x})\,,\end{aligned} \tag{E2}$$

QED.

F. Mathematical proof of the absence of symmetry in the orbital dynamics

Context & Notations. In Sec. IV D, we have seen on simulation results that, although the gravity field is exactly symmetric with respect to the line $\theta = 0$, the dynamics of a point-mass in orbit is not. Let us translate this statement into mathematical terms. Let $\{\mathbf{X}_*(t)\,,\ t > 0\}$ be a trajectory in phase space that is a solution of the ODE of interest, i.e. $\forall t > 0$, $\dot{\mathbf{X}}_*(t) = F(t, \mathbf{X}_*(t))$. At some point, the particle will pass over the mountain so that we can define t_m, the time at which $\theta(t_m) = 0$ for the

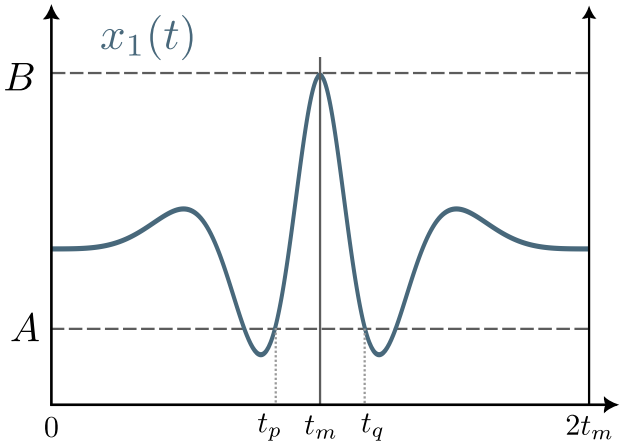


FIG. 19. Visual support for the proof.

first time. Demanding that the trajectory is symmetric with respect to $\theta = 0$ actually means

$$\forall t \in [0, t_m]\,,\ \mathbf{X}_*(t) = \mathbf{X}_*(2t_m - t)\,. \tag{F1}$$

The state vector $\mathbf{X}_*(t)$ has components

$$\begin{aligned}\mathbf{X}_*(t) &= \left[\delta r(t),\, \dot{\delta r}(t),\, \delta\theta(t),\, \delta L(t)\right] \\ &= \left[x_1(t),\, x_2(t),\, x_3(t),\, x_4(t)\right]\,.\end{aligned}$$

We furthermore recall that the vector field $F\colon (s, \mathbf{Y}) \in \mathbb{R}\times\mathbb{R}^4 \mapsto (F_1, F_2, F_3, F_4) \in \mathbb{R}^4$ is given by

$$\begin{aligned} F_1 &= y_2 \\ F_2 &= \frac{(L_0 + y_4)^2}{(a+y_1)^3} + g(a + y_1, \theta_0 + \omega s + y_3) \\ F_3 &= \frac{L_0 + y_4}{(a+y_1)^2} - \omega \\ F_4 &= h(a + y_1, \theta_0 + \omega s + y_3)\,. \end{aligned} \tag{F2}$$

In the above, functions g and h refer to the gravitational potential partial derivatives $-\partial_r U$ and $-\partial_\theta U$ respectively.

Theorem. The perturbed Keplerian problem with vector field (F2) is not symmetric on both sides of the mountain in the general case.

Proof by contradiction. Let us suppose that (F1) holds and derive a (necessary) condition on function g and h. Because x_1 is continuous, there exists $t_p \in [0, t_m[$ such that x_1 is monotonous over $[t_p, t_m]$. Letting $t_q = 2t_m - t_p$, x_1 is also monotonous over $[t_m, t_q]$ due to the symmetry (F1). We further set

$$\begin{cases} A &= x_1(t_p) = x_1(t_q) \\ B &= x_1(t_m) \end{cases}.$$

Notations introduced so far are shown in Fig. 19. Let us assume for now that $A \neq B$ so that x_1 is actually *strictly* monotonous over $[t_p, t_m]$ and $[t_m, t_q]$ respectively. Then $V \coloneqq [\min(A,B), \max(A,B)]$ is not a degenerate interval and we can define

$$\begin{aligned} x_1^p\colon [t_p, t_m] &\to V \\ t &\mapsto x_1(t) \end{aligned} \qquad \begin{aligned} x_1^q\colon [t_m, t_q] &\to V \\ t &\mapsto x_1(t) \end{aligned}$$

together with their respective inverse

$$\begin{aligned} z_1^p\colon V &\to [t_p, t_m] \\ u &\mapsto z_1^p(u) \end{aligned} \qquad \begin{aligned} z_1^q\colon V &\to [t_m, t_q]\,. \\ u &\mapsto z_1^q(u) \end{aligned}$$

We will make use of the following property on the inverse functions

$$\forall u \in V\,,\ z_1^p(u) = 2t_m - z_1^q(u)\,. \tag{F3}$$

Indeed, for $u \in V$, there exist two unique times $t_\alpha \in [t_p, t_m]$ and $t_\beta \in [t_m, t_q]$ such that $u = x_1^p(t_\alpha) = x_1^q(t_\beta)$. Reciprocally, $t_\alpha = z_1^p(u)$ and $t_\beta = z_1^q(u)$. Then

$$\begin{aligned} z_1^p(u) &= z_1^p(x_1^q(t_\beta)) = z_1^p(x_1(t_\beta)) = z_1^q(x_1(2t_m - t_\beta)) \\ &= z_1^p(x_1^p(2t_m - t_\beta)) = 2t_m - t_\beta \\ &= 2t_m - z_1^q(u)\,. \end{aligned}$$

We then compute the integral

$$I \coloneqq \int_{t_p}^{t_q} \frac{\mathrm{d}x_1}{\mathrm{d}s}(s)\, x_2(s)\, \mathrm{d}s$$

by two different ways. On the one hand,

$$I = \int_{t_p}^{t_q} \left|\frac{\mathrm{d}x_1}{\mathrm{d}s}(s)\right|^2 \mathrm{d}s \tag{F4}$$

because $\dot{x}_1(s) = x_2(s)$ along the trajectory. On the other hand,

$$I = \int_{t_p}^{t_m} \frac{\mathrm{d}x_1}{\mathrm{d}s}(s)\, x_2(s)\, \mathrm{d}s + \int_{t_m}^{t_q} \frac{\mathrm{d}x_1}{\mathrm{d}s}(s)\, x_2(s)\, \mathrm{d}s\,,$$

from which we can make the changes of variable $u = x_1^p(s)$ in the first integral and $u = x_1^q(s)$ in the second one, yielding

$$\begin{aligned} I &= \int_A^B x_2(z_1^p(u))\, \mathrm{d}u + \int_B^A x_2(z_1^q(u))\, \mathrm{d}u \\ &= \int_A^B x_2(z_1^p(u))\, \mathrm{d}u + \int_B^A x_2(2t_m - z_1^p(u))\, \mathrm{d}u \\ &= 0 \quad \text{because of symmetry (F1).} \end{aligned}$$

From Eq. (F4), we immediately deduce that $\dot{x}_1 \equiv 0$ on $[t_p, t_q]$. The fact that x_1 is constant contradicts our previous assumption that $A \neq B$. Therefore, A has to be equal to B. Put in perspective with the fact that x_1 is monotonous on $[t_p, t_m]$ and on $[t_m, t_q]$, we have

- x_1 is constant over $[t_p, t_q]$. Let H be this constant and define the radial distance $R \coloneqq a + H$.
- $\dot{x}_2 \equiv 0$ on $[t_p, t_q]$ as well.

The final stage of this demonstration follows from the specific form of the vector field F. Let $s \in [t_p, t_q]$. For convenience, we recall that $\theta = \theta_0 + \omega s + x_3(s)$ and we denote by ∂_1, ∂_2 the partial derivatives of a two-variable function with respect to the first and second variable respectively. Taking the derivative with respect to s in the second equation of the ODE system $\dot{\mathbf{X}_*}(s) = F(s, \mathbf{X}_*(s))$ yields

$$\frac{\mathrm{d}}{\mathrm{d}s}\{\dot{x}_2(s)\} = \frac{\mathrm{d}}{\mathrm{d}s}\left\{\frac{[L_0 + x_4(s)]^2}{R^3} + g(R,\theta)\right\} = 0$$
$$\iff 2\dot{x}_4(s)\frac{L_0 + x_4(s)}{R^3} + [\omega + \dot{x}_3(s)]\,\partial_2 g(R,\theta) = 0\,.$$

In this last equation, we can substitute $\dot{x}_4$ and $\dot{x_3}$ by the r.h.s. of the ODE, yielding

$$[L_0 + x_4(s)]\,[2h(R,\theta) + R\partial_2 g(R,\theta)] = 0\,. \tag{F5}$$

The total angular momentum $L(s) = L_0 + x_4(s) = R^2\dot{\theta}(s)$ cannot be zero, because otherwise the point-mass would be frozen it time. Moreover, the interval $J \coloneqq \{\theta_0 + \omega s + x_3(s)\,,\ s \in [t_p, t_q]\}$ is not a singleton because again, the point-mass cannot remain frozen in time. Therefore, Eq. (F5) leads straightforwardly to

$$\forall \theta \in J\,,\ 2h(R,\theta) + R\partial_2 g(R,\theta) = 0\,. \tag{F6}$$

Replacing functions g and h by their definition in relation to the gravitational potential U, we arrive at the final conclusion that $\forall \theta \in J$:

$$\begin{aligned} &\partial_\theta\,[2U(R,\theta) + R\partial_r U(R,\theta)] = 0 \\ \text{i.e.}\quad &2U(R,\theta) + R\partial_r U(R,\theta) = C^{st} \end{aligned} \tag{F7}$$

Condition (F7) is very restrictive on the form of admissible gravitational potentials and one can check that the potential of the {sphere + mountain} system that we have been using throughout this article does not satisfy this criterion.

Conclusion. We found a necessary condition on the gravitational potential (F7) for the dynamics of a point-mass in orbit to be symmetric with respect to $\theta = 0$. As a remark, the potential created by a perfect sphere trivially satisfies the criterion.

G. Orbital periods

In classical central force problems, the period T of a satellite in circular orbit around a planet can be expressed as a function of the distance r to the planet's center and the acceleration a it undergoes: $T = 2\pi\sqrt{r/a}$. A direct consequence of this formula is that the addition of the chameleon acceleration to the Newtonian one will slightly modify the orbital period. In Sec. IV D 5, we laid emphasis on the fact that the use of different altitudes was one possible way of circumventing the issue of model uncertainties. We can thus examine the difference in orbital period for the two gravity models, namely

$$\Delta T^{\mathrm{New}} = 2\pi\left(\sqrt{\frac{R_\oplus + h_1}{a_{\mathrm{New}}(r_1)}} - \sqrt{\frac{R_\oplus + h_2}{a_{\mathrm{New}}(r_2)}}\right),$$

$$\Delta T^{\mathrm{cham}} = 2\pi\left(\sqrt{\frac{R_\oplus + h_1}{(a_{\mathrm{New}} + a_{\mathrm{cham}})(r_1)}} - \sqrt{\frac{R_\oplus + h_2}{(a_{\mathrm{New}} + a_{\mathrm{cham}})(r_2)}}\right),$$

with $r = R_\oplus + h$ and $a_{\mathrm{New}}(r_1) = \mu_\oplus/r^2$.

Fig. 20 illustrates the difference $\mathfrak{D} = \Delta T^{\mathrm{New}} - \Delta T^{\mathrm{cham}}$ (expressed in seconds) in the (h_1, h_2)-plane, for several values of the parameter α. Equivalently, one can parameterize deviation from Newtonian gravity as $\Delta T^{\mathrm{New}} = \Delta T^{\mathrm{cham}}(1+\epsilon)$. Then we have $\epsilon = \Delta T^{\mathrm{New}}\mathfrak{D}$. In order to remain consistent with the rest of this article, we have fixed $\Lambda = \Lambda_{\mathrm{DE}}$ and $n = 1$. The setup being spherically symmetric, the numerical computation of the chameleon fifth force can be performed with 1D finite elements. We can see that the orbital period anomaly $\mathfrak{D}$ can be greater than 10^{-5} s. This effect scales linearly with the number of completed orbits: after 100 000 orbits, one can expect the anomaly to be of the order of a second — which would take approximately 20 years for a satellite orbiting 1000 km above the Earth surface (ignoring all the perturbing forces otherwise present in a realistic scenario).

[1] U.s. standard atmosphere, 1976. Technical report, U.S. Government Printing Office, Washington, D.C., Oct 1976.

[2] Abich, K. et al. In-orbit performance of the grace follow-on laser ranging interferometer. *Phys. Rev. Lett.*, 123:031101, Jul 2019.

[3] Allgower, E.L. and Georg, K. *Numerical Continuation Methods: An Introduction.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[4] Arnold, D. et al. Satellite laser ranging to low earth orbiters: orbit and network validation. *Journal of Geodesy*, 93(11):2315–2334, Nov 2019.

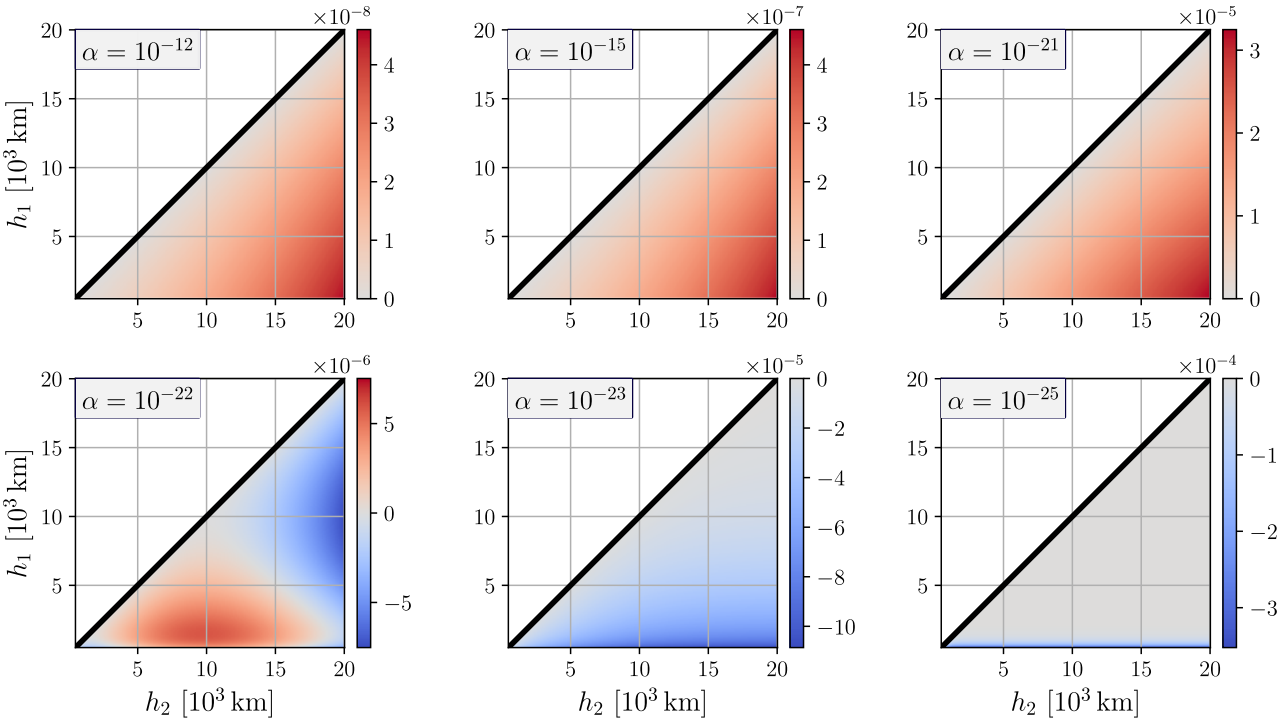


FIG. 20. $\mathfrak{D} = \Delta T^{\rm New} - \Delta T^{\rm cham}$ [s]. We set 2×10^4 km $\geq h_2 \geq h_1 \geq 5 \times 10^2$ km and the black solid line corresponds to $h_1 = h_2$. For $-\log_{10}\alpha \in \{12, 15, 21\}$, the anomaly $\mathfrak{D}$ is maximal when the two considered altitudes are far apart. The case $\alpha = 10^{-22}$ exhibits the transition between the latter regime and a new regime ($\alpha \leq 10^{-23}$) where h_1 has to remain low to produce a notable anomaly. This second regime arises because in the limit $\alpha \to 0$, the chameleon field quickly (i.e. at low altitude) gets close to its asymptotic value. The fifth force thus vanishes at higher altitudes, and so does the orbital period anomaly $\mathfrak{D}$.

[5] Babichev, E., Deffayet, C. and Ziour, R. k-mouflage gravity. *International Journal of Modern Physics D*, 18(14):2147–2154, 2009.

[6] Bergé, J. Microscope's view at gravitation. *Reports on Progress in Physics*, 86(6):066901, may 2023.

[7] Bergé, J. et al. Interpretation of geodesy experiments in non-newtonian theories of gravity. *Classical and Quantum Gravity*, 35(23):234001, nov 2018.

[8] Bertotti, B., Iess, L. and Tortora, P. A test of general relativity using radio links with the cassini spacecraft. *Nature*, 425(6956):374–376, Sep 2003.

[9] Bothwell, T. et al. Resolving the gravitational redshift across a millimetre-scale atomic sample. *Nature*, 602(7897):420–424, Feb 2022.

[10] Boulmezaoud, Tahar Zamène. Inverted finite elements: a new method for solving elliptic problems in unbounded domains. *ESAIM: M2AN*, 39(1):109–145, 2005.

[11] Brax, P., Burrage, C. and Davis, A.C. Screening fifth forces in k-essence and dbi models. *Journal of Cosmology and Astroparticle Physics*, 2013(01):020, jan 2013.

[12] Brax, P. et al. Testing screened modified gravity. *Universe*, 8(1), 2022.

[13] Brax, P. et al. Detecting dark energy in orbit: The cosmological chameleon. *Phys. Rev. D*, 70:123518, Dec 2004.

[14] Brax, P. et al. Detecting chameleons through casimir force measurements. *Phys. Rev. D*, 76:124034, Dec 2007.

[15] Brewer, S.M. et al. $^{27}al^+$ quantum-logic clock with a systematic uncertainty below 10^{-18}. *Phys. Rev. Lett.*, 123:033201, Jul 2019.

[16] Briddon, C. et al. Selcie: a tool for investigating the chameleon field of arbitrary sources. *Journal of Cosmology and Astroparticle Physics*, 2021(12):043, dec 2021.

[17] Burrage, C., Copeland, E.J. and Stevenson, J.A. Ellipticity weakens chameleon screening. *Phys. Rev. D*, 91:065030, Mar 2015.

[18] Burrage, C. and Sakstein, J. A compendium of chameleon constraints. *Journal of Cosmology and Astroparticle Physics*, 2016(11):045, nov 2016.

[19] Burrage, C. and Sakstein, J. Tests of chameleon gravity. *Living Reviews in Relativity*, 21(1):1, Mar 2018.

[20] Butler, R. and Tsuboi, S. Antipodal seismic reflections upon shear wave velocity structures within earth's inner core. *Physics of the Earth and Planetary Interiors*, 321:106802, 2021.

[21] Chouhan, A.K., Choudhury, P. and Pal, S.K. New evidence for a thin crust and magmatic underplating beneath the cambay rift basin, western india through modelling of eigen-6c4 gravity data. *Journal of Earth System Science*, 129(1):64, Feb 2020.

[22] Christophe, B. et al. A new generation of ultra-sensitive electrostatic accelerometers for grace follow-on and towards the next generation gravity missions. *Acta Astronautica*, 117:1–7, 2015.

[23] Cimrman, R., Lukeš, V. and Rohan, E. Multiscale finite element calculations in python using sfepy. *Advances in Computational Mathematics*, 45(4):1897–1921,

Aug 2019.

[24] Ciufolini, I. et al. A test of general relativity using the lares and lageos satellites and a grace earth gravity model. *The European Physical Journal C*, 76(3):120, Mar 2016.

[25] Ciufolini, I. et al. The lares 2 satellite, general relativity and fundamental physics. *The European Physical Journal C*, 83(1):87, Jan 2023.

[26] Dvali, G., Gruzinov, A. and Zaldarriaga, M. The accelerated universe and the moon. *Phys. Rev. D*, 68:024012, Jul 2003.

[27] Dziewonski, A.M. and Anderson, D.L. Preliminary reference earth model. *Physics of the Earth and Planetary Interiors*, 25(4):297–356, 1981.

[28] Everitt, C.W.F. et al. Gravity probe b: Final results of a space experiment to test general relativity. *Phys. Rev. Lett.*, 106:221101, May 2011.

[29] Frei, W. Load ramping of nonlinear problems. COMSOL Blog, Nov 2013. Accessed: May 26$^{\text{th}}$, 2023.

[30] Frei, W. Nonlinearity ramping for improving convergence of nonlinear problems. COMSOL Blog, Dec 2013. Accessed: May 26$^{\text{th}}$, 2023.

[31] Fukushima, T. Numerical computation of gravitational field for general axisymmetric objects. *Monthly Notices of the Royal Astronomical Society*, 462(2):2138–2176, 07 2016.

[32] Gabriel, G. et al. Anomalies of the Earth's total magnetic field in Germany – the first complete homogenous data set reveals new opportunities for multiscale geoscientific studies. *Geophysical Journal International*, 184(3):1113–1118, 03 2011.

[33] Geuzaine, C. and Remacle, J.F. Gmsh: A 3-d finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[34] Gu, J.A. and Lin, W.T. Solar-system constraints on f(r) chameleon gravity, 2011.

[35] Hairer, E., Wanner, G. and Lubich, C. *Conservation of First Integrals and Methods on Manifolds*, pages 97–142. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[36] Jaffe, M. et al. Testing sub-gravitational forces on atoms from a miniature in-vacuum source mass. *Nature Physics*, 13(10):938–942, Oct 2017.

[37] Jain, B., Vikram, V. and Sakstein, J. Astrophysical tests of modified gravity: Constraints from distance indicators in the nearby universe. *The Astrophysical Journal*, 779(1):39, nov 2013.

[38] Jones-Smith, K. and Ferrer, F. Detecting chameleon dark energy via an electrostatic analogy. *Phys. Rev. Lett.*, 108:221101, May 2012.

[39] Katsuragawa, T. et al. Gravitational waves in $f(r)$ gravity: Scalar waves and the chameleon mechanism. *Phys. Rev. D*, 99:124050, Jun 2019.

[40] Khoury, J. and Weltman, A. Chameleon cosmology. *Phys. Rev. D*, 69:044026, Feb 2004.

[41] Khoury, J. and Weltman, A. Chameleon fields: Awaiting surprises for tests of gravity in space. *Phys. Rev. Lett.*, 93:171104, Oct 2004.

[42] Kozak, A. and Wojnar, A. Earthquakes as probing tools for gravity theories, 2023.

[43] Kraiselburd, L. et al. Equivalence principle in chameleon models. *Phys. Rev. D*, 97:104044, May 2018.

[44] Landerer, F.W. et al. Extending the global mass change data record: Grace follow-on instrument and science data performance. *Geophysical Research Letters*, 47(12), 2020.

[45] Langtangen, H.P. *Computational Partial Differential Equations*. Springer Berlin Heidelberg, 2003.

[46] Lévy, H. Numerical investigation of screened scalar-tensor theories in space: Microscope and the future. In *57th Rencontres de Moriond 2023 [Gravitation]*, mar 2023.

[47] Lévy, H., Bergé, J. and Uzan, J.P. Solving nonlinear klein-gordon equations on unbounded domains via the finite element method. *Phys. Rev. D*, 106:124021, Dec 2022.

[48] Li, Q. et al. Measurements of the gravitational constant using two independent methods. *Nature*, 560(7720):582–588, Aug 2018.

[49] Lou, Y. et al. A review of real-time multi-gnss precise orbit determination based on the filter method. *Satellite Navigation*, 3(1):15, Jul 2022.

[50] Matoza, R.S. and Roman, D.C. One hundred years of advances in volcano seismology and acoustics. *Bulletin of Volcanology*, 84(9):86, Aug 2022.

[51] Montenbruck, O. and Gill, E. *Satellite Orbits*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[52] Moritz, H. *Classical Physical Geodesy*, pages 253–289. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[53] Mota, D.F. and Shaw, D.J. Evading equivalence principle violations, cosmological, and other experimental constraints in scalar field theories with a strong coupling to matter. *Phys. Rev. D*, 75:063501, Mar 2007.

[54] Musen, P. The influence of the solar radiation pressure on the motion of an artificial satellite. *Journal of Geophysical Research (1896-1977)*, 65(5):1391–1396, 1960.

[55] Nabizadeh, M.S., Ramamoorthi, R. and Chern, A. Kelvin transformations for simulations on infinite domains. *ACM Trans. Graph.*, 40(4), jul 2021.

[56] Nicolis, A., Rattazzi, R. and Trincherini, E. Galileon as a local modification of gravity. *Phys. Rev. D*, 79:064036, Mar 2009.

[57] Ogden, L. et al. Electrostatic analogy for symmetron gravity. *Phys. Rev. D*, 96:124029, Dec 2017.

[58] Oh, H.S., Jang, B. and Jou, Y. The weighted ritz-galerkin method for elliptic boundary value problems on unbounded domains. *Numerical Methods for Partial Differential Equations*, 19(3):301–326, 2003.

[59] Pavlis, N.K. et al. The development and evaluation of the earth gravitational model 2008 (egm2008). *Journal of Geophysical Research: Solid Earth*, 117(B4), 2012.

[60] Pearlman, M. et al. Laser geodetic satellites: a high-accuracy scientific tool. *Journal of Geodesy*, 93(11):2181–2194, Nov 2019.

[61] Pernot-Borràs, M. et al. General study of chameleon fifth force in gravity space experiments. *Phys. Rev. D*, 100:084006, Oct 2019.

[62] Pernot-Borràs, M. et al. Constraints on chameleon gravity from the measurement of the electrostatic stiffness of the *microscope* mission accelerometers. *Phys. Rev. D*, 103:064070, Mar 2021.

[63] Pernot-Borràs, M. *Testing gravity in space : towards a realistic treatment of chameleon gravity in the MICROSCOPE mission*. phdthesis, Sorbonne Université, November 2020.

[64] Pourhasan, R. et al. Chameleon gravity, electrostatics, and kinematics in the outer galaxy. *Journal of Cosmology and Astroparticle Physics*, 2011(12):005, dec 2011.

[65] Savalle, E. et al. Gravitational redshift test with the

future ACES mission. *Classical and Quantum Gravity*, 36, Nov 2019.

[66] Schärer, A. et al. Testing scalar-tensor theories and parametrized post-newtonian parameters in earth orbit. *Phys. Rev. D*, 90:123005, Dec 2014.

[67] Schreiner, P. et al. On precise orbit determination based on doris, gps and slr using sentinel-3a/b and -6a and subsequent reference frame determination based on doris-only. *Advances in Space Research*, 72:47–64, jul 2023.

[68] Tapley, B., Schutz, B. and Born, G. *Statistical Orbit Determination*. Academic Press, Burlington, 2004.

[69] Tiesinga, E. et al. Codata recommended values of the fundamental physical constants: 2018. *Rev. Mod. Phys.*, 93:025010, Jun 2021.

[70] Touboul, P. et al. Champ, grace, goce instruments and beyond. In Kenyon, S., Pacino, M.C. and Marti, U., editors, *Geodesy for Planet Earth*, pages 215–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[71] Touboul, P. et al. *microscope* mission: Final results of the test of the equivalence principle. *Phys. Rev. Lett.*, 129:121102, Sep 2022.

[72] Upadhye, A. Dark energy fifth forces in torsion pendulum experiments. *Phys. Rev. D*, 86:102003, Nov 2012.

[73] Vainshtein, A. To the problem of nonvanishing gravitation mass. *Physics Letters B*, 39(3):393–394, 1972.

[74] Vikram, V. et al. Astrophysical tests of modified gravity: Stellar and gaseous rotation curves in dwarf galaxies. *Phys. Rev. D*, 97:104055, May 2018.

[75] Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[76] Waterhouse, T.P. An introduction to chameleon gravity, 2006.

[77] Wieczorek, M.A. and Meschede, M. Shtools: Tools for working with spherical harmonics. *Geochemistry, Geophysics, Geosystems*, 19(8):2574–2592, 2018.

[78] Wilcox, H. et al. The XMM Cluster Survey: testing chameleon gravity using the profiles of clusters. *Monthly Notices of the Royal Astronomical Society*, 452(2):1171–1183, 07 2015.

[79] Will, C.M. The confrontation between general relativity and experiment. *Living Reviews in Relativity*, 17(1):4, Dec 2014.

[80] Williams, J.G., Turyshev, S.G. and Boggs, D.H. Progress in lunar laser ranging tests of relativistic gravity. *Phys. Rev. Lett.*, 93:261101, Dec 2004.

[81] Williams, J.G., Turyshev, S.G. and Boggs, D.H. Lunar laser ranging tests of the equivalence principle. *Classical and Quantum Gravity*, 29(18):184004, aug 2012.

[82] Yin, P. et al. Experiments with levitated force sensor challenge theories of dark energy. *Nature Physics*, 18(10):1181–1185, Oct 2022.

[83] Zhao, Y. et al. A brief review of magnetic anomaly detection. *Measurement Science and Technology*, 32(4):042002, February 2021. Publisher: IOP Publishing.

Chapter summary

We delved into the possibility of taking advantage of current space geodesy missions to provide constraints on scalar-tensor theories of gravity with screening mechanism. Focusing on the chameleon model, we shed new light on two long-standing issues, namely the influence of the atmosphere on the fifth force experienced by a spacecraft and the backreaction of the latter on the scalar field, and considered a deviation from spherical symmetry through the implementation of a true to scale mountain. Simulating the dynamics of the GRACE-FO pair of satellites which are assumed to be unscreened, we show that the anomaly brought about by the scalar fifth force is comfortably within the range of sensitivity offered by current space-born technology. However, the existence of uncertainties in the model, most notably the fact that the distribution of matter within the Earth is poorly known, greatly mitigates the constraining power of such tests. We explore one way around this deadlock which consists in performing the same experiment at different altitudes.

Chapter 6

Testing screened scalar-tensor theories with clocks

Outline of the current chapter

This last chapter further develops the idea put forward in the outlook section of our work [141], namely the potential possibility to test scalar-tensor theories of gravity by means of redshift experiments. Building on the theoretical aspects laid out in Chapt. 1, we derive the redshift expression in the framework of scalar-tensor models and single out the scalar contribution in the Newtonian limit. As in Chapt. 5, we focus our discussion on the chameleon model. Unlike fifth force effects, which are mainly dependent on the magnitude of the gradient of the scalar field, it appears that the scalar contribution to the total redshift depends, for the most part, on the field's value itself. We then endeavor to show that precise redshift measurements could reveal the presence of the scalar field. For this purpose, we imagine a thought experiment which guides us towards more realistic experimental setups, in the laboratory and in space.

	Einstein Frame $(g_{\mu\nu},\,\phi)$		**Jordan Frame** $(\tilde{g}_{\mu\nu},\,\varphi)$	
Field and geodesic Eqs.	$R_{\mu\nu}-\frac{1}{2}Rg_{\mu\nu}=\frac{1}{M_{\rm Pl}^2}\left(T_{\mu\nu}+T^{(\phi)}_{\mu\nu}\right)$	(6.2)	$F(\varphi)\left(\tilde{R}_{\mu\nu}-\frac{1}{2}\tilde{R}\tilde{g}_{\mu\nu}\right)=\frac{1}{M_{\rm Pl}^2}\left(\tilde{T}_{\mu\nu}+\tilde{T}^{(\varphi)}_{\mu\nu}\right)$	(6.3)
	$\Box\phi=\frac{\mathrm{d}V}{\mathrm{d}\phi}-\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}T$	(6.4)	$Z(\varphi)\tilde{\Box}\varphi=\frac{\mathrm{d}U}{\mathrm{d}\varphi}-\frac{M_{\rm Pl}^2}{2}\frac{\mathrm{d}F}{\mathrm{d}\varphi}\tilde{R}-\frac{1}{2}\frac{\mathrm{d}Z}{\mathrm{d}\varphi}\tilde{g}^{\alpha\beta}\partial_\alpha\varphi\,\partial_\beta\varphi$	(6.5)
	$u^\alpha\nabla_\alpha u^\mu=-\perp^{\mu\nu}\partial_\nu(\ln\Omega)$	(6.6)	$\tilde{u}^\alpha\tilde{\nabla}_\alpha\tilde{u}^\mu=0$	(6.7)
Newtonian limit	$ds^2=-(1+2\Phi)\mathrm{d}t^2+g_{ij}\mathrm{d}x^i\mathrm{d}x^j$	(6.8)	$d\tilde{s}^2=-\big(1+2\tilde{\Phi}\big)\mathrm{d}t^2+\tilde{g}_{ij}\mathrm{d}x^i\mathrm{d}x^j$	(6.9)
	$2M_{\rm Pl}^2\Delta\Phi=\rho-2V(\phi)$	(6.10)	$M_{\rm Pl}^2\Big[2F(\varphi)\tilde{\Delta}\tilde{\Phi}+\tilde{\Delta}F\Big]=\tilde{\rho}-2U(\varphi)$	(6.11)
	$\Delta\phi=\frac{\mathrm{d}V}{\mathrm{d}\phi}+\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\rho$	(6.12)	$Z(\varphi)\tilde{\Delta}\varphi=\frac{\mathrm{d}U}{\mathrm{d}\varphi}-\frac{1}{2}\frac{\mathrm{d}\ln F}{\mathrm{d}\varphi}\left[\tilde{\rho}+4U(\varphi)+3M_{\rm Pl}^2\tilde{\Delta}F\right]$	(6.13)
	$\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2}=-\partial_i\Phi-\frac{\mathrm{d}\ln\Omega}{\mathrm{d}\phi}\partial_i\phi$	(6.14)	$\frac{\mathrm{d}^2x^i}{\mathrm{d}t^2}=-\partial_i\tilde{\Phi}$	(6.15)

Table 6.1: Compilation of the relevant field equations from Sec. 1.1.2. The Jordan-frame metric is conformally related to the Einstein-frame metric through $\tilde{g}_{\mu\nu}=\Omega^2(\phi)g_{\mu\nu}$. The tensors $T^{(\phi)}_{\mu\nu}$ and $\tilde{T}^{(\varphi)}_{\mu\nu}$ are given by Eq. (1.44) and Eq. (1.55) respectively. We set $\perp^{\mu\nu}=g^{\mu\nu}+u^\mu u^\nu$. We have assumed that the conformal function Ω is close to unity, i.e. $\Omega(\phi)=1+\omega(\phi)$ with $|\omega(\phi)|\ll 1$ [otherwise the Newtonian limits in the Einstein and Jordan frames would not be consistent with one another, see Eqs. (1.67–1.68) and the associated discussion].

6.1 Gravitational redshift in scalar-tensor theories

One of the most conspicuous features of scalar-tensor theories of gravity, as presented in Sec. 1.1, are *fifth forces*. In theories with screening mechanisms (see Sec. 1.2), fifth forces can be greatly mitigated to pass existing tests of gravity without compromising the very *raisons d'être* of such theories (e.g., explain cosmic acceleration). Over the past two decades or so, a major research effort has been carried out to find new ways of constraining screened scalar-tensor models, almost exclusively relying on fifth force effects in the case of the chameleon — see e.g. the review articles [71, 151]. There, the 3-acceleration experienced by a test particle, owing to the presence of the scalar field only, is given by

$$\mathbf{a}_\phi=-\boldsymbol{\nabla}\big[\ln\Omega(\phi)\big]=-\frac{\beta}{M_{\rm Pl}}\boldsymbol{\nabla}\phi \tag{6.1}$$

in the Newtonian limit, see Sec. 1.1.3. This expression explicitly states the proportionality between the fifth force and the field's gradient. Consequently, experimental designs looking for such fifth forces generally try to make the field's gradient as large as possible [276], or look for ways to disentangle it from the Newtonian gravitational attraction [157, 292–294], through ingenious mass distributions.

Nonetheless, there are other physical effects studied in the literature that can be leveraged for testing this model aside from fifth force searches, notably:

- *scalar radiation*, in pulsating stars [182, 198, 295] or in compact binary systems [296, 297];
- *interaction with photons*, when one considers a non-zero coupling between the scalar field and photons [298, 299], but this is out of the scope of the framework laid out in Sec. 1.1.2.

Here, we want to assess whether *gravitational redshift* (or equivalently, gravitational time-dilation) can constitute yet another venue for testing screened scalar-tensor theories. This idea is not new. Ref. [300] establishes a rigorous derivation of the anomalous redshift arising from vector and scalar fields non-minimally coupled to matter in the Einstein frame. The seminal article on the chameleon model [115] mentions in Sec. VIII the possibility of deriving constraints from the Vessot–Levine bound [25]. There, the authors perform a few order-of-magnitude computations to argue that chameleons comfortably satisfy this bound. Ref. [125] also underlines the fact that scalar-tensor theories predict a measured redshift different from that given in GR.

This section aims at deriving, under very general assumptions, the correct expression of the redshift in scalar-tensor theories. Emphasis is laid on *measurable quantities* — in that respect, the expression thereby obtained is put into perspective with the usual form of parameterized redshift violations, see Eq. (1.30). Finally, we show that the chameleon model could be quite sensitive to redshift tests in some specific cases, precisely because of their inherent nonlinear character. In this endeavor, we frequently need to refer to equations that were derived back in Chapt. 1. For the sake of convenience, the relevant equations have been reproduced in Table 6.1.

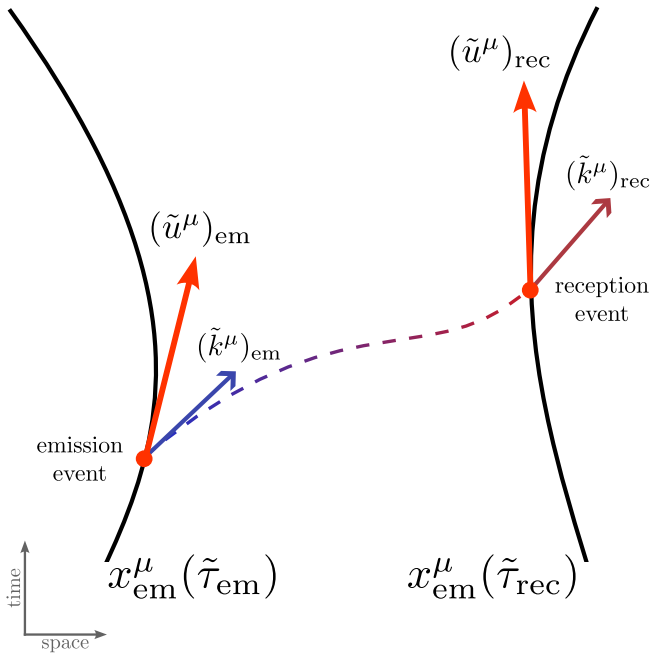


Figure 6.1: Notations associated with the redshift definition on a spacetime diagram. The solid black line labeled $x^\mu_{\rm em}$ (resp. $x^\mu_{\rm rec}$) is the worldline of the emitter (resp. receiver), parametrized by the proper time $\tilde{\tau}_{\rm em}$ (resp. $\tilde{\tau}_{\rm rec}$) and with 4-velocity $(\tilde{u}^\mu)_{\rm em}$ (resp. $(\tilde{u}^\mu)_{\rm rec}$). $(\tilde{k}^\mu)_{\rm em}$ (resp. $(\tilde{k}^\mu)_{\rm rec}$) denotes the photon 4-wave-vector at the emission event (resp. reception event). The dashed line corresponds to the photon's null geodesic between the two events.

6.1.1 Derivation of the redshift expression in scalar-tensor theories

Here we derive the redshift expression in the framework of scalar-tensor theories given by the action (1.33–1.34). One observer, the *emitter*, sends a photon to another observer, the *receiver*. The gravitational redshift, denoted by z, is *defined* as

$$z = \frac{E_{\rm em}}{E_{\rm rec}} - 1\,, \tag{6.16}$$

where $E_{\rm em}$ (resp. $E_{\rm rec}$) denotes the energy of the photon *measured* by the emitter (resp. by the receiver). The *physical* metric to be used in the subsequent calculations of these energies is the Jordan-frame metric $\tilde{g}_{\mu\nu}$. Indeed, it is the metric to which matter is universally coupled [see Eq. (1.11b)] and thus defines the lengths and times measured by material rods and clocks — see Refs. [70, 300] or the discussions we had back in Sec. 1.1. Consequently, we have

$$\frac{E_{\rm em}}{E_{\rm rec}} = \frac{\hbar(\tilde{u}_\mu \tilde{k}^\mu)_{\rm em}}{\hbar(\tilde{u}_\mu \tilde{k}^\mu)_{\rm rec}}\,, \tag{6.17}$$

where $(\tilde{u}^\mu)_{\rm em}$ (resp. $(\tilde{u}^\mu)_{\rm rec}$) denotes the 4-velocity of the emitter (resp. receiver) and $\tilde{k}^\mu$ represents the null tangent vector of a geodesic joining the emission and reception events [where the affine parameter is normalized so that $\tilde{k}^\mu$ coincides with the 4-wavevector, see Eq. (1.9)]. These quantities are defined in the Jordan frame, and normalized such that

$$\tilde{g}_{\mu\nu}\tilde{u}^\mu\tilde{u}^\nu = -1 \qquad \text{and} \qquad \tilde{g}_{\mu\nu}\tilde{k}^\mu\tilde{k}^\nu = 0\,. \tag{6.18}$$

For the sake of clarity, notations are illustrated on a spacetime diagram in Fig. 6.1.

Let us now make some additional assumptions:

- The spatial coordinates of the emitter on the one hand, and those of the receiver on the other hand, remain fixed throughout this experiment.

- The metric $\tilde{g}_{\mu\nu}$ is stationary, meaning that it does not depend upon the x^0 coordinate, i.e. $\partial_0 \tilde{g}_{\mu\nu} = 0$. Consequently, the scalar field ϕ will also be assumed stationary.

As a direct consequence, the relation $\tilde{g}_{\mu\nu}\tilde{u}^\mu\tilde{u}^\nu = -1$ leads to $\tilde{u}^0 = 1/\sqrt{-\tilde{g}_{00}}$ (since $\tilde{u}^i = 0$). Moreover, the fact that the metric is taken stationary implies that there exists a timelike Killing vector $\tilde{\boldsymbol{\xi}} = (1, 0, 0, 0)$ associated with the time translation symmetry. Denoting $\tilde{\lambda}$ the affine parameter of the photon geodesic, mathematical

properties of Killing vectors (see e.g. Ref. [301]) let us write

$$\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left(\tilde{\xi}^{\mu}\tilde{k}_{\mu}\right)=\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left(\tilde{\xi}_{\mu}\tilde{k}^{\mu}\right)=0\implies\frac{\mathrm{d}}{\mathrm{d}\tilde{\lambda}}\left(\tilde{k}_{0}\right)=0\,. \tag{6.19}$$

Thus, $\left(\tilde{k}_0\right)_{\rm em}=\left(\tilde{k}_0\right)_{\rm rec}$ and we eventually get

$$1+z=\sqrt{\frac{(\tilde{g}_{00})_{\rm rec}}{(\tilde{g}_{00})_{\rm em}}}=\frac{\Omega_{\rm rec}}{\Omega_{\rm em}}\sqrt{\frac{(g_{00})_{\rm rec}}{(g_{00})_{\rm em}}}\,. \tag{6.20}$$

This formula sheds light on the dependence of the redshift on the scalar field ϕ. On the one hand, the Einstein-frame metric coefficient g_{00} intricately depends on ϕ through Eq. (6.2). Taking the Newtonian limit of this equation helps clarify this dependence — Eq. (6.10) indeed shows that the potential $\Phi=-(g_{00}+1)/2$ obeys a modified Poisson equation where the scalar potential $V(\phi)$ is part of the source term alongside ρ. On the other hand, the presence of the conformal factor Ω is somewhat easier to interpret as it is merely a function of the scalar field [see Eqs. (1.110–1.112) for concrete examples of such functions]. Therefore, z is a *measurable* quantity that (a priori) depends on the field's amplitude at the emission and reception spacetime events. In Sec. 6.1.3, we shall expand Eq. (6.20) in the framework of the chameleon model and study the corresponding Newtonian limit.

As a side note, the above derivation of the redshift formula (6.20) in scalar-tensor theories also applies to the cosmological setting.[1] In particular, we retrieve Eq. (1.85) giving the redshift of a distant object in the sky (in scalar-tensor gravity). Therefore, the two following statements hold simultaneously:

1. Null geodesics are invariant under conformal transformations in a four-dimensional spacetime. Thus, light-like geodesics of the Jordan-frame metric $\tilde{g}_{\mu\nu}$ coincide with those of the Einstein-frame metric $g_{\mu\nu}$ and massless particles, such as photons, do not 'feel' any force from the scalar field.

2. The amount by which light emitted from distant objects gets redshifted (through the expansion of the universe) when it eventually gets to us depends explicitly on the scalar field and its cosmological evolution.

This source of confusion had to be clarified. For instance, authors of Ref. [302] use the wrong formula $1+z=a_{\rm rec}/a_{\rm em}$ in their study of the cosmological gravitational redshift in clusters of galaxies in the symmetron and Hu–Sawicky $f(R)$ models.[2]

6.1.2 Link with observable quantities

Parameterized redshift tests

Experiments that measure the gravitational frequency shift of light usually introduce a dimensionless parameter γ to quantify deviations from what is predicted by GR.[3] As such, γ is *defined* as

$$z_{12}=(1+\gamma)\Delta_{12}\mathfrak{U}, \tag{6.21}$$

where, for two locations, $\Delta_{12}\mathfrak{U}=\mathfrak{U}_2-\mathfrak{U}_1$. As mentioned in Sec. 1.1.1, testing the consistency of γ with 0 is a test of LPI — which is of course embedded in GR, but not exclusively. Current upper bounds on $|\gamma|$ are around 10^{-5} (see Sec. 6.2.1 thereafter). In Eq. (6.21), $\mathfrak{U}$ is either referred to as "the Newtonian potential" [115, 303], or as the "gravitational potential" [26, 27, 304–306]. The bothering issue with this designation is that it is unclear how one should actually define and measure it. Along the lines of Will's book [1], we define U as the "gravitational potential whose gradient is related to the test-body acceleration", i.e. in the Newtonian limit,

$$\mathbf{a}=-\boldsymbol{\nabla}\mathfrak{U}\,. \tag{6.22}$$

Acceleration $\boldsymbol{\nabla}\mathfrak{U}$ and redshift z can be measured with accelerometers and clocks respectively, and

$$\Delta_{12}\mathfrak{U}=\int_{\mathcal{C}}\mathbf{a}\cdot\mathrm{d}\mathbf{l}\,,\ \forall\mathcal{C}\text{ joining points 1 and 2}\,. \tag{6.23}$$

[1] One may rightly object that the FLRW metric (1.25) is not stationary and so the redshift derivation we just conducted does not apply. Nevertheless, this stationarity assumption is stronger than needed. It is in fact sufficient to note that ∂_η is a *conformal Killing vector* on the FLRW spacetime manifold, that is tangent to the world lines of the source and of the observer (comoving with the Hubble flow), so that Eq. (6.19) still applies. Here, η denotes the conformal time which is related to the coordinate time through $\mathrm{d}t=a(t)\mathrm{d}\eta$ and a is the scale factor. See e.g. Ref. [301] for more insights into these mathematical considerations.

[2] There are several factors which cause the measured gravitational redshift to be different from GR's prediction in this setup. In particular, fifth force effects result in different matter distribution within clusters compared to GR, which in turn alters the redshift. This is precisely the aspect that is being studied in Ref. [302], which is different in nature from the physical effect central to the present chapter.

[3] This parameter is often denoted by α in the literature. However, we use γ here in order to avoid confusion with the dimensionless parameter appearing in the dimensionless version of the chameleon Klein–Gordon equation (4.3).

Hence, definitions (6.21–6.23) are a check of the consistency between clock comparisons and acceleration measurements. If the separation between point 1 and point 2 is relatively small compared to the characteristic length scale of $\mathfrak{U}$-variations, Eq. (6.23) can be simplified to $\Delta_{12}\mathfrak{U} = \mathbf{g}\cdot\mathbf{r}_{12}$, where $\mathbf{g}$ is the gravitational field and $\mathbf{r}_{12}$ is the vector joining the two positions.[4] If LPI holds, then $\gamma = 0$. In particular $\gamma_{\rm GR} = 0$ (at the first post-Newtonian order).

Newtonian limit in the Jordan frame

Let us now show that scalar-tensor theories that fall into the class of models introduced in Sec. 1.1.2 (the so-called 'traditional' class) also verify $\gamma_{\rm ST} = 0$. In the Newtonian limit, the Jordan-frame metric can be put in the form (6.9) with $|\tilde{\Phi}| \ll 1$. Substituting this definition in Eq. (6.20) yields, at first order,

$$z = \sqrt{\frac{1+2\tilde{\Phi}_2}{1+2\tilde{\Phi}_1}} - 1 \simeq \tilde{\Phi}_2 - \tilde{\Phi}_1 = \Delta_{12}\tilde{\Phi}\,, \tag{6.24}$$

where we have further re-labeled by 1 and 2 the emission and reception events respectively. There only remains to check that $\mathfrak{U} \equiv \tilde{\Phi}$. This is done by looking at the geodesic equation in the Newtonian limit in the case of a stationary metric. The derivation of the latter follows from Eqs. (1.14–1.17), and we end up with

$$\frac{\mathrm{d}^2 x^i}{\mathrm{d}t^2} \simeq -\partial_i\tilde{\Phi} \implies \mathfrak{U} = \tilde{\Phi} \implies \gamma_{\rm ST} = 0\,. \tag{6.25}$$

A few remarks are in order. First, the Jordan frame scalar field φ does not appear explicitly Eqs. (6.9, 6.25). However, recall that the potential $\tilde{\Phi}$ is obtained through the field equations (6.11, 6.13), which of course depend on the scalar field. In other words, $\tilde{\Phi}$ cannot be considered as the "Newtonian potential" in the sense that it does not obey the usual Poisson's equation in the static regime. Instead, Eqs. (6.11, 6.13) remain coupled second-order partial differential equations. Second, finding $\gamma_{\rm ST} = 0$ should not come as a surprise at all. Indeed, conformally coupled scalar-tensor models belong to the wider class of *metric theories*, which all satisfy the three pillars of the EEP — namely the WEP, LLI and LPI (see Fig. 1.1). One may however raise the objection that there is no point in trying to use redshift measurements to constrain scalar-tensor gravity, since in particular the latter satisfies LPI and is thus consistent with all bounds on the parameter γ. In the face of this argument, we stress that it is not because a given theory satisfies LPI that it cannot be distinguished from GR in redshift experiments. This point will be made irrefutable in Sec. 6.2.2.

Newtonian limit in the Einstein frame

Similarly, we derive the Newtonian limit of the redshift expression (6.20) in the Einstein frame. For the Newtonian approximations in both frames to be consistent with one another, we need to further assume that the conformal function Ω is close to unity, see Sec. 1.1.2. We write this down as

$$\Omega(\phi) = 1 + \omega(\phi)\,, \qquad \text{with} \quad |\omega(\phi)| \ll 1\,. \tag{6.26}$$

On the other hand, the Einstein-frame metric is put in the Newtonian gauge (6.8) with $|\Phi| \ll 1$. Note that these assumptions allows for the identification $\tilde{\Phi} \simeq \Phi + \omega(\phi)$. We then immediately obtain

$$z = \Delta_{12}\big[\Phi + \omega(\phi)\big]\,. \tag{6.27}$$

The Newtonian limit of the geodesic equation in the Einstein frame is

$$\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}t^2} = -\boldsymbol{\nabla}\Phi - \boldsymbol{\nabla}\big[\ln\Omega(\phi)\big] \simeq -\boldsymbol{\nabla}\big[\Phi + \omega(\phi)\big]\,,$$

and so we recover the fact that $\mathfrak{U} = \Phi + \omega(\phi)$ and $\gamma_{ST} = 0$, as expected. Φ and ϕ are solutions to Eq. (6.10) and Eq. (6.12) respectively.

Again, let us make some remarks:

- The definition of the densities $\tilde{\rho}$ and ρ is explained in Box B. Here, assumption (6.26) allows us to approximate $\tilde{\rho} \simeq \rho$, terms of the form $\rho\omega(\phi)$ being considered as higher order terms in the field equations.
- All subsequent computations can be conducted in the Einstein frame. Essentially, in order to be able to discuss the redshift, we will have to solve the modified Poisson equation (6.10) and the Klein–Gordon equation (6.12). In particular, the latter does not depend on Φ and should thus be solved first, yielding ϕ.

[4] This approximation is performed for laboratory experiments on Earth, see e.g. Refs. [305, 307].

Only then can we tackle the modified Poisson equation, because the source term $V(\phi)$ is fully determined after completion of the first step.

6.1.3 Focus on the chameleon model

We now focus on the chameleon model with $n > 0$, given by the functions

$$\Omega(\phi) = \exp\left(\frac{\beta\phi}{M_{\rm Pl}}\right) \qquad \text{and} \qquad V(\phi) = \Lambda^4\left(\frac{\Lambda}{\phi}\right)^n . \tag{6.28}$$

In the study of fifth force effects, the bare potential function V could be defined up to an additive constant since it only played a role through its derivatives in computations. Things are different here since $V(\phi)$ appears as is in Eq. (6.10). As long as $\phi \ll M_{\rm Pl}/\beta$, assumption (6.26) holds and we get $\omega(\phi) = \beta\phi/M_{\rm Pl}$. To avoid constantly having to refer to Chapt. 1, we recall that the scalar field obeys a nonlinear Klein–Gordon equation

$$\Delta\phi = \frac{\beta}{M_{\rm Pl}}\rho - n\frac{\Lambda^{n+4}}{\phi^{n+1}} .$$

Moreover, the field's value that minimizes the effective potential together with the effective mass are given by

$$\phi_{\rm min}(\rho) = \left(M_{\rm Pl}\frac{n\Lambda^{n+4}}{\beta\rho}\right)^{\frac{1}{n+1}} \qquad \text{and} \qquad m_\phi^2(\rho) = n(n+1)\Lambda^{n+4}\left(\frac{\beta\rho}{nM_{\rm Pl}\Lambda^{n+4}}\right)^{\frac{n+2}{n+1}} . \tag{6.29}$$

The effective Compton wavelength λ_ϕ is related to the effective mass through $\lambda_\phi = m_\phi^{-1}$.

For this specific scalar-tensor model, the redshift expression (6.27) becomes

$$z = \Delta_{12}\left[\Phi + \frac{\beta\phi}{M_{\rm Pl}}\right] . \tag{6.30}$$

There, it is already interesting to note that, unlike the chameleonic force which is proportional to the gradient of the scalar field $\boldsymbol{\nabla}\phi$, part of the chameleon contribution to the total redshift is proportional to $\Delta_{12}\phi = \phi_2 - \phi_1$. This mere observation has important consequences in terms of choice of experimental designs when it comes to constraining such a model. Maximizing $\|\boldsymbol{\nabla}\phi\|$ is undeniably not the same thing as maximizing $\Delta_{12}\phi$.[5] The more intricate ϕ-dependence through Φ shall be examined in more details in Sec. 6.2.2.

For the scalar field to leave a measurable imprint on the total redshift (6.30), the scalar field must be able to vary significantly from one place to another. In that respect, the chameleon field specifically may actually be a very good candidate. Indeed, Eq. (6.29) highlights the fact that $\phi_{\rm min}(\rho) \propto 1/\rho^{1/(n+1)}$. In particular, $\phi_{\rm min}(\rho) \to +\infty$ as $\rho \to 0$. Moreover, we have seen in Sec. 1.2.2 that, deep inside a medium of constant density ρ, $\phi \sim \phi_{\rm min}(\rho)$ provided this medium occupies a large enough spatial region. All in all, this means that the chameleon field value should grow to very large values in vast-enough, low-density environments.

6.2 Thought experiment and orders of magnitude

So far, we have derived the redshift formula in scalar-tensor theories and showed its dependence on the scalar field. However, we have yet to show how to translate redshift measurements into actual constraints on the scalar-tensor model at stake. Following on from the previous section, this discussion is illustrated with the example of the chameleon field again. After briefly reviewing the current state of the art in atomic clocks, we propose a thought experiment, underlying more realistic experimental designs, for testing chameleon gravity.

6.2.1 State of the art in atomic clocks

Measuring the gravitational redshift effect on Earth is best achieved by atomic clocks. Indeed, these devices represent the pinnacle of precision timekeeping, playing a critical role in fundamental physics experiments and underlying the definition of the second in the International System of Units. They rely on the ultra-stable oscillations of atoms to measure time with unparalleled accuracy. Among the most advanced types are optical lattice clocks which probe the hyperfine transitions of trapped ions or atoms with laser light. They achieve relative frequency precisions of 10^{-18} and below [308–312], down to 7.6×10^{-21} [313].

These levels of instabilities and inaccuracies open the way to stringent tests of GR, notably by putting upper bounds on parametrized tests of gravitational redshift (see Secs. 1.1.1 and 6.1.2). In space, comparing the frequency of hydrogen masers onboard Galileo satellites with eccentric orbits have produced the strongest

[5] The mean value theorem nonetheless establishes a link between these two quantities.

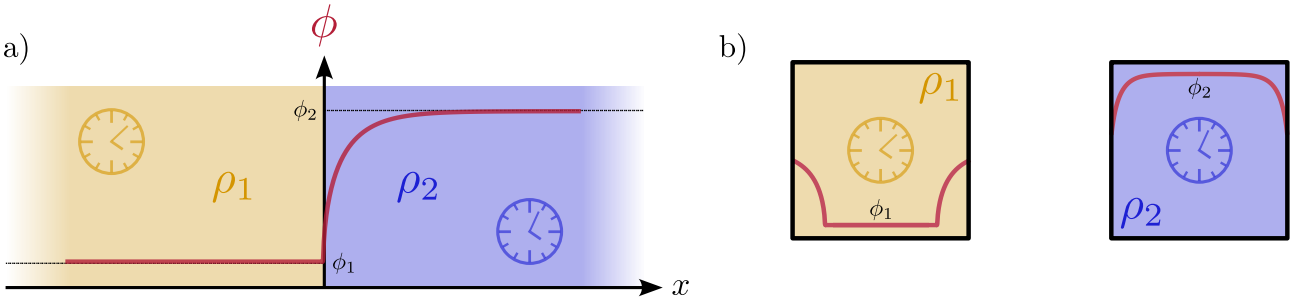


Figure 6.2: A redshift thought experiment with the chameleon model.

limits on deviations from the expected redshift — at the 10^{-5} level on γ defined by Eq. (6.21) [26, 27].[6] The ACES mission [303], to be launched in 2025, aims at improving that bound by one order of magnitude. On Earth, 18-digit-precision frequency comparison between optical lattice clocks produces nearly as tight bounds on this test [304] (see Ref. [307] for future prospects). Aside from testing LPI, atomic clocks underlie the field of relativistic geodesy as they can probe the geopotential at the sub-centimeter scale, see e.g. Refs. [309, 313–315].

The orders of magnitude mentioned in the above paragraphs will serve as benchmarks when we discuss redshift-based tests of the chameleon model, namely in Secs. 6.2.2 and 6.3. Specifically, we will assess how these precision levels translate into constraints in the parameter space of the model.

6.2.2 A first *Gedankenexperiment*

Here, we imagine a toy experiment to prove the point we made in Sec. 6.1.2, namely that it is possible to distinguish a scalar-tensor theory complying with LPI from GR by means of redshift measurements.

Setup a)

Let us start with a basic setup where space is divided into two regions separated by a plane. Region 1 is filled with a fluid of density ρ_1 while region 2 is filled with another fluid of density ρ_2. In such a configuration, the chameleon field is expected to vary significantly nearby the transition between the two regions, and relax to $\phi_{\rm min}(\rho_i) \equiv \phi_i$ in the $i^{\rm th}$ region, $i \in \{1, 2\}$. This configuration is depicted in the left-hand panel of Fig. 6.2, for $\rho_1 \gg \rho_2$. Suppose that we supplement this basic setup with two clocks, one in each of the two regions, placed sufficiently far away from the median plane so that the field does not vary much anymore. Put another way, clock 1 is immersed in a region of space where $\phi = \phi_1$ while clock 2 is immersed in a region of space where $\phi = \phi_2$. We further assume, for now, that the clocks are *perfect* in the sense that they do not perturb the scalar field profile at all.

In this scenario, the two clocks will tick at different frequencies, and the relative frequency shift is given by

$$\frac{f_1 - f_2}{f_2} = \frac{{\rm d}\tilde{\tau}_2}{{\rm d}\tilde{\tau}_1} - 1 = z = \frac{\Omega_2}{\Omega_1}\sqrt{\frac{(g_{00})_2}{(g_{00})_1}} - 1 \simeq \Delta_{12}\left(\Phi + \frac{\beta\phi}{M_{\rm Pl}}\right) , \tag{6.31}$$

where $\tilde{\tau}_i$ denotes the proper time experienced by the $i^{\rm th}$ clock. As discussed above, the scalar field has two contributions in this expression: an explicit one, with the $\beta\phi/M_{\rm Pl}$ term, and a hidden one, with the Φ term through Eq. (6.10). These two contributions are first discussed analytically. On the one hand, since the scalar field is given by Eq. (6.29) at the clocks' position (by hypothesis), we have

$$\Delta_{12}\left(\frac{\beta\phi}{M_{\rm Pl}}\right) = \left[\left(\frac{\beta}{M_{\rm Pl}}\right)^n n\Lambda^{n+4}\right]^{\frac{1}{n+1}} \Delta_{12}\left(\rho^{-\frac{1}{n+1}}\right) . \tag{6.32}$$

Eq. (6.32) encapsulates the central idea of this thought experiment. Indeed, it is easy to see that the scalar field contribution can virtually be made as large as one desires in the limit $\rho_2 \to 0$ (for any fixed density ρ_1). On the other hand, it is somewhat harder to get an analytical estimate of the scalar field contribution to the potential Φ. Worse, the left panel of Fig. 6.2 depicts two infinite half-spaces of constant density each, so that Eq. (6.10) boils down to

$$2M_{\rm Pl}^2\,\Phi''(x) = \begin{cases} \rho_1 - 2V\big(\phi(x)\big) & \text{if } x < 0\,, \\ \rho_2 - 2V\big(\phi(x)\big) & \text{if } x > 0\,. \end{cases}$$

The lack of obvious physical boundary conditions in this case makes this ODE problem ill-posed.

[6] Note that this new bound is approximately one order of magnitude lower than the Vessot-Levine experiment [25].

Setup b)

In order to circumvent this thorny issue, we consider a slightly more realistic experimental design where the two clocks are put into separate boxes filled with materials of density ρ_1 and ρ_2 (in an otherwise vacuum medium), as shown in the right panel of Fig. 6.2. Provided that the boxes are big enough for the field to reach $\phi_{\rm min}$ in their interior, the previously exposed qualitative arguments of this thought experiment shall remain valid. Back to the estimation of Φ, the linearity of Eq. (6.10) allows us to decompose this potential as $\Phi = \Phi_N + \delta\Phi_V$, where Φ_N and $\delta\Phi_V$ are solutions to

$$2M_{\rm Pl}^2\,\Delta\Phi_N = \rho \qquad \text{and} \qquad M_{\rm Pl}^2\,\Delta\delta\Phi_V = -V(\phi)\,, \tag{6.33}$$

respectively. By doing so, we can rewrite the total redshift z as $z = z_N + z_\phi$ with

$$z_N = \Delta_{12}\Phi_N \qquad \text{and} \qquad z_\phi = \Delta_{12}\left(\frac{\beta\phi}{M_{\rm Pl}}\right) + \Delta_{12}\delta\Phi_V\,. \tag{6.34}$$

This decomposition is convenient for physical interpretation because it separates the contribution of the chameleon field from that of the Newtonian potential Φ_N. Yet, we crucially need an estimate of $\Delta_{12}\delta\Phi_V$ for comparison against the contribution given by Eq. (6.32). To this end, let us make the simplifying assumptions that *(i)* the boxes are spherical with radius $R_{\rm box}$ and *(ii)* the boxes are *screened* and exhibit a *thin-shell* of negligible thickness — i.e. the scalar field sits at $\phi_{\rm min}$ in most of the spherical boxes. Then, using Green's function for the Laplacian at the geometrical center of the boxes yields

$$\Phi_N = -\left(\frac{R_{\rm box}}{2M_{\rm Pl}}\right)^2\rho_i \qquad \text{and} \qquad \delta\Phi_V = \frac{1}{2}\left(\frac{R_{\rm box}}{M_{\rm Pl}}\right)^2 V\big(\phi_{\rm min}(\rho_i)\big)\,, \quad i\in\{1,\,2\}\,. \tag{6.35}$$

With these approximations at hands, we get

$$z_N = -\left(\frac{R_{\rm box}}{2M_{\rm Pl}}\right)^2\Delta_{12}\rho\,, \tag{6.36}$$

$$z_\phi = M_{\rm Pl}^{-\frac{n}{n+1}}(n\beta^n\Lambda^{n+4})^{\frac{1}{n+1}}\Delta_{12}\Big(\rho^{-\frac{1}{n+1}}\Big) + \frac{R_{\rm box}^2}{2}M_{\rm Pl}^{-\frac{3n+2}{n+1}}(n\beta^n\Lambda^{n+4})^{\frac{1}{n+1}}\Delta_{12}\big(\rho^{\frac{n}{n+1}}\big)\,. \tag{6.37}$$

Recalling that $M_{\rm Pl}\simeq 2.4\times10^{27}\,$eV in natural units, we can safely assume that the contribution (6.32) is the dominant term in z_ϕ.[7] Consequently, the $\Delta_{12}\delta\Phi_V$ term is not retained in our subsequent analysis. Besides, the Newtonian contribution z_N to the total redshift is not expected to be overwhelmingly larger than z_ϕ, as (i) it is weighted by $M_{\rm Pl}^{-2}$, and (ii) the $\Delta_{12}\rho$ term cannot be made as large as $\Delta_{12}\rho^{-1/(n+1)}$ can be.[8]

Now, we have not yet explained how this kind of experiment could be translated into constraints on the chameleon model. Here is a proposal of a well-posed experiment:

1. We start with the two boxes filled with the same higher density material ρ_1. At first, there is no reason for the clocks to be synchronized as they could be at slightly different altitudes within the geopotential,[9] so we adjust their relative height so that $z\equiv 0$.

2. Then using pumps, we replace ρ_1 by ρ_2 (with $\rho_2\ll\rho_1$) in one of the two boxes.

3. The frequency shift between the two clocks is measured again, without moving the boxes. In pure GR, the removed mass from the box affects the redshift through its Newtonian potential, which can readily be estimated. In scalar-tensor gravity, one has to further take into account the scalar field contribution $z_\phi\propto\Delta_{12}\phi$. The measured redshift together with its uncertainty can be used to put upper bounds on $|z_\phi|$, which in turn constrains the underlying scalar-tensor model.

We stress the importance of precisely defining a protocol — altitudes, for instance, cannot be assumed to be known.

Optimal constraints

Given the above computations and discussion, we shall approximate the scalar contribution to the total redshift by Eq. (6.32), or put more simply $\beta\Delta_{12}\phi_{\rm min}/M_{\rm Pl}$. We consider three pairs of 'materials' to fill the boxes depicted

[7] This has also been numerically verified by computing the two contributions of Eq. (6.37) for the $(\rho_1,\,\rho_2)$ pairs that we consider thereafter.

[8] The densest materials we can find on Earth have density that do not exceed a few $10^4\,$kg/m^3. Conversely, we are able to achieve high vacuum levels in vacuum chambers.

[9] As a side note, nowadays we are able to resolve the gravitational potential of the Earth at the millimeter scale [313].

Material designation	Density (kg/m^3)	Density (eV^4)
Lead	11.4×10^3	4.9×10^{19}
Water	10^3	4.3×10^{18}
Air	1.225	5.3×10^{15}
UHV	10^{-10}	4.3×10^5
XHV	10^{-15}	4.3
IPM	10^{-20}	4.3×10^{-5}

Table 6.2: Typical materials together with their densities (in kg/m^3 and in eV^4) considered in Sec. 6.2.2. 'UHV' and 'XHV' stand for *ultra-high vacuum* and *extremely-high vacuum*, and can be produced in the laboratory using sophisticated vacuum chambers. 'IPM' stands for *interplanetary medium* and represents the thinly scattered matter that exists between the planets and other large bodies of the Solar system. To put Fig. 4.7 into perspective, the density at the geostationary altitude is roughly $\sim O(10^{-19}\,\text{kg/m}^3)$.

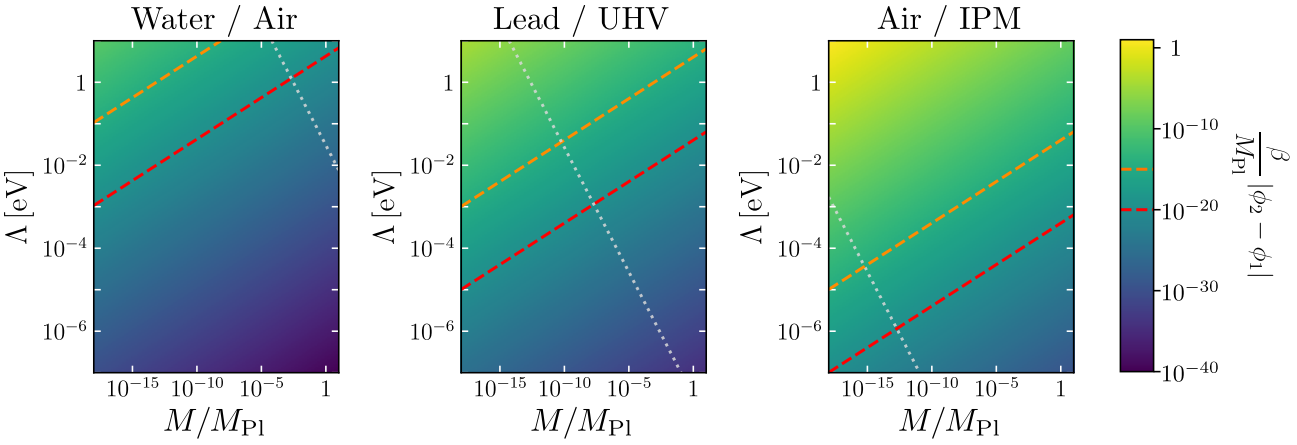


Figure 6.3: Expected redshift from the chameleon field contribution Eq. (6.32) for different pairs of materials. The chosen densities are reported in Table 6.2. 'UHV' stands for *ultra-high vacuum* and corresponds to the vacuum level reachable in vacuum cavities while 'IPM' stands for *interplanetary medium*. The orange and red dashed lines correspond to iso-redshifts at levels $\varepsilon_{\rm rel} = 10^{-15}$ and $\varepsilon_{\rm rel} = 10^{-20}$ respectively. The gray dotted line corresponds to $\lambda_c(\rho_2) = 1\,\text{m}$, where ρ_2 refers to the density of the less dense material of each pair: the Compton wavelength is larger (resp. smaller) than 1 meter above (resp. below) this line. The x-axis is $M/M_{\rm Pl} \equiv 1/\beta$ to be in line with the exclusion plots found in the literature, see e.g. Refs. [71, 150–152].

in the right panel of Fig. 6.2: (Water / Air), (Lead / UHV) and (Air / IPM). Here 'UHV' stands for *ultra-high vacuum* and corresponds to a vacuum level reachable in vacuum chambers, while 'IPM' stands for *interplanetary medium* [316]. The associated densities are reported in Table 6.2, in SI units (kg/m^3) and in natural units (eV^4).

In Fig. 6.3, we represent the scalar field contribution to the redshift, z_ϕ, in the (β^{-1}, Λ)-plane[10] ($n = 1$) for the three aforementioned pairs of materials. The bounds on (β, Λ) are chosen large enough for the redshift to cover many orders of magnitudes, ranging from ~ 1 to 10^{-40}. To put Sec. 6.2.1 into perspective, we depict by orange and red dashed lines the iso-redshift levels at $\varepsilon_{\rm rel} = 10^{-15}$ and $\varepsilon_{\rm rel} = 10^{-20}$ respectively, which correspond to achievable relative precisions with current atomic clocks on Earth. As it can be guessed from Eq. (6.32), these iso-levels map to straight lines in the parameter space with log-scaled axes. Unsurprisingly, the more precise the atomic clock is, the smaller the measurable redshift and so the tighter the potential constraints on the chameleon parameter space. Additionally, it is worth noting that high density materials (water, lead) on the one hand, and low density materials (UHV, IPM) on the other hand, do not play symmetrical roles at all. Because of the dependence $\phi_{\rm min} \propto \rho^{-1/(n+1)}$, the lower-density material has more weight on the redshift. In plain language, lowering ρ_2 by one order of magnitude at fixed ρ_1 results in an increase of the redshift much greater than if we were to increase ρ_1 by one order of magnitude at fixed ρ_2. As a matter of fact, replacing the (Air / IPM) pair by (Lead / IPM) would not have any visible effect on the right panel of Fig. 6.3.

To a certain extent, these order-of-magnitude forecasts justify the present study since it appears we can gain at least several orders of magnitude with respect to current constraints from laboratory experiments (see Fig. 4 from Ref. [152]). They can be considered as *optimal* constraints because for any two densities (ρ_1, ρ_2), $|\phi_{\rm min}(\rho_2) - \phi_{\rm min}(\rho_1)|$ represents an upper bound for $\Delta_{12}\phi$. Overall, the potential constraints outlined in Fig. 6.3 remain overly optimistic for several reasons:

[10] Most references in the literature do indeed use $M = M_{\rm Pl}/\beta$ in place of β in their exclusion plots, see e.g. Ref. [152].

1. they rely on the best atomic clocks ever built, which may not be well-suited for the experimental design one ends up choosing, which is all the more true if one thinks of space-borne experiments;
2. we have assumed vacuum levels hardly reachable on Earth (especially the IPM density, see Table 6.2);
3. we have assumed that the boxes were large enough in size for the scalar field to reach $\phi_{\rm min}$ at their center;
4. we have assumed that the atomic clock itself does not perturb the scalar field profile inside the box, which is not realistic.

The main goal of the remainder of this study is to take these points into account, and see whether some experimental design could realistically produce competitive constraints on the chameleon model. Specifically, points 2 and 3 are discussed next, while points 1 and 4 shall be addressed in Sec. 6.3.2 and Sec. 6.3.1 respectively, partly through numerical simulations.

Going to vacuum

Let us first comment the limit $\rho_2 \to 0$. Assuming that the chameleon field is indeed able to track the minimum of its effective potential, the scalar contribution to the redshift tends to infinity. In the face of this rather unphysical outcome, we have to take a closer look at the various assumptions that led to it.

First, no vacuum is truly perfect, thus the limit $\rho_2 \to 0$ should be replaced by $\rho_2 \to \rho_* > 0$. In the laboratory, vacuum is primarily measured by its absolute pressure, which can be translated into a density provided that other parameters (such as temperature or chemical composition) have been determined. Vacuum tubes typically have $\sim 10^8$ particles per cm^3, while cryopumped MBE[11] chambers can go down to densities as low as $\sim 10^5$ particles per cm^3.[12] Outer space gets even closer to 'true' vacuum. Far enough from the Earth, at the altitude of geostationary satellites, the density of residual atmosphere neighbors $4 \times 10^{-19}\,$kg/m^3. Density keeps decreasing as we go to interplanetary space (~ 11 molecules per cm^3), interstellar space (~ 1 particle per cm^3), and eventually intergalactic space ($\sim 10^{-6}$ particle per cm^3) [316, 317]. The latter is the closest physical approximation of a perfect vacuum, with a density of $\sim 10^{-27}\,$kg/m^3 assuming particles the mass of hydrogen. Ultimately, even if every matter particle could somehow be removed from a given volume, quantum fluctuations ensure that the energy contained in it is never quite zero, and so the chameleon does not diverge to $+\infty$.

Two remarks have to be made regarding the above:

1. Speaking of matter density on inter-galactic scales, the background value of the scalar field should match that predicted by its cosmological evolution. The latter can actually be smaller than $\phi_{\rm min}(\rho_{\rm vac})$, see Eq. (1.124) and the corresponding discussion there.
2. Considering such rarefied environments (e.g. few thousands of particles per cubic meter) raises the question of the legitimacy of *averaging* the density. Loosely speaking, does the chameleon field 'perceive' a collection of isolated N particles in the same way as a homogeneous medium? To our knowledge, the only work that examine this problem is Ref. [153]. There, the authors find, on the basis of analytical approximations, that the macroscopic Compton wavelength $\langle\lambda_\phi\rangle$ of the chameleon inside a screened body that is itself made of individual particles is
$$\langle\lambda_\phi\rangle = \max\Big(m_\phi^{-1}\big(\langle\rho\rangle\big),\, m_{\rm crit}^{-1}\Big)\,,$$
where $\langle\rho\rangle$ denotes the average density of the body at stake while $m_{\rm crit}$ is a quantity depending solely on its microscopic properties and n. However, no insight is provided regarding the mean value of the chameleon field. This question could be investigated in more details numerically with *femtoscope* by using a representative volume element as the simulation box, with periodic boundary conditions. Such an investigation is left for future work.

Second, we have to verify that the expansion of the conformal factor around 1 [Eq. (6.26)], which we have assumed in the derivation of the Newtonian limits of the redshift, holds. Since the maximum value of the chameleon field is given by $\phi_{\rm min}(\rho_{\rm min})$ [see Eq. (6.29)], where $\rho_{\rm min}$ denotes the minimum density in the spatial region of interest, the condition $\phi \ll M_{\rm Pl}/\beta$ translates to
$$\rho_{\rm min} \gg n\Lambda^{n+4}\left(\frac{\beta}{M_{\rm Pl}}\right)^n . \tag{6.38}$$
This condition is easily met for the $(\beta,\,\Lambda)$ ranges and material density considered in Fig. 6.3 — except in the top left corner of the parameter space ($\beta = 10^{18}$, $\Lambda = 10\,$eV) for which the rhs of Eq. (6.38) reaches $\sim 10^{-20}\,$kg/m^3,

[11] Molecular-beam epitaxy.

[12] Assuming air with an average molar mass of 29 g/mol, this corresponds to densities of $5 \times 10^{-12}\,$kg/m^3 for the vacuum tube and $5 \times 10^{-15}\,$kg/m^3 for the cryopumped MBE chamber. This is in line with Ref. [142] which assumes a density of $10^{-14}\,$kg/m^3 inside a vacuum chamber.

which corresponds to the IPM density. This zone of the parameter space is already well-constrained and thus not very relevant anyway. Finally, it should be reminded that this approximation was primarily performed for expanding the ratio of conformal factors $\Omega_{\rm rec}/\Omega_{\rm em}$ in Eq. (6.20). Without this approximation and all other things being equal, the successive implications

$$\rho_2 \to 0 \implies \phi_{\rm min}(\rho_2) \to +\infty \implies \frac{\Omega_2}{\Omega_1} \to +\infty$$

remain true.

Constraints with finite-size boxes

A more debatable hypothesis is the one which states that the scalar field indeed reaches $\phi_{\rm min}$ at the center of the box — as depicted in the right panel of Fig. 6.2. It is well-known that this situation does not arise when an object is *unscreened*. One relevant quantity to qualitatively assess whether the box is screened (as desired) or not is the Compton wavelength

$$\lambda_\phi(\rho) = m_\phi^{-1}(\rho) = \sqrt{\frac{1}{n(n+1)\Lambda^{n+4}}\left(M_{\rm Pl}\frac{n\Lambda^{n+4}}{\beta\rho}\right)^{\frac{n+2}{n+1}}}\,. \tag{6.39}$$

Typically, we would expect the box to have a radius at least a few Compton wavelengths in size, so that our assumption is fulfilled.[13] Yet, as $\rho_2 \to 0$, $\lambda_\phi(\rho_2) \to +\infty$, meaning that the chameleon field would not have enough space to reach $\phi_{\rm min}(\rho_2)$ within any finite-size box. In Fig. 6.3, we have represented in silver dotted lines the iso Compton wavelength $\lambda_\phi(\rho_2) = 1\,{\rm m}$ in the (β^{-1}, Λ)-plane, where ρ_2 refers to the density of the less dense material of each pair. Above (resp. below) this line, $\lambda_\phi(\rho_2) > 1\,{\rm m}$ (resp. $\lambda_\phi(\rho_2) < 1\,{\rm m}$). We chose to show the one meter reference as it corresponds to the typical size of objects found in the laboratory. As ρ_2 decreases from $\rho_{\rm air}$ to $\rho_{\rm UHV}$ and to $\rho_{\rm IPM}$, the portion of the parameter space for which the Compton wavelength is smaller than 1 m shrinks to the bottom-left corner. In other words, there is a trade-off to be made between (*i*) maximizing z_ϕ on the one hand, and (*ii*) making sure that the experiment is sensitive to a wide enough area of the parameter space on the other hand. The former condition is an incentive to aim for the best possible level of vacuum for ρ_2, while the latter condition requires the two boxes to be sufficiently dense for otherwise the field will not reach $\phi_{\rm min}$ at their center.

We thus need to revise the forecasts presented in Fig. 6.3 by accounting for the fact that the boxes containing the atomic clocks are finite in size. The best solution we found to this constrained optimization problem is to make ρ_2 vary continuously, and combine all the resulting constraints together. By doing so, we can derive the best constraints for the relevant β- and Λ-ranges. These *weaker* (but more realistic) forecasts for laboratory experiments are obtained by solving the algebraic system

$$\begin{cases} \lambda_\phi(\rho_2) = R_{\rm box} = 1\,{\rm m} \\ \beta\Delta_{12}(\phi_{\rm min})/M_{\rm Pl} = \varepsilon_{\rm rel} \end{cases}, \tag{6.40}$$

for (β, Λ), where $\varepsilon_{\rm rel} \in \{10^{-15}, 10^{-20}\}$ denotes the atomic clock relative precision. The resulting bounds, which turn out to be straight lines in log space, are shown in Fig. 6.4. These revised bounds exhibit a steeper slope, meaning that high-M (or equivalently, low-β) regions are more difficult to constrain than what Fig. 6.3 suggested.

6.3 Towards more realistic experimental designs

In the previous section, we imagined an idealized setup whereby atomic clocks are placed in different chameleon field backgrounds, which is achieved by adjusting the density of the medium in which they are immersed. The scalar field contribution z_ϕ to the total redshift between the two clocks is dominated by $\Delta_{12}(\beta\phi/M_{\rm Pl})$, while the Newtonian contribution $z_N = \Delta_{12}\Phi_N$ can readily be estimated (by calculation) since the mass content inside each box is assumed to be well-controlled. Using rough orders of magnitude, we evaluated the constraining power of such an experiment on the chameleon model.

The most controversial assumption we still have not discussed is the backreaction of the atomic clocks themselves on the scalar field profile. So far, we considered that the clocks were somehow 'transparent' to the gravitational fields (scalar and metric in the Einstein frame), in the sense that the former would not significantly perturb background values of the latter. In GR, this is most likely true as the geopotential is overwhelmingly

[13] Of course, the Compton wavelength alone is not sufficient to determine whether an object is screened or not. The density of the background medium in which it is embedded must also be taken into account — remark that the thin shell parameter (1.122) depends on $\phi_{\rm min}(\rho_{\rm background})$, or see e.g. Ref. [149] for a more accurate screening criterion.

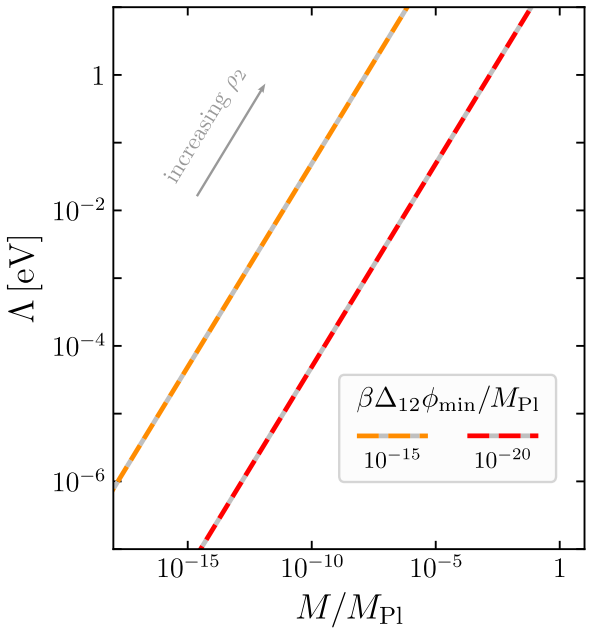


Figure 6.4: Expected constraints on the chameleon from redshift measurements for $\varepsilon_{\rm rel} \in \{10^{-15}, 10^{-20}\}$. This corresponds to revised bounds on parameters (M, Λ) compared to what is presented in Fig. 6.3 by accounting for the finite box sizes (1 m).

dominant over, say, laboratory-scale objects — this is because gravity is mediated by a *massless* spin-two particle. In chameleon gravity however, the nonlinearity and mass-changing properties of the scalar field mean that the atomic clocks can be screened in the setup described above. This issue is of the utmost importance as it is an experiment killer: in this scenario, the interior of the clocks, where atoms are being 'interrogated', becomes completely decoupled from the exterior, and therefore insensitive to the actual material filling the rest of the box. In that case, z_ϕ is expected to be essentially zero, meaning that the experiment cannot probe for the chameleon field.

In this section, we go a step further by taking these considerations into account. First, we examine more closely the implications of adding macroscopic atomic clocks to the model. We propose a more realistic redshift experiment in the laboratory that could be relevant for probing chameleons very strongly coupled to matter ($\beta \gtrsim 10^5$). Secondly, we revive the idea of space-based experiments as they could be sensitive to chameleons with gravitational strength coupling ($\beta \lesssim 10$).

6.3.1 Laboratory experiment [very high coupling]

Following on from Sec. 6.2.2, we study what happens to the forecasts outlined in Figs. 6.3 and 6.4.

Why the *Gedankenexperiment* does not work

An atomic clock relies on the interaction between two electron states in a given atom and some electromagnetic radiation. A group of atoms (e.g. cesium-133, rubidium-87 or strontium-88) is prepared in one energy state before being subjected to some monochromatic electromagnetic radiation, whose frequency is adjusted to match the targeted transition between the two energy states. Achieving this usually requires a whole apparatus, including a quartz crystal oscillator (in the case of microwave clocks), a frequency synthesizer, an atomic interrogation chamber, etc. The most precise clocks are therefore quite large objects in the laboratory,[14] where the meter is a good characteristic length scale. Nonetheless, the past two decades have witnessed the development of chip-scale atomic clocks, a few centimeters in size and demonstrating a fractional frequency instability of one part in 10^{13} [318].

However small the actual clocks used in our *gedankenexperiment*, they involve materials that are just too dense and too thick for it to be viable. Without going into too much details regarding the way an atomic clock is put together, it is conservative to assume that the average density of the apparatus is of the order of $\rho_{\rm water} = 10^3\,{\rm kg/m^3}$, with walls of thickness greater than 1 mm (even for the smaller chip-scale atomic clocks). In that respect, Fig. 6.5 provides insights into the various Compton wavelengths of the chameleon field involved in this experimental setup for the parameters $(M, \Lambda, n=1)$. First, we saw with Fig. 6.3 that the lower density material has to be such that $\rho_2 \lesssim \rho_{\rm UHV}$ for the experiment to yield interesting forecasts constraint-wise. At the

[14] Early on, they even used to be the size of an entire room!

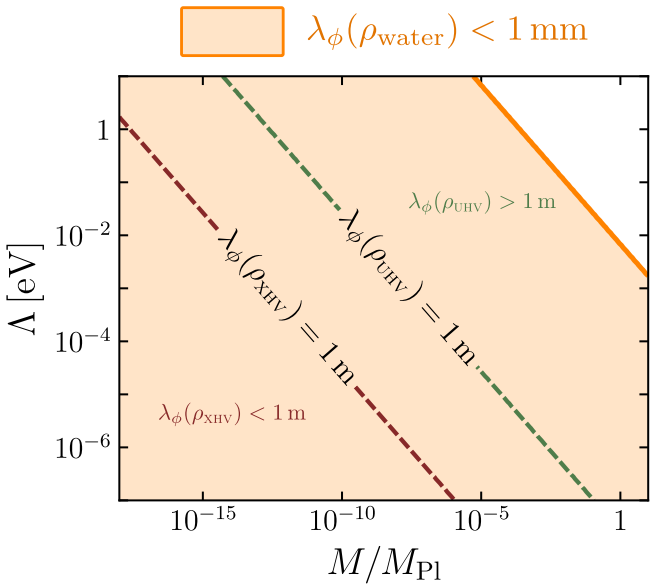


Figure 6.5: The problem of Compton wavelengths. In the chameleon parameter space (M, Λ, $n=1$), the two dashed lines represent the set of parameters that result in a one-meter Compton wavelength in the UHV and XHV vacua [see Eq. (6.39) and Table 6.2]. The orange shaded area maps to sub-millimeter Compton wavelength in water — $\rho_{\rm water} = 10^3\,{\rm kg/m^3}$ is representative of the typical density of materials found in the laboratory (including that of atomic clocks).

same time, the constraint of the scalar field reaching $\phi_{\rm min}(\rho_i)$, $i \in \{1, 2\}$ in finite-size boxes ($\sim 1\,{\rm m}$) considerably restricts the region of the parameter space that can actually be probed. The two dashed lines in Fig. 6.5 represent the set of parameters that result in a one-meter Compton wavelength for the UHV and XHV vacua, while the orange area maps to $\lambda_\phi(\rho_{\rm water}) < 1\,{\rm mm}$. An admissible region of the parameter space for the experiment to work would have to satisfy:

1. yield a Compton wavelength in the clock's walls greater than their thickness, i.e. outside the orange shaded area;

2. ensure that there is enough space in the box for the field to reach the value that minimizes the effective potential, i.e. below the dashed lines.

Unfortunately, the intersection of these two regions is empty. In the region where condition 2 holds, the clock is expected to be deeply screened, thereby jeopardizing the whole concept of the thought experiment. The lessons drawn from this first experimental concept will nevertheless prove to be useful for the following.

Alternative experimental design

There may be nonetheless ways to benefit from atomic clocks, if one agrees to modify the experimental setup initially envisioned. In Sec. 6.2.2, we insisted on the need for high vacuum levels — the less dense, the better — for probing yet-unconstrained regions of the chameleon parameter space. As it turns out, atomic clocks also require such ultra-high vacuum environments to operate in optimal conditions. Indeed, this reduces background gas collisions in the atomic interrogation chamber (where the atoms interact with the electromagnetic radiation), the latter being detrimental to frequency stability. This is true for both cesium / rubidium fountain clocks and for optical lattice clocks.

The idea is the following. We suppose that the science chamber is big enough for the chameleon field to reach $\phi_{\rm min}$ where the interrogated atoms sit. In order to modulate the scalar field they perceive, we cannot just increase the density inside the chamber as the atomic clock cannot operate correctly but in vacuum. Instead, we could imagine shrinking the chamber's size in order to bring its walls closer to atoms. The walls being dense and screened, this would effectively lower the chameleon field the atoms experience.

There are several shortcomings in this picture. A redshift measurement is a relative comparison between two frequencies, so we would need two clocks as before. One way to single out z_ϕ would be (*i*) to start with two identical clocks with a 'large' vacuum chamber, (*ii*) synchronize them by adjusting their relative height, and (*iii*) somehow shrink one clock's vacuum chamber and see how this affects the redshift. Having moving parts in a vacuum chamber, however, seems unfeasible. Actually, this is not needed. Since the regime we are probing here corresponds to high couplings of the scalar field to matter, any object with density $\sim 10^3\,{\rm kg/m^3}$ will be screened inside the vacuum chamber. Therefore, it is sufficient to bring such an object close enough to the atoms

being interrogated to significantly alter the chameleon field they experience, all other things remaining equal. Below the dashed line $\lambda_\phi(\rho_{\rm UHV}) = 1\,\mathrm{m}$ in Fig. 6.5, even an aluminum foil (which has a thickness of $\sim 0.2\,\mathrm{mm}$) would be comfortably more than enough to screen the field. The closer the foil is to the atoms, the better. The material of the foil to be chosen, as well as the minimum distance at which it can be placed without disturbing the measurement is beyond the scope of this chapter.

Orders of magnitude obtained from numerical computations

In order to get an estimate of the part of the parameter space that can be probed with this idea of using a thin foil, we conduct 1D radial numerical computations with *femtoscope* in three stages:

1. We compute the scalar field profile assuming a spherical vacuum chamber of radius $R_{\rm vc}$. Its walls are taken thick enough to be screened, so that the exterior environment has no influence whatsoever on the interior scalar field profile.

2. We then add the foil to the numerical domain, modeled as a spherical shell of density $\rho_{\rm water} = 10^3\,\mathrm{kg/m^3}$, thickness 1 mm and radius $R_{\rm foil}$, centered at the atoms' location. Note that the thickness parameter is not very relevant here since $\lambda_\phi(\rho_{\rm water})$ is smaller than the micrometer scale in the region $\lambda_\phi(\rho_{\rm UHV}) < R_{\rm vc}$ probed here.

3. Finally, we estimate z_ϕ as $\beta|\phi_{\text{with foil}} - \phi_{\text{without foil}}|/M_{\rm Pl}$.

Of course, in reality putting a spherical shell around the atoms is absurd since it would block the electromagnetic radiation with which they have to interact. Nonetheless, this is deemed a good enough first approximation and allows for a relatively cheap numerical exploration of the full parameter space (since simulations are conducted in 1D).

The results of this simple study are presented in Fig. 6.6. As in Fig. 6.3, we represent the scalar field contribution to the total redshift, z_ϕ, for four different sets of the relevant parameters, namely $R_{\rm vc}$, $R_{\rm foil}$ and the density inside the vacuum chamber (UHV or XHV). The iso-redshift contours, at 10^{-15} (orange dashed line) and 10^{-20} (red dashed line), exhibit a typical 'V' shape in the (M, Λ)-plane with log-scaled axes:

– *left branch of the 'V'* — In the lower left corner of each panel, the redshift suddenly drops to a very low level. This is due to fact that below a certain value of the α parameter [Eq. (4.3)], the Compton wavelength of the field in vacuum becomes smaller than $R_{\rm foil}$. As a result, the scalar field value at the atoms' location is the same with and without the foil, hence the vanishing redshift.

– *right branch of the 'V'* — This is more or less the same behavior as the one exhibited in Figs. 6.3 and 6.4, although the interpretation is slightly different. As we increase α [Eq. (4.3)], the Compton wavelength increases in the vacuum chamber. In either of the two configurations, the field has not enough space to reach $\phi_{\rm min}(\rho_{\rm vac})$ at its center, but takes nonetheless a higher value in the absence of the foil.

The *sweet spot* is the bottom of the 'V', that is when $\phi_{\text{without foil}} = \phi_{\rm min}(\rho_{\rm vac})$ but $\phi_{\text{with foil}} \ll \phi_{\rm min}(\rho_{\rm vac})$ — the foil playing its role in lowering the scalar field nearby the atoms. From the several sets of parameters tested (not all represented in Fig. 6.6), it appears that the best forecasts are obtained when the following three *rules of thumb* are met: *(i)* large vacuum chamber to give the field enough space to reach its highest value, (*ii*) high vacuum level for maximizing the latter, (*iii*) bringing the foil as close as possible to the atoms (without perturbing the measurement, which constitutes an open question).

Finally, it should be noted that the presence of the foil will also have an impact on the Newtonian potential, which in turn affects the total redshift through Eq. (6.34). The Newtonian potential at the center of a spherical shell of radius $R_{\rm foil}$, thickness $e \ll R_{\rm foil}$, and volumic density $\rho_{\rm foil}$ is simply

$$\Phi_{N,\,\text{foil}} \simeq -4\pi e G \rho_{\rm foil} R_{\rm foil}\,.$$

For all four cases considered in Fig. 6.6, the contribution of $\Phi_{N,\,\text{foil}}$ is many orders of magnitude below the sensitivity of the best atomic clocks, and one can thus ignore this term.

Challenges

We need to cast a critical eye on this setup idea. First, we have eluded the question of how to actually compare the scalar field value with and without the screened foil. One way to proceed, for instance, is to use a multiplexed optical lattice clock as in Refs. [314, 319], where two clouds of atoms are spatially separated in the same lattice and interrogated simultaneously by a shared clock laser and read-out in parallel. After performing a reference measurement of the gravitational redshift exactly as described in those references, the foil is added to surround only one of the two clouds of atoms and the measurement is repeated. The comparison of this second

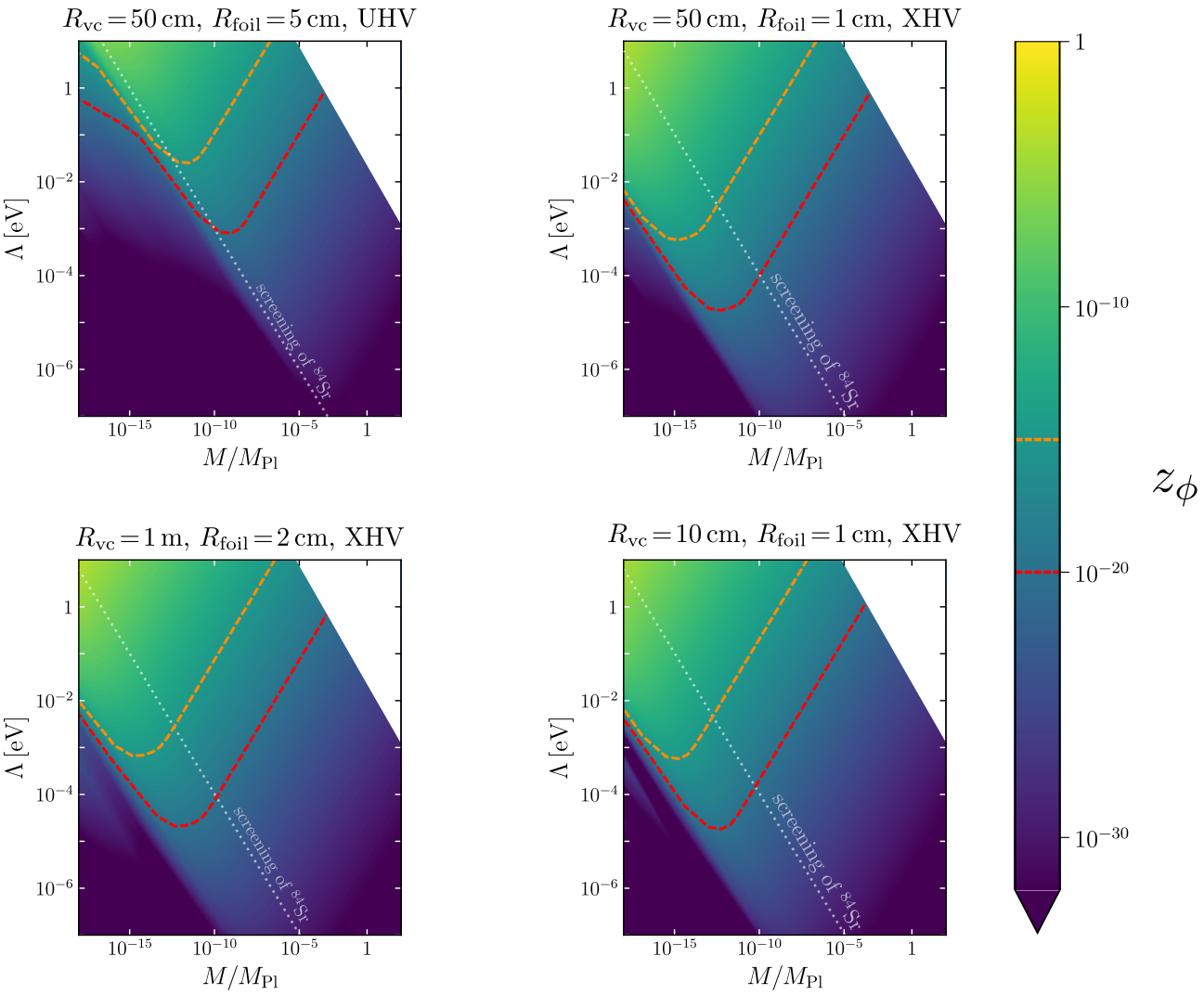


Figure 6.6: Expected redshift from the chameleon field contribution z_ϕ for three different sets of parameters — the size of the vacuum chamber, the distance from the atoms to the foil and the vacuum density (UHV or XHV, see Table 6.2) — in the chameleon parameter space ($n = 1$). The orange and red dashed lines correspond to the iso-redshift at $\varepsilon_{\rm rel} = 10^{-15}$ and $\varepsilon_{\rm rel} = 10^{-20}$ respectively. The white triangular mask in the top right corner of each panel correspond to a region of the parameter space where the foil is no longer screened (see Fig. 6.5), which was not probed in the numerical computations. Below the diagonal dotted line, strontium-84 atoms are screened [see Eq. (6.41)].

measurement against the reference one is then used to assert the consistency of the data within the chameleon model. Although such an experiment is more realistic than the first thought experiment exposed in Sec. 6.2.2, it remains overly simplistic and might turn out to be unfeasible in practice. Addressing the corresponding technical issues is beyond the scope of this chapter.

Moreover, we have assumed throughout this discussion that the clouds of atoms do not perturb the chameleon field inside the vacuum chamber. We examine this hypothesis now. To this end, we employ the following quite common criterion, which states that a spherical body of density ρ_i and radius R_i immersed in a background chameleon field $\phi_{\rm bg}$ will be screened if

$$\rho_i R_i^2 > 3\frac{M_{\rm Pl}}{\beta}\phi_{\rm bg} \iff \alpha < \frac{\tilde{\rho}_i \tilde{R}_i^2}{3\tilde{\phi}_{\rm bg}}\,, \tag{6.41}$$

where α is the dimensionless parameter given by Eq. (4.3) and the tilde notation is used to denote the associated dimensionless quantities. We evaluated the latter criterion in our numerical computations for the ^{84}Sr isotope, which is notably used in optical lattice redshift experiments [314, 319]. In this regard, the dotted line in each panel of Fig. 6.6 is where the transition from the unscreened to the the screened regime occurs in the parameter space.[15] More precisely, the nucleus of the strontium atoms is expected to have a thin shell below that line. For all forecasts assuming $\rho_{\rm XHV}$ inside the vacuum chamber, the aforementioned 'sweet spot' falls into regime where the atoms are screened. Further work is required to assess the consequences of this phenomenon on the actual energy transition probed by redshift measurements. Such questions are also the topic of Ref. [320–322] and the Appendix E of Ref. [142].

6.3.2 Going to space [gravitational strength coupling]

The above considerations showed that laboratory-based redshift experiments are at best sensitive to very strongly coupled scalar fields. Specifically, they are completely blind to the region $\beta \lesssim 10^3$, which turns out to the least constrained by experiments [71, 152]. Given the above, we identify two main reasons for these limitations. On the one hand, we saw with the orders of magnitude laid out in Fig. 6.3 that for $\Lambda \lesssim 10^{-3}\,$eV, the range $\beta \in [10^{-1}, 10^5]$ is only accessible in very low density environments. Yet, current vacuum technology has its limits. For instance, it does not allow us to reach the density levels found in the interplanetary medium. Nevertheless, laboratory experiments come with the constraint of *size*. If we want to dictate the chameleon field value in two nearby regions of space, it has to be very dynamical and to closely follow density variations — this requires the field to be strongly coupled. The downside is that most laboratory objects end up being screened, including the atomic clocks themselves.

On the screening of satellites

Going to space could precisely resolve these two limitations at once. Take a satellite in orbit around the Earth with an onboard atomic clock. Depending on its altitude, the background chameleon field in which it is immersed can be very high — see notably Chapt. 5. For instance, at geostationary altitude, the density is close to the 'IPM' value tabulated in Table 6.2 and the scalar neighbors $\phi_{\rm min}(\rho_{\rm IPM})$ (see e.g. Fig. 4.14).

The sine qua non condition for hoping to use atomic clocks in space for constraining the chameleon model is that the spacecraft must be unscreened. Otherwise, the onboard clock will not see $\phi \sim \phi_{\rm min}(\rho_{\rm IPM})$ but rather $\phi \sim 0$, and will therefore experience time as in GR. The assessment of whether a spacecraft orbiting the Earth is screened has been discussed several times in the literature — see e.g. Refs. [115, 136, 149]. The key point that is usually stressed is that objects which possess thin shells down on Earth may loose them when they are taken into space. This can be seen by referring to the approximate screening criterion (6.41): the low density environment offered by space, together with the large distance with respect to the Earth's surface, result in a higher background value for the scalar field $\phi_{\rm bg}$, which in turn means that the criterion is less easily satisfied. In Ref. [141] (Chapt. 5), we went beyond this qualitative criterion by computing the full chameleon field of the {Earth + satellite} system (without atmosphere though) using *femtoscope* for various satellite's density and size in LEO. This showed, in line with the aforementioned qualitative argument, that a spacecraft could be unscreened in relevant parts of the parameter space — namely for sufficiently large values of α (which, at fixed Λ, means small enough values of β).

Here however, we are interested in going to higher altitudes, farther away from the Earth's atmosphere where the vacuum is more pristine. In order to check whether a satellite is screened or not at such high altitudes, we use *femtoscope* to solve the Klein–Gordon equation governing the chameleon field of a ball immersed in a background medium of density $\rho_{\rm bg} \in \{10^{-12}\,{\rm kg/m^3}, 10^{-20}\,{\rm kg/m^3}\}$. Spherical symmetry allows for rapid radial computations, with the correct asymptotic boundary condition enabled. Specifically, we look for the value of the dimensionless parameter $\alpha_{\rm screened}$ below which the ball is screened (by dichotomy). Table 6.3 compiles such values for several

[15] Note that this is in line with what is shown in Fig. 1 of Ref. [142] for cesium and lithium atoms.

Satellite	Mass	Equivalent radius	Mean density	$\alpha_{\rm screened}$ $\rho_{\rm bg}=10^{-12}\,{\rm kg/m^3}$	$\alpha_{\rm screened}$ $\rho_{\rm bg}=10^{-20}\,{\rm kg/m^3}$
CubeSat	$1\,{\rm kg}$	$6.2\,{\rm cm}$	$10^3\,{\rm kg/m^3}$	2×10^{-6}	2×10^{-10}
MICROSCOPE	$330\,{\rm kg}$	$82\,{\rm cm}$	$1.4\times10^2\,{\rm kg/m^3}$	5×10^{-5}	5×10^{-9}
Galileo	$675\,{\rm kg}$	$95\,{\rm cm}$	$1.9\times10^2\,{\rm kg/m^3}$	9×10^{-5}	9×10^{-9}
HST	$1.1\times10^4\,{\rm kg}$	$5.6\,{\rm m}$	$15\,{\rm kg/m^3}$	3×10^{-4}	3×10^{-8}

Table 6.3: Screening of satellites in space. Determination of $\alpha_{\rm screened}$ for two typical background densities $\rho_{\rm bg}\in\{10^{-12}\,{\rm kg/m^3}, 10^{-20}\,{\rm kg/m^3}\}$ via 1D radial simulations performed with *femtoscope*. The equivalent radius is computed such that a sphere of that radius would have the same volume as the actual satellite at stake. 'Galileo' designate a satellite of the GNSS constellation of the same name, and 'HST' is the acronym of the Hubble Space Telescope. We set $L_0=1\,{\rm m}$, $\rho_0=1\,{\rm kg/m^3}$, $n=1$. Note that these specific characteristic scales lead to different α values compared to Fig. 4.15 or the study conducted in Chapt. 5 where we used $L_0=R_{\rm Earth}$.

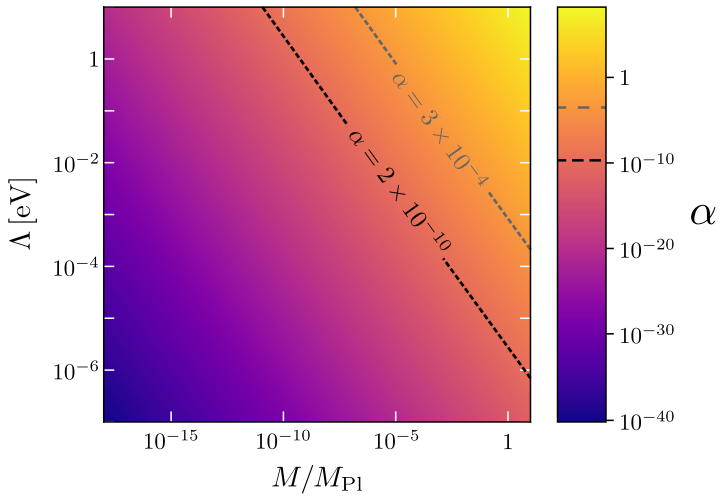


Figure 6.7: Mapping from the chameleon parameter space (M, Λ, $n=1$) to the dimensionless α parameter given by Eq. (4.3), which controls the behavior of the field up to a scaling. The iso lines $\alpha\in\{2\times10^{-10}, 3\times10^{-4}\}$ are highlighted by the black and gray dashed lines respectively. They correspond to the minimum and maximum $\alpha_{\rm screened}$ values reported in Table 6.3.

satellites which are characterized by their mass and dimensions. Whatever the actual shape of the satellite at stake, we model it as a ball of equivalent density. We consider two background densities: $\rho_{\rm bg}=10^{-12}\,{\rm kg/m^3}$ corresponds to the density found at an altitude of roughly $400\,{\rm km}$, while $\rho_{\rm bg}=10^{-20}\,{\rm kg/m^3}$ is the density representative of the IPM. The data obeys the scaling relation $\alpha_{\rm screened}\bowtie\tilde{\rho}\tilde{R}^2$, as expected from Eq. (6.41). Nonetheless, we lay emphasis that this short study, despite being numerical, remains a crude approximation: in reality the atmospheric density varies non-isotropically away from the satellite, and the proximity to Earth (which is deeply screened in the range of parameters considered in Table 6.3) further complicates the picture. Fig. 9 in Chapt. 5 clearly shows that, depending on both the actual atmospheric model and the altitude, the scalar field value can be either greater or lower than the value that minimizes the effective potential.

Fig. 6.7 further helps to get a sense of what are the corresponding actual chameleon parameters (β, Λ) that map to the various values of $\alpha_{\rm screened}$ tabulated in Table 6.3. Specifically, we isolated only the minimum and maximum values from this table and plotted the associated iso α lines (all other values fall within the narrow region delimited by these two lines). This must be put in perspective with Fig. 6.3, especially its left side panel, where we represented the theoretical upper bound on z_ϕ.

Redshift measurements in space and ideas

Best possible constraints How does this translate into constraints on the chameleon model? While obtaining bounds from redshift measurements requires specifying an actual experiment, we can readily derive the best possible constraints by comparing z_ϕ to $\epsilon_{\rm rel}$. The scalar field redshift contribution z_ϕ is maximal for an unscreened satellite orbiting the Earth from a very high altitude (geostationary and beyond), where the ambient density is the lowest (more or less representative of the interplanetary medium, see Table 6.2). In that case, an atomic clock onboard such a spacecraft would experience the very high value of the scalar field, which would be slightly lower but nonetheless close to $\phi_{\rm min}(\rho_{\rm IPM})$. Comparing time as measured by this onboard clock against a ground-based

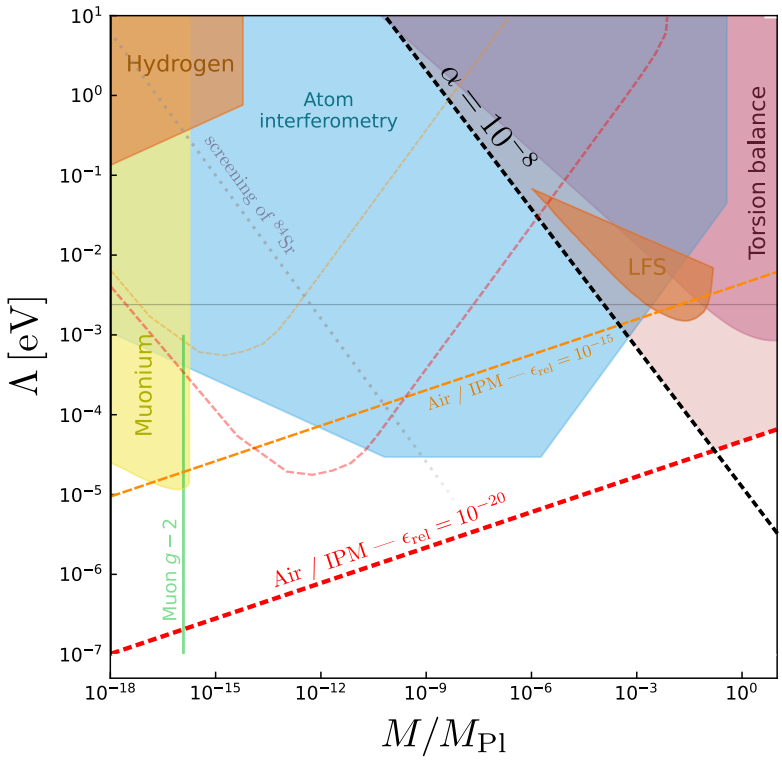


Figure 6.8: Compilation of the forecasts coming from the redshift experiments — in the laboratory and in space — on the chameleon parameter space with state-of-the-art constraints (adapted from Fig. 4 of Ref. [152]). The laboratory constraints ('V'-shaped) are those from Fig. 6.6 with parameters $\{R_{\rm vc} = 50\,{\rm cm},\ R_{\rm foil} = 1\,{\rm cm},\ {\rm XHV}\}$ (top right panel). The region below the dotted gray line is where ^{84}Sr nucleus starts to become screened and is thus, a priori, inaccessible to the atomic clock experiment described in Sec. 6.3.1. The shaded red area in the top right corner corresponds to the best possible constraints that could be set with a space-based experiment, regardless of the actual underlying mission concept. In this best case, highly optimistic scenario — small satellite beyond the geostationary altitude carrying a state-of-the-art atomic clock with a relative precision at $\varepsilon_{\rm rel} = 10^{-20}$ — one could access an unconstrained region of the chameleon parameter space, for gravitational strength couplings.

reference one (or alternatively, one onboard a screened satellite orbiting at lower altitudes where the atmospheric density is several orders of magnitude higher than $\rho_{\rm IPM}$) yields $z_\phi \sim \beta\phi_{\rm min}(\rho_{\rm IPM})/M_{\rm Pl}$. Such a value then constitutes a theoretical upper bound on z_ϕ in any realistic, well-defined experiment. In this perspective, Fig. 6.8 shows the associated best possible constraints in the chameleon parameter space for $\varepsilon_{\rm rel} \in \{10^{-15}, 10^{-20}\}$ (red shaded area), together with the current constraints from other experiments (adapted from Fig. 4 of Ref. [152]). In particular, it is interesting to notice that space-based redshift measurements could open a new window for testing chameleons coupled to matter with gravitational strength — for $\beta \lesssim 10^3$.

Note that redshift measurements involving satellites mean that we can no longer assume that the emitter and the receiver are not moving in relation to each other. Consequently, an additional Doppler term from special relativity must be added to the redshift formula (6.20). This Doppler effect can be larger than the 'pure gravitational redshift' one — this is the case, for instance, for the ACES mission [303] where the overall rate of the clock on-board the ISS will be slower than a static ground one. This further complicates the data analysis process.

Thoughts on an actual mission design In this penultimate paragraph, we very briefly speculate on the actual feasibility of such space missions. First, we lay emphasis on the fact that clocks suitable for flying in space are not nearly as good as the best optical clocks engineered in the laboratory (given all the constraints inherent to space-flight). For instance, passive hydrogen masers onboard Galileo satellites exhibit $O(10^{-14})$ stability levels at averaging times of $\sim 1000\,{\rm s}$, which is roughly one order of magnitude better than the cesium clocks onboard GLONASS satellites. The ACES mission, to be launched in 2025, will put a cold atom cesium clock (PHARAO) in the International Space Station, targeting a $O(10^{-16})$ precision [303].

Given these orders of magnitude, it appears that space-borne atomic clocks are not mature enough yet to probe still unexplored regions of the chameleon parameter space — see Fig. 6.8. Provided that they will continue to improve by a few orders of magnitude in the future, one could draw from past proposal for testing

the gravitational redshift effect in space [25–27, 306]. In particular, highly elliptical orbits are good candidates for such tests in several respects. Regardless of the model being tested, an elliptic orbit induces a periodic modulation of the gravitational redshift [323]. In chameleon gravity, one could imagine having the science payload screened at the perigee, where the atmosphere is still thick enough, yet unscreened at the apogee, where density drops below $10^{-19}\,\mathrm{kg/m^3}$. In this regard, the RadioAstron space mission has the orbital characteristics that we seek, with a perigee at 500 km above the Earth's surface, where the atmospheric density is roughly $10^{-12}\,\mathrm{kg/m^3}$, and very high apogee at an altitude of 350 000 km, where density is that of the interplanetary medium. As reported in Table 6.3 and Fig. 6.7, there is a band in the parameter space for which a spacecraft would possess a thin shell at the perigee but would be unscreened at the apogee. This would leave a chameleon-modulated signal in the redshift data.

Closing remarks While Fig. 6.8 optimistically summarizes this whole section, in practice, putting competitive constraints on the chameleon model via redshift measurements appears to be currently out of reach. Whether the experiments are conducted in the laboratory or in outer space, the greatly idealized bounds we derived have no chance to hold in a realistic scenario.

Additionally, we focused on the very specific case of the chameleon model with exponent $n = 1$. Because $\phi_{\min}(\rho) \propto \rho^{-1/(n+1)}$, models with $n < 0$ do not exhibit the crucial property that $\phi_{\min}(\rho) \to +\infty$ as $\rho \to 0$, on which the whole idea developed in this chapter holds. In the symmetron model [see e.g. Eqs. (1.110, 4.5)], the scalar field has a non-zero VEV in very low density environments, reading

$$\phi_{\min}(\rho) = \pm\frac{\mu}{\sqrt{\lambda}}\sqrt{1-\frac{\rho}{\mu^2 M^2}} \simeq \pm\frac{\mu}{\sqrt{\lambda}} \implies \Omega(\phi_{\min}) \simeq 1 + \frac{\mu^2}{2\lambda M^2}\,.$$

In this model, the dimensionless parameter λ can be very low (values considered in the literature go down to $\lambda \lesssim 10^{-60}$), while $\log_{10}\mu$ is generally taken between -3 and $+3$ and $\log_{10}(M/\mathrm{GeV})$ between -10 and $+20$. A rough order-of-magnitude computation seems to indicate that redshift experiments could yield interesting bounds on this model. The careful analysis of this question is left for future work.

Chapter summary

This chapter clarified, in theoretical terms, the predicted outcome of redshift measurements in the framework of scalar-tensor theories of gravity. Despite satisfying LPI, such theories can nonetheless be distinguished from GR in redshift experiments. Focusing on the chameleon model with positive exponent, we are able to single out the scalar field contribution to the total redshift (in the Newtonian limit), which beyond improving readability, allows us to imagine a first *gedankenexperiment* involving atomic clocks and aimed at either detecting or constraining the model at stake. The orders of magnitude derived from this idealized yet well-defined experiment are translated into constraints in the parameter space of the chameleon, given the current state of the art in atomic clock technology. It appears that these 'optimal' theoretical constraints are competitive with current bounds, which is why we go a step further and assess whether they hold in a more realistic scenario. In the laboratory, modular atomic clocks could be sensitive to chameleons very strongly coupled to matter ($\beta \gg 10^5$), although (*i*) the feasibility of the setup we describe is quite uncertain and (*ii*) the screening of individual atomic nuclei beyond a certain threshold mitigates our forecasts. In space, the very low-density environment found in high-altitude orbits allows spacecraft to be unscreened in some regions of the parameter space — most notably for gravitational strength couplings. However, the clocks onboard satellites cannot be as precise as their ground-based counterparts, due to the constraints inherent to space-flight. We find that the current level of precision exhibited by the best clocks in space is still a few orders of magnitude too low to yield interesting constraints. Finally, while we focused on the chameleon model, the idea of using redshift measurements to test modified gravity could potentially be applied to other scalar-tensor theories with screening mechanisms. In particular, the symmetron is expected to be also quite sensitive to the kind of experiments we described in this chapter, the full study of which is left for future work.

Conclusion and prospects

Once you eliminate the impossible, whatever remains,
no matter how improbable, must be the truth.

Sherlock Holmes

Unveiling the true nature of gravity is a vast research effort that requires the harmonious integration of theoretical insights on the one hand, and experimental data on the other hand. For this program to move forward, it is of the utmost importance to establish quantitative contact between abstract physical models and actual observations. Precisely, making the most of the constraining power of the latter requires accurate modeling of both standard gravity (GR) and alternative theories. This PhD work is part of this overall perspective.

In particular, I focused on scalar-tensor theories of gravity — where the gravitational interaction is mediated by both a rank-2 tensor field and a scalar field — as they arguably constitute one of the most compelling, resilient, and phenomenologically-rich alternatives to GR. In this framework, physically-relevant models can remain viable — notwithstanding the stringent bounds on allowed deviations from GR at the Solar system scale and below — by means of screening mechanisms. The latter have attracted a great deal of attention over the past two decades or so, initially advocating for space-based tests of gravity. Subsequently, the recent attempt to set constraints on the chameleon model using data from the MICROSCOPE space mission mitigated these earlier speculations. Most importantly, one key takeaway from this study is that the testability of such screened fifth forces is crucially dependent on the development of new numerical tools for a realistic modeling of their features.

This PhD work aims at filling this gap, namely to obtain well-posed and versatile numerical solutions of scalar-tensor theories with screening mechanisms. The foreseen benefits of which are twofold. On the one hand, such a numerical tool allows for the translation of experimental data into constraints on modified gravity models in a reliable and accurate way. On the other hand and at earlier stages, it can also prove very useful, if not indispensable, for assessing the relevance of envisioned experimental setups, or for gaining insights into physical scenarios that could not be explored beforehand. In this regard, this thesis further aims at addressing open questions in connection with the ability of space-based mission to efficiently constrain screened scalar fifth forces.

This research work brings out three main results which can be summarized as follows.

The first phase of this work was devoted to the development of *femtoscope*,[16] a PYTHON code relying on the finite element method for solving the semi-linear partial differential equations that arise in scalar-tensor models with screening mechanisms. The use of non-uniform meshes allows one to deal with arbitrarily complex geometries and multi-scale problems, notably through h-adaptivity. Nonlinear PDEs arising e.g. in the chameleon or symmetron models are handled via Newton's method by iteratively solving a sequence of linearized problems. Notably, the Newton-linearized Klein–Gordon equation associated with the chameleon field as well as the Poisson equation, on a bounded domain $\Omega \subset \mathbb{R}^3$, both have a unique weak solution in the usual Sobolev space $H^1(\Omega)$ when supplemented with Dirichlet boundary conditions. The software also allows for dimensional reduction when the sought solution enjoys a particular continuous symmetry, which greatly alleviate the resulting computational burden compared to solving the full 3D problem. This is made possible through the implementation of the underlying weak formulations not only in Cartesian coordinates, but also in spherical and cylindrical coordinates. Specifically, we prove that the well-posedness of the latter is implied by the well-posedness of the former.

There remains the issue of asymptotic boundary conditions (Ω unbounded), which naturally arise in the context of scalar-tensor gravity when the value taken by the scalar field of interest is not known anywhere near the matter sources, but infinitely far away from them. For the two aforementioned problems, the asymptotic behavior is imposed by looking for weak solutions in adequate weighted Sobolev spaces, for which Lax–Milgram hypotheses apply. Unbounded domains cannot be meshed as in standard FEM, which is why we explored several techniques based on compactification transforms. In particular, we successfully leveraged the Kelvin inversion by implementing the inverted finite element method (*ifem*) in *femtoscope*. Moreover, building on top of *ifem* and a specific domain decomposition scheme, we proposed a novel method for solving elliptic PDEs on the whole space,

[16] Open-source code publicly available at https://github.com/onera/femtoscope.

which we call the alternate inverted finite element method (*a-ifem*).

In a second phase, we made use of *femtoscope* in order to quantitatively assess the testability of a chameleonic fifth force in Earth orbit by means of space geodesy techniques. The PREM and US76 models are used to assign densities to the Earth's interior and atmosphere respectively, and we introduce a small deviation from spherical symmetry embodied by a true-to-scale mountain. Then, solving for the Newtonian potential and the chameleon field gives us access, upon computing their gradient, to both standard and modified gravity. In viable regions of the chameleon parameter space, the Earth is screened, which means that the corresponding fifth force is sourced only by its outermost layers, at odds with the inverse-square law. In this modified gravity setting and in the absence of atmosphere, the chameleon therefore leaves a distinctive signature on the Earth's overall gravitational potential, as brought to the fore by its multipolar expansion.

In order to assess whether this signature can be detected, we first tackled the issue related to the influence of the atmosphere on the fifth force. Fixing the parameters (Λ, n) and gradually increasing the coupling β brings out four regimes: (i) for low values of β, the atmosphere is *transparent* to the fifth force, (ii) above a certain threshold, it acts as an attenuator, effectively reducing the chameleon acceleration, (iii) for even stronger couplings, any non-radial dependence of the scalar field vanishes so that the mountain is plainly *invisible*, and (iv) the atmosphere itself eventually becomes screened.

We also addressed the question of the backreaction of an object as small as a satellite on the background chameleon field. For the first time, we went beyond the various approximations found in the literature by computing the full $\{\text{Earth} + \text{satellite}\}$ system. We showed that the transition from the unscreened to the screened regime occurs over a very narrow band in the chameleon parameter space. In the latter regime, the resulting fifth force acting on the satellite is suppressed extremely efficiently.

Finally, we selected a 'best-case scenario', with no atmosphere, and simulated the dynamics of the GRACE-FO pair of satellites, treated as point masses, in the $\{\text{sphere} + \text{mountain}\}$ system, with and without the putative chameleonic force. We showed that the anomaly brought about by the scalar fifth force is well above the sensitivity range offered by current space-borne technology. However, the existence of uncertainties in the model, most notably the fact that the distribution of matter within the Earth is poorly known, greatly mitigates the constraining power of such tests. This degeneracy can in principle be lifted by performing this kind of space geodesy experiment at more than one altitude. However, given the orders of magnitude involved and the optimistic model underlying them, the take-home message is that space geodesy is not likely to result in competitive constraints on the chameleon model in the near future.

In a third and final phase, we explored the possibility of testing screened scalar-tensor theories by means of redshift experiments, which are radically different in nature from fifth force searches. We derived the expression of the measured redshift in this context and isolated the scalar field contribution in the Newtonian limit. Despite satisfying local position invariance, such theories can nonetheless be distinguished from GR in redshift experiments. Focusing on the chameleon model, we derived the optimal bounds that could be put on the model's parameters given the state of the art in atomic clock technology, which turned out to be competitive with current bounds by several orders of magnitude. We then considered more realistic scenarios. In the laboratory, we find that modular atomic clocks could be sensitive to chameleons very strongly coupled to matter, although (i) the feasibility of the setup we describe is quite uncertain and (ii) the screening of individual atomic nuclei beyond a certain threshold mitigates our forecasts. In space, the very low-density environment found in high-altitude orbits allows spacecraft to be unscreened in some regions of the parameter space — most notably for gravitational strength couplings. However, the clocks onboard satellites cannot be as precise as their ground-based counterparts, due to the constraints inherent to space-flight. As a result, we found that the current level of precision exhibited by the best clocks in space is still a few orders of magnitude too low to yield interesting constraints. Be it as it may, it is not inconceivable that the desired levels of precision will be achieved in the future given the recent progress made in the field of optical clocks.

This PhD work has implications that go beyond the main results synthesized above. First of all, we lay emphasis on the fact that, while it has been used primarily to study scalar-tensor theories of gravity, *femtoscope* is a general-purpose PYTHON software based on FEM for solving nonlinear elliptic PDEs on unbounded domains. To the best of our knowledge, there are no other publicly available codes with this set of specifications. Its versatility means that it can readily be utilized to study completely different physics whose governing equations fit into this framework. This includes, but is far from being limited to, Laplace equation for the electrostatic potential in free space around perfect conductors, stationary states to nonlinear heat or wave equations that take the generic form $-\Delta u = f(\mathbf{x}, u)$ in an infinite media, the elliptic sine-Gordon equation on the whole space which appears in the study of surfaces of constant negative curvature, or the quasi-linear p-Laplacian equation arising in non-Newtonian fluid dynamics in the whole space.

The *femtoscope* software could also prove useful in the design of MICROSCOPE-2, the envisioned successor the MICROSCOPE space mission that will aim to reach a 10^{-17} precision on the Eötvös parameter for the WEP test.

In this context, *femtoscope* could be an elementary building block of a larger, more comprehensive end-to-end simulator of all the physics taking place in the instrument onboard the satellite, including putative WEP-violating fifth forces stemming e.g. from scalar-tensor models with non-universal couplings.

Still on the subject of fundamental physics experiments in space, our conclusions from Chapts. 4 and 5 overlap with those that were drawn from the previous study on the testability of the chameleon model with MICROSCOPE. Particularly, our results confirm, in the light of more comprehensive numerical computations, that the latter turns out to be screened at LEO orbits — where the atmosphere is still relatively dense compared to the interplanetary medium — for the largest part of yet unconstrained regions of the parameter space. To be clear, this greatly reduces the hope of detecting WEP violations stemming from a non-universally coupled chameleon to matter with MICROSCOPE-like experiment.

Another important implication of this research work is the potential of redshift-based experiments to competitively constrain scalar-tensor models with screening mechanisms. To the best of our knowledge, the possibility of leveraging this effect (exhibited by all metric theories) has not yet been considered in the literature, and opens the way to more discussions with experimental physicists to assess the feasibility of the ideas laid out in this Chapt. 6. Regarding space-based tests however, one should not expect the soon-to-be-launched ACES mission to be able to put constraints on the chameleon through its measurement of gravitational redshift, mainly due to the fact that the ISS is screened for still relevant model parameters.

Several areas of development can be foreseen to extend this research work. Regarding *femtoscope*, we have mentioned several times in this manuscript the possibility to extend its scope to time-dependent problems. While the way this can be achieved in the finite element framework has been explained in detail in Chapt. 2, this feature has yet to be implemented on top of the current version of the software. Qualitatively, the validity of the quasi-static assumption hinges on the comparison of two time scales. On the one hand, the chameleon field adapts to changes in the density distribution ρ with a characteristic time $\tau_1 \sim \lambda_\phi(\rho)$ (in natural units), while on the other hand, the moving objects under study exhibit a time scale $\tau_2 \sim L_0/v$, where L_0 and v are typical length scale and velocity. For most situations, $\tau_1 \ll \tau_2$, but when this ceases to be the case, time-dependent simulations become necessary to obtain physically meaningful results. This would represent a further challenge in terms of computational cost — at each time step, one has to solve a nonlinear problem which in turn translates to solving a sequence of linear problems (see Fig. 4.5) — and would thus most likely require to leverage high-performance computing techniques.

In Chapt. 4, we only scratched the surface of the two-body problem in the framework of chameleon gravity. For one thing, the draft study we presented on the shifting of equilibrium points with respect to Newtonian gravity should be further expanded to include experimental insights, the lack of which currently prevents us from setting out a well-defined experiment. In an astrophysical context, the study of binary neutron star inspirals is a priori within our reach, since the weak field and quasi-static regimes do apply in this phase preceding the merger. Given the scalar field profiles computed with *femtoscope* for a given range of distances between the two compact bodies, one could address the question of the influence of the scalar field on the lifetime of such binary systems and compare it against GR's prediction.

Besides, our preliminary study on redshift-based tests of scalar-tensor theories of gravity with screening mechanisms was illustrated only on the chameleon model with positive exponent ($n>0$) due to its ability to reach very high values in low-density media. We noticed that the symmetron field also exhibits very large VEVs in vacuum in some parts of its parameter space. A work similar to what has been done in Chapt. 6 but for the symmetron is necessary to further assess the possibility of putting constraints on this model with atomic clocks. As as additional remark, we stress that the validity of treating screened scalar-tensor theories of gravity as effective field theories is not guaranteed at the particle physics level. The extent to which this could impact the discussions undertaken in Chapt. 6 remains to be estimated.

As closing words, scalar-tensor theories of gravity acquired a new lease on life with the introduction of screening mechanisms. While the latter were engineered precisely to escape from our grasp, a global research effort has been at play for the past decades to try and rule-out ever widening regions of their respective parameter spaces. Taking a step back, it is quite amazing to look at the great variety of tests that have been proposed as attempts to 'outsmart' the scalars in this game of hide-and-seek. This research work was not intended to analyze real data, so there was no question of imposing new constraints on these models. Instead, we have provided new numerical means — which have the potential to benefit a broad spectrum of scientific communities — to help and guide this effort. Indeed, we have reached a stage where making further progress calls for elaborate numerical modeling. The versatility of *femtoscope* makes it a tool of choice for probing gravity in a wide range of physical contexts: from the laboratory (e.g. atom interferometry or torsion pendulum experiments), to space-based missions [e.g. MICROSCOPE(-2)] and scenarios involving astrophysical objects (e.g. binary neutron star systems). Even gravitational physics happening at different scales can be encompassed through h-adaptivity. In that sense, my research work has opened new exciting avenues for gaining valuable insights into alternative theories of gravity.

Taking another step back, scalar-tensor theories are one way of extending GR among many others. Looking at the whole zoo of modified gravity models may justifiably be overwhelming, and one may feel lost amidst it. However, in the absence of deeper and more profound guiding principles, testing all possibilities and ruling them out one by one becomes an essential strategy for forging ahead in our understanding of gravity.

Appendix A

Natural units

Natural units have been introduced and briefly discussed in Chapt. 1. This Appendix provide a more thorough review of this widely used convention in theoretical physics — especially in quantum field theory and particle physics. Choosing to translate the equations of GR (or alternative theories) from SI units to natural units might seem a bit superfluous. After all, the physics is already there. Yet, this has a genuine added value in several respects. Beyond making equations slightly simpler by removing occurrences of c and $\hbar$, the use of natural units allows for a more direct comparison of *scales*. Indeed, all kinematical quantities can be expressed in powers of electron-volt. Moreover, if we think of GR or scalar-tensor theories as field theories, it becomes easier to appreciate the connection between such gravitational theories and, say, particle physics.

However, employing natural units comes at the cost of loosing dimensional clarity. Unlike with SI units, physical quantities are expressed without reference to their dimensions (such as meter for length, kilogram for mass and second for time), which makes it more difficult to keep track of the physical meaning of quantities. This Appendix aims at laying out the algorithm for switching from SI unit to natural unit, and vice-versa. It is then showcased on the equations arising in scalar-tensor theories written throughout this manuscript.

A.1 Conversion between SI units and natural units

A.1.1 Definition of natural units

The speed of light c and the reduced Planck constant $\hbar = h/2\pi$ are fundamental constants in physics, whose values in SI units are $c = 299\,792\,458\,\mathrm{m\,s^{-1}}$ and $2\pi\hbar = 6.626\,070\,15 \times 10^{-34}\,\mathrm{kg\,m^2\,s^{-1}}$. Natural units are defined as the system of units in which c and $\hbar$ are equal to one. The most sensible way to look at this definition is to regard energy E, velocity V and angular momentum A as new base quantities, rather than length L, mass M and time T. Any kinematical variable can be equivalently expressed in both base quantities since

$$\begin{cases} \mathrm{V} = \mathrm{L\,T^{-1}} \\ \mathrm{E} = \mathrm{M\,L^2\,T^{-2}} \\ \mathrm{A} = \mathrm{M\,L^2\,T^{-1}} \end{cases} \quad \Longleftrightarrow \quad \begin{cases} \mathrm{M} = \mathrm{E\,V^{-2}} \\ \mathrm{L} = \mathrm{A\,V\,E^{-1}} \\ \mathrm{T} = \mathrm{A\,E^{-1}} \end{cases} . \tag{A.1}$$

This is illustrated heuristically in Fig. A.1. Note that using energy as our new base quantity is somewhat arbitrary as we could e.g. have chosen mass to assume this role.

Now let us look at a few examples. In the original base quantities, a force has dimension $\mathrm{M\,L\,T^{-2}}$ with SI units $\mathrm{kg\,m\,s^{-2}}$. In the new base quantities, it has dimension $\mathrm{E^2\,A^{-1}\,V^{-1}}$ with units $\mathrm{eV^2}\,(\text{unit of } c)^{-1}\,(\text{unit of } \hbar)^{-1}$. In practice, we do not bother writing "(unit of c)" or "(unit of $\hbar$)". Therefore, a force may be expressed in $\mathrm{eV^2}$. Likewise, a mass can be expressed in eV, a length or a time in $\mathrm{eV^{-1}}$ and an energy density in $\mathrm{eV^4}$.

At this point, several remarks are in order. First, using the electron-volt as our energy unit is a *choice* — any other unit of energy would do the job. Second, the consequence of these choices is that any kinematical physical quantity can be expressed in eV^α, for some relevant exponent α depending on the original SI units of the quantity at stake. This remaining 'uncollapsed' dimension still allows one to perform basic dimensional analysis. Third, it is no surprise that the equations of the gravity models studied in this PhD work involve the gravitational constant G (rather than $\hbar$). A perfectly sensible choice is to set $c = G = 1$ (leaving out of the picture the Planck constant), which results in so-called *geometrized units* (e.g. used in the 'bible' of GR [324]). However, going further and setting $c = \hbar = G = 1$ would result in a purely natural system of units which has all of its dimensions collapsed. Leaving herewith no dimensional quantities is typically a bad idea as it would imply the definitive abandon of dimensional analysis, which has proven to be a powerful tool in physics!

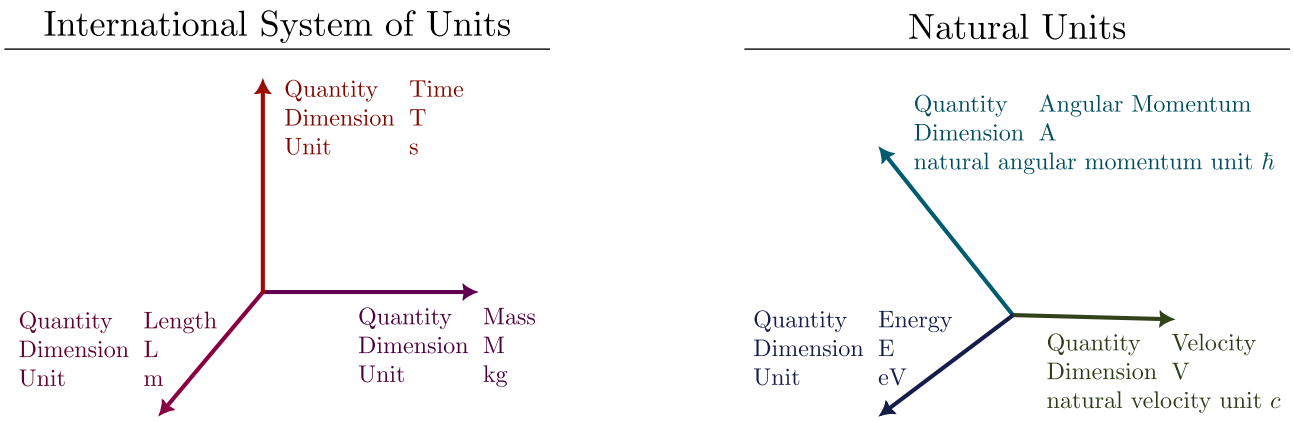


Figure A.1: Heuristic illustration of SI units *vs* natural units. For the record, the 'new' international system of units (2018) fixes the numerical value of c and $\hbar$ to those provided in Eq. (A.2).

A.1.2 Conversion algorithm

Consider a kinematical variable X with SI units $\mathrm{kg}^\alpha\,\mathrm{m}^\beta\,\mathrm{s}^\gamma$, with numerical value x in this system of units. Given Eq. (A.1), X can be expressed in $\mathrm{eV}^{\alpha-\beta-\gamma}$, with a different numerical value χ. Of course, what we would like to know is the conversion factor to go from x to χ. This conversion factor involves powers of the numbers

$$\mathcal{C} = 299\,792\,458\ , \quad \mathcal{H} = 6.626\,070\,15\times 10^{-34}\ , \quad \mathcal{E} = 1.602\,176\,634\times 10^{-19} \tag{A.2}$$

(intentionally written without their usual accompanying units), reading

$$\frac{x}{\chi} = \mathcal{H}^{\beta+\gamma}\,\mathcal{C}^{\beta-2\alpha}\,\mathcal{E}^{\alpha-\beta-\gamma} \tag{A.3}$$

We also set the dimensionful quantity $e = \mathcal{E}\,\mathrm{J/eV}$. Table A.1 compiles the conversion factor x/χ for various physical entities that are often encountered in gravitational physics. To go from natural units to SI units: multiply by the conversion factor. To go from SI units to natural units: divide by the conversion factor.

A.2 Dimensional analysis

Let us perform the dimensional analysis of the Einstein's field equations (1.12) and show that, as claimed at the beginning of Sec. 1.1.1, the numerical value of the reduced Planck constant $\hbar$ does not really appear. For the record, the field's equations, in natural units, read

$$M_{\rm Pl}^2\left[R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\right] = T_{\mu\nu}\ . \tag{A.4}$$

The homogeneity of such expression is readily checked: $M_{\rm Pl}^2$ as units of eV^2 and so does $R_{\mu\nu}$, R, while the energy-momentum tensor $T_{\mu\nu}$ is expressed in eV^4.

In order to go back to SI units, we can apply the rules set out above in Sec. A.1.2 and use the conversion Table A.1. In particular

Variable	**SI Unit**	**Natural Unit**	x/χ **factor**
mass	kg	eV	$\mathcal{E}\,\mathcal{C}^{-2}$
length	m	eV^{-1}	$\mathcal{H}\,\mathcal{C}\,\mathcal{E}^{-1}$
time	s	eV^{-1}	$\mathcal{H}\,\mathcal{E}^{-1}$
energy	$\mathrm{kg\,m^2\,s^{-2}}$	eV	$\mathcal{E}$
velocity	$\mathrm{m\,s^{-1}}$	1	$\mathcal{C}$
gravitational potential	$\mathrm{m^2\,s^{-2}}$	1	$\mathcal{C}^2$
matter density	$\mathrm{kg\,m^{-3}}$	eV^4	$\mathcal{H}^{-3}\,\mathcal{C}^{-5}\,\mathcal{E}^4$
acceleration	$\mathrm{m\,s^{-2}}$	eV	$\mathcal{H}^{-1}\,\mathcal{C}\,\mathcal{E}$

Table A.1: Conversion form (SI units $\longleftrightarrow$ natural units) for various physical quantities encountered in gravitational physics.

- M_{Pl}^2 can be converted back to kg^2 through the factor $\mathcal{C}^{-4}\mathcal{E}^2$;
- $R_{\mu\nu}$ and R can be converted back to m^{-2} through the factor $(\mathcal{H}\mathcal{C})^{-2}\mathcal{E}^2$;
- $T_{\mu\nu}$ can be converted back to $\mathrm{kg\,m^{-1}\,s^{-2}}$ through the factor $\mathcal{H}^{-3}\mathcal{C}^{-3}\mathcal{E}^4$.

Therefore, multiplying both sides of Eq. (A.4) by $e^4(\hbar c)^{-3}$ results in the SI-units version of the Einstein's field equations, reading

$$\frac{c^4}{8\pi G}\left(R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\right) = T_{\mu\nu}\,, \tag{A.5}$$

for which it is clear that the numerical value of $\hbar$ does not play any role.

Appendix B

Mapping of $f(R)$ and extra-dimensional theories to scalar-tensor models

B.1 $f(R)$ theory

In $f(R)$ models, the Ricci scalar R in the Einstein–Hilbert action (1.11b) is replaced by $f(R)$, i.e.

$$S = \frac{M_{\rm Pl}^2}{2}\int \mathrm{d}^4x\sqrt{-g}f(R) + S_{\rm mat}[g_{\mu\nu}]\,, \tag{B.1}$$

where f designates an arbitrary real function. It turns out that such theories can be recast in the framework of scalar-tensor theories. This can be shown in three steps:

1. Define $\chi(x^\mu)$ a dynamical scalar field such that $f(R) = f(\chi) - f'(\chi)(R-\chi)$.
2. The variation of the action (B.1) with respect to χ yields $f''(\chi)(R-\chi) \equiv 0$. If f is chosen such that $f''(\chi(x^\mu)) \neq 0$ for all spacetime coordinates x^μ but a set of negligible measure, this implies $\chi \equiv R$. Note that $\delta S_{\rm mat}/\delta\chi = 0$.
3. We can then define a new scalar field $\varphi(x^\mu) = -f'(\chi(x^\mu))$ so that the action (B.1) now reads

$$S = \frac{M_{\rm Pl}^2}{2}\int \mathrm{d}^4x\sqrt{-g}\left\{\varphi R - \left[\varphi\chi(\varphi) - f\big(\chi(\varphi)\big)\right]\right\} + S_{\rm mat}[g_{\mu\nu}]\,. \tag{B.2}$$

   This corresponds to the action of a scalar-tensor model in the Jordan frame [Eq. (1.41)] where

$$F(\varphi) = \varphi\,, \quad Z(\varphi) = 0\,, \quad U(\varphi) = \frac{M_{\rm Pl}^2}{2}\Big[\varphi\chi(\varphi) - f\big(\chi(\varphi)\big)\Big].$$

   This is a well-known procedure, see e.g. Refs. [325, 326]

B.2 Extra-dimensional Kaluza–Klein theory

This discussion of how theories with extra dimension can be made to look like traditional scalar-tensor theories is based on Ref. [327].

We consider Kaluza–Klein theories for which a $(4+d)$-dimensional spacetime is equipped with a metric G_{ab} and a set of coordinates $\{X^a\}_{0\leq a\leq d+3}$. We assume that the line element can be put in the following form

$$ds^2 = G_{ab}\mathrm{d}X^a\mathrm{d}X^b = g_{\mu\nu}(x)\mathrm{d}x^\mu\mathrm{d}x^\nu + b^2(x)\gamma_{ij}(y)\mathrm{d}y^i\mathrm{d}y^j\,. \tag{B.3}$$

In this splitting of the metric G_{ab}, we introduced

- $\{x^\mu\}$ a set of coordinates in the 4-dimensional spacetime together with a metric on that manifold $g_{\mu\nu}$ which depends only on these coordinates;
- $\{y^i\}$ a set of coordinates in the d-dimensional manifold (which is assumed to be maximally symmetric for the sake of simplicity) and equipped with the metric γ_{ij} which depends only on the latter coordinates;

– $b(x)$ a scale factor.

The action of the theory under consideration is

$$S = \frac{1}{16\pi G_{4+d}} \int \mathrm{d}^{4+d}X \sqrt{-G} R[G_{ab}] + \int \mathrm{d}^{4+d}X \sqrt{-G} \mathcal{L}_{\mathrm{mat}} \,, \tag{B.4}$$

where G is the determinant of the metric G_{ab}, $R[G_{ab}]$ is the Ricci scalar constructed from this metric and G_{4+d} is a constant of the theory.

The idea is then to recover a 4-dimensional theory by integrating out the extra d dimensions in the action (B.4). To do so, we rely on the fact that

$$\sqrt{-G} = b^d \sqrt{-g}\sqrt{\gamma} \quad \text{and} \quad R[G_{ab}] = R[g_{\mu\nu}] + b^{-2} R[\gamma_{ij}] - 2db^{-1} g^{\mu\rho}\nabla_\mu\nabla_\rho b - d(d-1)b^{-2} g^{\mu\sigma}\nabla_\mu b \nabla_\sigma b \,, \tag{B.5}$$

where ∇_μ denotes the usual covariant derivative constructed from the metric $g_{\mu\nu}$. Newton's constant G_4 is obtained through

$$G_4 = \frac{G_{4+d}}{\mathcal{V}} \,, \qquad \text{where} \quad \mathcal{V} = \int \mathrm{d}^d y \sqrt{\gamma} \tag{B.6}$$

is the volume of the extra-dimensional space when $b \equiv 1$. We set $M_{\mathrm{Pl}}^2 = 1/8\pi G_4$. Using Eqs. (B.5, B.6), the integration over the extra dimensions in the action yields

$$\begin{aligned} S &= \frac{1}{16\pi G_{4+d}} \int \mathrm{d}^4 x \sqrt{-g} b^d \int \mathrm{d}^d y \sqrt{\gamma} \Big(R[G_{ab}] + \mathcal{L}_{\mathrm{mat}} \Big) \\ &= \frac{M_{\mathrm{Pl}}^2}{2} \int \mathrm{d}^4 x \sqrt{-g} b^{d-2} \Big\{ b^2 R[g_{\mu\nu}] + d(d-1) g^{\mu\nu} \nabla_\mu b \nabla_\nu b + d(d-1)\kappa \Big\} + \int \mathrm{d}^4 x \sqrt{-g} \mathcal{V} b^d \mathcal{L}_{\mathrm{mat}} \,. \end{aligned} \tag{B.7}$$

In the above, we have set $R[\gamma_{ij}] = d(d-1)\kappa$ (remember that the extra d-dimensional manifold is taken to maximally symmetric). We start to see a resemblance with the framework of scalar-tensor theories exposed in Chapt. 1, Sec. 1.1.2, where the scale factor b plays the role of a scalar field. It can be shown (see Ref. [327]) that the changes of variables

$$\beta(x) = \ln b \,, \quad g^*_{\mu\nu} = \mathrm{e}^{d\beta} g_{\mu\nu} \,, \quad \phi = \sqrt{\frac{d(d+2)}{2}} M_{\mathrm{Pl}} \beta \tag{B.8}$$

lead to the action

$$\begin{aligned} S = &\int \mathrm{d}^4 x \sqrt{-g_*} \left\{ \frac{M_{\mathrm{Pl}}^2}{2} R_* - \frac{1}{2} g_*^{\mu\nu} \nabla^*_\mu \phi \nabla^*_\nu \phi + \frac{\kappa}{2M_{\mathrm{Pl}}^2} d(d-1) \exp\left(-\sqrt{\frac{2(d+2)}{d}} \frac{\phi}{M_{\mathrm{Pl}}} \right) \right\} \\ &+ \int \mathrm{d}^4 x \sqrt{-g_*} \mathcal{V} \exp\left(-\sqrt{\frac{2(d+2)}{d}} \frac{\phi}{M_{\mathrm{Pl}}} \right) \mathcal{L}_{\mathrm{mat}} \,, \end{aligned} \tag{B.9}$$

where quantities with a star $*$ are derived from the metric $g^*_{\mu\nu}$. Eq. (B.9) makes it clear that the original action S given by Eq. (B.4) can be put (under some assumptions) in the form of a traditional scalar-tensor model with non-trivial potential and conformal factor function. This scalar field ϕ is usually referred to as the *dilaton* or the *radion*. It is related to the size of the extra-dimensional manifold.

Appendix C

On the existence of solutions to semi-linear PDEs

This appendix is dedicated to the mathematical study of the nonlinear PDEs that are discussed throughout this manuscript, in particular the chameleon and symmetron field equations. On the basis of known results in the field of semi-linear PDE analysis, we strive to provide answers to the two key questions:

1. Do the field equations at stake have solutions?
2. If so, are they unique?

The existence and uniqueness of solutions to semi-linear PDE problems is the subject of an entire field of research which dates back to the early XX$^{\text{th}}$ century, with significant contributions notably from Hadamard and Fréchet. Today, we have a whole arsenal of more or less sophisticated techniques at our disposal to tackle these questions: maximum principles [205], the method of sub- and super-solutions [328], variational methods and critical point theory [329, 330], including minimax procedures [331], the *mountain pass* and *saddle point* theorems [332, 333], etc.

The starting point of this analysis of well-posedness are the dimensionless equations (4.4, 4.6), which can both be cast into the generic form

$$-\alpha\Delta u = f(\mathbf{x}, u) \quad \text{in } \Omega \subseteq \mathbb{R}^3\,,\ \alpha > 0\,. \tag{C.1}$$

Unfortunately, there is no single 'great theorem' straightforwardly applicable to the generic case (C.1), but rather a plethora of results scattered across research articles, which are restricted to specific forms of the rhs function $f(\mathbf{x}, u)$ together with precise sets of assumptions. The aim of the present appendix is to pinpoint, when possible, the theorems that apply to the cases of interest and verify that their assumptions hold.

C.1 Chameleon field equation

In the case of the chameleon model, the rhs function featured in Eq. (C.1) reads $f(\mathbf{x}, s) = s^{-m} - \rho(\mathbf{x})$ for some $m \in \mathbb{Z}$ and $0 \leq \rho(\mathbf{x}) \leq \rho_{\max}$. We first study the case of chameleon models with positive exponents, i.e. $m > 1$. As for the negative exponent case, we only focus on the $m = -3$ example. Refer to Sec. 1.2.2 for a summary of admissible exponents in the framework of the chameleon model.

C.1.1 Positive exponent, bounded domain

The theorem we leverage here was first proven in Refs. [334, 335]. In particular, it provides us with the existence and uniqueness of a *classical* solution to our boundary value problem. It is to be noted that having $m > 1$ results in $f(\mathbf{x}, s) \to +\infty$ as $s \to 0^+$. Problems of this form go under the name of *singular semi-linear elliptic problems* and are the subject of a whole strand of literature [334–338]. In particular, we shall seek positive solutions to such a problem.

Let $m > 1$ and Ω be a bounded open connected subset of $\mathbb{R}^3$, whose boundary $\Gamma = \bar{\Omega} \setminus \Omega$ is assumed to be a surface of class $C^{2,\gamma}$ for some $\gamma \in]0, 1[$.[1] Define the rhs

$$\begin{aligned} f\colon \bar{\Omega} \times]0, +\infty[&\to \mathbb{R} \\ \mathbf{x},\, s &\mapsto s^{-m} - \rho(\mathbf{x})\,, \end{aligned} \tag{C.2}$$

[1] The precise meaning of a $C^{2,\gamma}$ boundary is given in Ref. [339], page 6. Note that this condition is satisfied if $\Omega = \mathcal{B}(R)$ for some $R > 0$.

where $\rho\colon \bar{\Omega} \to \mathbb{R}_+$ is assumed to be Lipschitz continuous.

Theorem C.1. Given $u_D \in C^{2,\gamma}(\Gamma)$ with $u_D(\mathbf{x}) > 0$ for all $\mathbf{x} \in \Gamma$, the PDE problem

$$\begin{cases} -\alpha\Delta u(\mathbf{x}) = f(\mathbf{x}, s) & \text{for } \mathbf{x} \in \Omega \\ u(\mathbf{x}) = u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma \end{cases} \tag{C.3}$$

has a unique *classical* positive solution $u \in C^{2,\beta}(\bar{\Omega})$, for some $\beta \in]0, \gamma[$.

Proof. The proof of this theorem is a direct consequence of theorems 2.1 and 4.3 from Ref. [335]. The existence part requires verification of hypotheses H1), H2)' and H3) laid out in this work, which all relate to the function f defined by Eq. (C.2).

H1) For any $\mathbf{x}$, $\mathbf{y} \in \bar{\Omega}$ and s, $r \in \mathbb{R}_+^*$, we have $|f(\mathbf{x}, s) - f(\mathbf{y}, r)| \leq |s^{-m} - r^{-m}| + |\rho(\mathbf{x}) - \rho(\mathbf{y})|$. The fact that the function ρ is Lipschitz continuous on $\bar{\Omega}$ and that the real function : $\mathbb{R}_+^* \ni t \mapsto t^{-m}$ is locally Lipschitz continuous imply the Hölder continuity of f on each compact subset of $\bar{\Omega} \times \mathbb{R}_+^*$.

H2)' Trivially, $s^{-1}|f(\mathbf{x}, s)| \to 0$ as $s \to +\infty$ uniformly for $\mathbf{x} \in \bar{\Omega}$. Moreover, there exist constants $\beta > 0$ and $A > 0$ such that $f(\mathbf{x}, s) \geq \beta$ for all $\mathbf{x} \in \bar{\Omega}$ and all $s \in]0, A[$. Indeed, one can simply take $\beta = 1$ and $A = (1 + \rho_{\max})^{-1/m}$, so that

$$f(\mathbf{x}, s) = s^{-m} - \rho(\mathbf{x}) \geq A^{-m} - \rho_{\max} = 1\,.$$

H3) Let $r > 0$ and $\mathbf{x} \in \bar{\Omega}$. For $p > q \geq r$, $f(\mathbf{x}, p) - f(\mathbf{x}, q) = p^{-m} - q^{-m} \leq 0$. Moreover, the mean value theorem yields

$$\left| \frac{p^{-m} - q^{-m}}{p - q} \right| \leq m r^{-(m+1)} =: M(r)\,,$$

so that $-M(r)(p - q) \leq f(\mathbf{x}, p) - f(\mathbf{x}, q) \leq 0$.

□

Ref. [337] provides a uniqueness criterion, which is automatically satisfied given that : $\mathbb{R}_+^* \ni s \mapsto f(\mathbf{x}, s)$ is non-increasing for each $\mathbf{x} \in \Omega$.

C.1.2 Negative exponent, bounded domain

In the case $m = -3$ (which corresponds to $n = -4$ with the conventions of Sec. 1.2.2), the PDE problem reads

$$\begin{cases} -\alpha\Delta u(\mathbf{x}) = u^3(\mathbf{x}) - \rho(\mathbf{x}) & \text{for } \mathbf{x} \in \Omega \\ u(\mathbf{x}) = u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma \end{cases}. \tag{C.4}$$

This type of PDE problem falls into the category of *super-linear* elliptic problems *with perturbed symmetry*,[2] for which most mathematical results concern the existence and uniqueness of *weak solutions*. They principally involve techniques related to the calculus of variations and critical point theory — see e.g. Refs. [329, 330, 340–342]; in particular Theorem C.1 does not apply in this case. Ref. [343] reviews a number of known results regarding super-linear elliptic problems. Specifically, the problem

$$-\Delta u = |u|^{p-2}u + h(\mathbf{x})\,, \quad u \in H_0^1(\Omega)$$

has infinitely many solutions provided that $h \in L^2(\Omega)$ and $2 < p < (2N - 2)/(N - 2)$, where $N \geq 3$ is the dimension of the problem — $p < 4$ in the three-dimensional case [344]. Noting that $u^3 = |u|^{4-2}u$, we see that this result does not cover problem (C.4). On top of that, the result from Ref. [344], as the vast majority of results found in the literature [329, 343], only applies to the case of homogeneous Dirichlet boundary conditions, i.e. $u_D \equiv 0$. In the case of linear PDEs, we showed in Sec.2.1.2 how non-homogeneous case can be reduced to homogeneous conditions. This procedure cannot be readily transposed to the nonlinear case, where other techniques must be employed [340, 345]. In particular, Ref. [340] devised a new approach for dealing with problems of the form $-\Delta u = |u|^{p-2}u + h(\mathbf{x})$ with $u = u_D$ on Γ which only applies to $p < 2N/(N - 1)$, i.e. $p < 3$ for the case $N = 3$ we are interested in. These are, to the best of our knowledge, the state-of-the-art results the closest to, but unfortunately not including, problem (C.4).

[2] Specifically, this perturbation from symmetry can be attributed to the presence of the rhs function h and non-homogeneous Dirichlet boundary conditions on Γ.

C.1.3 The case $\Omega = \mathbb{R}^3$

In this PhD work, we are also interested in the unbounded case where $\Omega = \mathbb{R}^3$. In that respect, some works deal with the case of singular / super-linear elliptic PDEs posed on the whole space, see e.g. Refs. [329, 342, 346, 347]. However, we did not find any result relevant to the two forms of PDE dealt with in Secs. C.1.1 and C.1.2.

C.2 Symmetron field equation

In the case of the symmetron model, the rhs function featured in Eq. (C.1) reads $f(\mathbf{x}, s) = [\beta^2 - \rho(\mathbf{x})]s - s^3$, for some constant $\beta \in \mathbb{R}$. Here, it is interesting to note the 's^3' term in this definition of f has switched sign with respect to chameleon case discussed in Sec. C.1.2. This makes the study of the corresponding PDE tremendously simpler, as we shall see in the following. Specifically, we tackle the question of well-posedness using well-known variational methods, which are exposed in the first two chapters of Ref. [329].

C.2.1 Bounded domain

Let Ω be an open bounded subset of $\mathbb{R}^3$. The PDE problem we are interested in is

$$\begin{cases} -\alpha\Delta u + q(\mathbf{x})u = -u^3 & \text{for } \mathbf{x} \in \Omega \\ u(\mathbf{x}) = u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma \end{cases}, \tag{C.5}$$

where $q(\mathbf{x}) = [\rho(\mathbf{x}) - \beta^2]$ and the boundary data $u_D \in H^{1/2}(\Gamma)$ is assumed to be bounded. As mentioned above, all the results presented in Ref. [329] are restricted to the homogeneous case $u_D \equiv 0$. From Sec. 2.1.2, we know that the linear boundary value problem

$$\begin{cases} \Delta u = 0 & \text{for } \mathbf{x} \in \Omega \\ u(\mathbf{x}) = u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \Gamma \end{cases} \tag{C.6}$$

has a unique weak solution in $H^1(\Omega)$, that we denote by u_*. Therefore, the study of problem (C.5) is equivalent to the study of

$$\begin{cases} -\alpha\Delta u + q(\mathbf{x})u = (u + u_*)^3 + h(\mathbf{x}) & \text{in } \Omega \\ u \equiv 0 & \text{on } \Gamma \end{cases}, \tag{C.7}$$

where $h \coloneqq -qu_* \in L^2(\Omega)$ [since $q \in L^\infty(\Omega)$ and $u_* \in H^1(\Omega) \subset L^2(\Omega)$].

From there, one can define the functional

$$\begin{aligned} I &: H_0^1(\Omega) \to \mathbb{R} \\ u &\mapsto I(u) = \frac{\alpha}{2}\int_\Omega \|\boldsymbol{\nabla} u\|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int_\Omega q(\mathbf{x})u^2 \,\mathrm{d}\mathbf{x} + \frac{1}{4}\int_\Omega (u + u_*)^4 \,\mathrm{d}\mathbf{x} - \int_\Omega h(\mathbf{x})u \,\mathrm{d}\mathbf{x}\,. \end{aligned} \tag{C.8}$$

Such a functional, often called the *energy functional*, is well-defined since $q \in L^\infty(\Omega)$, $h \in L^2(\Omega)$, and the Sobolev embedding theorem guarantees that $(u + u_*) \in H^1(\Omega) \subseteq L^4(\Omega)$. It can further be shown that I is Fréchet-differentiable on $H_0^1(\Omega)$ — see Examples 1.3.17 and 1.3.20 from Ref. [329] — and its differential at $u \in H_0^1(\Omega)$ reads

$$I'(u)v = \alpha\int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + \int_\Omega q(\mathbf{x})uv \,\mathrm{d}\mathbf{x} + \int_\Omega (u + u_*)^3 v \,\mathrm{d}\mathbf{x} - \int_\Omega h(\mathbf{x})v \,\mathrm{d}\mathbf{x}\,, \quad v \in H_0^1(\Omega) \tag{C.9}$$

Here, one must realize that $u \in H_0^1(\Omega)$ is a weak solution of the PDE problem (C.7) if and only if it is a *critical point* of the energy functional (C.8), i.e. $I'(u)v = 0$ for all $v \in H_0^1(\Omega)$. It follows therefrom that studying the existence and uniqueness of weak solutions of the initial problem (C.5) can be addressed by concentrating on the properties of the functional I given by Eq. (C.8).

Usually, the existence of critical points of the energy functional is established by proving that *(i)* I is coercive, and *(ii)* I is weakly lower semi-continuous on $H_0^1(\Omega)$ so that the infimum of I is attained. However, we can do better here thanks to the following theorem.

Theorem C.2. Let λ_1 be the smallest eigenvalue of the operator $-\alpha\Delta + q(\mathbf{x})$ under homogeneous Dirichlet boundary conditions. If λ_1 is strictly positive, then the energy functional I defined by Eq. (C.8) has a unique critical point, i.e. the PDE problem (C.5) has a unique weak solution in $H_0^1(\Omega)$.

Proof. The proof is based on the two following assertions:

1. I is a continuous coercive functional;

2. I is strictly convex.

Then, by virtue of Theorems 1.5.6 and 1.5.8 from Ref. [329], I has exactly one minimum point (which is thus global). I being Fréchet-differentiable, this minimum is the only critical point in $H^1_0(\Omega)$.

1. Since $\lambda_1 > 0$, the quantity

$$(u \,|\, v) \coloneqq \alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}\mathbf{x} + \int_\Omega q(\mathbf{x}) uv \,\mathrm{d}\mathbf{x}\,, \quad u,\, v \in H^1_0(\Omega)$$

defines a scalar product on $H^1_0(\Omega)$, which induces a norm that is equivalent to the usual H^1-norm. In particular, there exists a constant $C > 0$ such that for all $u \in H^1_0(\Omega)$,

$$\begin{aligned} I(u) &\geq C \,\|u\|^2_{H^1} + \frac{1}{4}\int_\Omega (u + u_*)^4 \,\mathrm{d}\mathbf{x} - \int_\Omega h(\mathbf{x}) u \,\mathrm{d}\mathbf{x} \\ &\geq C \,\|u\|^2_{H^1} - \|h\|_{L^2} \,\|u\|_{H^1}\,, \end{aligned} \tag{C.10}$$

where we have used the Cauchy–Schwarz inequality and the fact that $\|u\|_{L^2} \leq \|u\|_{H^1}$. The latter inequality shows that the energy functional is coercive, i.e. for every sequence $(u_k)_{k\in\mathbb{N}} \in \big(H^1_0(\Omega)\big)^{\mathbb{N}}$, $\|u_k\|_{H^1} \to +\infty$ implies $I(u_k) \to +\infty$ (as $k \to +\infty$).

2. To show that I is strictly convex, let $u,\, v \in H^1_0(\Omega)$ and compute

$$\begin{aligned} \big(I'(u) - I'(v)\big)(u - v) &= \alpha \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla}(u - v) \,\mathrm{d}\mathbf{x} + \int_\Omega q(\mathbf{x}) u(u - v) \,\mathrm{d}\mathbf{x} + \int_\Omega (u + u_*)^3 (u - v) \,\mathrm{d}\mathbf{x} - \int_\Omega h(\mathbf{x})(u - v) \,\mathrm{d}\mathbf{x} \\ &\quad - \alpha \int_\Omega \boldsymbol{\nabla} v \cdot \boldsymbol{\nabla}(u - v) \,\mathrm{d}\mathbf{x} + \int_\Omega q(\mathbf{x}) v(u - v) \,\mathrm{d}\mathbf{x} + \int_\Omega (v + u_*)^3 (u - v) \,\mathrm{d}\mathbf{x} - \int_\Omega h(\mathbf{x})(u - v) \,\mathrm{d}\mathbf{x} \\ &= \big((u - v) \,|\, (u - v)\big) + \int_\Omega (u - v) \Big[(u + u_*)^3 - (v + u_*)^3\Big] \,\mathrm{d}\mathbf{x}\,. \end{aligned}$$

For arbitrary real numbers a, b, c, we have

$$(a - b)\Big[(a + c)^3 - (b + c)^3\Big] = (a - b)^2 \Big[(a + c)^2 + (b + c)^2 + (a + c)(b + c)\Big] \geq 0\,,$$

since for any $x,\, y \in \mathbb{R}$, $x^2 + y^2 \geq xy$. This proves that $\big(I'(u) - I'(v)\big)(u - v) \geq C \,\|u - v\|^2_{H^1}$, where C is same strictly positive constant as in Eq. (C.10). Ergo, I is strictly coercive, see e.g. Proposition 1.5.10 from Ref. [329]. □

C.2.2 Discussion of the case $\Omega = \mathbb{R}^3$

When Ω is equal to the whole space $\mathbb{R}^3$, Theorem C.2 does not apply directly. In particular, the definition domain of the functional I has to be replaced by some adequate space of (generalized) functions defined on $\mathbb{R}^3$. The specification of such a functional space is to be done according to physical insights into the symmetron model. For instance, assuming that the density becomes homogeneous in all directions far away from the system of interest, i.e. $\rho(\mathbf{x}) \to \rho_\infty$ as $\|\mathbf{x}\| \to +\infty$, one can expect the scalar field to stabilize towards some constant value u_∞ at spatial infinity. From the symmetron PDE, we get

$$u_\infty \Big[\big(\rho_\infty - \beta^2\big) + u_\infty^2\Big] = 0 \iff \begin{cases} u_\infty = 0\,, \text{ or} \\ u_\infty^2 = \beta^2 - \rho_\infty\,. \end{cases}$$

The case $u_\infty = 0$ is perhaps the simplest to discuss. Indeed, this special case makes it possible to look for weak solutions to the symmetron PDE in $H^1(\mathbb{R}^3)$. The energy functional I [Eq. (C.8)] is to be replaced by

$$\begin{aligned} &J\colon H^1(\mathbb{R}^3) \to \mathbb{R} \\ &\qquad u \mapsto J(u) = \frac{\alpha}{2}\int\limits_{\mathbb{R}^3} \|\boldsymbol{\nabla} u\|^2 \,\mathrm{d}\mathbf{x} + \frac{1}{2}\int\limits_{\mathbb{R}^3} q(\mathbf{x}) u^2 \,\mathrm{d}\mathbf{x} + \frac{1}{4}\int\limits_{\mathbb{R}^3} u^4 \,\mathrm{d}\mathbf{x} - \int\limits_{\mathbb{R}^3} h(\mathbf{x}) u \,\mathrm{d}\mathbf{x}\,, \end{aligned}$$

which is well-defined given the embedding $H^1(\mathbb{R}^3) \hookrightarrow L^q(\mathbb{R}^3)$ for every $q \in [2, 6]$. Furthermore, it is still Fréchet-differentiable (see e.g. Example 1.3.21 from Ref. [329]). Following the same steps as in Sec. C.2.1, one

can show that J has a single critical point in $H^1(\mathbb{R}^3)$, i.e. the problem associated with the symmetron model posed on the whole space has a unique weak solution in $H^1(\mathbb{R}^3)$ that uniformly decays to zero at spatial infinity.

However, the scenario in which $u_\infty = 0$ is not always the most physically-relevant one, especially if ρ_∞ is to be interpreted as a vacuum density (i.e. $\rho_\infty \sim 0$). To understand why, one should examine the symmetron field effective potential V_{eff} (see Box C) which, in its dimensionless form, reads

$$V_{\text{eff}}(u) = \frac{1}{2}\big[\rho(\mathbf{x}) - \beta^2\big]u^2 + \frac{1}{4}u^4\,.$$

A sketch of this $\mathbb{Z}_2$-symmetric effective potential is provided e.g. in Ref. [71], Fig. 2. In particular, the coefficient of the quadratic term, $[\rho(\mathbf{x}) - \beta^2]/2$ can be either positive or negative depending on the ambient density. In particular, it becomes negative in low density environment, where $\rho < \beta^2$, giving rise to a so-called *symmetry breaking transition* where the field can roll into one of the two minima $\pm\sqrt{\beta^2 - \rho}$. The scalar field is no longer trapped at zero, which has become a local maximum and thus constitutes an unstable point of equilibrium. In light of this physical insights, it is clear $u(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to +\infty$ is not a suitable asymptotic boundary condition for the symmetron field when $\rho_\infty < \beta^2$ and one should instead consider $u_\infty^2 = \beta^2 - \rho_\infty$. This leads to several issues:

- First, the aforementioned symmetry breaking transition can lead to the formation of so-called *domain walls*, which are the interface between adjoining regions of space where the scalar field has picked different vacuum expectation values (VEVs) — positive in some regions and negative in some others. Consequently, there is no reason why the field should relax towards the same VEV in all directions.
- Given that the scalar field does not go to zero at spatial infinity, it makes no sense to look for weak solutions in $H^1(\mathbb{R}^3)$, and other functional spaces should be considered instead. In this respect, the weighted Sobolev spaces introduced in Chapt. 3 are good candidates.
- The fact that $q(\mathbf{x})$ can become negative for large $\|\mathbf{x}\|$ threatens the coercivity of the energy functional.

Addressing all these points is beyond the scope of this appendix.

Appendix D

Mathematical proof of the vanishing gradient

Consider the following radial ODE problem

$$\forall r > 0\,, \Delta_r \phi = \rho(r) - \phi^{-(n+1)} \qquad \text{with} \quad \begin{cases} \phi'(r=0) = 0 \\ \rho(r) \underset{r\to+\infty}{\longrightarrow} \rho_{\text{vac}} \\ \phi(r) \underset{r\to+\infty}{\longrightarrow} \phi_{\text{vac}} \end{cases}, \tag{D.1}$$

where Δ_r refers to the radial part of the Laplacian expressed in spherical coordinates in 3 dimensions

$$\Delta_r f = \frac{1}{r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2 \frac{\mathrm{d}f}{\mathrm{d}r}\right) = \frac{\mathrm{d}^2 f}{\mathrm{d}r^2} + \frac{2}{r}\frac{\mathrm{d}f}{\mathrm{d}r}\,, \text{ for any } f \in \mathcal{C}^2(\mathbb{R}_+, \mathbb{R}) \tag{D.2}$$

and

$$\phi_{\text{vac}} = (\rho_{\text{vac}})^{-\frac{1}{n+1}}\,. \tag{D.3}$$

The asymptotic values of ρ and ϕ are such that the rhs of the ODE (D.1) vanishes at infinity, which readily implies that $\Delta_r \phi \to 0$ as $r \to \infty$. The goal of this appendix is to show that the solution of Eq. (D.1) — provided it exists and is unique — is such that

$$\phi'(r) \underset{r\to+\infty}{\longrightarrow} 0\,, \tag{D.4}$$

i.e. the radial gradient also vanishes at infinity.

D.1 Proof that $\phi''(r) \to 0$ as $r \to +\infty$

The asymptotic condition on the radial part of the Laplacian [Eq. (D.2)] may be reformulated as:

$$\text{there exists a function } \epsilon : \mathbb{R}_+^* \to \mathbb{R} \text{ such that } \begin{cases} \phi''(r) + \dfrac{2}{r}\phi'(r) = \epsilon(r) \\ \epsilon(r) \underset{r\to+\infty}{\longrightarrow} 0 \end{cases}. \tag{D.5}$$

The above is nothing but a second-order linear ordinary differential equation (ODE) which can be solved via the method of variation of parameters. The general solution of the homogeneous equation can be expressed as $-A/r + B$, with $A, B \in \mathbb{R}$. Then a particular solution of the inhomogeneous equation is sought in the form $\phi(r) = -A(r)/r + B(r)$, with A and B two real functions satisfying the system

$$\begin{cases} -A'(r)/r + B'(r) = 0 \\ A'(r)/r^2 + 0 = \epsilon(r) \end{cases} \iff \begin{cases} A'(r) = r^2\epsilon(r) \\ B'(r) = r\epsilon(r) \end{cases}. \tag{D.6}$$

Therefore, a particular solution of the ODE on $\mathbb{R}_+^*$ is

$$\phi(r) = -\frac{1}{r}\int_1^r s^2\epsilon(s)\,\mathrm{d}s + \int_1^r s\epsilon(s)\,\mathrm{d}s\,. \tag{D.7}$$

The general solution then reads

$$\phi(r) = -\frac{1}{r}\left[\int_1^r s^2\epsilon(s)\,\mathrm{d}s + A\right] + \int_1^r s\epsilon(s)\,\mathrm{d}s + B \quad , \quad A, B \in \mathbb{R}\,. \tag{D.8}$$

From there, we can compute the second order derivative as

$$\phi''(r) = -\frac{2}{r^3}\left[\int_1^r s^2\epsilon(s)\,\mathrm{d}s + A\right] + \epsilon(r) \tag{D.9}$$

and the proof boils down to showing that

$$\frac{1}{r^3}\int_1^r s^2\epsilon(s)\,\mathrm{d}s \underset{r\to+\infty}{\longrightarrow} 0\,. \tag{D.10}$$

Let $\delta > 0$, $\epsilon(r) \underset{r\to+\infty}{\longrightarrow} 0$ hence there exists $R_\delta > 0$ such that for all $r \geq R_\delta$, $|\epsilon(r)| < \delta$. Let us introduce

$$M \coloneqq \max_{s\in[1,+\infty[} |\epsilon(s)| \quad \text{and} \quad R_* \coloneqq \frac{R_\delta M}{\delta}\,. \tag{D.11}$$

For $r \geq \max(R_*, R_\delta) =: R_\mathrm{m}$, we get:

$$\begin{aligned}
|I(r)| \coloneqq \left|\frac{1}{r^3}\int_1^r s^2\epsilon(s)\,\mathrm{d}s\right| &= \left|\frac{1}{r}\int_1^r \underbrace{\left(\frac{s}{r}\right)^2}_{\leq 1}\epsilon(s)\,\mathrm{d}s\right| \leq \frac{1}{r}\int_1^r |\epsilon(s)|\,\mathrm{d}s \\
&\leq \frac{1}{r}\int_1^{R_\delta} |\epsilon(s)|\,\mathrm{d}s + \frac{1}{r}\int_{R_\delta}^r |\epsilon(s)|\,\mathrm{d}s \\
&\leq \frac{1}{r}\int_1^{R_\delta} \max_{s\in[1,R_\delta]} |\epsilon(s)|\,\mathrm{d}s + \frac{1}{r}\int_{R_\delta}^r \delta\,\mathrm{d}s \\
&\leq \frac{R_\delta - 1}{r}\max_{s\in[1,R_\delta]} |\epsilon(s)| + \frac{r - R_\delta}{r}\delta \\
&\leq \frac{R_\delta M}{r} + \delta \leq \frac{R_\delta M}{R_*} + \delta \leq \delta + \delta \leq 2\delta\,.
\end{aligned} \tag{D.12}$$

We have shown that $\forall\delta > 0\ ,\ \exists R_\mathrm{m} > 0\ /\ \forall r > R_\mathrm{m}\ , |I(r)| \leq \delta$, which is the exact definition of $I(r) \underset{r\to+\infty}{\longrightarrow} 0$ and concludes the first part of the proof.

D.2 Proof that $\phi'(r) \to 0$ as $r \to +\infty$

Let $f \in \mathcal{C}^2(\mathbb{R}_+, \mathbb{R})$ be such that

$$\begin{cases} f \text{ has a limit } l \text{ as } x \text{ approaches } +\infty \\ f'' \text{ goes to 0 as } x \text{ approaches } +\infty \end{cases}\,. \tag{D.13}$$

These two hypotheses can be rewritten in a more mathematical formalism as

$$[f'' \text{ goes to } 0] \quad \forall\epsilon > 0\ ,\ \exists M \in \mathbb{R}_+\ /\ \forall x \geq M\ ,\ |f''(x)| \leq \epsilon\,, \tag{D.14}$$

$$[f \text{ goes to } l] \quad \forall\epsilon > 0\ ,\ \exists M \in \mathbb{R}_+\ /\ \forall x \geq M\ ,\ |f(x) - l| \leq \epsilon\,. \tag{D.15}$$

The fact that f converges allows us to write a third proposition that slightly differs from (D.15)

$$[f \text{ converges}] \quad \forall\epsilon > 0\ ,\ \exists M \in \mathbb{R}_+\ /\ \forall x_1, x_2 \geq M\ ,\ |f(x_1) - f(x_2)| \leq \epsilon\,. \tag{D.16}$$

Strategy: We develop a proof by contradiction. To that end, let us suppose that f' does not go to 0 at $+\infty$, that is

$$\exists\delta > 0\ /\ \forall A \in \mathbb{R}_+\ ,\ \exists x \geq A\ /\ |f'(x)| > \delta\,. \tag{D.17}$$

Property (D.17) provides us with $\delta > 0$. Even if it means redefining $f \leftarrow -f$, one can get rid of the absolute value in (D.17) so that

$$\forall A \in \mathbb{R}_+ \ , \ \exists x \geq A \ / \ f'(x) > \delta \, . \tag{D.18}$$

Note that this potential change of sign does not change in any way the asymptotic behavior of f' and f''. From here, the proof follows the subsequent steps.

1. f' reaches δ for arbitrarily large x.

More precisely, let us demonstrate that $\forall A > 0 \ , \ \exists x \geq A \ / \ f'(x) = \delta$. Let $A > 0$, according to (D.18), there exists $x_m \geq A$ such that $f'(x_m) > \delta$. We employ reductio ad absurdum, assuming that for all $x \geq x_m$, $f'(x) \neq \delta$. Because f' is continuous over $\mathbb{R}_+$, this implies that $\forall x \geq x_m$, $f'(x) > \delta$. This statement is in contradiction with the convergence of f. Indeed, let $\epsilon > 0$ and get $M \in \mathbb{R}_+$ given by property (D.16). We set

$$x_1 \coloneqq \max(x_m, M) \quad \text{and} \quad x_2 \coloneqq x_1 + \frac{2}{\delta}\epsilon \, . \tag{D.19}$$

On the one hand,

$$|f(x_1) - f(x_2)| \leq \epsilon \quad \text{because } x_1, x_2 \geq M \, , \tag{D.20}$$

and on the other hand, $\forall x \in [x_1, x_2]$, $f'(x) \geq \delta$ so that the mean value inequality gives

$$\int_{x_1}^{x_2} f'(x) \, \mathrm{d}x \geq \int_{x_1}^{x_2} \delta \, \mathrm{d}x \quad \text{thus} \quad |f(x_1) - f(x_2)| \geq f(x_2) - f(x_1) \geq \delta |x_2 - x_1| = \delta \frac{2}{\delta}\epsilon = 2\epsilon > 0 \, . \tag{D.21}$$

The contradiction is now clear.

2. f' reaches $\delta/2$ for arbitrarily large x.

Using the exact sames arguments as above, one proves that $\forall A > 0 \ , \ \exists x \geq A \ / \ f'(x) = \delta/2$. Before going any further, we define two sets:

$$E_\delta \coloneqq \{x \in \mathbb{R}_+ \text{ such that } f'(x) = \delta\} \quad \text{and} \quad E_{\delta/2} \coloneqq \left\{ x \in \mathbb{R}_+ \text{ such that } f'(x) = \frac{\delta}{2} \right\} . \tag{D.22}$$

We have just shown that these two sets are infinite and that they contain arbitrarily large values of x.

3. Construction of the interval sequence $(I_n)_{n\in\mathbb{N}}$.

The aim of this part is to show that f'-values stay between $\delta/2$ and δ on arbitrarily large intervals. To that extent, we construct a sequence of disjoint intervals $(I_n)_{n\in\mathbb{N}}$ such that f' falls between $\delta/2$ and δ on each I_n:

- For I_0, we set $x_{0,\delta}$ in E_δ and $x_{0,\delta/2}$ in $E_{\delta/2}$ such that $x_{0,\delta} < x_{0,\delta/2}$ and $\forall x \in [x_{0,\delta}, x_{0,\delta/2}]$, $f'(x) \in [\delta/2, \delta]$.
- For I_1, we choose $x_{1,\delta}$ in $E_\delta \cap]x_{0,\delta} + \infty]$ and $x_{1,\delta/2}$ in $E_{\delta/2} \cap]x_{0,\delta/2} + \infty]$ such that $x_{1,\delta} < x_{1,\delta/2}$ and $\forall x \in [x_{1,\delta}, x_{1,\delta/2}]$, $f'(x) \in [\delta/2, \delta]$. By construction, I_1 and I_0 are indeed disjoints.
- For I_2, we choose $x_{2,\delta}$ in $E_\delta \cap]x_{1,\delta} + \infty]$ and $x_{2,\delta/2}$ in $E_{\delta/2} \cap]x_{1,\delta/2} + \infty]$ such that ...
- etc.

This construction is illustrated on Fig. D.1. We now demonstrate that

$$\forall X, A > 0, \exists I \in (I_n)_{n\in\mathbb{N}} \text{ such that } \begin{cases} \inf(I) \geq X \\ |I| \geq A \end{cases} . \tag{D.23}$$

Let $X, A > 0$ and set $\epsilon = A^{-1}$. We make use of the fact that f'' goes to 0 by applying property (D.14) for $\epsilon\delta/2 > 0$. Let us denote $M \geq 0$ the constant provided with this property and set $R \coloneqq \max(X, M)$. According to what has been shown in the previous point, one can choose an element $I = [a, b]$ of the sequence $(I_n)_{n\in\mathbb{N}}$ such that $I \subset [R, +\infty[$. The hypotheses of the mean value inequality are verified, namely:

- f' is continuous over $[a, b]$;
- f' is differentiable over $]a, b[$;
- for all $x \in]a, b[$, $f''(x) \leq \epsilon\delta/2$ (since $x \geq M$);

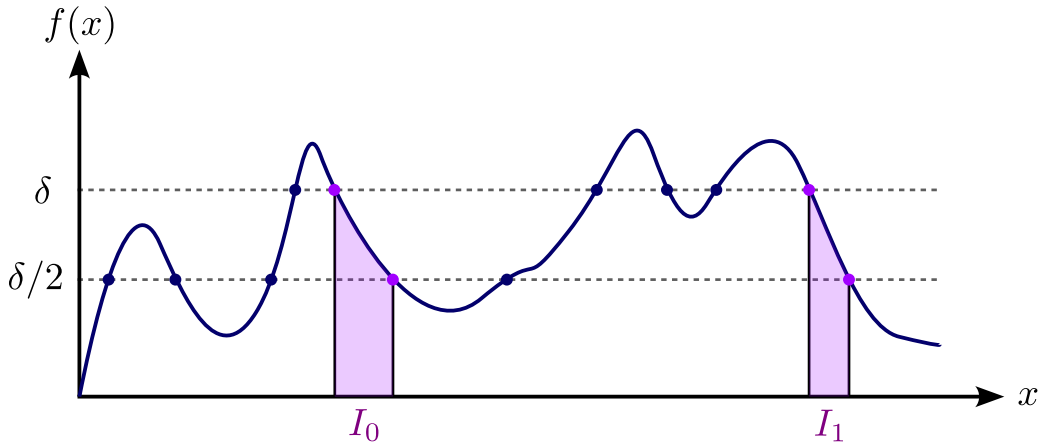


Figure D.1: Construction of the $(I_n)_{n\in\mathbb{N}}$ sequence.

so that

$$\left|\frac{f'(b)-f'(a)}{b-a}\right| \leq \frac{\delta}{2}\epsilon\,. \tag{D.24}$$

Yet, by definition of I, $f'(a)=\delta$, $f'(b)=\delta/2$ and $(b-a)=|I|$. The above inequality therefore boils down to

$$\frac{\delta-\frac{\delta}{2}}{|I|} \leq \frac{\delta}{2}\epsilon \quad\iff\quad \frac{1}{|I|} \leq \epsilon \iff |I| \geq A\,, \tag{D.25}$$

which concludes the proof.

4. Contradiction.

Finally, we use the convergence of f to bring out a contradiction. Let $\epsilon>0$ and $M\geq 0$ the constant associated to property (D.16). According to the previous point, there exists $I\in(I_n)_{n\in\mathbb{N}}$ such that

$$\begin{cases} I \subset [M,+\infty[\\ |I| \geq \dfrac{4}{\delta}\epsilon \end{cases}. \tag{D.26}$$

Let us denote $[a,b]\coloneqq I$. On the one hand, the convergence of f provides the inequality

$$|f(b)-f(a)| \leq \epsilon \quad \text{because } a,b\geq M\ , \tag{D.27}$$

and on the other hand, $\forall x\in[a,b]$, $f'(x)\geq\delta/2$ so that the mean value inequality gives

$$\int_a^b f'(x)\,\mathrm{d}x \geq \int_a^b \frac{\delta}{2}\,\mathrm{d}x \text{ hence } |f(b)-f(a)| \geq f(b)-f(a) \geq \frac{\delta}{2}|I| \geq \frac{\delta}{2}\frac{4}{\delta}\epsilon = 2\epsilon > 0\,. \tag{D.28}$$

The contradiction is clear. Q.E.D.

Appendix E

Solving ordinary differential equation with projection on constraint space

In Chapt. 5, we solved the ODE system governing the dynamics of a spacecraft whose orbit lies in a plane, where the conservation of energy was numerically enforced through the use of a *projection technique*. This appendix is devoted to the presentation of such a technique and provides an example of implementation in PYTHON. It is mainly based on Refs. [348, 349].

E.1 Statement of the problem

Differential Algebraic Equations (DAE) are a generalization of Ordinary Differential Equations (ODE). Schematically, one has

$$\dot{y} = f(t, y) \text{ for and ODE,} \quad \text{and} \quad g(t, y, \dot{y}) = 0 \text{ for a DAE,}$$

where $f\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and $g\colon \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$, with $m, n \in \mathbb{N}^*$. DAE arise in physics when we want to account for conservation laws, in which case the traditional equations of dynamics (which are generally second-order differential equations) are supplemented with one or several algebraic equations. Take for example the spring-mass system. The dynamics boil down to the familiar harmonic oscillator equation

$$\begin{cases} m\ddot{x} = kx \\ x(t=0) = x_0 \\ \dot{x}(t=0) = v_0 \end{cases} \tag{E.1}$$

The mechanical energy is conserved along the trajectory of the mass, that is

$$E = \frac{1}{2}\left(m\dot{x}^2 + kx^2\right) = C^{st}\,.$$

Of course, energy conservation can be derived from Eq. (E.1) by multiplying both sides of the equation by $\dot{x}$ and integrating. The energy of a conservative system is an example of a first integral. The two formulations are equivalent.

Yet, from a numerical perspective, there is no reason for our favorite ODE solver to preserve energy conservation when solving the ODE (E.1). In some applications, we could be interested in ensuring that the energy of the system remains constant over long periods of time (e.g. in celestial mechanics). In fact, most of the time, doing nothing particular to ensure this condition will result in either a steady increase of the energy over time or a dissipation, both phenomena originating from the discrete numerical scheme employed (see Fig. E.1, left column). We thus may be tempted to solve both problems at once, that is

$$\begin{cases} m\ddot{x} = kx \\ \frac{1}{2}\left(m\dot{x}^2 + kx^2\right) - E_0 = 0 \end{cases} \quad \text{with} \quad \begin{cases} x(t=0) = x_0 \\ \dot{x}(t=0) = v_0 \end{cases} . \tag{E.2}$$

This is an example of an over-determined DAE. Indeed, setting $y = (x, \dot{x})$, we can recast this system in the form described above $g(t, y, \dot{y}) = 0$ with $g\colon \mathbb{R} \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^3$: we have two degrees of freedom and three equations to satisfy...

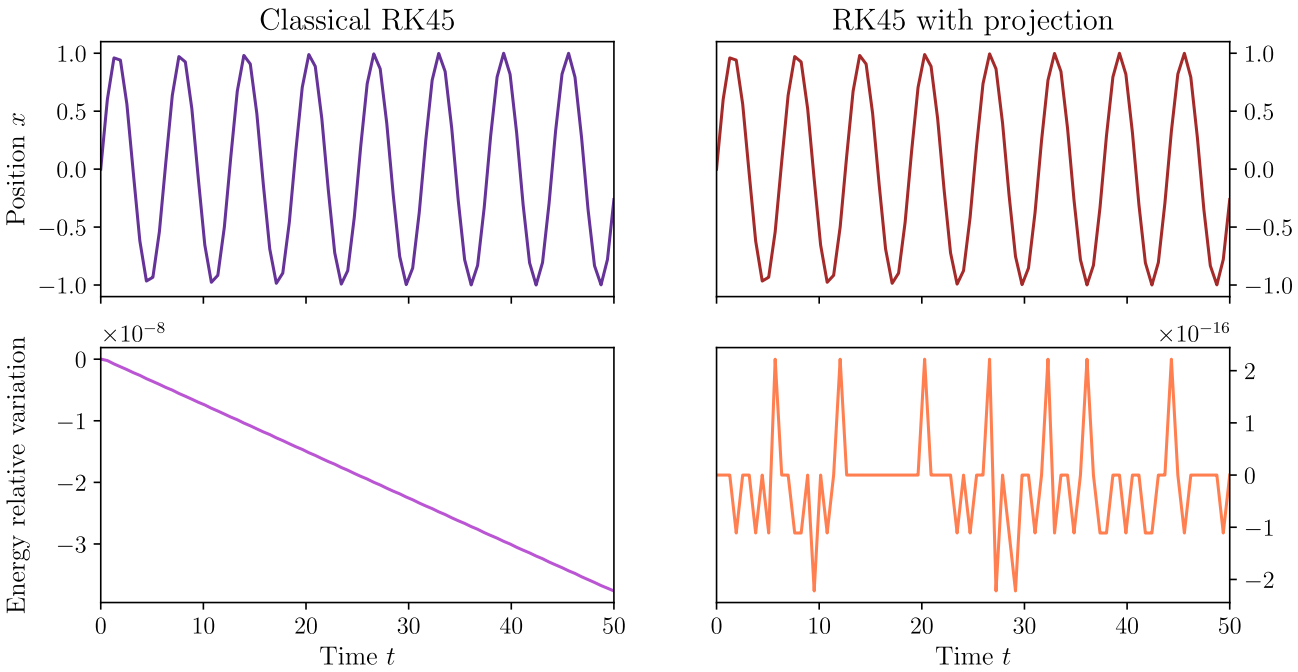


Figure E.1: Numerical solution of the spring-mass system for $(x_0, v_0) = (0, 1)$ (first row) together with the relative variation of the energy $(E(t) - E_0)/E_0$ (second row). The left column represents results obtained with the traditional *RK45* method, available with `scipy.integrate.solve_ivp`. The right column is obtained by modifying that same routine to enable the projection of the state onto the constraint manifold at each time step. Despite the graphs of $x(t)$ looking similar to the naked eye, the energy is not preserved with the RK45 scheme (numerical dissipation) whereas the projection technique allows the energy to remain constant to within 2 parts in 10^{16} (basically, what numerical precision allows).

E.2 Imposing constraints through projection

Projection is a straight forward way to preserve a given first integral. Formally, the constraint can be thought of as a manifold $\mathcal{M} = \{y \in \mathbb{R}^n\,,\ I(y) = 0\}$, where $I\colon \mathbb{R}^n \to \mathbb{R}$. Projection is going to be applied at the end of each discrete time step of the arbitrary numerical scheme employed. Let us denote by y_k the numerical approximation at time t_k. Going from y_k to y_{k+1} takes two steps:

1. Use the set of ODE $\dot{y} = f(t, y)$ and some numerical solver (e.g. RK4) to compute an approximation $\tilde{y}_{k+1}$ at time t_{k+1}.
2. Project $\tilde{y}_{k+1}$ onto the constraint sub-manifold $\mathcal{M}$ using some projector $P\colon \mathbb{R}^n \to \mathbb{R}^n$ to be defined, giving $y_{k+1} = P(\tilde{y}_{k+1})$.

For the first step, there is nothing particular to be discussed as numerical solvers are widely available across all languages. In Python, one can use `scipy.integrate.solve_ivp` which implements few *classics* such as `RK45`, `Radau`, `BDF`, etc. When it comes to the projection itself, there seem to be really few publicly available codes, in spite of the wide range of potential applications. We thus dedicate the remainder of this note to the computation of a well-suited projector and its implementation in an existing solver.

E.2.1 Projection techniques

Orthogonal projection

The first idea that comes to mind is to compute $y \in \mathcal{M}$ such that it minimizes the quantity $\|y - \tilde{y}\|^2$ ($\|\cdot\|$ is the 2-norm in $\mathbb{R}^n$ here). This is equivalent to introducing a *Lagrange multiplier* $\lambda \in \mathbb{R}$ such that

$$\begin{cases} y = \tilde{y} + \lambda \nabla I(y) \\ I(y) = 0 \end{cases}. \tag{E.3}$$

We commonly refer to this projector $P\colon \tilde{y} \mapsto y$ as the *orthogonal projector*, because we are looking for the nearest point on $\mathcal{M}$ that satisfies the constraint. However, problem (E.3) is implicit which makes it difficult to solve at low computational cost.

Oblique projection

Oblique projection relaxes the constraint of finding *the* nearest point on the sub-manifold $\mathcal{M}$. Instead, we look for *a* projector that will give us a small enough deviation $\delta y \coloneqq P(\tilde{y}) - \tilde{y}$ with respect to the size of the solver's step $\tilde{y}_{k+1} - y_k$. A *natural* way to achieve this is to approximate the gradient of I at point y in (E.3) by the gradient of I at point $\tilde{y}$. That is we look for $\lambda \in \mathbb{R}$ such that

$$\begin{cases} y = \tilde{y} + \lambda \nabla I(\tilde{y}) \\ I(y) = 0 \end{cases} . \tag{E.4}$$

This projection technique, because it is slightly different from the orthogonal projection presented above, is referred to as an *oblique projection*. The problem is no longer implicit. Indeed, we can substitute the first equation of (E.4) into the second, yielding $I(\tilde{y} + \lambda \nabla I(\tilde{y})) = 0$, which is a nonlinear equation in λ. It can be solved with some *root finding* algorithm.

It is possible to go one step further. Letting $I_s \colon \lambda \in \mathbb{R} \mapsto I(\tilde{y} + \lambda \nabla I(\tilde{y})) \in \mathbb{R}$, we have an explicit form for the derivative of I_s:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} I_s(\lambda) = \nabla I(y) \cdot \nabla I(\tilde{y}) \ , \quad \text{with} \quad y = \tilde{y} + \lambda \nabla I(\tilde{y}) \, .$$

This is a valuable expression as it lets us employ the Newton method as our root finding algorithm, converging very quickly towards the actual root (at most two iterations were needed for the spring-mass system, see Fig. E.1).

E.2.2 Implementation

As we saw above, the projection technique is minimally intrusive as it only requires to change the state vector *after* the solver's step. This means that it can be plugged virtually on top of any general purpose ODE solver. We have successfully implemented the oblique projection by redefining the Runge–Kutta solver at `scipy.integrate._ivp.rk`.

```
    def cons_proj(fun, t, y, cons, grad_cons):
    """
    Project the current state `y` onto the constraint manifold using an oblique
    projection technique.

    Parameters
    ----------
    fun : callable
        Right-hand side of the system.
    t : float
        Current time.
    y : ndarray, shape (n,)
        Current state.
    cons : callable
        Function defining the constraint I(t, X) = 0.
    grad_cons : callable
        Function defining the gradient of the constraint function at a given
        position X, i.e. $\nabla_{X} I$.

    Returns
    -------
    y_new : ndarray, shape (n,)
        Solution at t + h computed with a higher accuracy.
    f_new : ndarray, shape (n,)
        Derivative ``fun(t + h, y_new)``.

    """

    y_lbd = lambda lbd : y + lbd * grad_cons(t, y)
    cons_lbd = lambda lbd : cons(t, y_lbd(lbd))
    cons_lbd_prime = lambda lbd : np.dot(grad_cons(t, y_lbd(lbd)), grad_cons(t, y))
    sol = root_scalar(cons_lbd, x0=0, fprime=cons_lbd_prime, method='newton')
    lbd = sol.root
    y_new = y_lbd(lbd)
    f_new = fun(t, y_new)
    return y_new, f_new
```

```
class RungeKuttaProj(OdeSolver):
r"""
Modified class for explicit Runge-Kutta methods. This portion of code
originate from `scipy.integrate._ivp.rk` and has been modified to allow for
a projection of the current state onto some constraint manifold.

New arguments:
--------------
cons : callable
    Function defining the constraint I(t, X) = 0.
grad_cons : callable
    Function defining the gradient of the constraint function at a given
    position X, i.e. $\nabla_{X} I$.
"""
# ============================
# Skipping untouched code...
# ============================
def __init__(self, fun, t0, y0, t_bound,
             max_step=np.inf, rtol=1e-3, atol=1e-6, vectorized=False,
             first_step=None, **extraneous):
    super().__init__(fun, t0, y0, t_bound, vectorized,
                     support_complex=True)
    cons = extraneous.get('cons', None)
    grad_cons = extraneous.get('grad_cons', None)
    if (cons is None) or (grad_cons is None):
        raise ValueError(
            "User must provide functions `cons` and `grad_cons`!")
    self.cons = cons
    self.grad_cons = grad_cons
    # ...

def _step_impl(self):
    t = self.t
    y = self.y
    step_accepted = False
    # ...

    while not step_accepted:
        t_new = t + h
        # ...

    # After the step is accepted, we project the state onto the
    # constraint's manifold

    y_new, f_new = cons_proj(self.fun, t_new, y_new, self.cons, self.grad_cons)
    # ...
```

The full code, together with an example script demonstrating its use on the spring-mass system, are available as supplementary material of Ref. [141].

Bibliography

[1] C. M. Will. *Theory and Experiment in Gravitational Physics*. Cambridge: Cambridge University Press, 2018. DOI: https://doi.org/10.1017/9781316338612.

[2] A. Einstein. "Die Grundlage der allgemeinen Relativitätstheorie". In: *Annalen der Physik* 354.7 (1916), pp. 769–822. DOI: https://doi.org/10.1002/andp.19163540702.

[3] C. M. Will. "The Confrontation between General Relativity and Experiment". In: *Living Reviews in Relativity* 17.1 (Dec. 2014), p. 4. ISSN: 1433-8351. DOI: 10.12942/lrr-2014-4.

[4] P. Peter and J.-P. Uzan. *Primordial Cosmology*. Oxford: Oxford University Press, Feb. 2013. ISBN: 9780199665150.

[5] G. Nordström. In: *Physikalische Zeitschrift* 13 (1912), p. 1126.

[6] G. Nordström. "Zur Theorie der Gravitation vom Standpunkt des Relativitätsprinzips". In: *Annalen der Physik* 347.13 (1913), pp. 533–554. DOI: https://doi.org/10.1002/andp.19133471303.

[7] G. Nordström. "Träge und schwere Masse in der Relativitätsmechanik". In: *Annalen der Physik* 345.5 (1913), pp. 856–878. DOI: https://doi.org/10.1002/andp.19133450503.

[8] T. Kaluza. "Zum Unitätsproblem der Physik". In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften* (Jan. 1921), pp. 966–972.

[9] O. Klein. "Quantentheorie und fünfdimensionale Relativitätstheorie". In: *Zeitschrift für Physik* 37.12 (Dec. 1926), pp. 895–906. ISSN: 0044-3328. DOI: 10.1007/BF01397481.

[10] O. Klein. "The Atomicity of Electricity as a Quantum Theory Law". In: *Nature* 118.2971 (Oct. 1926), pp. 516–516. ISSN: 1476-4687. DOI: 10.1038/118516a0.

[11] H. Weyl. "Eine neue Erweiterung der Relativitätstheorie". In: *Annalen der Physik* 364.10 (1919), pp. 101–133. DOI: https://doi.org/10.1002/andp.19193641002.

[12] S. M. Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2019.

[13] R. M. Wald. *General relativity*. Chicago, IL: Chicago University Press, 1984. URL: https://cdn.preterhuman.net/texts/science%5C_and%5C_technology/physics/General%5C_Relativity%5C_Theory/General%20Relativity%20-%20R.%20Wald.pdf.

[14] S. Weinberg. *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*. New York: John Wiley and Sons, 1972. ISBN: 978-0-471-92567-5.

[15] J. Perez. *Théorie des champs. Les équations de la physique*. Collection "Les cours". 828, Boulevard des maréchaux, 91120 Palaiseau: Les Presses de l'ENSTA, 2022. ISBN: 978-2-7225-0962-7. URL: https://physique.ensta-paris.fr/PAT/tdc_troisieme_edition.pdf.

[16] G. Efstathiou and S. Gratton. "The evidence for a spatially flat Universe". In: *Monthly Notices of the Royal Astronomical Society: Letters* 496.1 (May 2020), pp. L91–L95. ISSN: 1745-3925. DOI: 10.1093/mnrasl/slaa093. eprint: https://academic.oup.com/mnrasl/article-pdf/496/1/L91/56979863/mnrasl_496_1_l91.pdf.

[17] P. Collaboration. "Planck 2018 results. VI. Cosmological parameters". In: *Astronomy & Astrophysics* 641, A6 (Sept. 2020), A6. DOI: 10.1051/0004-6361/201833910.

[18] S. Chandrasekhar. "The general theory of relativity: Why "It is probably the most beautiful of all existing theories"". In: *Journal of Astrophysics and Astronomy* 5.1 (Mar. 1984), pp. 3–11. DOI: 10.1007/BF02714967.

[19] S. Schlamminger et al. "Test of the Equivalence Principle Using a Rotating Torsion Balance". In: *Phys. Rev. Lett.* 100 (4 Jan. 2008), p. 041101. DOI: 10.1103/PhysRevLett.100.041101.

[20] T. A. Wagner et al. "Torsion-balance tests of the weak equivalence principle". In: *Classical and Quantum Gravity* 29.18 (Aug. 2012), p. 184002. DOI: 10.1088/0264-9381/29/18/184002.

[21] S. M. Merkowitz. "Tests of Gravity Using Lunar Laser Ranging". In: *Living Reviews in Relativity* 13.1 (Nov. 2010), p. 7. ISSN: 1433-8351. DOI: 10.12942/lrr-2010-7.

[22] S. Baeßler et al. "Improved Test of the Equivalence Principle for Gravitational Self-Energy". In: *Phys. Rev. Lett.* 83 (18 Nov. 1999), pp. 3585–3588. DOI: 10.1103/PhysRevLett.83.3585.

[23] P. Touboul et al. "MICROSCOPE Mission: Final Results of the Test of the Equivalence Principle". In: *Phys. Rev. Lett.* 129 (12 Sept. 2022), p. 121102. DOI: 10.1103/PhysRevLett.129.121102.

[24] P. Touboul et al. "Result of the MICROSCOPE weak equivalence principle test". In: *Classical and Quantum Gravity* 39.20 (Sept. 2022), p. 204009. DOI: 10.1088/1361-6382/ac84be.

[25] R. F. C. Vessot et al. "Test of Relativistic Gravitation with a Space-Borne Hydrogen Maser". In: *Phys. Rev. Lett.* 45 (26 Dec. 1980), pp. 2081–2084. DOI: 10.1103/PhysRevLett.45.2081.

[26] S. Herrmann et al. "Test of the Gravitational Redshift with Galileo Satellites in an Eccentric Orbit". In: *Phys. Rev. Lett.* 121 (23 Dec. 2018), p. 231102. DOI: 10.1103/PhysRevLett.121.231102.

[27] P. Delva et al. "Gravitational Redshift Test Using Eccentric Galileo Satellites". In: *Phys. Rev. Lett.* 121 (23 Dec. 2018), p. 231101. DOI: 10.1103/PhysRevLett.121.231101.

[28] K. Nordtvedt. "Equivalence Principle for Massive Bodies. I. Phenomenology". In: *Phys. Rev.* 169 (5 May 1968), pp. 1014–1016. DOI: 10.1103/PhysRev.169.1014.

[29] K. Nordtvedt. "Equivalence Principle for Massive Bodies. II. Theory". In: *Phys. Rev.* 169 (5 May 1968), pp. 1017–1025. DOI: 10.1103/PhysRev.169.1017.

[30] K. Nordtvedt. "Testing Relativity with Laser Ranging to the Moon". In: *Phys. Rev.* 170 (5 June 1968), pp. 1186–1187. DOI: 10.1103/PhysRev.170.1186.

[31] V. Viswanathan et al. "The new lunar ephemeris INPOP17a and its application to fundamental physics". In: *Monthly Notices of the Royal Astronomical Society* 476.2 (Jan. 2018), pp. 1877–1888. ISSN: 0035-8711. DOI: 10.1093/mnras/sty096. eprint: https://academic.oup.com/mnras/article-pdf/476/2/1877/24327360/sty096.pdf.

[32] A. S. Konopliv et al. "Mars high resolution gravity fields from MRO, Mars seasonal gravity, and other dynamical parameters". In: *Icarus* 211.1 (2011), pp. 401–428. ISSN: 0019-1035. DOI: https://doi.org/10.1016/j.icarus.2010.10.004.

[33] E. V. Pitjeva. "Relativistic effects and solar oblateness from radar observations of planets and spacecraft". In: *Astronomy Letters* 31.5 (May 2005), pp. 340–349. ISSN: 1562-6873. DOI: 10.1134/1.1922533.

[34] N. Deruelle. "Nordström's scalar theory of gravity and the equivalence principle". In: *General Relativity and Gravitation* 43.12 (Dec. 2011), pp. 3337–3354. ISSN: 1572-9532. DOI: 10.1007/s10714-011-1247-x.

[35] C. Will. *Theory and Experiment in Gravitational Physics*. Cambridge University Press, 1993. ISBN: 9780521439732. URL: https://books.google.fr/books?id=BhnUITA7sDIC.

[36] B. Bertotti, L. Iess, and P. Tortora. "A test of general relativity using radio links with the Cassini spacecraft". In: *Nature* 425.6956 (Sept. 2003), pp. 374–376. ISSN: 1476-4687. DOI: 10.1038/nature01997.

[37] E. Fomalont et al. "Progress in Measurements of the Gravitational Bending of Radio Waves using the VLBA". In: *The Astrophysical Journal* 699.2 (June 2009), p. 1395. DOI: 10.1088/0004-637X/699/2/1395.

[38] Lambert, S. B. and Le Poncin-Lafitte, C. "Improved determination of γ by VLBI". In: *Astronomy & Astrophysics* 529 (2011), A70. DOI: 10.1051/0004-6361/201016370.

[39] A. Fienga et al. "The INPOP10a planetary ephemeris and its applications in fundamental physics". In: *Celestial Mechanics and Dynamical Astronomy* 111.3 (Sept. 2011), p. 363. ISSN: 1572-9478. DOI: 10.1007/s10569-011-9377-8.

[40] A. Genova et al. "Solar system expansion and strong equivalence principle as seen by the NASA MESSENGER mission". In: *Nature Communications* 9.1 (Jan. 2018), p. 289. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02558-1.

[41] L. Shao and N. Wex. "New limits on the violation of local position invariance of gravity". In: *Classical and Quantum Gravity* 30.16 (July 2013), p. 165020. DOI: 10.1088/0264-9381/30/16/165020.

[42] L. Shao and N. Wex. "New tests of local Lorentz invariance of gravity with small-eccentricity binary pulsars". In: *Classical and Quantum Gravity* 29.21 (Oct. 2012), p. 215018. DOI: 10.1088/0264-9381/29/21/215018.

[43] J. F. Bell and T. Damour. "A new test of conservation laws and Lorentz invariance in relativistic gravity". In: *Classical and Quantum Gravity* 13.12 (Dec. 1996), p. 3121. DOI: 10.1088/0264-9381/13/12/003.

[44] I. H. Stairs et al. "Discovery of Three Wide-Orbit Binary Pulsars: Implications for Binary Evolution and Equivalence Principles". In: *The Astrophysical Journal* 632.2 (Oct. 2005), p. 1060. DOI: 10.1086/432526.

[45] C. M. Will. "Is Momentum Conserved? A Test in the Binary System PSR 1913+16". In: *The Astrophysical Journal Letters* 393 (July 1992), p. L59. DOI: 10.1086/186451.

[46] B. P. Abbott et al. "Observation of Gravitational Waves from a Binary Black Hole Merger". In: *Phys. Rev. Lett.* 116 (6 Feb. 2016), p. 061102. DOI: 10.1103/PhysRevLett.116.061102.

[47] N. V. Krishnendu and F. Ohme. "Testing General Relativity with Gravitational Waves: An Overview". In: *Universe* 7.12 (2021). ISSN: 2218-1997. DOI: 10.3390/universe7120497.

[48] B. P. Abbott et al. "Tests of general relativity with the binary black hole signals from the LIGO-Virgo catalog GWTC-1". In: *Phys. Rev. D* 100 (10 Nov. 2019), p. 104036. DOI: 10.1103/PhysRevD.100.104036.

[49] B. P. Abbott et al. "Multi-messenger Observations of a Binary Neutron Star Merger". In: *The Astrophysical Journal Letters* 848.2 (Oct. 2017). DOI: 10.3847/2041-8213/aa91c9.

[50] T. Baker et al. "Strong Constraints on Cosmological Gravity from GW170817 and GRB 170817A". In: *Phys. Rev. Lett.* 119 (25 Dec. 2017), p. 251301. DOI: 10.1103/PhysRevLett.119.251301.

[51] P. Creminelli and F. Vernizzi. "Dark Energy after GW170817 and GRB170817A". In: *Phys. Rev. Lett.* 119 (25 Dec. 2017), p. 251302. DOI: 10.1103/PhysRevLett.119.251302.

[52] J. M. Ezquiaga and M. Zumalacárregui. "Dark Energy After GW170817: Dead Ends and the Road Ahead". In: *Phys. Rev. Lett.* 119 (25 Dec. 2017), p. 251304. DOI: 10.1103/PhysRevLett.119.251304.

[53] J. Sakstein and B. Jain. "Implications of the Neutron Star Merger GW170817 for Cosmological Scalar-Tensor Theories". In: *Phys. Rev. Lett.* 119 (25 Dec. 2017), p. 251303. DOI: 10.1103/PhysRevLett.119.251303.

[54] H. Ding et al. "The Orbital-decay Test of General Relativity to the 2% Level with 6 yr VLBA Astrometry of the Double Neutron Star PSR J1537+1155". In: *The Astrophysical Journal Letters* 921.1 (Nov. 2021), p. L19. DOI: 10.3847/2041-8213/ac3091.

[55] E. Vishniac, ed. *The Astrophysical Journal Letters* 875.1 (Apr. 2019). URL: https://iopscience.iop.org/issue/2041-8205/875/1.

[56] E. H. T. Collaboration et al. "First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way". In: *The Astrophysical Journal Letters* 930.2 (May 2022), p. L12. DOI: 10.3847/2041-8213/ac6674.

[57] M. Ishak. "Testing general relativity in cosmology". In: *Living Reviews in Relativity* 22.1 (Dec. 2018), p. 1. ISSN: 1433-8351. DOI: 10.1007/s41114-018-0017-4.

[58] A. H. Guth. "Inflationary universe: A possible solution to the horizon and flatness problems". In: *Phys. Rev. D* 23 (2 Jan. 1981), pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

[59] F. Zwicky. "Die Rotverschiebung von extragalaktischen Nebeln". In: *Helvetica Physica Acta* 6 (Jan. 1933), pp. 110–127.

[60] K. C. Freeman. "On the Disks of Spiral and S0 Galaxies". In: *The Astrophysical Journal* 160 (June 1970), p. 811. DOI: 10.1086/150474.

[61] L. Perivolaropoulos and F. Skara. "Challenges for ΛCDM: An update". In: *New Astronomy Reviews* 95 (2022), p. 101659. ISSN: 1387-6473. DOI: https://doi.org/10.1016/j.newar.2022.101659.

[62] A. Joyce, L. Lombriser, and F. Schmidt. "Dark Energy Versus Modified Gravity". In: *Annual Review of Nuclear and Particle Science* 66.Volume 66, 2016 (2016), pp. 95–122. ISSN: 1545-4134. DOI: https://doi.org/10.1146/annurev-nucl-102115-044553.

[63] J. Bergé. "MICROSCOPE's view at gravitation". In: *Reports on Progress in Physics* 86.6 (May 2023). DOI: 10.1088/1361-6633/acd203.

[64] D. Lovelock. "The Einstein Tensor and Its Generalizations". In: *Journal of Mathematical Physics* 12.3 (Mar. 1971), pp. 498–501. ISSN: 0022-2488. DOI: 10.1063/1.1665613.

[65] D. Lovelock. "The Four-Dimensionality of Space and the Einstein Tensor". In: *Journal of Mathematical Physics* 13.6 (June 1972), pp. 874–876. ISSN: 0022-2488. DOI: 10.1063/1.1666069.

[66] T. Clifton et al. "Modified gravity and cosmology". In: *Physics Reports* 513.1 (2012). Modified Gravity and Cosmology, pp. 1–189. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2012.01.001.

[67] D. Langlois and K. Noui. "Degenerate higher derivative theories beyond Horndeski: evading the Ostrogradski instability". In: *Journal of Cosmology and Astroparticle Physics* 2016.02 (Feb. 2016), p. 034. DOI: 10.1088/1475-7516/2016/02/034.

[68] D. Langlois and K. Noui. "Hamiltonian analysis of higher derivative scalar-tensor theories". In: *Journal of Cosmology and Astroparticle Physics* 2016.07 (July 2016), p. 016. DOI: 10.1088/1475-7516/2016/07/016.

[69] G. W. Horndeski. "Second-order scalar-tensor field equations in a four-dimensional space". In: *International Journal of Theoretical Physics* 10.6 (Sept. 1974), pp. 363–384. ISSN: 1572-9575. DOI: 10.1007/BF01807638.

[70] G. Esposito-Farèse. "Motion in Alternative Theories of Gravity". In: *Mass and Motion in General Relativity*. Ed. by L. Blanchet, A. Spallicci, and B. Whiting. Dordrecht: Springer Netherlands, 2011, pp. 461–489. ISBN: 978-90-481-3015-3. DOI: 10.1007/978-90-481-3015-3_17.

[71] C. Burrage and J. Sakstein. "Tests of chameleon gravity". In: *Living Reviews in Relativity* 21.1 (Mar. 2018), p. 1. ISSN: 1433-8351. DOI: 10.1007/s41114-018-0011-x.

[72] L. Hui, A. Nicolis, and C. W. Stubbs. "Equivalence principle implications of modified gravity models". In: *Phys. Rev. D* 80 (10 Nov. 2009), p. 104002. DOI: 10.1103/PhysRevD.80.104002.

[73] J.-P. Uzan. "Varying Constants, Gravitation and Cosmology". In: *Living Reviews in Relativity* 14.1 (Mar. 2011), p. 2. ISSN: 1433-8351. DOI: 10.12942/lrr-2011-2.

[74] T. Damour and A. Polyakov. "The string dilation and a least coupling principle". In: *Nuclear Physics B* 423.2 (1994), pp. 532–558. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(94)90143-0.

[75] T. Damour and A. M. Polyakov. "String theory and gravity". In: *General Relativity and Gravitation* 26.12 (Dec. 1994), pp. 1171–1176. ISSN: 1572-9532. DOI: 10.1007/BF02106709.

[76] T. Damour and J. F. Donoghue. "Equivalence principle violations and couplings of a light dilaton". In: *Phys. Rev. D* 82 (8 Oct. 2010), p. 084033. DOI: 10.1103/PhysRevD.82.084033.

[77] C. Pitrou and J.-P. Uzan. "Hubble Tension as a Window on the Gravitation of the Dark Matter Sector". In: *Phys. Rev. Lett.* 132 (19 May 2024), p. 191001. DOI: 10.1103/PhysRevLett.132.191001.

[78] L. Amendola and V. Pettorino. "Beyond self-acceleration: Force- and fluid-acceleration". In: *Physics Letters B* 802 (2020), p. 135214. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2020.135214.

[79] R. H. Dicke. "Mach's Principle and Invariance under Transformation of Units". In: *Phys. Rev.* 125 (6 Mar. 1962), pp. 2163–2167. DOI: 10.1103/PhysRev.125.2163.

[80] N. Deruelle and M. Sasaki. "Conformal Equivalence in Classical Gravity: the Example of "Veiled" General Relativity". In: *Cosmology, Quantum Vacuum and Zeta Functions*. Ed. by S. D. Odintsov, D. Sáez-Gómez, and S. Xambó-Descamps. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 247–260. ISBN: 978-3-642-19760-4.

[81] R. Catena, M. Pietroni, and L. Scarabello. "Einstein and Jordan frames reconciled: A frame-invariant approach to scalar-tensor cosmology". In: *Phys. Rev. D* 76 (8 Oct. 2007), p. 084039. DOI: 10.1103/PhysRevD.76.084039.

[82] R. Catena, M. Pietroni, and L. Scarabello. "Local transformations of units in scalar–tensor cosmology". In: *Journal of Physics A: Mathematical and Theoretical* 40.25 (June 2007), p. 6883. DOI: 10.1088/1751-8113/40/25/S34.

[83] G. Esposito-Farèse and D. Polarski. "Scalar-tensor gravity in an accelerating universe". In: *Phys. Rev. D* 63 (6 Feb. 2001), p. 063504. DOI: 10.1103/PhysRevD.63.063504.

[84] K. Falls and M. Herrero-Valea. "Frame (in)equivalence in quantum field theory and cosmology". In: *The European Physical Journal C* 79.7 (July 2019), p. 595. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-7070-3.

[85] A. Y. Kamenshchik and C. F. Steinwachs. "Question of quantum equivalence between Jordan frame and Einstein frame". In: *Phys. Rev. D* 91 (8 Apr. 2015), p. 084033. DOI: 10.1103/PhysRevD.91.084033.

[86] S. Chakraborty, A. Mazumdar, and R. Pradhan. "Distinguishing Jordan and Einstein frames in gravity through entanglement". In: *Phys. Rev. D* 108 (12 Dec. 2023), p. L121505. DOI: 10.1103/PhysRevD.108.L121505.

[87] E. J. Copeland, M. Sami, and S. Tsujikawa. "Dynamics of Dark Energy". In: *International Journal of Modern Physics D* 15.11 (2006), pp. 1753–1935. DOI: 10.1142/S021827180600942X. eprint: https://doi.org/10.1142/S021827180600942X.

[88] B. Ratra and P. J. E. Peebles. "Cosmological consequences of a rolling homogeneous scalar field". In: *Phys. Rev. D* 37 (12 June 1988), pp. 3406–3427. DOI: 10.1103/PhysRevD.37.3406.

[89] S. Perlmutter et al. "Measurements of Ω and Λ from 42 High-Redshift Supernovae". In: *The Astrophysical Journal* 517.2 (June 1999), p. 565. DOI: 10.1086/307221.

[90] A. G. Riess et al. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". In: *The Astronomical Journal* 116.3 (Sept. 1998), p. 1009. DOI: 10.1086/300499.

[91] M. Crisostomi and K. Koyama. "Self-accelerating universe in scalar-tensor theories after GW170817". In: *Phys. Rev. D* 97 (8 Apr. 2018), p. 084004. DOI: 10.1103/PhysRevD.97.084004.

[92] L. Lombriser and N. A. Lima. "Challenges to self-acceleration in modified gravity from gravitational waves and large-scale structure". In: *Physics Letters B* 765 (2017), pp. 382–385. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2016.12.048.

[93] L. Amendola and V. Pettorino. "Beyond self-acceleration: Force- and fluid-acceleration". In: *Physics Letters B* 802 (2020), p. 135214. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2020.135214.

[94] J. Wang, L. Hui, and J. Khoury. "No-Go Theorems for Generalized Chameleon Field Theories". In: *Phys. Rev. Lett.* 109 (24 Dec. 2012), p. 241301. DOI: 10.1103/PhysRevLett.109.241301.

[95] C. Brans and R. H. Dicke. "Mach's Principle and a Relativistic Theory of Gravitation". In: *Phys. Rev.* 124 (3 Nov. 1961), pp. 925–935. DOI: 10.1103/PhysRev.124.925.

[96] E. Mach. *Die Mechanik in ihrer Entwickelung historisch-kritisch dargestellt*. Internationale wissenschaftliche Bibliothek. F.A. Brockhaus, 1883. URL: https://books.google.fr/books?id=taJMxPYNeBAC.

[97] M. Rossi et al. "Cosmological constraints on post-Newtonian parameters in effectively massless scalar-tensor theories of gravity". In: *Phys. Rev. D* 100 (10 Nov. 2019), p. 103524. DOI: 10.1103/PhysRevD.100.103524.

[98] J. H. Gundlach and S. M. Merkowitz. "Measurement of Newton's Constant Using a Torsion Balance with Angular Acceleration Feedback". In: *Phys. Rev. Lett.* 85 (14 Oct. 2000), pp. 2869–2872. DOI: 10.1103/PhysRevLett.85.2869.

[99] Q. Li et al. "Measurements of the gravitational constant using two independent methods". In: *Nature* 560.7720 (Aug. 2018), pp. 582–588. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0431-5.

[100] C. H. Brans. *The roots of scalar-tensor theory: an approximate history*. 2005. arXiv: gr-qc/0506063 [gr-qc].

[101] C. H. Brans. "Scalar-tensor Theories of Gravity: Some personal history". In: *AIP Conference Proceedings* 1083.1 (Dec. 2008), pp. 34–46. ISSN: 0094-243X. DOI: 10.1063/1.3058577. eprint: https://pubs.aip.org/aip/acp/article-pdf/1083/1/34/12118774/34_1_online.pdf.

[102] T. Damour and G. Esposito-Farèse. "Tensor-multi-scalar theories of gravitation". In: *Classical and Quantum Gravity* 9.9 (Sept. 1992), p. 2093. DOI: 10.1088/0264-9381/9/9/015.

[103] T. Damour and K. Nordtvedt. "Tensor-scalar cosmological models and their relaxation toward general relativity". In: *Phys. Rev. D* 48 (8 Oct. 1993), pp. 3436–3450. DOI: 10.1103/PhysRevD.48.3436.

[104] J. Bergé et al. "Interpretation of geodesy experiments in non-Newtonian theories of gravity". In: *Classical and Quantum Gravity* 35.23 (Nov. 2018), p. 234001. DOI: 10.1088/1361-6382/aae9a1.

[105] J. Bergé et al. "MICROSCOPE's constraint on a short-range fifth force". In: *Classical and Quantum Gravity* 39.20 (Sept. 2022), p. 204010. DOI: 10.1088/1361-6382/abe142.

[106] E. Adelberger, B. Heckel, and A. Nelson. "Tests of the gravitational inverse-square law". In: *Annual Review of Nuclear and Particle Science* 53.Volume 53, 2003 (2003), pp. 77–121. ISSN: 1545-4134. DOI: https://doi.org/10.1146/annurev.nucl.53.041002.110503.

[107] D. J. Kapner et al. "Tests of the Gravitational Inverse-Square Law below the Dark-Energy Length Scale". In: *Phys. Rev. Lett.* 98 (2 Jan. 2007), p. 021101. DOI: 10.1103/PhysRevLett.98.021101.

[108] D. Benisty. "Testing modified gravity via Yukawa potential in two body problem: Analytical solution and observational constraints". In: *Phys. Rev. D* 106 (4 Aug. 2022), p. 043001. DOI: 10.1103/PhysRevD.106.043001.

[109] G. D'Amico, M. Kamionkowski, and K. Sigurdson. "Dark Matter Astrophysics". In: *Dark Matter and Dark Energy: A Challenge for Modern Cosmology*. Ed. by S. Matarrese et al. Dordrecht: Springer Netherlands, 2011, pp. 241–272. ISBN: 978-90-481-8685-3. DOI: 10.1007/978-90-481-8685-3_5.

[110] L. Hui et al. "Ultralight scalars as cosmological dark matter". In: *Phys. Rev. D* 95 (4 Feb. 2017), p. 043541. DOI: 10.1103/PhysRevD.95.043541.

[111] A. Joyce et al. "Beyond the cosmological standard model". In: *Physics Reports* 568 (2015). Beyond the cosmological standard model, pp. 1–98. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2014.12.002.

[112] K. Hinterbichler and J. Khoury. "Screening Long-Range Forces through Local Symmetry Restoration". In: *Phys. Rev. Lett.* 104 (23 June 2010), p. 231301. DOI: 10.1103/PhysRevLett.104.231301.

[113] K. Hinterbichler et al. "Symmetron cosmology". In: *Phys. Rev. D* 84 (10 Nov. 2011), p. 103521. DOI: 10.1103/PhysRevD.84.103521.

[114] P. Brax et al. "Nonlinear structure formation with the environmentally dependent dilaton". In: *Phys. Rev. D* 83 (10 May 2011), p. 104026. DOI: 10.1103/PhysRevD.83.104026.

[115] J. Khoury and A. Weltman. "Chameleon cosmology". In: *Phys. Rev. D* 69 (4 Feb. 2004), p. 044026. DOI: 10.1103/PhysRevD.69.044026.

[116] J. Khoury and A. Weltman. "Chameleon Fields: Awaiting Surprises for Tests of Gravity in Space". In: *Phys. Rev. Lett.* 93 (17 Oct. 2004), p. 171104. DOI: 10.1103/PhysRevLett.93.171104.

[117] P. Brax, C. Burrage, and A.-C. Davis. "Screening fifth forces in k-essence and DBI models". In: *Journal of Cosmology and Astroparticle Physics* 2013.01 (Jan. 2013), p. 020. DOI: 10.1088/1475-7516/2013/01/020.

[118] E. Babichev, C. Deffayet, and R. Ziour. "k-mouflage gravity". In: *International Journal of Modern Physics D* 18.14 (2009), pp. 2147–2154. DOI: 10.1142/S0218271809016107. eprint: https://doi.org/10.1142/S0218271809016107.

[119] C. Burrage and J. Khoury. "Screening of scalar fields in Dirac-Born-Infeld theory". In: *Phys. Rev. D* 90 (2 July 2014), p. 024001. DOI: 10.1103/PhysRevD.90.024001.

[120] P. Brax and P. Valageas. "K-mouflage cosmology: The background evolution". In: *Phys. Rev. D* 90 (2 July 2014), p. 023507. DOI: 10.1103/PhysRevD.90.023507.

[121] A. Vainshtein. "To the problem of nonvanishing gravitation mass". In: *Physics Letters B* 39.3 (1972), pp. 393–394. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(72)90147-5.

[122] C. Deffayet et al. "Nonperturbative continuity in graviton mass versus perturbative discontinuity". In: *Phys. Rev. D* 65 (4 Jan. 2002), p. 044026. DOI: 10.1103/PhysRevD.65.044026.

[123] E. Babichev, C. Deffayet, and R. Ziour. "Recovery of general relativity in massive gravity via the Vainshtein mechanism". In: *Phys. Rev. D* 82 (10 Nov. 2010), p. 104008. DOI: 10.1103/PhysRevD.82.104008.

[124] P. Brax et al. "Detecting dark energy in orbit: The cosmological chameleon". In: *Phys. Rev. D* 70 (12 Dec. 2004), p. 123518. DOI: 10.1103/PhysRevD.70.123518.

[125] O. Minazzoli and A. Hees. "Intrinsic Solar System decoupling of a scalar-tensor theory with a universal coupling between the scalar field and the matter Lagrangian". In: *Phys. Rev. D* 88 (4 Aug. 2013), p. 041504. DOI: 10.1103/PhysRevD.88.041504.

[126] O. Minazzoli and A. Hees. "Late-time cosmology of a scalar-tensor theory with a universal multiplicative coupling between the scalar field and the matter Lagrangian". In: *Phys. Rev. D* 90 (2 July 2014), p. 023017. DOI: 10.1103/PhysRevD.90.023017.

[127] T. Damour, F. Piazza, and G. Veneziano. "Runaway Dilaton and Equivalence Principle Violations". In: *Phys. Rev. Lett.* 89 (8 Aug. 2002), p. 081601. DOI: 10.1103/PhysRevLett.89.081601.

[128] T. Damour, F. Piazza, and G. Veneziano. "Violations of the equivalence principle in a dilaton-runaway scenario". In: *Phys. Rev. D* 66 (4 Aug. 2002), p. 046007. DOI: 10.1103/PhysRevD.66.046007.

[129] P. Brax, C. Burgess, and F. Quevedo. "Axio-Chameleons: a novel string-friendly multi-field screening mechanism". In: *Journal of Cosmology and Astroparticle Physics* 2024.03 (Mar. 2024), p. 015. DOI: 10.1088/1475-7516/2024/03/015.

[130] C. Burgess and F. Quevedo. "Axion homeopathy: screening dilaton interactions". In: *Journal of Cosmology and Astroparticle Physics* 2022.04 (Apr. 2022), p. 007. DOI: 10.1088/1475-7516/2022/04/007.

[131] P. Brax and A. Ouazzani. "Two-field screening and its cosmological dynamics". In: *Phys. Rev. D* 108 (6 Sept. 2023), p. 063517. DOI: 10.1103/PhysRevD.108.063517.

[132] W. Hu and I. Sawicki. "Models of $f(R)$ cosmic acceleration that evade solar system tests". In: *Phys. Rev. D* 76 (6 Sept. 2007), p. 064004. DOI: 10.1103/PhysRevD.76.064004.

[133] D. F. Mota and H. A. Winther. "Cosmology of chameleons with power-law couplings". In: *The Astrophysical Journal* 733.1 (Apr. 2011), p. 7. DOI: 10.1088/0004-637X/733/1/7.

[134] A. N. Ivanov et al. "Influence of the chameleon field potential on transition frequencies of gravitationally bound quantum states of ultracold neutrons". In: *Phys. Rev. D* 87 (10 May 2013), p. 105013. DOI: 10.1103/PhysRevD.87.105013.

[135] S. Schlögel, S. Clesse, and A. Füzfa. "Probing modified gravity with atom-interferometry: A numerical approach". In: *Phys. Rev. D* 93 (10 May 2016), p. 104036. DOI: 10.1103/PhysRevD.93.104036.

[136] M. Pernot-Borràs et al. "General study of chameleon fifth force in gravity space experiments". In: *Phys. Rev. D* 100 (8 Oct. 2019), p. 084006. DOI: 10.1103/PhysRevD.100.084006.

[137] H. Lévy, J. Bergé, and J.-P. Uzan. "Solving nonlinear Klein-Gordon equations on unbounded domains via the finite element method". In: *Phys. Rev. D* 106 (12 Dec. 2022), p. 124021. DOI: 10.1103/PhysRevD.106.124021.

[138] M. Pernot-Borràs. "Testing gravity in space : towards a realistic treatment of chameleon gravity in the MICROSCOPE mission". Theses. Sorbonne Université, Nov. 2020. URL: https://theses.hal.science/tel-03333888.

[139] A.-C. Davis et al. "Modified gravity makes galaxies brighter". In: *Phys. Rev. D* 85 (12 June 2012), p. 123006. DOI: 10.1103/PhysRevD.85.123006.

[140] P. Brax et al. "Unified description of screened modified gravity". In: *Phys. Rev. D* 86 (4 Aug. 2012), p. 044015. DOI: 10.1103/PhysRevD.86.044015.

[141] H. Lévy, J. Bergé, and J.-P. Uzan. "What to expect from scalar-tensor space geodesy". In: *Phys. Rev. D* 109 (8 Apr. 2024), p. 084009. DOI: 10.1103/PhysRevD.109.084009.

[142] C. Burrage, E. J. Copeland, and E. Hinds. "Probing dark energy with atom interferometry". In: *Journal of Cosmology and Astroparticle Physics* 2015.03 (Mar. 2015), p. 042. DOI: 10.1088/1475-7516/2015/03/042.

[143] C. Burrage, E. J. Copeland, and J. A. Stevenson. "Ellipticity weakens chameleon screening". In: *Phys. Rev. D* 91 (6 Mar. 2015), p. 065030. DOI: 10.1103/PhysRevD.91.065030.

[144] P. Brax et al. "Detecting chameleons through Casimir force measurements". In: *Phys. Rev. D* 76 (12 Dec. 2007), p. 124034. DOI: 10.1103/PhysRevD.76.124034.

[145] A. Hees and A. Füzfa. "Combined cosmological and solar system constraints on chameleon mechanism". In: *Phys. Rev. D* 85 (10 May 2012), p. 103005. DOI: 10.1103/PhysRevD.85.103005.

[146] T. Katsuragawa and S. Matsuzaki. "Cosmic history of chameleonic dark matter in $F(R)$ gravity". In: *Phys. Rev. D* 97 (6 Mar. 2018), p. 064037. DOI: 10.1103/PhysRevD.97.064037.

[147] H. Chen et al. "Big Bang Nucleosynthesis hunts chameleon dark matter". In: *Journal of High Energy Physics* 2020.2 (Feb. 2020), p. 155. ISSN: 1029-8479. DOI: 10.1007/JHEP02(2020)155.

[148] R.-G. Cai and S.-J. Wang. "Higgs chameleon". In: *Phys. Rev. D* 103 (2 Jan. 2021), p. 023502. DOI: 10.1103/PhysRevD.103.023502.

[149] D. F. Mota and D. J. Shaw. "Evading equivalence principle violations, cosmological, and other experimental constraints in scalar field theories with a strong coupling to matter". In: *Phys. Rev. D* 75 (6 Mar. 2007), p. 063501. DOI: 10.1103/PhysRevD.75.063501.

[150] C. Burrage and J. Sakstein. "A compendium of chameleon constraints". In: *Journal of Cosmology and Astroparticle Physics* 2016.11 (Nov. 2016), p. 045. DOI: 10.1088/1475-7516/2016/11/045.

[151] H. Fischer, C. Käding, and M. Pitschmann. "Screened Scalar Fields in the Laboratory and the Solar System". In: *Universe* 10.7 (2024). ISSN: 2218-1997. DOI: 10.3390/universe10070297.

[152] P. Brax, A.-C. Davis, and B. Elder. "Casimir tests of scalar-tensor theories". In: *Phys. Rev. D* 107 (8 Apr. 2023), p. 084025. DOI: 10.1103/PhysRevD.107.084025.

[153] D. F. Mota and D. J. Shaw. "Strongly Coupled Chameleon Fields: New Horizons in Scalar Field Theory". In: *Phys. Rev. Lett.* 97 (15 Oct. 2006), p. 151102. DOI: 10.1103/PhysRevLett.97.151102.

[154] C. W. F. Everitt et al. "Gravity Probe B: Final Results of a Space Experiment to Test General Relativity". In: *Phys. Rev. Lett.* 106 (22 May 2011), p. 221101. DOI: 10.1103/PhysRevLett.106.221101.

[155] I. Ciufolini et al. "The LARES 2 satellite, general relativity and fundamental physics". In: *The European Physical Journal C* 83.1 (Jan. 2023), p. 87. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-023-11230-6.

[156] *Classical and Quantum Gravity* 39.20 (Oct. 2022): *The MICROSCOPE space mission: the first test of the equivalence principle in a space laboratory*. ISSN: 0264-9381. URL: https://iopscience.iop.org/issue/0264-9381/39/20.

[157] A. Upadhye. "Dark energy fifth forces in torsion pendulum experiments". In: *Phys. Rev. D* 86 (10 Nov. 2012), p. 102003. DOI: 10.1103/PhysRevD.86.102003.

[158] G. M. Tino. "Testing gravity with cold atom interferometry: results and prospects". In: *Quantum Science and Technology* 6.2 (Mar. 2021). DOI: 10.1088/2058-9565/abd83e.

[159] A. J. Sanders and W. E. Deeds. "Proposed new determination of the gravitational constant G and tests of Newtonian gravitation". In: *Phys. Rev. D* 46 (2 July 1992), pp. 489–504. DOI: 10.1103/PhysRevD.46.489.

[160] A. J. Sanders et al. "Project SEE (Satellite Energy Exchange): an international effort to develop a space-based mission for precise measurements of gravitation". In: *Classical and Quantum Gravity* 17.12 (June 2000), p. 2331. DOI: 10.1088/0264-9381/17/12/305.

[161] R. Reinhard, P. Worden, and C. Everitt. "STEP: A Satellite Test of the Equivalence Principle". en. In: *Europhysics News* 22.11 (1991), pp. 216–218. ISSN: 0531-7479, 1432-1092. DOI: 10.1051/epn/19912211216.

[162] J. Mester et al. "The STEP mission: principles and baseline design". In: *Classical and Quantum Gravity* 18.13 (July 2001), p. 2475. DOI: 10.1088/0264-9381/18/13/310.

[163] J. Overduin et al. "STEP and fundamental physics". In: *Classical and Quantum Gravity* 29.18 (Aug. 2012), p. 184012. DOI: 10.1088/0264-9381/29/18/184012.

[164] A. M. Nobili et al. "'Galileo Galilei' (GG) small-satellite project: an alternative to the torsion balance for testing the equivalence principle on Earth and in space". In: *Classical and Quantum Gravity* 17.12 (June 2000), p. 2347. DOI: 10.1088/0264-9381/17/12/306.

[165] G. Galilei. *Discourses and Mathematical Demonstrations Relating to Two New Sciences*. Leiden: Louis Elsevier, 1638.

[166] M. Rodrigues et al. "MICROSCOPE mission scenario, ground segment and data processing". In: *Classical and Quantum Gravity* 39.20 (Sept. 2022), p. 204004. DOI: 10.1088/1361-6382/ac4b9a.

[167] M. Rodrigues et al. "MICROSCOPE: systematic errors". In: *Classical and Quantum Gravity* 39.20 (Sept. 2022), p. 204006. DOI: 10.1088/1361-6382/ac49f6.

[168] P. Fayet. "Extra U(1)'s and new forces". In: *Nuclear Physics B* 347.3 (1990), pp. 743–768. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(90)90381-M.

[169] P. Fayet. "The light U boson as the mediator of a new force, coupled to a combination of Q, B, L and dark matter". In: *The European Physical Journal C* 77.1 (Jan. 2017), p. 53. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4568-9.

[170] J. Bergé et al. "MICROSCOPE Mission: First Constraints on the Violation of the Weak Equivalence Principle by a Light Scalar Dilaton". In: *Phys. Rev. Lett.* 120 (14 Apr. 2018), p. 141101. DOI: 10.1103/PhysRevLett.120.141101.

[171] J. Bergé et al. "MICROSCOPE's constraint on a short-range fifth force". In: *Classical and Quantum Gravity* 39.20 (Sept. 2022), p. 204010. DOI: 10.1088/1361-6382/abe142.

[172] T. Damour and J. F. Donoghue. "Equivalence principle violations and couplings of a light dilaton". In: *Phys. Rev. D* 82 (8 Oct. 2010), p. 084033. DOI: 10.1103/PhysRevD.82.084033.

[173] T. Damour and J. F. Donoghue. "Phenomenology of the equivalence principle with light scalars". In: *Classical and Quantum Gravity* 27.20 (Aug. 2010), p. 202001. DOI: 10.1088/0264-9381/27/20/202001.

[174] T. Bouley, P. Sørensen, and T.-T. Yu. "Constraints on ultralight scalar dark matter with quadratic couplings". In: *Journal of High Energy Physics* 2023.3 (Mar. 2023), p. 104. ISSN: 1029-8479. DOI: 10.1007/JHEP03(2023)104.

[175] C. J. A. P. Martins and L. Vacher. "Astrophysical and local constraints on string theory: Runaway dilaton models". In: *Phys. Rev. D* 100 (12 Dec. 2019), p. 123514. DOI: 10.1103/PhysRevD.100.123514.

[176] H. P.-l. Bars et al. "New Test of Lorentz Invariance Using the MICROSCOPE Space Mission". In: *Phys. Rev. Lett.* 123 (23 Dec. 2019), p. 231102. DOI: 10.1103/PhysRevLett.123.231102.

[177] P. Fayet. "MICROSCOPE limits for new long-range forces and implications for unified theories". In: *Phys. Rev. D* 97 (5 Mar. 2018), p. 055039. DOI: 10.1103/PhysRevD.97.055039.

[178] P. Fayet. "$MICROSCOPE$ limits on the strength of a new force with comparisons to gravity and electromagnetism". In: *Phys. Rev. D* 99 (5 Mar. 2019), p. 055043. DOI: 10.1103/PhysRevD.99.055043.

[179] M. Pernot-Borràs et al. "Constraints on chameleon gravity from the measurement of the electrostatic stiffness of the $MICROSCOPE$ mission accelerometers". In: *Phys. Rev. D* 103 (6 Mar. 2021), p. 064070. DOI: 10.1103/PhysRevD.103.064070.

[180] M. Pernot-Borràs et al. "Fifth force induced by a chameleon field on nested cylinders". In: *Phys. Rev. D* 101 (12 June 2020), p. 124056. DOI: 10.1103/PhysRevD.101.124056.

[181] L. Ogden et al. "Electrostatic analogy for symmetron gravity". In: *Phys. Rev. D* 96 (12 Dec. 2017), p. 124029. DOI: 10.1103/PhysRevD.96.124029.

[182] A. Dima, M. Bezares, and E. Barausse. "Dynamical chameleon neutron stars: Stability, radial oscillations, and scalar radiation in spherical symmetry". In: *Phys. Rev. D* 104 (8 Oct. 2021), p. 084017. DOI: 10.1103/PhysRevD.104.084017.

[183] E. Babichev and D. Langlois. "Relativistic stars in $f(R)$ and scalar-tensor theories". In: *Phys. Rev. D* 81 (12 June 2010), p. 124051. DOI: 10.1103/PhysRevD.81.124051.

[184] P. Hamilton et al. "Atom-interferometry constraints on dark energy". In: *Science* 349.6250 (2015), pp. 849–851. DOI: 10.1126/science.aaa8883.

[185] S. Schlögel, S. Clesse, and A. Füzfa. "Probing modified gravity with atom-interferometry: A numerical approach". In: *Phys. Rev. D* 93 (10 May 2016), p. 104036. DOI: 10.1103/PhysRevD.93.104036.

[186] B. Elder et al. "Chameleon dark energy and atom interferometry". In: *Phys. Rev. D* 94 (4 Aug. 2016), p. 044051. DOI: 10.1103/PhysRevD.94.044051.

[187] A. Upadhye, S. S. Gubser, and J. Khoury. "Unveiling chameleon fields in tests of the gravitational inverse-square law". In: *Phys. Rev. D* 74 (10 Nov. 2006), p. 104024. DOI: 10.1103/PhysRevD.74.104024.

[188] V. Vardanyan and D. J. Bartlett. "Modeling and Testing Screening Mechanisms in the Laboratory and in Space". In: *Universe* 9.7 (2023). ISSN: 2218-1997. DOI: 10.3390/universe9070340.

[189] C. Burrage et al. "The shape dependence of chameleon screening". In: *Journal of Cosmology and Astroparticle Physics* 2018.01 (Jan. 2018), p. 056. DOI: 10.1088/1475-7516/2018/01/056.

[190] C. Briddon et al. "SELCIE: a tool for investigating the chameleon field of arbitrary sources". In: *Journal of Cosmology and Astroparticle Physics* 2021.12 (Dec. 2021), p. 043. DOI: 10.1088/1475-7516/2021/12/043.

[191] K. Clements et al. "Detecting dark domain walls through their impact on particle trajectories in tailored ultrahigh vacuum environments". In: *Phys. Rev. D* 109 (12 June 2024), p. 123023. DOI: 10.1103/PhysRevD.109.123023.

[192] B. Elder et al. "Classical symmetron force in Casimir experiments". In: *Phys. Rev. D* 101 (6 Mar. 2020), p. 064065. DOI: 10.1103/PhysRevD.101.064065.

[193] J. Braden et al. "φenics: Vainshtein screening with the finite element method". In: *Journal of Cosmology and Astroparticle Physics* 2021.03 (Mar. 2021), p. 010. DOI: 10.1088/1475-7516/2021/03/010.

[194] C. Llinares. "Simulation techniques for modified gravity". In: *International Journal of Modern Physics D* 27.15 (2018), p. 1848003. DOI: 10.1142/S0218271818480036. eprint: https://doi.org/10.1142/S0218271818480036.

[195] B. Li et al. "ECOSMOG: an Efficient COde for Simulating MOdified Gravity". In: *Journal of Cosmology and Astroparticle Physics* 2012.01 (Jan. 2012), p. 051. DOI: 10.1088/1475-7516/2012/01/051.

[196] Llinares, Claudio, Mota, David F., and Winther, Hans A. "ISIS: a new N-body cosmological code with scalar fields based on RAMSES - Code presentation and application to the shapes of clusters". In: *Astronomy & Astrophysics* 562 (2014), A78. DOI: 10.1051/0004-6361/201322412.

[197] E. Puchwein, M. Baldi, and V. Springel. "Modified-Gravity-gadget: a new code for cosmological hydrodynamical simulations of modified gravity models". In: *Monthly Notices of the Royal Astronomical Society* 436.1 (Sept. 2013), pp. 348–360. ISSN: 0035-8711. DOI: 10.1093/mnras/stt1575. eprint: https://academic.oup.com/mnras/article-pdf/436/1/348/18497113/stt1575.pdf.

[198] A. Silvestri. "Scalar Radiation from Chameleon-Shielded Regions". In: *Phys. Rev. Lett.* 106 (25 June 2011), p. 251101. DOI: 10.1103/PhysRevLett.106.251101.

[199] M. Bezares et al. "No Evidence of Kinetic Screening in Simulations of Merging Binary Neutron Stars beyond General Relativity". In: *Phys. Rev. Lett.* 128 (9 Mar. 2022), p. 091103. DOI: 10.1103/PhysRevLett.128.091103.

[200] A. Quarteroni. "Mathematical Aspects of Domain Decomposition Methods". In: *First European Congress of Mathematics Paris, July 6–10, 1992: Vol. II: Invited Lectures (Part 2)*. Ed. by A. Joseph et al. Basel: Birkhäuser Basel, 1994, pp. 355–379. ISBN: 978-3-0348-9112-7. DOI: 10.1007/978-3-0348-9112-7_15.

[201] V. Dolean, P. Jolivet, and F. Nataf. *An Introduction to Domain Decomposition Methods*. Philadelphia, PA 19104-2688 USA: Society for Industrial and Applied Mathematics, Feb. 2016. DOI: 10.1137/1.9781611974065. eprint: https://hal.science/cel-01100932v5.

[202] M. J. Turner et al. "Stiffness and Deflection Analysis of Complex Structures". In: *Journal of the Aeronautical Sciences* 23.9 (Sept. 1956), pp. 805–823. DOI: 10.2514/8.3664.

[203] W. K. Liu, S. Li, and H. S. Park. "Eighty Years of the Finite Element Method: Birth, Evolution, and Future". In: *Archives of Computational Methods in Engineering* 29.6 (Oct. 1, 2022), pp. 4431–4453. ISSN: 1886-1784. DOI: 10.1007/s11831-022-09740-9.

[204] J. Hadamard. "Sur les problèmes aux dérivés partielles et leur signification physique". In: *Princeton University Bulletin* 13 (1902), pp. 49–52. URL: https://babel.hathitrust.org/cgi/pt?id=chi.095582186&view=1up&seq=65.

[205] L. C. Evans. *Partial Differential Equations*. 2nd. Graduate Studies in Mathematics: V. 19. New Delhi: American Mathematical Society, 2010.

[206] R. G. McClarren. "Chapter 16 - Gauss Quadrature and Multi-dimensional Integrals". In: *Computational Nuclear Engineering and Radiological Science Using Python*. Ed. by R. G. McClarren. Academic Press, 2018, pp. 287–299. ISBN: 978-0-12-812253-2. DOI: https://doi.org/10.1016/B978-0-12-812253-2.00018-2.

[207] J. Cea. "Approximation variationnelle des problèmes aux limites". fr. In: *Annales de l'Institut Fourier* 14.2 (1964), pp. 345–444. DOI: 10.5802/aif.181.

[208] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. New York, NY: Springer New York, 1994. ISBN: 978-1-4757-4338-8. DOI: https://doi.org/10.1007/978-1-4757-4338-8.

[209] J. Argyris and D. Scharpf. "Finite elements in time and space". In: *Nuclear Engineering and Design* 10.4 (1969), pp. 456–464. ISSN: 0029-5493. DOI: https://doi.org/10.1016/0029-5493(69)90081-8.

[210] I. FRIED. "Finite-element analysis of time-dependent phenomena." In: *AIAA Journal* 7.6 (1969), pp. 1170–1173. DOI: 10.2514/3.5299. eprint: https://doi.org/10.2514/3.5299.

[211] J. T. Oden. "A general theory of finite elements. II. Applications". In: *International Journal for Numerical Methods in Engineering* 1.3 (1969), pp. 247–259. DOI: https://doi.org/10.1002/nme.1620010304. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nme.1620010304.

[212] G. M. Hulbert and T. J. Hughes. "Space-time finite element methods for second-order hyperbolic equations". In: *Computer Methods in Applied Mechanics and Engineering* 84.3 (1990), pp. 327–348. ISSN: 0045-7825. DOI: https://doi.org/10.1016/0045-7825(90)90082-W.

[213] S. Banach. "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales". In: *Fundamenta Mathematicae* 3 (1922), pp. 133–181. DOI: 10.4064/fm-3-1-133-181.

[214] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1995. DOI: 10.1137/1.9781611970944. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611970944.

[215] H. P. Langtangen. *Computational Partial Differential Equations*. Springer Berlin Heidelberg, 2003. ISBN: 978-3-642-55769-9. DOI: 10.1007/978-3-642-55769-9.

[216] E. L. Allgower and K. Georg. *Numerical Continuation Methods: An Introduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1990. ISBN: 978-3-642-61257-2. DOI: 10.1007/978-3-642-61257-2.

[217] W. Frei. *Nonlinearity Ramping for Improving Convergence of Nonlinear Problems*. COMSOL Blog. Accessed: May 26th, 2023. Dec. 2013. URL: https://www.comsol.com/blogs/nonlinearity-ramping-improving-convergence-nonlinear-problems.

[218] W. Frei. *Load Ramping of Nonlinear Problems*. COMSOL Blog. Accessed: May 26th, 2023. Nov. 2013. URL: https://www.comsol.com/blogs/load-ramping-nonlinear-problems.

[219] A. Bayliss, M. Gunzburger, and E. Turkel. "Boundary Conditions for the Numerical Solution of Elliptic Equations in Exterior Regions". In: *SIAM Journal on Applied Mathematics* 42.2 (1982), pp. 430–451. DOI: 10.1137/0142032.

[220] C. I. Goldstein. "Finite element method with nonuniform mesh sizes for unbounded domains". In: *Math. Comput.; (United States)* 36 (Apr. 1981). DOI: 10.1090/S0025-5718-1981-0606503-5.

[221] H. Han and W. Bao. "The discrete artificial boundary condition on a polygonal artificial boundary for the exterior problem of Poisson equation by using the direct method of lines". In: *Computer Methods in Applied Mechanics and Engineering* 179.3 (1999), pp. 345–360. ISSN: 0045-7825. DOI: https://doi.org/10.1016/S0045-7825(99)00046-8.

[222] T. Hagstrom and H. Keller. "Asymptotic boundary conditions and numerical methods for nonlinear elliptic problems on unbounded domains". In: *Mathematics of Computation* 48 (1987), pp. 449–470. DOI: https://doi.org/10.1090/S0025-5718-1987-0878684-5.

[223] T. Hagstrom and H. B. Keller. "Exact Boundary Conditions at an Artificial Boundary for Partial Differential Equations in Cylinders". In: *SIAM Journal on Mathematical Analysis* 17.2 (1986), pp. 322–341. DOI: 10.1137/0517026. eprint: https://doi.org/10.1137/0517026.

[224] J. J. Shirron and I. Babuška. "A comparison of approximate boundary conditions and infinite element methods for exterior Helmholtz problems". In: *Computer Methods in Applied Mechanics and Engineering* 164.1 (1998). Exterior Problems of Wave Propagation, pp. 121–139. ISSN: 0045-7825. DOI: https://doi.org/10.1016/S0045-7825(98)00050-4.

[225] B. Engquist and A. Majda. "Absorbing boundary conditions for numerical simulation of waves". In: *Proceedings of the National Academy of Sciences* 74.5 (1977), pp. 1765–1766. DOI: 10.1073/pnas.74.5.1765.

[226] J.-P. Berenger. "A perfectly matched layer for the absorption of electromagnetic waves". In: *Journal of Computational Physics* 114.2 (1994), pp. 185–200. ISSN: 0021-9991. DOI: https://doi.org/10.1006/jcph.1994.1159.

[227] N. Frédéric. "Absorbing boundary conditions and perfectly matched layers in wave propagation problems". In: *Direct and Inverse Problems in Wave Propagation and Applications*. Ed. by I. Graham et al. Berlin, Boston: De Gruyter, 2013, pp. 219–232. ISBN: 9783110282283. DOI: doi:10.1515/9783110282283.219.

[228] M. Costabel. "Principles of boundary element methods". In: *Computer Physics Reports* 6.1 (1987), pp. 243–274. ISSN: 0167-7977. DOI: https://doi.org/10.1016/0167-7977(87)90014-1.

[229] B. Gernot, S. Ian M., and D. Christian. *The Boundary Element Method with Programming: For engineers and scientists*. Vienna: Springer Vienna, 2008. ISBN: 978-3-211-71576-5. DOI: https://doi.org/10.1007/978-3-211-71576-5.

[230] P. Bettess. “Infinite elements”. In: *International Journal for Numerical Methods in Engineering* 11.1 (1977), pp. 53–64. DOI: https://doi.org/10.1002/nme.1620110107.

[231] G. Beer and J. L. Meek. “‘Infinite domain’ elements”. In: *International Journal for Numerical Methods in Engineering* 17.1 (1981), pp. 43–52. DOI: https://doi.org/10.1002/nme.1620170104.

[232] K. Gerdes and L. Demkowicz. “Solution of 3D-Laplace and Helmholtz equations in exterior domains using hp-infinite elements”. In: *Computer Methods in Applied Mechanics and Engineering* 137.3 (1996), pp. 239–273. ISSN: 0045-7825. DOI: https://doi.org/10.1016/0045-7825(95)00987-6.

[233] K. Ishioka. “A Spectral Method for Unbounded Domains and Its Application to Wave Equations in Geophysical Fluid Dynamics”. In: *IUTAM Symposium on Computational Physics and New Perspectives in Turbulence*. Ed. by Y. Kaneda. Dordrecht: Springer Netherlands, 2008, pp. 291–296. ISBN: 978-1-4020-6472-2.

[234] J. Shen and L.-L. Wang. “Some Recent Advances on Spectral Methods for Unbounded Domains”. In: *Communications in Computational Physics* 5.2-4 (2009), pp. 195–241. ISSN: 1991-7120. DOI: https://doi.org/.

[235] T. Chou, S. Shao, and M. Xia. “Adaptive Hermite spectral methods in unbounded domains”. In: *Applied Numerical Mathematics* 183 (2023), pp. 201–220. ISSN: 0168-9274. DOI: https://doi.org/10.1016/j.apnum.2022.09.003.

[236] H. N. Gharti and J. Tromp. *A spectral-infinite-element solution of Poisson’s equation: an application to self gravity*. 2017. arXiv: 1706.00855 [physics.geo-ph].

[237] C. E. Grosch and S. A. Orszag. “Numerical solution of problems in unbounded regions: Coordinate transforms”. In: *Journal of Computational Physics* 25.3 (1977), pp. 273–295. ISSN: 0021-9991. DOI: https://doi.org/10.1016/0021-9991(77)90102-4.

[238] A. Zenginoğlu. “Hyperboloidal layers for hyperbolic equations on unbounded domains”. In: *Journal of Computational Physics* 230.6 (2011), pp. 2286–2302. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2010.12.016.

[239] T. Chernogorova, I. Dimov, and L. Vulkov. “Coordinate Transformation Approach for Numerical Solution of Environmental Problems”. In: *Mathematical Problems in Meteorological Modelling*. Ed. by A. Bátkai et al. Cham: Springer International Publishing, 2016, pp. 117–127. ISBN: 978-3-319-40157-7.

[240] H.-S. Oh, B. Jang, and Y. Jou. “The weighted Ritz-Galerkin method for elliptic boundary value problems on unbounded domains”. In: *Numerical Methods for Partial Differential Equations* 19.3 (2003), pp. 301–326. DOI: https://doi.org/10.1002/num.10049. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/num.10049.

[241] H.-S. Oh, J.-H. Yun, and B. S. Jang. “Finite element solutions for three-dimensional elliptic boundary value problems on unbounded domains”. In: *Numerical Methods for Partial Differential Equations* 22.6 (2006), pp. 1418–1437. DOI: https://doi.org/10.1002/num.20151. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/num.20151.

[242] Boulmezaoud, Tahar Zamène. “Inverted finite elements: a new method for solving elliptic problems in unbounded domains”. In: *ESAIM: M2AN* 39.1 (Mar. 2005), pp. 109–145. DOI: 10.1051/m2an:2005001.

[243] T. Z. Boulmezaoud, S. Mziou, and T. Boudjedaa. “Numerical Approximation of Second-Order Elliptic Problems in Unbounded Domains”. In: *Journal of Scientific Computing* 60.2 (Aug. 2014), pp. 295–312. ISSN: 1573-7691. DOI: 10.1007/s10915-013-9798-5.

[244] T. Z. Boulmezaoud et al. “Inverted finite elements for degenerate and radial elliptic problems in unbounded domains”. In: *Japan Journal of Industrial and Applied Mathematics* 32.1 (Mar. 2015), pp. 237–261. ISSN: 1868-937X. DOI: 10.1007/s13160-015-0169-5.

[245] S. K. Bhowmik et al. “Solving two dimensional second order elliptic equations in exterior domains using the inverted finite elements method”. In: *Computers & Mathematics with Applications* 72.9 (2016), pp. 2315–2333. ISSN: 0898-1221. DOI: https://doi.org/10.1016/j.camwa.2016.08.030.

[246] T. Z. Boulmezaoud, K. Kaliche, and N. Kerdid. “Inverted finite elements for div-curl systems in the whole space”. In: *Advances in Computational Mathematics* 43.6 (Dec. 2017), pp. 1469–1489. ISSN: 1572-9044. DOI: 10.1007/s10444-017-9532-1.

[247] T. Z. Boulmezaoud and K. Kaliche. “Stray field computation by inverted finite elements: a new method in micromagnetic simulations”. In: *Advances in Computational Mathematics* 50.3 (May 2024), p. 44. ISSN: 1572-9044. DOI: 10.1007/s10444-024-10139-2.

[248] M. S. Nabizadeh, R. Ramamoorthi, and A. Chern. “Kelvin transformations for simulations on infinite domains”. In: *ACM Trans. Graph.* 40.4 (July 2021). ISSN: 0730-0301. DOI: 10.1145/3450626.3459809.

[249] V. Gol’dshtein and A. Ukhlov. “Weighted Sobolev spaces and embedding theorems”. en. In: *Transactions of the American Mathematical Society* 361.07 (July 2009), pp. 3829–3829. ISSN: 0002-9947. DOI: 10.1090/S0002-9947-09-04615-7.

[250] A. C. Cavalheiro. “Weighted Sobolev Spaces and Degenerate Elliptic Equations”. en. In: *Boletim da Sociedade Paranaense de Matemática* 26.1-2 (June 2008), pp. 117–132. ISSN: 2175-1188, 0037-8712. DOI: 10.5269/bspm.v26i1-2.7415.

[251] A. Kufner and B. Opic. “How to define reasonably weighted Sobolev spaces”. eng. In: *Commentationes Mathematicae Universitatis Carolinae* 025.3 (1984), pp. 537–554. URL: http://eudml.org/doc/17341.

[252] J. Giroire. “Etude de quelques problèmes aux limites extérieurs et résolution par équations intégrales”. Thèse de doctorat dirigée par Raviart, Pierre-Arnaud Sciences. Mathématiques Paris 6 1987. PhD thesis. Paris 6, 1987, 1 vol. (VII–515 p.) URL: http://www.theses.fr/1987PA066399.

[253] B. Hanouzet. “Espaces de Sobolev avec poids. Application au problème de Dirichlet dans un demi espace”. fr. In: *Rendiconti del Seminario Matematico della Università di Padova* 46 (1971), pp. 227–272. URL: http://www.numdam.org/item/RSMUP_1971__46__227_0/.

[254] T. Z. Boulmezaoud. “On the Laplace operator and on the vector potential problems in the half-space: an approach using weighted spaces”. In: *Mathematical Methods in the Applied Sciences* 26.8 (2003), pp. 633–669. DOI: https://doi.org/10.1002/mma.369.

[255] C. Amrouche, V. Girault, and J. Giroire. “Weighted Sobolev spaces for Laplace’s equation in Rn”. In: *Journal de Mathématiques Pures et Appliquées* 73 (1994), pp. 579–606.

[256] C. Amrouche, V. Girault, and J. Giroire. “Dirichlet and neumann exterior problems for the n-dimensional laplace operator an approach in weighted sobolev spaces”. In: *Journal de Mathématiques Pures et Appliquées* 76.1 (1997), pp. 55–81. ISSN: 0021-7824. DOI: https://doi.org/10.1016/S0021-7824(97)89945-X.

[257] F. Alliot. “Etude des équations stationnaires de Stokes et Navier-Stokes dans des domaines extérieurs”. PhD thesis. Ecole des Ponts ParisTech, July 1998. URL: https://pastel.hal.science/tel-00005589.

[258] G. Hardy, J. Littlewood, and G. Pólya. *Inequalities*. Cambridge Mathematical Library. Cambridge University Press, 1952. ISBN: 9780521358804. URL: https://books.google.fr/books?id=t1RCSP8YKt8C.

[259] R. Penrose. “Asymptotic Properties of Fields and Space-Times”. In: *Phys. Rev. Lett.* 10 (2 Jan. 1963), pp. 66–68. DOI: 10.1103/PhysRevLett.10.66.

[260] A. Zenginoğlu. “Hyperboloidal layers for hyperbolic equations on unbounded domains”. In: *Journal of Computational Physics* 230.6 (2011), pp. 2286–2302. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2010.12.016.

[261] M. S. Nabizadeh, R. Ramamoorthi, and A. Chern. “Kelvin transformations for simulations on infinite domains”. In: *ACM Trans. Graph.* 40.4 (July 2021). ISSN: 0730-0301. DOI: 10.1145/3450626.3459809.

[262] Y. Sun and C. Westphal. “An adaptively weighted Galerkin finite element method for boundary value problems”. In: *Communications in Applied Mathematics and Computational Science* 10.1 (Mar. 2015). Publisher: Mathematical Sciences Publishers, pp. 27–41. ISSN: 2157-5452. DOI: 10.2140/camcos.2015.10.27.

[263] I. Ergatoudis, B. Irons, and O. Zienkiewicz. “Curved, isoparametric, “quadrilateral” elements for finite element analysis”. In: *International Journal of Solids and Structures* 4.1 (1968), pp. 31–42. ISSN: 0020-7683. DOI: https://doi.org/10.1016/0020-7683(68)90031-0.

[264] T. Hughes, J. Cottrell, and Y. Bazilevs. “Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement”. In: *Computer Methods in Applied Mechanics and Engineering* 194.39 (2005), pp. 4135–4195. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2004.10.008.

[265] L. D. Marini and A. Quarteroni. “A relaxation procedure for domain decomposition methods using finite elements”. In: *Numerische Mathematik* 55.5 (Sept. 1989), pp. 575–598. ISSN: 0945-3245. DOI: 10.1007/BF01398917.

[266] D. Funaro, A. Quarteroni, and P. Zanolli. “An Iterative Procedure with Interface Relaxation for Domain Decomposition Methods”. In: *SIAM Journal on Numerical Analysis* 25.6 (1988), pp. 1213–1236. DOI: 10.1137/0725069. eprint: https://doi.org/10.1137/0725069.

[267] P. E. Bjørstad and O. B. Widlund. “Iterative Methods for the Solution of Elliptic Problems on Regions Partitioned into Substructures”. In: *SIAM Journal on Numerical Analysis* 23.6 (1986), pp. 1097–1120. DOI: 10.1137/0723075. eprint: https://doi.org/10.1137/0723075.

[268] Q. Deng. “Timely Communicaton: An Analysis for a Nonoverlapping Domain Decomposition Iterative Procedure”. In: *SIAM Journal on Scientific Computing* 18.5 (1997), pp. 1517–1525. DOI: 10.1137/S1064827595286797. eprint: https://doi.org/10.1137/S1064827595286797.

[269] J. L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications*. Berlin, Heidelberg: Springer, 1972. ISBN: 978-3-642-65161-8. DOI: https://doi.org/10.1007/978-3-642-65161-8.

[270] J. L. Lions. “Contribution à un problème de M. M. Picone”. In: *Annali di Matematica Pura ed Applicata* 41.1 (Dec. 1956), pp. 201–219. ISSN: 1618-1891. DOI: 10.1007/BF02411668.

[271] J. Nečas. *Direct Methods in the Theory of Elliptic Equations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-10455-8. DOI: https://doi.org/10.1007/978-3-642-10455-8.

[272] *Poincaré-Steklov’s Operators and Domain Decomposition Methods in Finite Dimensional Spaces*. Proceedings of the first International Conference on Domain Decomposition Methods in Paris, France. Jan. 1987. URL: http://www.ddm.org/DD01/Poincare-Steklovs_Operators_and_Domain_Decomposition_Methods_in_Finite_Dimensional_Spaces.pdf.

[273] R. Cimrman, V. Lukeš, and E. Rohan. “Multiscale finite element calculations in Python using SfePy”. In: *Advances in Computational Mathematics* 45.4 (Aug. 2019), pp. 1897–1921. ISSN: 1572-9044. DOI: 10.1007/s10444-019-09666-0.

[274] C. Geuzaine and J.-F. Remacle. "Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities". In: *International Journal for Numerical Methods in Engineering* 79.11 (2009), pp. 1309–1331. DOI: https://doi.org/10.1002/nme.2579.

[275] C. Canuto et al. *Spectral Methods: Fundamentals in Single Domains*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 3–37. ISBN: 978-3-540-30726-6. DOI: 10.1007/978-3-540-30726-6.

[276] C. Briddon et al. "Using machine learning to optimise chameleon fifth force experiments". In: *Journal of Cosmology and Astroparticle Physics* 2024.02 (Feb. 2024), p. 011. DOI: 10.1088/1475-7516/2024/02/011.

[277] A. Vaschy. "Sur les lois de similitude en physique". In: *Annales Télégraphiques* 19 (Jan.–Feb. 1892), pp. 25–28.

[278] E. Buckingham. "On Physically Similar Systems; Illustrations of the Use of Dimensional Equations". In: *Phys. Rev.* 4 (4 Oct. 1914), pp. 345–376. DOI: 10.1103/PhysRev.4.345.

[279] A. M. Dziewonski and D. L. Anderson. "Preliminary reference Earth model". In: *Physics of the Earth and Planetary Interiors* 25.4 (1981), pp. 297–356. ISSN: 0031-9201. DOI: https://doi.org/10.1016/0031-9201(81)90046-7.

[280] Work of the US Government. *U.S. Standard Atmosphere, 1976*. Tech. rep. NASA, 1976. URL: https://ntrs.nasa.gov/citations/19770009539.

[281] P. Virtanen et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python". In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[282] C. Maclaurin. *A treatise of fluxions. In two books*. Vol. 1. Edinburgh: T.W. and T. Ruddimans, 1742. ISBN: 9780598423467.

[283] N. Deruelle and J.-P. Uzan. "Chapt. 15 Self-gravitating fluids". In: *Relativity in Modern Physics*. Oxford University Press, Aug. 2018. ISBN: 9780198786399. DOI: 10.1093/oso/9780198786399.003.0015.

[284] S. Chandrasekhar. *Ellipsoidal figures of equilibrium*. Vol. 10. New Haven: Yale University Press, 1969.

[285] M. Hvoždara and I. Kohút. "Gravity field due to a homogeneous oblate spheroid: Simple solution form and numerical calculations". In: *Contributions to Geophysics and Geodesy* 41.4 (Dec. 2011), pp. 307–327. DOI: 10.2478/v10126-011-0013-0.

[286] A. Tamosiunas et al. "Chameleon screening depends on the shape and structure of NFW halos". In: *Journal of Cosmology and Astroparticle Physics* 2022.04 (Apr. 2022), p. 047. DOI: 10.1088/1475-7516/2022/04/047.

[287] R. P. Kornfeld et al. "GRACE-FO: The Gravity Recovery and Climate Experiment Follow-On Mission". In: *Journal of Spacecraft and Rockets* 56.3 (2019), pp. 931–951. DOI: 10.2514/1.A34326.

[288] O. Montenbruck and E. Gill. *Satellite Orbits*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000. ISBN: 978-3-540-67280-7. DOI: 10.1007/978-3-642-58351-3.

[289] M. Ahlers et al. "Alpenglow: A signature for chameleons in axionlike particle search experiments". In: *Phys. Rev. D* 77 (1 Jan. 2008), p. 015018. DOI: 10.1103/PhysRevD.77.015018.

[290] L. Kraiselburd et al. "Thick shell regime in the chameleon two-body problem". In: *Phys. Rev. D* 99 (8 Apr. 2019), p. 083516. DOI: 10.1103/PhysRevD.99.083516.

[291] J. Betz et al. "Searching for Chameleon Dark Energy with Mechanical Systems". In: *Phys. Rev. Lett.* 129 (13 Sept. 2022), p. 131302. DOI: 10.1103/PhysRevLett.129.131302.

[292] J.-P. Uzan, M. Pernot-Borràs, and J. Bergé. "Effects of a scalar fifth force on the dynamics of a charged particle as a new experimental design to test chameleon theories". In: *Phys. Rev. D* 102 (4 Aug. 2020), p. 044059. DOI: 10.1103/PhysRevD.102.044059.

[293] K. Jones-Smith and F. Ferrer. "Detecting Chameleon Dark Energy via an Electrostatic Analogy". In: *Phys. Rev. Lett.* 108 (22 May 2012), p. 221101. DOI: 10.1103/PhysRevLett.108.221101.

[294] C. Burrage, E. J. Copeland, and J. A. Stevenson. "A proposed experimental search for chameleons using asymmetric parallel plates". In: *Journal of Cosmology and Astroparticle Physics* 2016.08 (Aug. 2016), p. 070. DOI: 10.1088/1475-7516/2016/08/070.

[295] A. Upadhye and J. H. Steffen. *Monopole radiation in modified gravity*. June 2013. arXiv: 1306.6113 [astro-ph.CO]. URL: https://arxiv.org/abs/1306.6113.

[296] X. Zhang, T. Liu, and W. Zhao. "Gravitational radiation from compact binary systems in screened modified gravity". In: *Phys. Rev. D* 95 (10 May 2017), p. 104027. DOI: 10.1103/PhysRevD.95.104027.

[297] X. Zhang. "Tests of gravitational scalar polarization and constraints of chameleon $f(R)$ gravity from comprehensive analysis of binary pulsars". In: *Phys. Rev. D* 106 (2 July 2022), p. 024010. DOI: 10.1103/PhysRevD.106.024010.

[298] A. S. Chou et al. "Search for Chameleon Particles Using a Photon-Regeneration Technique". In: *Phys. Rev. Lett.* 102 (3 Jan. 2009), p. 030402. DOI: 10.1103/PhysRevLett.102.030402.

[299] P. Brax and K. Zioutas. "Solar chameleons". In: *Phys. Rev. D* 82 (4 Aug. 2010), p. 043007. DOI: 10.1103/PhysRevD.82.043007.

[300] R. J. Hughes. "Constraints on new macroscopic forces from gravitational redshift experiments". In: *Phys. Rev. D* 41 (8 Apr. 1990), pp. 2367–2373. DOI: 10.1103/PhysRevD.41.2367.

[301] A. Harvey, E. L. Schucking, and E. J. Surowitz. "Redshifts and Killing vectors". In: *American Journal of Physics* 74.11 (Nov. 2006), pp. 1017–1024. ISSN: 0002-9505. DOI: 10.1119/1.2338544.

[302] Gronke, Max B., Llinares, Claudio, and Mota, David F. "Gravitational redshift profiles in the f(R) and symmetron models". In: *Astronomy & Astrophysics* 562 (Jan. 2014), A9. DOI: 10.1051/0004-6361/201322403.

[303] E. Savalle et al. "Gravitational redshift test with the future ACES mission". In: *Classical and Quantum Gravity* 36.24 (Nov. 2019), p. 245004. DOI: 10.1088/1361-6382/ab4f25.

[304] M. Takamoto et al. "Test of general relativity by a pair of transportable optical lattice clocks". In: *Nature Photonics* 14.7 (July 2020), pp. 411–415. ISSN: 1749-4893. DOI: 10.1038/s41566-020-0619-8.

[305] S. Manly and E. Page. "Experimental feasibility of measuring the gravitational redshift of light using dispersion in optical fibers". In: *Phys. Rev. D* 63 (6 Feb. 2001), p. 062003. DOI: 10.1103/PhysRevD.63.062003.

[306] N. V. Nunes et al. "Gravitational redshift test of EEP with RadioAstron from near Earth to the distance of the Moon". In: *Classical and Quantum Gravity* 40.17 (July 2023), p. 175005. DOI: 10.1088/1361-6382/ace609.

[307] J. Liu et al. "Test of the gravitational redshift with single-photon-based atomic clock interferometers". In: *Quantum Frontiers* 3.1 (Feb. 2024), p. 2. ISSN: 2731-6106. DOI: 10.1007/s44214-024-00049-1.

[308] N. Huntemann et al. "Single-Ion Atomic Clock with 3×10^{-18} Systematic Uncertainty". In: *Phys. Rev. Lett.* 116 (6 Feb. 2016), p. 063001. DOI: 10.1103/PhysRevLett.116.063001.

[309] W. F. McGrew et al. "Atomic clock performance enabling geodesy below the centimetre level". In: *Nature* 564.7734 (Dec. 2018), pp. 87–90. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0738-2.

[310] T. Bothwell et al. "JILA SrI optical lattice clock with uncertainty of 2.0×10^{-18}". In: *Metrologia* 56.6 (Oct. 2019), p. 065004. DOI: 10.1088/1681-7575/ab4089.

[311] S. M. Brewer et al. "$^{27}Al^{+}$ Quantum-Logic Clock with a Systematic Uncertainty below 10^{-18}". In: *Phys. Rev. Lett.* 123 (3 July 2019), p. 033201. DOI: 10.1103/PhysRevLett.123.033201.

[312] N. Ohmae et al. "Transportable Strontium Optical Lattice Clocks Operated Outside Laboratory at the Level of 10^{-18} Uncertainty". In: *Advanced Quantum Technologies* 4.8 (2021), p. 2100015. DOI: https://doi.org/10.1002/qute.202100015.

[313] T. Bothwell et al. "Resolving the gravitational redshift across a millimetre-scale atomic sample". In: *Nature* 602.7897 (Feb. 2022), pp. 420–424. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04349-7.

[314] X. Zheng et al. "A lab-based test of the gravitational redshift with a miniature clock network". In: *Nature Communications* 14.1 (Aug. 2023), p. 4886. ISSN: 2041-1723. DOI: 10.1038/s41467-023-40629-8.

[315] P. Delva, H. Denker, and G. Lion. "Chronometric Geodesy: Methods and Applications". In: *Relativistic Geodesy: Foundations and Applications*. Ed. by D. Puetzfeld and C. Lämmerzahl. Cham: Springer International Publishing, 2019, pp. 25–85. ISBN: 978-3-030-11500-5. DOI: 10.1007/978-3-030-11500-5_2.

[316] L. F. Bularga. *Micro-scale structures in the interplanetary medium*. Technical report NASA-TM-X-55995. NASA Goddard Space Flight Center, Sept. 1967. URL: https://ntrs.nasa.gov/api/citations/19680000537/downloads/19680000537.pdf.

[317] A. Chambers. *Modern Vacuum Physics*. Ed. by D. S. Betts. Masters series in physics and astronomy. Chapman & Hall/CRC, 2004. ISBN: 0-8493-2438-6.

[318] Z. L. Newman et al. "Architecture for the photonic integration of an optical atomic clock". In: *Optica* 6.5 (May 2019), pp. 680–685. DOI: 10.1364/OPTICA.6.000680.

[319] X. Zheng et al. "Differential clock comparisons with a multiplexed optical lattice clock". In: *Nature* 602.7897 (Feb. 2022), pp. 425–430. ISSN: 1476-4687. DOI: 10.1038/s41586-021-04344-y.

[320] P. Brax and C. Burrage. "Chameleon induced atomic afterglow". In: *Phys. Rev. D* 82 (9 Nov. 2010), p. 095014. DOI: 10.1103/PhysRevD.82.095014.

[321] P. Brax and C. Burrage. "Atomic precision tests and light scalar couplings". In: *Phys. Rev. D* 83 (3 Feb. 2011), p. 035020. DOI: 10.1103/PhysRevD.83.035020.

[322] S. S. Muñoz. *A particle's perspective on screening mechanisms*. 2024. arXiv: 2407.08779 [hep-ph]. URL: https://arxiv.org/abs/2407.08779.

[323] N. Ashby. "Relativity in the Global Positioning System". In: *Living Reviews in Relativity* 6.1 (Jan. 2003), p. 1. ISSN: 1433-8351. DOI: 10.12942/lrr-2003-1.

[324] C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. San Francisco: W. H. Freeman, 1973. ISBN: 9780716703440.

[325] D. Wands. "Extended gravity theories and the Einstein–Hilbert action". In: *Classical and Quantum Gravity* 11.1 (Jan. 1994), p. 269. DOI: 10.1088/0264-9381/11/1/025.

[326] G. Magnano and L. M. Sokołowski. "Physical equivalence between nonlinear gravity theories and a general-relativistic self-gravitating scalar field". In: *Phys. Rev. D* 50 (8 Oct. 1994), pp. 5039–5059. DOI: 10.1103/PhysRevD.50.5039.

[327] S. M. Carroll et al. "Classical stabilization of homogeneous extra dimensions". In: *Phys. Rev. D* 66 (2 July 2002), p. 024036. DOI: 10.1103/PhysRevD.66.024036.

[328] M. Montenegro and A. C. Ponce. "The sub-supersolution method for weak solutions". In: *Proceedings of the American Mathematical Society* 136.07 (Feb. 2008), pp. 2429–2438. ISSN: 0002-9939. DOI: 10.1090/S0002-9939-08-09231-9.

[329] M. Badiale and E. Serra. *Semilinear Elliptic Equations for Beginners. Existence Results via the Variational Approach*. 1st ed. London: Springer, 2011. ISBN: 978-0-85729-226-1. DOI: 10.1007/978-0-85729-227-8.

[330] M. Struwe. *Variational Methods. Applications to Nonlinear Partial Differential Equations and Hamiltonian Systems*. Fourth. Vol. 34. Berlin, Heidelberg: Springer, 2008. ISBN: 978-3-540-74012-4. DOI: 10.1007/978-3-540-74013-1.

[331] M. Willem. *Minimax Theorems*. 1st ed. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Boston, MA, 1996. ISBN: 978-0-8176-3913-6. DOI: 10.1007/978-1-4612-4146-1.

[332] A. Ambrosetti and P. H. Rabinowitz. "Dual variational methods in critical point theory and applications". In: *Journal of Functional Analysis* 14.4 (1973), pp. 349–381. ISSN: 0022-1236. DOI: https://doi.org/10.1016/0022-1236(73)90051-7.

[333] P. H. Rabinowitz. "Some critical point theorems and applications to semilinear elliptic partial differential equations". en. In: *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze* Ser. 4, 5.1 (1978), pp. 215–223. URL: http://www.numdam.org/item/ASNSP%5C_1978%5C_4%5C_5%5C_1%5C_215%5C_0.

[334] W. Fulks and J. S. Maybee. "A singular non-linear equation". In: *Osaka Mathematical Journal* 12.1 (1960), pp. 1–19.

[335] C. A. Stuart. "Existence and approximation of solutions of non-linear elliptic equations". In: *Mathematische Zeitschrift* 147.1 (Feb. 1976), pp. 53–63. ISSN: 1432-1823. DOI: 10.1007/BF01214274.

[336] P. H. R. M. G. Crandall and L. Tartar. "On a dirichlet problem with a singular nonlinearity". In: *Communications in Partial Differential Equations* 2.2 (1977), pp. 193–222. DOI: 10.1080/03605307708820029.

[337] H. Amann. "A uniqueness theorem for nonlinear elliptic boundary value problems". In: *Archive for Rational Mechanics and Analysis* 44.3 (Jan. 1972), pp. 178–181. ISSN: 1432-0673. DOI: 10.1007/BF00250776.

[338] A. Lazer and P. McKenna. "A singular elliptic boundary value problem". In: *Applied Mathematics and Computation* 65.1 (1994), pp. 183–194. ISSN: 0096-3003. DOI: https://doi.org/10.1016/0096-3003(94)90175-9.

[339] O. A. Ladyzhenskaya and N. N. Ural'tseva. *Linear and quasilinear elliptic equations*. Ed. by R. Bellman. Vol. 46. University of Southern California: Elsevier, 1968. ISBN: 978-0-12-432850-1. DOI: https://doi.org/10.1016/S0076-5392(08)62571-0.

[340] P. Bolle, N. Ghoussoub, and H. Tehrani. "The multiplicity of solutions in non-homogeneous boundary value problems". In: *manuscripta mathematica* 101.3 (Mar. 2000), pp. 325–350. ISSN: 1432-1785. DOI: 10.1007/s002290050219.

[341] T. Bartsch, Z.-Q. Wang, and M. Willem. "Chapter 1 - The Dirichlet Problem for Superlinear Elliptic Equations". In: *Stationary Partial Differential Equations*. Ed. by M. Chipot and P. Quittner. Vol. 2. Handbook of Differential Equations: Stationary Partial Differential Equations. North-Holland, 2005, pp. 1–55. DOI: https://doi.org/10.1016/S1874-5733(05)80009-9.

[342] M. Schechter. "Superlinear elliptic boundary value problems". In: *manuscripta mathematica* 86.1 (Dec. 1995), pp. 253–265. ISSN: 1432-1785. DOI: 10.1007/BF02567993.

[343] M. Ramos. *On elliptic equations with superlinear nonlinearities*. Featured article of the 24th Bulletin of the International Center for Mathematics. June 2008. URL: http://www.cim.pt/magazines/bulletin/16/article/133/pdf.

[344] A. Bahri and P. L. Lions. "Morse index of some min-max critical points. I. Application to multiplicity results". In: *Communications on Pure and Applied Mathematics* 41.8 (Dec. 1988), pp. 1027–1037. DOI: https://doi.org/10.1002/cpa.3160410803.

[345] A. Salvatore. "Multiple solutions for some elliptic equations with non-homogeneous boundary conditions". In: *Nonlinear Analysis: Theory, Methods & Applications* 47.3 (Aug. 2001). Proceedings of the Third World Congress of Nonlinear Analysts, pp. 1593–1604. ISSN: 0362-546X. DOI: https://doi.org/10.1016/S0362-546X(01)00293-0.

[346] I. Kuzin and S. Pohozaev. *Entire Solutions of Semilinear Elliptic Equations*. 1st ed. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Basel, 1997. ISBN: 978-3-7643-5323-0. DOI: 10.1007/978-3-0348-9250-6.

[347] H. Brezis. "Semilinear equations in $\mathbb{R}^N$ without condition at infinity". In: *Applied Mathematics and Optimization* 12.1 (Oct. 1984), pp. 271–282. ISSN: 1432-0606. DOI: 10.1007/BF01449045.

[348] J. E. Frank. *Constrained Dynamics*. Lecture notes, Numerical Modelling of Dynamical Systems, Utrecht University. Nov. 5, 2008. URL: https://webspace.science.uu.nl/~frank011/Classes/numwisk/ch15.pdf.

[349] E. Hairer, G. Wanner, and C. Lubich. "Conservation of First Integrals and Methods on Manifolds". In: *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 97–142. ISBN: 978-3-540-30666-5. DOI: 10.1007/3-540-30666-8_4.