Reliable estimations of a function and its derivatives & Theoretical and numerical study of a shape from shading problem

Jean-Francois Abadie




Sorbonne Université

École doctorale **Sciences Mathématiques de Paris Centre (ED 386)**

Unité de recherche **Laboratoire Jacques-Louis LIONS**

Thèse présentée par **Jean-François ABADIE**

Soutenue le **18 octobre 2019**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Analyse numérique**

Estimations fiables d'une fonction et de ses dérivées & Étude théorique et numérique d'un problème de « shape from shading »

Thèse dirigée par **Yvon MADAY**

Composition du jury :

Rapporteurs	Martin J. GANDER	Professeur à l'Université de Genève
	Bertrand MAURY	Professeur à l'Université Paris-Sud
Examinateurs	Edwige GODLWESKI	Professeur émérite à Sorbonne Université
	Madalina-Elena PETCU	Maître de conférences à l'Université de Poitiers
	Didier SMETS	Professeur à Sorbonne Université
Directeur de thèse	Yvon MADAY	Professeur à Sorbonne Université

Keywords

Differentiable functions, bounded functions, certified bounds, Taylor-Lagrange formula, optimisation, real time computations, Hamilton-Jacobi equations, finite differences schemes, shape from shading.

Mots-clés

Fonctions différentiables, fonctions bornées, bornes certifiées, formule de Taylor-Lagrange, optimisation, calculs en temps réel, équations de Hamilton-Jacobi, schémas aux différences finies, problème de « shape from shading ».

This thesis has been prepared at:
Cette thèse a été préparée au :

Laboratoire Jacques-Louis Lions

4, place Jussieu
75005 PARIS
FRANCE

https://www.ljll.math.upmc.fr/

À mon grand-père, parti bien trop tôt.

"*Il n'est pas nécessaire d'aller vite, le tout est de ne pas s'arrêter*"
Confucius

"*La seule chose que je sais, c'est que je ne sais rien*"
Socrate

"*Science sans conscience n'est que ruine de l'âme*"
François RABELAIS

Abstract

The work presented in the first part of this thesis is the result of a collaboration between Alstom and the RATP. We present various models and algorithms that can be used to bound a real-valued function f defined on an interval I and its $(d-1)$ first derivatives by knowing reliable bounds on f in some discrete points and global bounds on its d^{th} derivative. These results are applied to a situation inspired by the railway world. Finally, we present various extensions of our work, and we explain how the previous models can be easily generalized to vector-valued applications defined on an interval.
The second part of this thesis is dedicated to the theoretical and numerical study of a *shape from shading* problem, which consists in a surface reconstitution from a black and white picture, by knowing only the shades of gray and the altitude of the surface at some points. We remind how the viscosity solutions framework allows us to obtain a well-posed formulation of this problem. Then we expose an explicit expression of an approximation scheme associated to this problem, and we propose a significant optimization of some algorithms used to solve numerically such a problem.
In the future, the works presented in the two parts of the thesis could be coupled to allow a real-time guidance of flying objects like drones over a given region.

Résumé

Les travaux présentés dans la première partie de ce manuscrit de thèse sont le fruit d'une collaboration entre Alstom et la RATP. Nous y présentons différents modèles et algorithmes permettant de borner une fonction réelle f définie sur un intervalle I et ses $(d-1)$ premières dérivées à partir de bornes sur f en certains points et de bornes globales sur la dérivée d-ième de f. Nous appliquons cela à une situation inspirée du monde ferroviaire. Enfin, nous présentons diverses extensions de nos travaux, et nous montrons comment les résultats précédents peuvent se généraliser à des applications définies sur un intervalle I et à valeurs vectorielles.
La seconde partie de ce manuscrit est consacrée à l'étude théorique et numérique d'un problème de « *shape from shading* », qui consiste à reconstituer une surface représentée sur une image en noir et blanc, par la seule connaissance des nuances de gris et d'altitudes en certains points. Nous y rappelons comment le cadre des solutions de viscosité permet d'obtenir une formulation mathématique bien posée de ce problème. Nous donnons ensuite une formulation explicite d'un schéma d'approximation associé à ce problème, et nous proposons une optimisation notable d'algorithmes permettant de résoudre numériquement un tel problème.
À terme, l'ensemble des travaux présentés dans ce manuscrit pourraient être couplés pour permettre le guidage en temps réel de mobiles volants, tels que des drones, au dessus d'une région donnée.

Remerciements

Nous vivons, à mon sens, une bien drôle d'époque, où l'on ne voit pas toujours le temps passer, où l'on ne prend pas forcément le temps de profiter de la vie, des êtres ou des choses qui nous entourent, et où il est donc facile de sombrer dans l'indifférence. Ceci peut avoir pour fâcheuse conséquence de nous amener à ne considérer ou ne dire que ce qui ne va pas, et à ne faire que peu de cas du reste. C'est pourquoi je tiens à prendre le temps, à travers ces quelques lignes, quitte à bousculer les habitudes et au risque de paraître un peu long ou pompeux, de remercier ou évoquer celles et ceux que j'ai associés, de près ou de loin, et pour des raisons que je tenterai d'exposer du mieux que je le peux, à la réalisation et à l'aboutissement des travaux de thèse que je présente dans ce manuscrit.

Mes premiers remerciements iront tout naturellement à la personne sans qui cette aventure n'aurait pas été possible : il s'agit bien entendu de mon directeur, Yvon MADAY. Je voudrais tout d'abord le remercier pour la confiance, dont j'espère avoir été digne, qu'il m'accorde depuis notre première rencontre qui, sauf erreur de ma part, remonte au mardi 16 septembre 2014, à cette époque où tout juste agrégé, mais hésitant à faire carrière dans l'enseignement (je me trouvais encore trop jeune pour cela), je souhaitais plutôt faire « quelque chose en lien avec les transports », si possible ferrés. C'est alors qu'il m'a parlé du partenariat qu'il avait monté avec quelques personnes du LJLL et la co-entreprise Metrolab, détenue à parts égales par Alstom et la RATP, et de la possibilité de rejoindre son équipe à travers un stage de fin d'études. Stage qui, suite quelques batailles administrativo-juridiques, a permis le lancement des travaux de thèse rassemblés ici-même. À travers cette aventure de presque cinq ans maintenant, je souhaiterais également le remercier pour son soutien à bien des égards. Si l'on sait son emploi du temps à l'image de son bureau, à savoir extrêmement chargé, et donc ce monsieur (parfois ?) difficile à intercepter, je voudrais tout de même préciser qu'Yvon a toujours répondu présent lorsque je me suis retrouvé désespéré, que ce soit sur le plan scientifique, administratif ou personnel. En bref, je tiens à le remercier très sincèrement d'avoir rendu possible cette aventure et de m'avoir soutenu tout au long de celle-ci ; étudier à ses côtés durant ces cinq restera, pour moi, une grande fierté.

Vient ensuite et tout naturellement Didier SMETS qui, au début de l'année 2017, alors que la collaboration entamée en 2014 avec Metrolab s'est interrompue, m'a proposé un nouveau sujet d'étude. Merci à lui de m'avoir guidé et épaulé tout au long de cette nouvelle étude qui, je l'avoue, m'a donné (voire me donne toujours) bien du fil à retordre. Et pour laquelle, je pense, il y a encore beaucoup à (re)faire.

Je souhaiterais également adresser des remerciements chaleureux à Martin GANDER et Bertrand MAURY, qui ont accepté de bien vouloir rapporter mes travaux de thèse, et pour les différentes remarques qu'ils ont formulées au sujet de ce manuscrit. Mes remerciements iront également à Madalina PETCU et Edwige GODLEWSKI, qui ont toutes deux accepté de prendre part à mon jury de thèse.

Côté laboratoire, il convient de remercier vivement toute l'équipe administrative du LJLL pour sa gentillesse et sa réactivité face à mes diverses sollicitations, allant de l'emprunt de la clé de l'armoire à fournitures à la gestion de mes déplacements professionnels, la répartition des doctorants ou bien encore des réservations de salles : Nathalie BERGHAME, Catherine DROUET, Malika LARCHER et Salima

LOUNICI. Merci également à Kashayar DADRAS et les personnes qui se sont succédées à ses côtés pour s'occuper du parc informatique des laboratoires de Mathématiques du campus. Merci aussi à celles et ceux qui font du laboratoire un lieu convivial et chaleureux. À titre personnel, outre Yvon MADAY que j'ai largement mentionné précédemment, je voudrais plus spécialement (re-)mentionner Fatiha ALABAU, Laurent BOUDIN, Jean-Yves CHEMIN ou Edwige GODLEWSKI, pour leur écoute et leur soutien à différents égards. Enfin, mes différents collègues de bureaux et les (post-)doctorants avec qui j'ai vécu, de plus ou moins près, cette aventure : je leur dédierai un paragraphe spécial dans ce qui suit.

Avant cela, je voudrais redire tout le plaisir que j'ai pris à assurer mes T.D. d'algèbre linéaire (2M371) et de topologie et de calcul différentiel (3M260), en espérant qu'il aura été partagé par celles et ceux qui y ont assisté. À cette occasion, je souhaiterais remercier tous les collègues avec qui j'ai pu interagir à travers ces enseignements, comme par exemple Laurent BOUDIN, Muriel BOULAKIA, Bruno DESPRÉS, Cindy GUICHARD, Pierre-Antoine GUIHÉNEUF ou Didier SMETS, mais aussi Claire DAVID qui m'a accordé sa confiance en me confiant « ses » étudiants de REO en 2017. Il convient également de ne pas oublier Murielle THICOT et Julie CAMONIN, du côté des secrétariats pédagogiques, qui ont toujours donné bonne suite à mes diverses demandes. Et bien-sûr, un grand merci à tous les étudiants qu'il m'a été permis de rencontrer à travers ces enseignements pour leur implication et leur gentillesse, avec des remerciements peut-être encore plus appuyés pour ceux de la promotion PIMA 2017 — les tout premiers que j'ai eus — que j'ai suivis, tout au long de l'année 2016, en 2M371 puis en 3M260 (avec un clin d'œil amical à l'équipe « Bayonne 2017 », qui se reconnaîtra). Ces premières expériences heureuses dans l'enseignement m'ont convaincu de poursuivre dans cette voie.

Cela dit, je ne serais pas honnête si je limitais ce goût pour l'enseignement aux quelques propos qui précèdent. En effet, celui-ci a été cultivé, tout comme mon goût pour les Mathématiques, par certains enseignants que j'ai pu avoir dès mon plus jeune âge. À cet égard, je souhaiterais évoquer ceux que j'ai eu la chance d'avoir au lycée, en Mathématiques : Robert COSYNS en seconde, Roland CHARNOLÉ en première S, Christine MOURIER pour l'enseignement obligatoire de terminale S, et enfin Stéphane GOICHON pour l'enseignement de spécialité de terminale S, avec qui j'ai conservé un lien d'amitié depuis mon passage au lycée. Mention spéciale à mon enseignant de Physique-Chimie de seconde, Stéphane LAMAISON, dont les qualités pédagogiques continuent de m'inspirer aujourd'hui. Viennent ensuite Julien DUMONT et Gil GUIBERT, qui m'ont respectivement enseigné la Physique-Chimie et les Mathématiques lors de mon court passage en MPSI, et avec qui je reste en contact régulier depuis. Je n'oublierai pas non plus Bernadette DESHOMMES qui, outre son accueil et son soutien chaleureux lors de mon arrivée à l'Université de Poitiers, a toujours remarquablement encadré ses étudiants en T.D. Ni Patrice TAUVEL, dont les enseignements et les polycopiés ont largement inspirés tant ceux que j'ai pu préparer pour mes T.D. que la structure de ce manuscrit de thèse. Je terminerai enfin par mon grand-père paternel, Jean-Louis, avec qui j'ai bien souvent étudié les Mathématiques dans le bureau du sous-sol, parfois en compagnie de la petite Gribouille (scènes que seule la famille est à même de se représenter).

Comme promis, retour au laboratoire, côté (post-)doctorants cette fois-ci. Tout d'abord, pour saluer mes différents collègues de bureau et les remercier pour leur patience, eux qui ont dû bien souvent supporter mes nombreux jurons (par exemple quand mon code ne marchait pas !), comme Nora AÏSSIOUENE, Cuc BUI ou Tommaso TADDEI en 15-25-320, puis Yangyang CAO, Pierre MARCHAND, Jean RAX, ou plus partiellement Ziad KOBEISSI et Marc PEGON (les squatteurs de P7), suite à ma migration en 15-25-324. Mais aussi Marcella BONAZZOLI, Nicolas CLOZEAU, Léo GIRARDIN, Ludovic GODARD-CADILLAC, Olivier GRAF, Élise GROSJEAN, Anne-Françoise de GUERNY, Gabirela LOPÉZ RUIZ, Lise MAURIN, Lucile MÉGRET, David MICHEL, Anouk NICOLOPOULOS, Jules PERTINAND, Teddy PICHARD, Antonin PRUNET, Guillaume SALL, Cécile TAING ou, là-encore, Pierre MARCHAND, Marc PEGON et Jean RAX, pour les soirées, repas du midi et/ou pauses que j'ai pu partager à leurs côtés. Côté CROUS, pour les repas du midi, salutations chaleureuses à la caissière de la salle Seine, Fatou, pour son incroyable

sympathie. Et plus largement, pour reprendre une phrase répandue mais pourtant bien vraie, toutes celles et ceux que j'ai pu oublier et qui contribuent à la bonne entente entre les différents (post-)doctorants.

J'aimerais également mentionner Bertrand THIERRY, avec qui j'ai lancé, au début de l'année 2018, le séminaire « infomath » ayant pour but de présenter quelques outils ou bonnes pratiques informatiques aux mathématiciens que nous sommes (ou tentons d'être ?). Mais aussi Pierre MARCHAND (oui, encore lui !) et Alexandre POULAIN, qui sont venus nous prêter main forte à partir de l'année 2018/2019. Et enfin Daphné GIORGI (du LPSM) pour son implication à nos côtés nous permettant d'élargir notre public, et plus généralement tous ceux qui prennent désormais part à l'organisation de ce séminaire.

Sur le plan scientifique, toujours, je voudrais avoir une pensée amicale pour mon cher collègue de la SEME Lyon 2017, Florent DEWEZ, avec qui je garde contact depuis cette aventure, avouons-le, sans doute peu conventionnelle et parfois même un peu folle… Et puis aussi pour notre chef d'orchestre durant cette SEME, Christophe PICARD, de MicrodB (précision utile pour éviter toute confusion).

Je tiens également à remercier chaleureusement les enseignants du DAPS, et plus spécialement Jérôme GORI, Jean LOUCHU (qui coule désormais, je l'espère, des jours heureux), Serge MAURO (le premier que j'ai rencontré sur place), Didier NGUYEN et William PETIT, avec qui j'ai pu partager certains cours de musculation ou de P.P.G. à l'université, histoire de (tenter de ?) conserver « un esprit sain dans un corps sain ». Si j'exécrais par dessus tout les cours d'E.P.S. lorsque j'étais collégien ou lycéen, ils ont tous, à leur manière, réussi à me réconcilier avec l'activité physique, et même donné l'envie d'en pratiquer une régulièrement. Un salut amical à Soliman qui, de son cagibi qu'il assimile à une loge, est le gardien nocturne des salles de sport (et de mes lunettes). Et puis, une pensée pour Victor COURTIN, Adrien DELORO (dont j'aurais aussi pu mentionner les qualités pédagogiques), Anne-Françoise de GUERNY, Xavier LHEBRARD, Pierre MARCHAND (qui est décidément partout), Teddy PICHARD et Alexandre POULAIN pour les quelques séances, essentiellement de P.P.G. orchestrées par Serge, que nous avons partagées. Séances durant lesquelles avons fait « hurler nos muscles » (« ischios », « psoas », etc.), toujours « plus vite », en « augmentant les répétitions » et en « diminuant les temps de pause » (« aïe, aïe, aïe ! »), tout cela en sachant que, dans la salle d'agrès-musculation, « le sol est brûlant ».

Je prendrais aussi le temps de glisser un petit mot à mes amis, extérieurs au laboratoire et/ou à l'université, avec qui j'ai partagé quelques moments (discussions, restaurants, soirées ou vacances) en parallèle de cette aventure. Alexis PUYGRENIER et sa petite famille, depuis notre rencontre à Poitiers où nous cherchions des salles de cours avec de beaux tableaux pour réviser. Mounir MECHROUK et Sohrab SAMIMI, que j'ai connus en M1. Maxime CELHAY et Sébastien VILLETTE, avec qui nous avons usé les chaises de la B.U. ou de la salle 15-25-104 (puis plus tard, celles des bars) lors de (ou après) notre préparation à l'agrégation. Et puis Aline POTIRON, que je côtoie plus ou moins régulièrement depuis le lycée.

Enfin, place à ma famille, qui mérite sans doute bien plus que les modestes mots de remerciements que je peux lui offrir à travers ces quelques lignes, pour son soutien sans faille tant sur le plan moral, affectif que financier. Merci à mes parents, mes grands parents, ma sœur, mes cousins et cousine, mes oncles et tantes d'avoir été à mes côtés, chacune et chacun à leur manière, tout au long de ma vie. De m'avoir transmis des valeurs essentielles pour faire de moi un être responsable, telles que le respect, la tolérance, l'écoute ou la persévérance. À ma mère, pour les nombreux sacrifices qu'elle a pu faire dans le seul but de me construire et de m'assurer un avenir à la hauteur de ses espérances, et dont j'espère avoir été digne. Un grand merci aussi à ma femme, avec qui j'ai partagé une année extrêmement mouvementée et riche en émotions. Nous avons fait face, ensemble, à bon nombre d'épreuves, que nous avons toujours su surmonter, preuve que l'amour permet sans doute bien plus qu'on ne peut le soupçonner. À titre personnel, son amour m'a permis de tenir bon à des moments où je n'avais plus la force de poursuivre jusqu'à mes travaux de thèse. Je souhaiterais donc la remercier du fond de mon cœur d'être là pour moi,

de me soutenir, peut-être même de me supporter, de me redonner la confiance qui me fait souvent défaut, et lui dire que je lui dois beaucoup plus qu'elle ne le pense pour cette thèse et à bien d'autres égards. Et puis, je voudrais aussi avoir un petit mot pour notre enfant à venir (du moins, à l'instant où j'achève ces remerciements), une fille semble-t-il, dont nous avons appris l'existence à un mois de la soumission de ce manuscrit.

Pour conclure, je souhaiterais avoir une pensée pour mon grand-père maternel, Jean-Philippe, qui, je pense, aurait sincèrement souhaité connaître le dénouement de cette aventure. Malheureusement, le sort en a décidé autrement, et la maladie a eu raison de lui au début de l'année 2017 ; étrange coïncidence, quinze ans jour pour jour après la disparition de son propre père. J'aimerais donc rappeler que je ne l'oublierai jamais, et lui dédier ce manuscrit puis, je l'espère aussi, mon titre de docteur.

Table des matières

Avant-propos

Cette thèse est le fruit de travaux menés à l'Université Pierre et Marie Curie (UPMC), devenue, depuis 2018, Sorbonne Université, au laboratoire Jacques-Louis Lions (LJLL), sous la direction d'Yvon MADAY (rattaché au LJLL), dans le cadre d'un contrat doctoral s'étalant du 1er janvier 2016 au 31 août 2019.

Les premiers travaux de cette thèse résultent d'une collaboration entre le LJLL et Metrolab, co-entreprise détenue à parts égales par Alstom et la régie autonome des transports parisiens (RATP). Une des études menées à travers cette collaboration avait pour thématique l'*odométrie*, à savoir l'estimation de la position d'un véhicule en mouvement, et pour objectif de proposer différents modèles et algorithmes mathématiques permettant une meilleure estimation des position, vitesse et accélération d'un train, en temps réel, et selon un très haut niveau de garantie. La partie I de cette thèse présente la plupart des résultats obtenus dans le cadre de cette étude, débutée à l'automne 2014 et ayant pris fin à l'automne 2016.

Les travaux présentés dans la partie II de ce manuscrit, ont quant à eux été entamés au début de l'année 2017 et menés conjointement avec Didier SMETS (rattaché au LJLL) et Yvon MADAY. Ils portent sur la résolution d'un problème de « *shape from shading* », qui consiste en la reconstitution d'une surface représentée sur une image en noir et blanc à partir des nuances de gris et, bien entendu, d'altitudes connues en certains endroits.

Ce manuscrit a été rédigé dans l'état d'esprit suivant : apporter suffisamment d'éléments au lecteur pour lui permettre de comprendre les modèles ou algorithmes qui y sont décrits d'une part, et de reproduire les simulations qui y sont présentées d'autre part. Si Alstom et la RATP sont propriétaires du code `MATLAB` ayant permis la réalisation des simulations de la partie I, et n'ont pas souhaité que celui-ci soit diffusé, le code `python` associé à celles de la partie II sera regroupé dans un notebook `jupyter`, en cours de conception, destiné à être librement diffusé.

I

Reliable estimations of a function and its derivatives

Introduction

Notations et conventions spécifiques

Dans cette partie :

- I désignera un intervalle de $\mathbb{R}$ d'intérieur non vide.
- N désignera un entier supérieur ou égal à 2, et t_1, t_2, …, t_N désigneront N points de I tels que $t_1 < t_2 < \cdots < t_N$, non nécessairement régulièrement espacés.
- Si x est un nombre réel, nous noterons $\sigma(x)$ son signe, qui correspondra au symbole "$+$" lorsque $x \geqslant 0$, et au symbole "$-$" lorsque $x < 0$. En particulier, si $k \in \mathbb{N}$, le signe de $(-1)^k$ sera noté $\sigma\langle k \rangle$.
- Toutes les simulations numériques ont été réalisées sous MATLAB, et ce même si la plupart des sorties graphiques ont été reproduites, pour des raisons essentiellement esthétiques, sous LaTeX.

Contexte et problématique

Pour un exploitant de transports en commun ferrés, la bonne localisation des trains en temps réel répond à deux enjeux majeurs. Le premier est d'ordre sécuritaire : pour prévenir tout risque de collision, et donc d'accident grave de voyageurs, il est important de maintenir les trains suffisamment éloignés les uns des autres. Sur le plan commercial, pour éviter la saturation des espaces voyageurs (qui peut d'ailleurs s'avérer source d'accidents…), l'exploitant doit être en mesure de faire circuler, de la manière la plus fluide possible, un maximum de trains sur son réseau. Au final, il s'agit de maintenir les trains en ligne suffisamment proches les uns des autres sans que cela ait un quelconque impact sur la sécurité.
Notre étude, menée pour les besoins d'Alstom et de la RATP, rentre dans le cadre de ce problème. Notre objectif était de proposer des modèles et algorithmes mathématiques qui permettent d'estimer les position, vitesse et accélération d'un train, en temps réel, et selon un très haut niveau de garantie. Le comportement d'un train est modélisé par des lois physiques usuelles : la dérivée de sa position correspond à sa vitesse, la dérivée de sa vitesse à son accélération, etc. Le problème est qu'il est extrêmement difficile, voire impossible, de quantifier ces grandeurs, puisqu'ils résultent de processus physiques tels qu'accélérations, décélérations ou glissements que l'on ne sait pas modéliser ou estimer de façon fiable. Pour cette raison, nous pouvions seulement nous fier à des mesures de la position du train données en des temps discrets (par des équipements placés à bord du train ou sur la voie), et en des bornes certifiées sur l'accélération du train et son *jerk*, *i.e.* la dérivée de son accélération.
Exprimons tout ceci en termes mathématiques. Supposons disposer d'une fonction $f : I \to \mathbb{R}$ soumise aux hypothèses suivantes :

(A_1) Pour tout $i \in \{1, \ldots, N\}$, il est possible de déterminer des nombres réels $f_-(t_i)$ et $f_+(t_i)$ tels que :

$$f_-(t_i) \leqslant f(t_i) \leqslant f_+(t_i),$$

(A_2) Il existe (au moins) un entier $d \in \mathbb{N}^*$ pour lequel f admet une dérivée d-ième bornée, autrement dit il est possible de déterminer deux constantes réelles f_-^* and f_+^* telles que pour tout $\theta \in I$:

$$f_-^* \leqslant f^{(d)}(\theta) \leqslant f_+^*,$$

En pratique, la fonction f sera bien évidemment inconnue, et l'entier d égal à 2 ou 3. Plus précisément, pour un réel t de I donné, $f(t)$ correspondra à la *position* du train (sur la voie) à l'instant t, $f'(t)$ à sa vitesse, $f''(t)$ à son *accélération* et $f'''(t)$ à son *jerk*. En utilisant ces notations, nous souhaitons donc :

- déterminer, pour tous $k \in \{1, \ldots, d-1\}$ et $i \in \{1, \ldots, N\}$, des quantités $f_-^{(k)}(t_i)$ et $f_+^{(k)}(t_i)$ telles que $f_-^{(k)}(t_i) \leqslant f^{(k)}(t_i) \leqslant f_+^{(k)}(t_i)$ et que l'écart $f_+^{(k)}(t_i) - f_-^{(k)}(t_i)$ soit le plus petit possible,
- mettre en cohérence, pour tous les indices $k \in \{1, \ldots, d-1\}$ et $i \in \{1, \ldots, N\}$, l'ensemble des bornes $f_-^{(k)}(t_i)$ et $f_+^{(k)}(t_i)$ ainsi disponibles.

Principaux outils

Conformément au très haut niveau de sécurité requis, nous avons décidé de ne développer *que des modèles ou algorithmes déterministes*. Par conséquent, la fiabilité des données qui en sont issues sera la même que celle des données d'entrée. En particulier, toute incohérence des données de sortie sera fatalement imputable aux données d'entrée. Par ailleurs, l'utilisation de modèles probabilistes, comme les filtres de Kalman utilisés dans ce genre de cas, par exemple évoqués par R. FARAGER dans [8], nous a paru inappropriée, les défaillances étant alors localisées dans les queues des distributions de probabilité.
L'approche que nous proposons ici est, à notre connaissance, totalement nouvelle. Elle est basée sur les deux résultats suivants, le premier d'entre eux étant la formule de Taylor-Lagrange, dont nous rappelons un énoncé :

Théorème. (*Formule de Taylor-Lagrange*)
Soient $m \in \mathbb{N}^*$, g une fonction de I dans $\mathbb{R}$, $x \in I$ et $h \in \mathbb{R}^*$ tel que $x + h \in I$. On suppose que f est $(m-1)$ fois dérivable en tout point compris entre x et $x+h$, et m fois dérivable en tout point strictement compris entre x et $x + h$. Il existe alors ξ strictement compris entre x et $x + h$ tel que :

$$g(x+h) = \left(\sum_{k=0}^{m-1} \frac{h^k}{k!}\, g^{(k)}(x) \right) + \frac{h^m}{m!}\, g^{(m)}(\xi_h).$$

Le second, dont la preuve est immédiate, est le suivant :

Proposition. (*Principe de sélection*)
Soit A un nombre réel, k un entier plus grand que 1, et $(a_i^-)_{1 \leqslant i \leqslant k}$ and $(a_i^+)_{1 \leqslant i \leqslant k}$ deux familles de nombres réels. Si $a_i^- \leqslant A \leqslant a_i^+$ pour tout $i \in \{1, \ldots, k\}$, alors :

$$A_- \leqslant A \leqslant A_+ \quad \text{où} \quad \left\{ \begin{array}{rcl} A^- & = & \max\{a_1^-, \ldots, a_m^-\}, \\ A^+ & = & \min\{a_1^+, \ldots, a_m^+\}. \end{array} \right.$$

De ce fait, à l'exception des techniques proposées par L. JAULIN dans [11], que nous avions considéré en début d'étude mais qui nous ont paru difficiles à mettre en œuvre, et le papier [13] de S. K. LELE sur les schémas compacts, que nous avions utilisé pour tenter d'optimiser certains modèles que nous proposons, aucune référence bibliographique n'accompagne les travaux présentés dans cette première partie de thèse. Ceux-ci résultent donc exclusivement des échanges que nous avons eu avec des industriels d'Alstom (I. BAINS, R. CADOT et J.-M. KLUTH) et de la RATP (S. FIORONI et F. JASMIN), et certains chercheurs du LJLL (J. GARNIER, Y. PRIVAT, E. TRÉLAT, et surtout Y. MADAY).

Organisation

Cette première partie de manuscrit s'organise comme suit. Au chapitre 1, nous présentons différentes techniques permettant de borner les dérivées d'une fonction ; pour ce faire, nous utiliserons des méthodes proches de celles traditionnellement utilisées pour approcher les dérivées intervenant dans une équation différentielle ordinaire ou aux dérivées partielles. Au chapitre 2, nous montrons comment mettre en cohérences différentes bornes sur une fonction et ses dérivées. Même si nous verrons par la suite comment les combiner, les théories de ces deux chapitres peuvent être utilisées indépendamment l'une de l'autre. Une résolution numérique du problème présenté dans cette introduction sera proposée au chapitre 3. Pour finir, le chapitre 4 regroupe différents résultats ou idées qui pourraient être utilisés pour améliorer les modèles présentés dans les chapitres précédents, mais que nous n'avons pas testés et/ou étudiés en détail. En particulier, nous montrerons dans ce dernier chapitre comment notre travail, essentiellement unidimensionnel, peut se généraliser de manière immédiate à des fonctions définies sur I et à valeurs dans $\mathbb{R}^n$, pour un entier n aussi grand que souhaité.

Chapter 1

How to bound the derivatives of a function?

Introduction

To solve the problem exposed in introduction, we first need to see how to bound the derivatives of a function. In order to do it, we fix in this chapter $t \in I$ and $\psi : I \to \mathbb{R}$ a(n unknown) function for which:

- for all $y \in I$, we can determine (known) quantities $\psi_-(y)$ and $\psi_+(y)$ such that:

$$\psi_-(y) \leqslant \psi(y) \leqslant \psi_+(y). \tag{1.1}$$

- there exists $p \in \mathbb{N}^*$ such that ψ admits a p^{th} bounded derivative, *i.e.* there exists (known) real constants ψ_-^* and ψ_+^* such that for all $\xi \in I$:

$$\psi_-^* \leqslant \psi^{(p)}(\xi) \leqslant \psi_+^*. \tag{1.2}$$

Then our goal consists in finding some techniques to bound $\psi^{(k)}(t)$, for all $k \in \{1, \dots, p\}$.
We will not consider the function f used in introduction for the two following reasons. The first one is that in practice, ψ will correspond to f as well as f'. The second one is because in practice, we will just have bounds on $\psi = f$ or $\psi = f'$ in $t_1, \dots, t_N$, and not overall I.
As mentioned in introduction, we will limit our study to cases $p = 2$ or $p = 3$, even if by doing heavier computations, most of the models presented here could be extended. From section 1.1 to section 1.9, we will present various ways that permits to bound a k^{th} derivative of ψ, with $0 < k < p$. For each one of them, we will proceed as follow:

1) finding generic formulas that enable to bound $\psi^{(k)}(t)$,

2) choosing bounds on $\psi^{(k)}(t)$ as close as possible.

More precisely in section 1.1 we will start with the simpler case $p = 2$, and so $k = 1$. Then until section 1.7, we will suppose $p = 3$. From sections 1.2 to 1.4 we will see how to bound $\psi'(t)$ without ψ'', and how to bound $\psi''(t)$ without ψ' from sections 1.5 to 1.7. In sections 1.8 and 1.9, we will present alternative ways to bound $\psi'(t)$ or $\psi''(t)$ by using bounds on ψ'' or ψ'. After that we will explain how to adapt the models presented in this chapter when ψ is just bounded in some discrete points (as function f from introduction is), and we will end by doing some general remarks about the models presented in this chapter.

1.1 A way to bound $\psi'(t)$ when $p = 2$

1.1.1 — In this section, we assume that $p = 2$. Thus our goal consists in bounding $\psi'(t)$ by using bounds on ψ from (1.1) and ψ'' from (1.2). In order to do this, we can apply the Taylor-Lagrange formula that insures, for all $\delta > 0$ satisfying $t - \delta \in I$, the existence of $\xi \in]t - \delta, t[$ such that

$$\psi(t - \delta) \;=\; \psi(t) - \delta\psi'(t) + \frac{\delta^2}{2}\,\psi''(\xi).$$

Therefore

$$\psi'(t) \;=\; \frac{\psi(t) - \psi(t - \delta)}{\delta} + \frac{\delta}{2}\,\psi''(\xi).$$

Finally for all $\delta > 0$ such that $t - \delta \in I$, by using (1.1) and (1.2), we obtain

$$\psi'_-(t\,;\delta) \;\leqslant\; \psi'(t) \;\leqslant\; \psi'_+(t\,;\delta) \qquad \text{with} \qquad \left\{ \begin{array}{rcl} \psi'_-(t\,;\delta) & = & \dfrac{\psi_-(t) - \psi_+(t - \delta)}{\delta} + \dfrac{\delta\psi^*_-}{2}, \\ \psi'_+(t\,;\delta) & = & \dfrac{\psi_+(t) - \psi_-(t - \delta)}{\delta} + \dfrac{\delta\psi^*_+}{2}. \end{array} \right. \tag{1.3}$$

1.1.2 — Let $I_t = \{\delta > 0 \mid t - \delta \in I\}$. Since inequalities from (1.3) are available for all $\delta \in I_t$, we need to choose the ones that are as close as possible. To do it, we introduce the associated *diameter function*

$$\operatorname{diam}\psi'_t \;:\; I_t \to \mathbb{R},\; \delta \mapsto \psi'_+(t\,;\delta) - \psi'_-(t\,;\delta)$$

which measures according to δ the diameter of the interval $[\psi'_-(t\,;\delta),\ \psi'_+(t\,;\delta)]$ to which $\psi'(t)$ belongs, *i.e.* the distance between $\psi'_-(t\,;\delta)$ the lower and $\psi'_+(t\,;\delta)$ the upper bounds on $\psi'(t)$ obtained in (1.3). For all $\delta \in I_t$, we clearly have

$$\operatorname{diam}\psi'_t(\delta) \;=\; \frac{[\psi_+(t) - \psi_-(t)] + [\psi_+(t - \delta) - \psi_-(t - \delta)]}{\delta} + \frac{\delta(\psi^*_+ - \psi^*_-)}{2}. \tag{1.4}$$

According to (1.1) and (1.2), differences $\psi_+(y) - \psi_-(y)$, for all $y \in I$, and $\psi^*_+ - \psi^*_-$ are positive. Consequently $\operatorname{diam}\psi'_t$ is a positive function and:

- If $\psi_+(y) - \psi_-(y) = 0$ for all $y \in I$, then $\operatorname{diam}\psi'_t(\delta)$ is minimal when δ is close to 0.
- If $\psi^*_+ - \psi^*_- = 0$, then $\operatorname{diam}\psi'_t(\delta)$ is minimal when δ is as huge as possible.
- In the other cases, we have $\operatorname{diam}\psi'_t(\delta) \to +\infty$ when $\delta \to 0$, but also when $\delta \to +\infty$ if $\inf I = -\infty$, hence the existence of a $\delta^* \in I_t$ that minimises the diameter function $\operatorname{diam}\psi'_t$.

1.1.3 — We now suppose that difference $\psi^*_+ - \psi^*_-$ is positive, and that differences $\psi_+(y) - \psi_-(y)$ are positive and do not depend on $y \in I$. According to (1.4), for all $\delta \in I_t$ we have

$$\operatorname{diam}\psi'_t(\delta) \;=\; \frac{\alpha}{\delta} + \beta\delta,$$

where α and β are two non-negative constants given by

$$\alpha \;=\; 2[\psi_+(t) - \psi_-(t)] \qquad \text{and} \qquad \beta \;=\; \frac{\psi^*_+ - \psi^*_-}{2}.$$

When the interval I (where ψ is defined) is sufficiently large, the minimizer of $\operatorname{diam}\psi'_t$ corresponds to the minimizer of the convex function

$$\varepsilon \,:\, \mathbb{R}_+^* \to \mathbb{R}_+,\ \delta \mapsto \frac{\alpha}{\delta} + \beta\delta,$$

that has an unique global minimizer δ^* on $\mathbb{R}_+^*$ satisfying

$$\delta^* \;=\; \sqrt{\frac{\alpha}{\beta}} \quad \text{and} \quad \varepsilon(\delta^*) \;=\; 2\sqrt{\alpha\beta}.$$

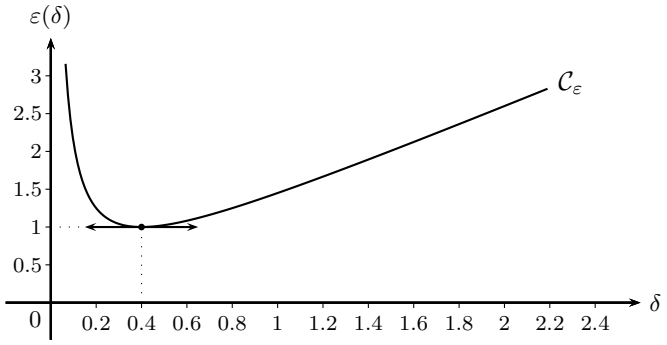

Figure 1 – Representative curve $\mathcal{C}_\varepsilon$ of the function ε of proposition 1.1.3 when $\alpha = 0.2$ and $\beta = 1.25$

1.2 Generic expressions of $\psi'(t)$ when $p = 3$

1.2.1 — From now on and until section 1.9, we will suppose that $p = 3$. According to (1.1) and (1.2), if we want to bound $\psi'(t)$, we can just use the various bounds on ψ and ψ'''. In order to do this, we can start from the Taylor expansions, available for all δ_1 and δ_2 in $\mathbb{R}^*$ such that $t + \delta_1 \in I$ and $t + \delta_2 \in I$,

$$\psi(t+\delta_1) \;=\; \psi(t) + \delta_1\psi'(t) + \frac{\delta_1^2}{2}\,\psi''(t) + \frac{\delta_1^3}{6}\,\psi'''(\xi_1), \tag{1.5}$$

$$\psi(t+\delta_2) \;=\; \psi(t) + \delta_2\psi'(t) + \frac{\delta_2^2}{2}\,\psi''(t) + \frac{\delta_2^3}{6}\,\psi'''(\xi_2), \tag{1.6}$$

where ξ_j is strictly between t and $t+\delta_j$, for all $j \in \{1, 2\}$. For all $\alpha_1, \alpha_2 \in \mathbb{R}$, by doing $\alpha_1(1.5)+\alpha_2(1.6)$ and by ordering $\alpha_1\delta_1 + \alpha_2\delta_2 \neq 0$ and $\alpha_1\delta_1^2 + \alpha_2\delta_2^2 = 0$, it is possible to express $\psi'(t)$ just by using various evaluations of ψ and ψ''', but not of ψ'', as follows:

$$\psi'(t) \;=\; \frac{\alpha_1\psi(t+\delta_1) + \alpha_2\psi(t+\delta_2) - (\alpha_1+\alpha_2)\psi(t)}{\alpha_1\delta_1 + \alpha_2\delta_2} - \frac{\alpha_1\delta_1^3\psi'''(\xi_1)}{6(\alpha_1\delta_1 + \alpha_2\delta_2)} - \frac{\alpha_2\delta_2^3\psi'''(\xi_2)}{6(\alpha_1\delta_1 + \alpha_2\delta_2)}.$$

Observing that we can multiply α_1 and α_2 by a same non-null quantity without changing this expression of $\psi'(t)$, we can set $\alpha_1 = \delta_2^2$, and so $\alpha_2 = -\delta_1^2$ (since $\alpha_1\delta_1^2 + \alpha_2\delta_2^2 = 0$). Moreover, to guarantee $0 \neq \alpha_1\delta_1 + \alpha_2\delta_2 = \delta_1\delta_2(\delta_2 - \delta_1)$, we need to impose $\delta_1 \neq \delta_2$. In the end, for all $\delta_1, \delta_2 \neq 0$ such that $t + \delta_1 \in I$, $t + \delta_2 \in I$ and $\delta_1 \neq \delta_2$, we have obtained

$$\psi'(t) \;=\; \frac{\delta_2^2\psi(t+\delta_1) - \delta_1^2\psi(t+\delta_2) - (\delta_2^2 - \delta_1^2)\psi(t)}{\delta_1\delta_2(\delta_2 - \delta_1)} + \frac{\delta_1\delta_2^2\psi'''(\xi_2)}{6(\delta_2 - \delta_1)} - \frac{\delta_2\delta_1^2\psi'''(\xi_1)}{6(\delta_2 - \delta_1)}. \tag{1.7}$$

1.2.2 — *Centred formulas.*

We suppose here that δ_1 and δ_2 have opposite signs. Then at the risk of exchanging the roles of δ_1 and δ_2, we can suppose that $\delta_1 > 0$ and $\delta_2 < 0$. Thus by setting $\delta_+ = \delta_1 > 0$, $\delta_- = -\delta_2 > 0$, $\xi_+ = \xi_1$ and $\xi_- = \xi_2$, formula (1.7) can be rewritten

$$\psi'(t) \;=\; \frac{\delta_-^2\psi(t+\delta_+) - \delta_+^2\psi(t-\delta_-) + (\delta_+^2-\delta_-^2)\psi(t)}{\delta_-\delta_+(\delta_+ + \delta_-)} - \frac{\delta_+\delta_-^2\psi'''(\xi_+)}{6(\delta_- + \delta_+)} - \frac{\delta_-\delta_+^2\psi'''(\xi_-)}{6(\delta_- + \delta_+)}. \quad (1.8)$$

This exact expression of $\psi'(t)$ is called a *centred formula* on $\psi'(t)$. But even if this denomination could be confusing, it does not mean that $\delta_- = \delta_+$.

1.2.3 — *Decentred formulas.*

With notations from 1.2.1, we now suppose that δ_1 and δ_2 have the same sign, that means $\delta_1 < 0$ and $\delta_2 < 0$, or $\delta_1 > 0$ and $\delta_2 > 0$. But since in practice ψ will be a function of a time variable, we will only treat the case where $\delta_1 < 0$ and $\delta_2 < 0$, that permits to bound $\psi'(t)$ just by using past data (useful to perform real time analyses). Obviously the interesting reader will also be able to find analogous results when $\delta_1 > 0$ and $\delta_2 > 0$.
Thus when $\delta_1 < 0$ and $\delta_2 < 0$ by setting $\delta_1' = -\delta_1$ and $\delta_2' = -\delta_2$, we note that we can work with positive δ_j in formula (1.7) by replacing δ_j by $-\delta_j$, for all $j \in \{1,2\}$. Therefore and at the risk of exchanging the roles of δ_1 and δ_2, for all $\delta_1, \delta_2 > 0$ such that $t - \delta_1 \in I$, $t - \delta_2 \in I$ and $\delta_1 < \delta_2$, formula (1.7) is equivalent to

$$\psi'(t) \;=\; \frac{\delta_1^2\psi(t-\delta_2) + (\delta_2^2-\delta_1^2)\psi(t) - \delta_2^2\psi(t-\delta_1)}{\delta_1\delta_2(\delta_2-\delta_1)} + \frac{\delta_1\delta_2^2\psi'''(\xi_2)}{6(\delta_2-\delta_1)} - \frac{\delta_2\delta_1^2\psi'''(\xi_1)}{6(\delta_2-\delta_1)}. \quad (1.9)$$

This exact expression of $\psi'(t)$ is called a *decentred formula* on $\psi'(t)$.

1.3 Centred bounds on $\psi'(t)$ when $p = 3$

1.3.1 — Let

$$P_t \;=\; \{(\delta_-,\delta_+) \in \mathbb{R}_+^* \mid t-\delta_- \in I, t+\delta_+ \in I\}.$$

For all $(\delta_-,\delta_+) \in P_t$, by setting

$$\begin{cases} \psi'_-(t\,;\delta_-,\delta_+) \;=\; \dfrac{\delta_-^2\psi_-(t+\delta_+) + (\delta_+^2-\delta_-^2)\psi_{-\sigma(\delta_+-\delta_-)}(t) - \delta_+^2\psi_+(t-\delta_-)}{\delta_-\delta_+(\delta_-+\delta_+)} - \dfrac{\delta_-\delta_+\psi_+^*}{6}, \\ \psi'_+(t\,;\delta_-,\delta_+) \;=\; \dfrac{\delta_-^2\psi_+(t+\delta_+) + (\delta_+^2-\delta_-^2)\psi_{\sigma(\delta_+-\delta_-)}(t) - \delta_+^2\psi_-(t-\delta_-)}{\delta_-\delta_+(\delta_-+\delta_+)} - \dfrac{\delta_-\delta_+\psi_-^*}{6}, \end{cases}$$

where the sign $\sigma(\delta_+ - \delta_-)$ of $\delta_+ - \delta_-$ also corresponds to the sign of $\delta_+^2 - \delta_-^2 = (\delta_+ - \delta_-)(\delta_+ + \delta_-)$. Since δ_- and δ_+ are positive, (1.1), (1.2) and the centred formula (1.8) imply:

$$\psi'_-(t\,;\delta_-,\delta_+) \;\leqslant\; \psi'(t) \;\leqslant\; \psi'_+(t\,;\delta_-,\delta_+). \quad (1.10)$$

$\psi'_-(t\,;\delta_-,\delta_+)$ and $\psi'_+(t\,;\delta_-,\delta_+)$ are respectively called lower and upper *centred bounds* on $\psi'(t)$.

1.3.2 — We now introduce the *diameter function* associated to bounds from (1.10) by setting

$$\operatorname{diam}\psi'_t \;:\; P_t \to \mathbb{R},\; (\delta_-,\delta_+) \mapsto \psi'_+(t\,;\delta_-,\delta_+) - \psi'_-(t\,;\delta_-,\delta_+).$$

For all $(\delta_-, \delta_+) \in P_t$, we have:

$$\operatorname{diam}\psi'_t(\delta_-,\delta_+) \;=\; \frac{\delta_-^2[\psi_+(t+\delta_+)-\psi_-(t+\delta_+)]+\delta_+^2[\psi_+(t-\delta_-)-\psi_-(t-\delta_-)]}{\delta_-\delta_+(\delta_-+\delta_+)} + \frac{\big|\delta_+^2-\delta_-^2\big|\,[\psi_+(t)-\psi_-(t)]}{\delta_-\delta_+(\delta_-+\delta_+)} + \frac{\delta_-\delta_+(\psi_+^*-\psi_-^*)}{6}.$$

According to (1.1) and (1.2), differences $\psi_+(y)-\psi(y)$, for all $y \in I$, and $\psi_+^*-\psi_-^*$ are non-negative. Therefore $\operatorname{diam}\psi'_t$ is also non-negative, and:

- If $\psi_+(y)-\psi(y)=0$ for all $y \in I$, then $\operatorname{diam}\psi'_t(\delta_-,\delta_+)$ is minimal when (δ_-,δ_+) is close to $(0,0)$.
- If $\psi_+^*-\psi_-^*=0$, then $\operatorname{diam}\psi_t(\delta_-,\delta_+)$ is minimal when δ_- and δ_+ are as huge as possible.
- In the other case, since $\operatorname{diam}\psi'_t(\delta_-,\delta_+) \to +\infty$ when $(\delta_-,\delta_+) \to (0,0)$, but also when $\delta_- \to +\infty$ or $\delta_+ \to +\infty$ if P_t (and thus I) is not bounded, we can conclude that there exists a couple $(\delta_-^*,\delta_+^*) \in P_t$ that minimizes the diameter function $\operatorname{diam}\psi'_t$.

1.3.3 — Until the end of this section, we suppose that difference $\psi_+^*-\psi_-^*$ is positive, and that differences $\psi_+(y)-\psi_-(y)$ are positive and do not depend on $y \in I$. Then for all $(\delta_-,\delta_+) \in P_t$,

$$\operatorname{diam}\psi'_t(\delta_-,\delta_+) \;=\; \frac{\alpha\big(\delta_-^2+\delta_+^2+\big|\delta_+^2-\delta_-^2\big|\big)}{\delta_-\delta_+(\delta_-+\delta_+)} + \beta\delta_-\delta_+,$$

where α and β are two non-negative constants given by

$$\alpha \;=\; \psi_+(t)-\psi_-(t) \qquad \text{and} \qquad \beta \;=\; \frac{\psi_+^*-\psi_-^*}{6}.$$

1.3.4 — Lemma. *Under the assumptions done in 1.3.3, for all $(\delta_-,\delta_+) \in P_t$, we have*

$$\operatorname{diam}\psi'_t(\delta_-,\delta_+) \;\geqslant\; \operatorname{diam}\psi'_t(\delta,\delta) \qquad \textit{with} \qquad \delta \;=\; \min(\delta_-,\delta_+).$$

Proof.— To start, we can note that for all $\delta > 0$ such that $t-\delta \in I$ and $t+\delta \in I$,

$$\operatorname{diam}\psi'_t(\delta,\delta) \;=\; \frac{\alpha}{\delta}+\beta\delta^2.$$

Let $(\delta_-,\delta_+) \in P_t$ such that $\delta_- \leqslant \delta_+$, and $c=\delta_+-\delta_-$. Thus $c \geqslant 0$ and $\delta_+=\delta_-+c$. Therefore

$$\begin{aligned}\operatorname{diam}\psi'_t(\delta_-,\delta_+) \;=\; \operatorname{diam}\psi'_t(\delta_-,\,\delta_-+c) \;&=\; \frac{\alpha(2\delta_-^2+4\delta_-c+2c^2)}{\delta_-(2\delta_-^2+3\delta_-c+c^2)}+\beta(\delta_-^2+\delta_-c)\\ &=\; \frac{\alpha}{\delta_-}+\beta\delta_-^2+\frac{\alpha(\delta_-c+c^2)}{\delta(2\delta_-^2+3\delta_-c+c^2)}+\beta\delta_-c\\ &=\; \operatorname{diam}\psi'_t(\delta_-,\delta_-)+\frac{\alpha(\delta_-c+c^2)}{\delta(2\delta_-^2+3\delta_-c+c^2)}+\beta\delta_-c.\end{aligned}$$

And since c, δ_-, α and β are non-negative, we obtain $\operatorname{diam}\psi'_t(\delta_-,\delta_+) \geqslant \operatorname{diam}\psi'_t(\delta_-,\delta_-)$. Otherwise since when $\delta_- > \delta_+$, we get $\operatorname{diam}\psi'_t(\delta_-,\delta_+) \geqslant \operatorname{diam}\psi'_t(\delta_+,\delta_+)$ by the same way, the result follows. □

1.3.5 — Let

$$I_t \;=\; \{\delta > 0 \mid t-\delta \in I,\; t+\delta \in I\} \qquad \text{and} \qquad D_t \;=\; \{(\delta,\delta) \mid \delta \in I_t\}.$$

Under the assumptions from 1.3.3, lemma 1.3.4 implies that the minimum of $\operatorname{diam}\psi'_t$ is reached on D_t. Therefore the research of a couple that minimizes the two-variables diameter function $\operatorname{diam}\psi'_t$ on P_t becomes a single-variable problem. In practice, we can explicitly determine a pair $(\delta^*, \delta^*) \in D_t$ that minimizes $\operatorname{diam}\psi'_t$ on P_t knowing that the function

$$\varepsilon \,:\, \mathbb{R}_+^* \to \mathbb{R},\ \delta \mapsto \frac{\alpha}{\delta} + \beta\delta^2,$$

which satisfies $\operatorname{diam}\psi'_{t\,|D_t} = \varepsilon_{|I_t}$, is decreasing on $]0, \delta^*]$ and increasing on $[\delta^*, +\infty[$, with

$$\delta^* \;=\; \left(\frac{\alpha}{2\beta}\right)^{1/3} \qquad \text{and} \qquad \varepsilon(\delta^*) = 3\left(\frac{\alpha}{2}\right)^{2/3}\beta^{1/3}.$$

1.4 Decentred bounds on $\psi'(t)$ when $p = 3$

1.4.1 — Let

$$P_t \;=\; \{(\delta_1, \delta_2) \in \mathbb{R}_+^* \mid t - \delta_1 \in I,\ t - \delta_2 \in I,\ \delta_1 < \delta_2\}.$$

For all $(\delta_1, \delta_2) \in P_t$, by setting

$$\begin{cases} \psi'_-(t\,;\delta_1,\delta_2) \;=\; \dfrac{\delta_1^2\psi_-(t-\delta_2) + (\delta_2^2-\delta_1^2)\psi_-(t) - \delta_2^2\psi_+(t-\delta_1)}{\delta_1\delta_2(\delta_2-\delta_1)} + \dfrac{\delta_1\delta_2^2\psi_-^*}{6(\delta_2-\delta_1)} - \dfrac{\delta_2\delta_1^2\psi_+^*}{6(\delta_2-\delta_1)}, \\ \psi'_+(t\,;\delta_1,\delta_2) \;=\; \dfrac{\delta_1^2\psi_+(t-\delta_2) + (\delta_2^2-\delta_1^2)\psi_+(t) - \delta_2^2\psi_-(t-\delta_1)}{\delta_1\delta_2(\delta_2-\delta_1)} + \dfrac{\delta_1\delta_2^2\psi_+^*}{6(\delta_2-\delta_1)} - \dfrac{\delta_2\delta_1^2\psi_-^*}{6(\delta_2-\delta_1)}, \end{cases}$$

we deduce from (1.1), (1.2) and the decentred formula (1.9) that

$$\psi'_-(t\,;\delta_1,\delta_2) \;\leqslant\; \psi'(t) \;\leqslant\; \psi'_+(t\,;\delta_1,\delta_2). \tag{1.11}$$

$\psi'_-(t\,;\delta_1,\delta_2)$ and $\psi'_+(t\,;\delta_1,\delta_2)$ are respectively called lower and upper *decentred bounds* on $\psi'(t)$.

1.4.2 — We now introduce the *diameter function* associated to bounds from (1.11)

$$\operatorname{diam}\psi'_t \,:\, P_t \to \mathbb{R},\ (\delta_1,\delta_2) \mapsto \psi'_+(t\,;\delta_1,\delta_2) - \psi'_-(t\,;\delta_1,\delta_2).$$

For all $(\delta_1, \delta_2) \in P_t$ we have

$$\begin{aligned} \operatorname{diam}\psi'_t(\delta_1,\delta_2) \;=\; & \frac{\delta_1^2[\psi_+(t-\delta_2) - \psi_-(t-\delta_2)] + \delta_2^2[\psi_+(t-\delta_1) - \psi_-(t-\delta_1)]}{\delta_1\delta_2(\delta_2-\delta_1)} \\ & + \frac{(\delta_2^2-\delta_1^2)[\psi_+(t) - \psi_-(t)]}{\delta_1\delta_2(\delta_2-\delta_1)} + \frac{(\delta_1\delta_2^2 + \delta_2\delta_1^2)(\psi_+^* - \psi_-^*)}{6(\delta_2-\delta_1)}. \end{aligned}$$

Thanks to (1.1) and (1.2), differences $\psi_+(y) - \psi(y)$, for all $y \in I$, and $\psi_+^* - \psi_-^*$ are non-negative. Therefore $\operatorname{diam}\psi'_t$ is also non-negative, and:

- If $\psi_+(y) - \psi(y) = 0$ for all $y \in I$, then $\operatorname{diam}\psi'_t(\delta_1,\delta_2)$ is minimal when (δ_1,δ_2) is close to $(0,0)$.
- If $\psi_+^* - \psi_-^* = 0$, then $\operatorname{diam}\psi_t(\delta_1,\delta_2)$ is minimal when δ_1 and δ_2 are as huge as possible.
- In the other case, since $\operatorname{diam}\psi'_t(\delta_1,\delta_2) \to +\infty$ when $(\delta_1,\delta_2) \to (0,0)$, but also when $\delta_1 \to +\infty$ if P_t (and thus I) is not bounded, we can conclude that there exists a couple $(\delta_1^*, \delta_2^*) \in P_t$ that minimizes the diameter function $\operatorname{diam}\psi'_t$.

1.4.3 — When difference $\psi_+^* - \psi_-^*$ is positive, and differences $\psi_+(y) - \psi_-(y)$ are positive and independent of $y \in I$, then for all $(\delta_1, \delta_2) \in P_t$,

$$\operatorname{diam} \psi_t'(\delta_1, \delta_2) \;=\; \frac{\alpha\delta_2}{\delta_1(\delta_2 - \delta_1)} + \frac{\beta(\delta_1\delta_2^2 + \delta_2\delta_1^2)}{\delta_2 - \delta_1},$$

where α and β are two positive constants given by

$$\alpha = 2[\psi_+(t) - \psi_-(t)] \quad \text{and} \quad \beta \;=\; \frac{\psi_+^* - \psi_-^*}{6}.$$

Therefore to determine a global minimizer of $\operatorname{diam} \psi_t'$, it can be useful to consider the function

$$\varepsilon \;:\; U \to \mathbb{R}_+,\; (\delta_1, \delta_2) \mapsto \frac{\alpha\delta_2}{\delta_1(\delta_2 - \delta_1)} + \frac{\beta(\delta_1\delta_2^2 + \delta_2\delta_1^2)}{\delta_2 - \delta_1},$$

where U is the set of couples $(\delta_1, \delta_2) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ such that $\delta_1 < \delta_2$. Since ε is positive and continue on U with $\varepsilon(\delta_1, \delta_2) \to +\infty$ when $\delta_2 \to 0$ or $\delta_2 - \delta_1 \to 0$ or $\delta_1 \to +\infty$, this function has actually a global minimizer. Unfortunately we have not found the exact expression of such a global minimizer. But in practice, it will be sufficient to compute an approximation of it.

REMARK. Such a function ε seems graphically convex. Since each critical point of a convex function is a global minimizer of it, the research of a global minimizer of ε could probably be limited to the research of one of its critical point.

1.5 Generic expressions of $\psi''(t)$ when $p = 3$

1.5.1 — To determine generic expressions of $\psi''(t)$, we will proceed as in section 1.2 by using linear combinations of relations (1.5) and (1.6) to express $\psi''(t)$ by using various evaluations of ψ and ψ''', but not of ψ'. So let δ_1, δ_2 as in 1.2.1. For all $\alpha_1, \alpha_2 \in \mathbb{R}$, by ordering here $\alpha_1\delta_1^2 + \alpha_2\delta_2^2 = 0$ and $\alpha_1\delta_1 + \alpha_2\delta_2 \neq 0$, and by doing $\alpha_1(1.5) + \alpha_2(1.6)$, we get

$$\psi''(t) \;=\; \frac{2[\alpha_1\psi(t+\delta_1) + \alpha_2\psi(t+\delta_2) - (\alpha_1+\alpha_2)\psi(t)]}{\alpha_1\delta_1^2 + \alpha_2\delta_2^2} - \frac{\alpha_1\delta_1^3\psi'''(\xi_1)}{3(\alpha_1\delta_1^2 + \alpha_2\delta_2^2)} - \frac{\alpha_2\delta_2^3\psi'''(\xi_2)}{3(\alpha_1\delta_1^2 + \alpha_2\delta_2^2)}.$$

Since we can multiply α_1 and α_2 by a same non-null quantity without changing this expression of $\psi''(t)$, we can choose $\alpha_1 = \delta_2$ and so $\alpha_2 = -\delta_1$ (since $\alpha_1\delta_1 + \alpha_2\delta_2 = 0$). Moreover in order to guarantee $0 \neq \alpha_1\delta_1^2 + \alpha_2\delta_2^2 = \delta_1\delta_2(\delta_1 - \delta_2)$, we need to impose $\delta_1 \neq \delta_2$. In the end, for all $\delta_1, \delta_2 > 0$ such that $t - \delta_1 \in I$, $t - \delta_2 \in I$ and $\delta_1 \neq \delta_2$, we get

$$\psi''(t) \;=\; \frac{2[\delta_1\psi(t+\delta_2) + (\delta_2 - \delta_1)\psi(t) - \delta_2\psi(t+\delta_1)]}{\delta_1\delta_2(\delta_2 - \delta_1)} + \frac{\delta_1^2\psi'''(\xi_1)}{3(\delta_2 - \delta_1)} - \frac{\delta_2^2\psi'''(\xi_2)}{3(\delta_2 - \delta_1)}. \tag{1.12}$$

1.5.2 — *Centred formulas.*

We suppose here that δ_1 and δ_2 have opposite signs. Then at the risk of exchanging the roles of δ_1 and δ_2, we can assume that $\delta_1 > 0$ and $\delta_2 < 0$. Thus by setting $\delta_+ = \delta_1 > 0$, $\delta_- = -\delta_2 > 0$, and $\xi_+ = \xi_1$, $\xi_- = \xi_2$, formula (1.12) can be rewritten

$$\psi''(t) \;=\; \frac{2[\delta_-\psi(t+\delta_+) + \delta_+\psi(t-\delta_-) - (\delta_- + \delta_+)\psi(t)]}{\delta_-\delta_+(\delta_- + \delta_+)} + \frac{\delta_-^2\psi'''(\xi_-)}{3(\delta_- + \delta_+)} - \frac{\delta_+^2\psi'''(\xi_-)}{3(\delta_- + \delta_+)}. \tag{1.13}$$

This expression of $\psi''(t)$ is called a *centred formula* on $\psi''(t)$, which still does not mean that $\delta_- = \delta_+$.

1.5.3 — *Decentred formulas.*

With notations from 1.5.1, we now suppose that δ_1 and δ_2 have the same sign. For the same reasons as those from paragraph 1.2.3, we will only consider that the case $\delta_1 < 0$ and $\delta_2 < 0$.
Then by setting $\delta'_1 = -\delta_1$ and $\delta'_2 = -\delta_2$, we note that we can work with positive δ_j in formula (1.7) by replacing δ_j by $-\delta_j$, for all $j \in \{1, 2\}$. Therefore, and at the risk of exchanging the roles of δ_1 and δ_2, for all $\delta_1, \delta_2 > 0$ such that $t - \delta_1 \in I$, $t - \delta_2 \in I$ and $\delta_1 < \delta_2$, we obtain the following *decentred formula* on $\psi''(t)$, which is just a reformulation of (1.12),

$$\psi''(t) \;=\; \frac{2[\delta_1\psi(t-\delta_2) + (\delta_2-\delta_1)\psi(t) - \delta_2\psi(t-\delta_1)]}{\delta_1\delta_2(\delta_2-\delta_1)} + \frac{\delta_2^2\psi'''(\xi_2)}{3(\delta_2-\delta_1)} - \frac{\delta_1^2\psi'''(\xi_1)}{3(\delta_2-\delta_1)}. \tag{1.14}$$

1.6 Centred bounds on $\psi''(t)$ when $p = 3$

1.6.1 — Let

$$P_t \;=\; \{(\delta_-, \delta_+) \in \mathbb{R}_+^* \mid t - \delta_- \in I, t + \delta_+ \in I\}.$$

For all $(\delta_-, \delta_+) \in P_t$, by setting

$$\begin{cases} \psi''_-(t\,;\delta_-,\delta_+) \;=\; \dfrac{2[\delta_-\psi_-(t+\delta_+) + \delta_+\psi_-(t-\delta_-) - (\delta_+ + \delta_-)\psi_+(t)]}{\delta_-\delta_+(\delta_- + \delta_+)} + \dfrac{\delta_-^2\psi_-^*}{3(\delta_- + \delta_+)} - \dfrac{\delta_+^2\psi_+^*}{3(\delta_- + \delta_+)}, \\ \psi''_+(t\,;\delta_-,\delta_+) \;=\; \dfrac{2[\delta_-\psi_+(t+\delta_+) + \delta_+\psi_+(t-\delta_-) - (\delta_+ + \delta_-)\psi_-(t)]}{\delta_-\delta_+(\delta_- + \delta_+)} + \dfrac{\delta_-^2\psi_+^*}{3(\delta_- + \delta_+)} - \dfrac{\delta_+^2\psi_-^*}{3(\delta_- + \delta_+)}, \end{cases}$$

we deduce from (1.1), (1.2) and the centred formula (1.13) that

$$\psi''_-(t\,;\delta_-,\delta_+) \;\leqslant\; \psi''(t) \;\leqslant\; \psi''_+(t\,;\delta_-,\delta_+). \tag{1.15}$$

$\psi''_-(t\,;\delta_-,\delta_+)$ and $\psi''_+(t\,;\delta_-,\delta_+)$ are respectively called lower and upper *centred bounds* on $\psi''(t)$.

1.6.2 — We now introduce the *diameter function* associated to bounds from (1.15)

$$\operatorname{diam}\psi''_t \;:\; P_t \to \mathbb{R},\; (\delta_-,\delta_+) \mapsto \psi''_+(t\,;\delta_-,\delta_+) - \psi''_-(t\,;\delta_-,\delta_+).$$

For all $(\delta_-, \delta_+) \in P_t$, we have

$$\begin{aligned} \operatorname{diam}\psi''_t(\delta_-,\delta_+) \;=\; & \frac{2\big(\delta_-[\psi_+(t+\delta_+) - \psi_-(t+\delta_+)] + \delta_+[\psi_+(t-\delta_-) - \psi_-(t-\delta_-)]\big)}{\delta_-\delta_+(\delta_- + \delta_+)} \\ & + \frac{2[\psi_+(t) - \psi_-(t)]}{\delta_-\delta_+} + \frac{(\delta_-^2 + \delta_+^2)(\psi_+^* - \psi_-^*)}{3(\delta_- + \delta_+)}. \end{aligned}$$

By (1.1) and (1.2), differences $\psi_+(y) - \psi(y)$, for all $y \in I$, and $\psi_+^* - \psi_-^*$ are non-negative. Consequently $\operatorname{diam}\psi'_t$ is also non-negative, and:

- If $\psi_+(y) - \psi(y) = 0$ for all $y \in I$, then $\operatorname{diam}\psi''_t(\delta_-,\delta_+)$ is minimal when (δ_-,δ_+) is close to $(0,0)$.
- If $\psi_+^* - \psi_-^* = 0$, then $\operatorname{diam}\psi''_t(\delta_-,\delta_+)$ is minimal when δ_- and δ_+ are as huge as possible.
- In the other cases, since $\operatorname{diam}\psi''_t(\delta_-,\delta_+) \to +\infty$ when $(\delta_-,\delta_+) \to (0,0)$, but also when $\delta_- \to +\infty$ or $\delta_+ \to +\infty$ when P_t (and thus I) is not bounded, we can conclude that there exists a couple (δ_-^*, δ_+^*) in P_t that minimizes the diameter function $\operatorname{diam}\psi''_t$.

1.6.3 — Until the end of this section, we assume that difference $\psi_+^* - \psi_-^*$ is positive and that differences $\psi_+(y) - \psi_-(y)$ are positive and do not depend on $y \in I$. Then for all $(\delta_-, \delta_+) \in P_t$,

$$\operatorname{diam} \psi_t''(\delta_-, \delta_+) \;=\; \frac{\alpha}{\delta_-\delta_+} + \frac{\beta(\delta_-^2 + \delta_+^2)}{(\delta_- + \delta_+)},$$

where α and β are two non-negative constants given by

$$\alpha \;=\; 4[\psi_+(t) - \psi_-(t)] \qquad \text{and} \qquad \beta \;=\; \frac{\psi_+^* - \psi_-^*}{3}.$$

1.6.4 — Lemma. *Under the assumptions from 1.6.3:*

(*i*) *For all $\delta \in \mathbb{R}_+^*$ such that $t - \delta \in I$ and $t + \delta \in I$, we have*

$$\operatorname{diam} \psi_t''(\delta, \delta) \;=\; \frac{\alpha}{\delta^2} + \beta\delta.$$

(*ii*) *For all $(\delta_-, \delta_+) \in P_t$ such that $t - \delta_- \in I$ and $t + \delta_+ \in I$, we have*

$$\operatorname{diam} \psi_t''(\delta_-, \delta_+) \;\geqslant\; \operatorname{diam} \psi_t''(\delta, \delta) \qquad \textit{with} \qquad \delta \;=\; \frac{\delta_- + \delta_+}{2}.$$

Proof.— We immediately obtain (*i*) by doing easy calculations. Let us prove (*ii*). In order to do this, we fix $(\delta_-, \delta_+) \in P_t$ with $t - \delta_- \in I$ and $t + \delta_+ \in I$.

– Thanks to (*i*), we immediatly obtain $\operatorname{diam} \psi_t'' \left(\frac{\delta_- + \delta_+}{2}, \frac{\delta_- + \delta_+}{2} \right) = \frac{4\alpha}{(\delta_- + \delta_+)^2} + \frac{\beta(\delta_- + \delta_+)}{2}$.

– Using the inequality $\delta_-\delta_+ \leqslant \left(\frac{\delta_- + \delta_+}{2} \right)^2 = \frac{(\delta_- + \delta_+)^2}{4}$, we get $\frac{1}{\delta_-\delta_+} \geqslant \frac{4}{(\delta_- + \delta_+)^2}$.

– Finally we also have

$$\frac{(\delta_-^2 + \delta_+^2)}{(\delta_- + \delta_+)} \;=\; \frac{2(\delta_-^2 + \delta_+^2)}{2(\delta_- + \delta_+)} \;=\; \frac{(\delta_- + \delta_+)^2 + (\delta_- - \delta_+)^2}{2(\delta_- + \delta_+)} \;\geqslant\; \frac{\delta_- + \delta_+}{2}.$$

In the end since α and β are non-negative, we obtain

$$\begin{aligned} \operatorname{diam} \psi_t''(\delta_-, \delta_+) \;&=\; \frac{\alpha}{\delta_-\delta_+} + \frac{\beta(\delta_-^2 + \delta_+^2)}{(\delta_+ + \delta_-)} \\ &\geqslant\; \frac{4\alpha}{(\delta_- + \delta_+)^2} + \frac{\beta(\delta_- + \delta_+)}{2} \;=\; \operatorname{diam} \psi_t'' \left(\frac{\delta_- + \delta_+}{2}, \frac{\delta_- + \delta_+}{2} \right), \end{aligned}$$

which ends the proof of this lemma. □

1.6.5 — Let

$$I_t \;=\; \{\delta > 0 \mid t - \delta \in I,\; t + \delta \in I\} \qquad \text{and} \qquad D_t \;=\; \{(\delta, \delta) \mid \delta \in I_t\}.$$

Under the assumptions from 1.6.3, when the interval I (where function ψ is defined) is sufficiently large, lemma 1.6.4 implies that $\operatorname{diam} \psi_t''$ reaches its minimum on D_t. In this case, to minimize $\operatorname{diam} \psi_t''$, we can use the fact that the function

$$\varepsilon \;:\; \mathbb{R}_+^* \to \mathbb{R}_+,\; \delta \mapsto \frac{\alpha}{\delta^2} + \beta\delta$$

as an unique global minimizer δ^* which satisfies

$$\delta^* \;=\; \left(\frac{2\alpha}{\beta} \right)^{1/3} \qquad \text{and} \qquad \varepsilon(\delta^*) = 3\alpha^{1/3} \left(\frac{\beta}{2} \right)^{2/3}.$$

1.7 Decentred bounds on $\psi''(t)$ when $p = 3$

1.7.1 — Let

$$P_t \;=\; \{(\delta_1,\delta_2)\in\mathbb{R}_+^* \mid t-\delta_1\in I,\; t-\delta_2\in I,\; \delta_1<\delta_2\}.$$

For all $(\delta_1,\delta_2)\in P_t$, by setting

$$\begin{cases} \psi''_-(t\,;\delta_1,\delta_2) \;=\; \dfrac{2[\delta_1\psi_-(t-\delta_2)+(\delta_2-\delta_1)\psi_-(t)-\delta_2\psi_+(t-\delta_1)]}{\delta_1\delta_2(\delta_2-\delta_1)}+\dfrac{\delta_2^2\psi_-^*}{3(\delta_2-\delta_1)}-\dfrac{\delta_1^2\psi_+^*}{3(\delta_2-\delta_1)},\\[2ex] \psi''_+(t\,;\delta_1,\delta_2) \;=\; \dfrac{2[\delta_1\psi_+(t-\delta_2)+(\delta_2-\delta_1)\psi_+(t)-\delta_2\psi_-(t-\delta_1)]}{\delta_1\delta_2(\delta_2-\delta_1)}+\dfrac{\delta_2^2\psi_+^*}{3(\delta_2-\delta_1)}-\dfrac{\delta_1^2\psi_-^*}{3(\delta_2-\delta_1)}, \end{cases}$$

we deduce from (1.1), (1.2) and the decentred formula (1.14) that

$$\psi''_-(t\,;\delta_1,\delta_2) \;\leqslant\; \psi''(t) \;\leqslant\; \psi''_+(t\,;\delta_1,\delta_2). \tag{1.16}$$

$\psi''_-(t\,;\delta_-,\delta_+)$ and $\psi''_+(t\,;\delta_-,\delta_+)$ are respectively called lower and upper *decentred bounds* on $\psi''(t)$.

1.7.2 — We now introduce the diameter function associated to bounds from (1.16):

$$\operatorname{diam}\psi''_t \;:\; P_t\to\mathbb{R},\; (\delta_1,\delta_2)\mapsto \psi''_+(t\,;\delta_1,\delta_2)-\psi''_-(t\,;\delta_1,\delta_2).$$

For all $(\delta_1,\delta_2)\in P_t$,

$$\begin{aligned}\operatorname{diam}\psi''_t(\delta_1,\delta_2) \;=\; &\frac{2\big(\delta_1[\psi_+(t-\delta_2)-\psi_-(t-\delta_2)]+\delta_2[\psi_+(t-\delta_1)-\psi_-(t-\delta_1)]\big)}{\delta_1\delta_2(\delta_2-\delta_1)}\\ &+\frac{2[\psi_+(t)-\psi_-(t)]}{\delta_1\delta_2}+\frac{(\delta_1^2+\delta_2^2)(\psi_+^*-\psi_-^*)}{3(\delta_2-\delta_1)}.\end{aligned}$$

By (1.1) and (1.2), differences $\psi_+(y)-\psi(y)$, for all $y\in I$, and $\psi_+^*-\psi_-^*$ are non-negative. Consequently $\operatorname{diam}\psi'_t$ is non-negative, and:

- If $\psi_+(y)-\psi(y)=0$ for all $y\in I$, then $\operatorname{diam}\psi''_t(\delta_1,\delta_2)$ is minimal when (δ_1,δ_2) is close to $(0,0)$.
- If $\psi_+^*-\psi_-^*=0$, then $\operatorname{diam}\psi''_t(\delta_1,\delta_2)$ is minimal when δ_1 and δ_2 are as huge as possible.
- In the other case, since $\operatorname{diam}\psi''_t(\delta_1,\delta_2)\to+\infty$ when $(\delta_1,\delta_2)\to(0,0)$, but also when $\delta_1\to+\infty$ if P_t (and thus I) is not bounded, we can conclude that there exists a couple $(\delta_1^*,\delta_2^*)\in P_t$ that minimizes the diameter function $\operatorname{diam}\psi''_t$.

1.7.3 — When difference $\psi_+^*-\psi_-^*$ is positive, and when differences $\psi_+(y)-\psi_-(y)$ are positive and independent of $y\in I$, then for all $(\delta_1,\delta_2)\in P_t$,

$$\operatorname{diam}\psi''_t(\delta_1,\delta_2) \;=\; \frac{\alpha}{\delta_1(\delta_2-\delta_1)}+\frac{\beta(\delta_1^2+\delta_2^2)}{(\delta_2-\delta_1)},$$

where α and β are two positive constants given by

$$\alpha \;=\; 4[\psi_+(t)-\psi_-(t)] \qquad\text{and}\qquad \beta \;=\; \frac{\psi_+^*-\psi_-^*}{3}.$$

Therefore by setting $U=\{(\delta_1,\delta_2)\in\mathbb{R}_+^*\times\mathbb{R}_+^* \mid \delta_1<\delta_2\}$, it can be useful to determine the minimizer(s) of the function:

$$\varepsilon \;:\; U\to\mathbb{R}_+,\; (\delta_1,\delta_2)\mapsto \frac{\alpha}{\delta_1(\delta_2-\delta_1)}+\frac{\beta(\delta_1^2+\delta_2^2)}{(\delta_2-\delta_1)}.$$

As for the decentred bounds on $\psi'(t)$, we can prove that this function has a global minimizer even if we have not found its exact expression, but it will be sufficient to compute an approximation of it in practice.

1.8 Crossed bounds on $\psi'(t)$ when $p = 3$

1.8.1 — In this section, we suppose that we are able determine real numbers $\psi''_-(t)$ and $\psi''_+(t)$ such that

$$\psi''_-(t) \ \leqslant \ \psi''(t) \ \leqslant \ \psi''_+(t). \tag{1.17}$$

Let us precise that it is always possible, for instance by using the theories from sections 1.6 or 1.7. Then the idea consists in using this information to determine other bounds on $\psi'(t)$. Let

$$I_t \ = \ \{\delta > 0 \mid t - \delta \in I\}.$$

By the Taylor-Lagrange formula, we know that for all $\delta \in I_t$, there exists $\xi \in]t - \delta, t[$ such that

$$\psi(t - \delta) \ = \ \psi(t) + \delta\psi'(t) + \frac{\delta^2}{2}\,\psi''(t) + \frac{\delta^3}{6}\,\psi'''(\xi), \tag{1.18}$$

that can be rewritten

$$\psi'(t) \ = \ \frac{\psi(t) - \psi(t - \delta)}{\delta} + \frac{\delta\psi''(t)}{2} - \frac{\delta^2\psi'''(\xi)}{6}.$$

Therefore by setting

$$\left\{\begin{array}{rcl} \psi'_-(t\,;\delta) & = & \dfrac{\psi_-(t) - \psi_+(t - \delta)}{\delta} + \dfrac{\delta\psi''_-(t)}{2} - \dfrac{\delta^2\psi^*_+}{6}, \\ \psi'_+(t\,;\delta) & = & \dfrac{\psi_+(t) - \psi_-(t - \delta)}{\delta} + \dfrac{\delta\psi''_+(t)}{2} - \dfrac{\delta^2\psi^*_-}{6}, \end{array}\right.$$

inequalities (1.1), (1.2) and (1.17) imply

$$\psi'_-(t\,;\delta) \ \leqslant \ \psi'(t) \ \leqslant \ \psi'_+(t\,;\delta). \tag{1.19}$$

Quantities $\psi'_-(t\,;\delta)$ and $\psi'_+(t\,;\delta)$ are respectively called lower and upper *crossed bounds* on $\psi'(t)$.

1.8.2 — Let us introduce the *diameter function* associated to the crossed bounds (1.19):

$$\operatorname{diam}\psi'_t \ : \ I_t \to \mathbb{R}, \ \delta \mapsto \psi'_+(t\,;\delta) - \psi'_-(t\,;\delta).$$

For all $\delta \in I_t$,

$$\operatorname{diam}\psi'_t(\delta) \ = \ \frac{[\psi_+(t) - \psi_-(t)] + [\psi_+(t - \delta) - \psi_-(t - \delta)]}{\delta} + \frac{\delta[\psi''_+(t) - \psi''_-(t)]}{2} + \frac{\delta^2(\psi^*_+ - \psi^*_-)}{6}.$$

By (1.1), (1.2) and (1.17), differences $\psi_+(y) - \psi_-(y)$, for all $y \in I$, $\psi^*_+ - \psi^*_-$ and $\psi''_+(t) - \psi''_-(t)$ are non-negative. Thus $\operatorname{diam}\psi'_t$ is also non-negative, and:

- If $\psi_+(y) - \psi_-(y) = 0$ for all $y \in I$, $\operatorname{diam}\psi'_t(\delta)$ is minimal when δ is close to 0.
- If $\psi''_+(t) - \psi''_-(t) = \psi^*_+ - \psi^*_- = 0$, then $\operatorname{diam}\psi'_t(\delta)$ is minimal when δ is as huge as possible.
- In the other cases, since $\operatorname{diam}\psi'_t(\delta) \to +\infty$ when $\delta \to 0$, but also when $\delta \to +\infty$ if $\inf I = -\infty$, we can conclude that there exists a $\delta^* \in I_t$ that minimizes the diameter function $\operatorname{diam}\psi'_t$.

1.8.3 — We now assume that differences $\psi_+(y) - \psi_-(y)$ are independent of $y \in I$. Then for all $\delta \in I_t$

$$\operatorname{diam} \psi'_t(\delta) \;=\; \frac{\alpha}{\delta} + \beta\delta + \gamma\delta^2,$$

where α, β and γ are three non-negative constants given by

$$\alpha \;=\; 2[\psi_+(t) - \psi_-(t)], \quad \beta \;=\; \frac{\psi''_+(t) - \psi''_-(t)}{2} \quad \text{and} \quad \gamma \;=\; \frac{\psi^*_+ - \psi^*_-}{6}.$$

In paragraph 1.8.2 we have explained how to minimize $\operatorname{diam} \psi'_t$ when $\alpha = 0$ or when $\beta = \gamma = 0$. In the other cases:

- When $\alpha > 0$, $\beta = 0$ and $\gamma > 0$, we can refer to paragraph 1.3.5.
- When $\alpha > 0$, $\beta > 0$ and $\gamma = 0$, we can refer to paragraph 1.1.3.

In the end, we just need to explain how to minimize $\operatorname{diam} \psi'_t$ when α, β and γ are all three positive.

1.8.4 — Lemma. *Let $\lambda, \mu, \nu > 0$ and*

$$u \;:\; \mathbb{R}^*_+ \to \mathbb{R},\; \delta \mapsto \lambda\delta^3 + \mu\delta^2 - \nu.$$

There exists an unique $\delta^* \in \mathbb{R}_+$ such that $u(\delta^*) = 0$, and u is negative on $]0, \delta^*]$, positive on $[\delta^*, +\infty[$.

Proof.— The function u is clearly differentiable and for all $\delta \in \mathbb{R}^*_+$

$$u'(t) \;=\; 3\lambda\delta^2 + 2\mu\delta \;=\; 3\lambda\delta\left(\delta + \frac{2\mu}{3\lambda}\right).$$

Since λ and μ are positive, it implies $u'(\delta) > 0$ for all $\delta \in \mathbb{R}^*_+$. Therefore u is non-decreasing on $\mathbb{R}^*_+$, and since $u(0) = -\mu < 0$ by hypothesis, $u(\delta) \to +\infty$ when $\delta \to +\infty$, we can easily conclude by using the continuity of u. □

1.8.5 — Taking the notations from 1.8.3 back, we suppose that α, β and γ are all three positive. Let

$$\varepsilon \;:\; \mathbb{R}^*_+ \to \mathbb{R}_+,\; \delta \mapsto \frac{\alpha}{\delta} + \beta\delta + \gamma\delta^2.$$

Functions $\operatorname{diam} \psi'_t$ and ε are obviously linked by the relation $\operatorname{diam} \psi'_t = \varepsilon_{|I_t}$. Thus to determine the minimizers of $\operatorname{diam} \psi'_t$, it could be useful to determine the ones of ε. For all $\delta \in \mathbb{R}^*_+$,

$$\varepsilon'(\delta) \;=\; \frac{2\gamma\delta^3 + \beta\delta^2 - \alpha}{\delta^2}.$$

Therefore lemma 1.8.4 applied with $\lambda = 2\gamma$, $\mu = \beta$ and $\nu = \alpha$ implies that ε has an unique minimizer $\delta^* \in \mathbb{R}^*$ satisfying

$$2\gamma(\delta^*)^3 + \beta(\delta^*)^2 - \alpha \;=\; 0.$$

As a root of a real polynomial of the third degree, it is possible to determine the exact expression of δ^*, but the corresponding formula is so complicated that we prefer to not explicit it. In practice, we suggest the use of an algorithm (e.g. Newton's method especially adapted to polynomial functions) to approach the exact value of δ^* the global minimizer of the function ε.

1.9 Crossed bounds on $\psi''(t)$ when $p = 3$

1.9.1 — In this section, we will lead an analogous analysis with $\psi''(t)$ as the one of section 1.8 with $\psi'(t)$. Thus we suppose here that we are able to determine quantities $\psi'_-(t)$ and $\psi'_+(t)$ such that

$$\psi'_-(t) \leqslant \psi'(t) \leqslant \psi'_+(t). \tag{1.20}$$

Let us take back the notation of I_t from paragraph 1.8.1. According to relation (1.18), we have

$$\psi''(t) = \frac{2[\psi(t-\delta)-\psi(t)]}{\delta^2} + \frac{2\psi'(t)}{\delta} + \frac{\delta\psi'''(\xi)}{3}.$$

Therefore by setting

$$\begin{cases} \psi''_-(t\,;\delta) = \dfrac{2[\psi_-(t-\delta)-\psi_+(t)]}{\delta^2} + \dfrac{2\psi'_-(t)}{\delta} + \dfrac{\delta\psi^*_-}{3}, \\ \psi''_+(t\,;\delta) = \dfrac{2[\psi_+(t-\delta)-\psi_-(t)]}{\delta^2} + \dfrac{2\psi'_+(t)}{\delta} + \dfrac{\delta\psi^*_+}{3}, \end{cases}$$

inequalities (1.1), (1.2) and (1.20) imply

$$\psi''_-(t\,;\delta) \leqslant \psi''(t) \leqslant \psi''_+(t\,;\delta). \tag{1.21}$$

Quantities $\psi''_-(t\,;\delta)$ and $\psi''_+(t\,;\delta)$ are respectively called lower and upper *crossed bounds* on $\psi''(t)$.

1.9.2 — Let us introduce the *diameter function* associated to the crossed bounds (1.21):

$$\operatorname{diam}\psi''_t \;:\; I_t \to \mathbb{R},\ \delta \mapsto \psi''_+(t\,;\delta) - \psi''_-(t\,;\delta).$$

For all $\delta \in I_t$, we have

$$\operatorname{diam}\psi''_t(\delta) = \frac{2\big[[\psi_+(t-\delta)-\psi_-(t-\delta)]+[\psi_+(t)-\psi_-(t)]\big]}{\delta^2} + \frac{2[\psi'_+(t)-\psi'_-(t)]}{\delta} + \frac{\delta(\psi^*_+-\psi^*_-)}{3}.$$

Thanks to (1.1), (1.2) and (1.20), differences $\psi_+(y)-\psi_-(y)$, for all $y \in I$, $\psi^*_+ - \psi^*_-$ and $\psi'_+(t)-\psi'_-(t)$ are non-negative. Thus $\operatorname{diam}\psi''_t$ is also non-negative, and:

- If $\psi_+(y)-\psi_-(y) = \psi'_+(t)-\psi'_-(t) = 0$ for all $y \in I$, then $\operatorname{diam}\psi''_t(\delta)$ is minimal when δ is close to 0.
- If $\psi^*_+ - \psi^*_- = 0$, then $\operatorname{diam}\psi''_t(\delta)$ is minimal when δ is as huge as possible.
- In the other cases, since $\operatorname{diam}\psi''_t(\delta) \to +\infty$ when $\delta \to 0$, but also when $\delta \to +\infty$ if $\inf I = -\infty$, we can conclude that there exists a $\delta^* \in I_t$ that minimizes the diameter function $\operatorname{diam}\psi''_t$.

1.9.3 — We now assume that differences $\psi_+(y)-\psi_-(y)$ are independent of $y \in I$. Then for all $\delta \in I_t$

$$\operatorname{diam}\psi''_t(\delta) = \frac{\alpha}{\delta^2} + \frac{\beta}{\delta} + \gamma\delta,$$

where α, β and γ are three non-negative constants given by

$$\alpha = 4[\psi_+(t)-\psi_-(t)], \quad \beta = 2[\psi'_+(t)-\psi'_-(t)] \quad \text{and} \quad \gamma = \frac{\psi^*_+-\psi^*_-}{3}.$$

In paragraph 1.9.2 we have explained how to minimize $\operatorname{diam}\psi'_t$ when $\alpha = \beta = 0$, or when $\gamma = 0$. In the other cases:

- When $\alpha = 0$, $\beta > 0$ and $\gamma = 0$, we can refer to proposition 1.1.3.
- When $\alpha > 0$, $\beta = 0$ and $\gamma > 0$, we can refer to proposition 1.6.5.

In the end, we just need to explain how to minimize $\operatorname{diam} \psi_t'$ when α, β and γ are all three positive.

1.9.4 — Lemma. *Let $\lambda, \mu, \nu > 0$ and*

$$u \ : \ \mathbb{R}_+^* \to \mathbb{R}, \ \delta \mapsto \lambda\delta^3 - \mu\delta - \nu.$$

There exists an unique $\delta \in \mathbb{R}_+^*$ such that $u(\delta^*) = 0$, u is negative on $]0, \delta^*[$ and u is positive on $]\delta^*, +\infty[$.

Proof.— The function u is clearly differentiable and for all $\delta \in \mathbb{R}_+^*$

$$u'(\delta) \ = \ 3\lambda\delta^2 - \mu.$$

Since λ and μ are non-negative, setting down $\rho = \sqrt{\mu/(3\lambda)}$, we can observe that u is non-increasing on $]0, \rho]$ and non-decreasing on $[\rho, +\infty[$. Consequently, having $u(0) = -\nu < 0$, u is negative on $]0, \rho]$. Moreover, having $u(\delta) \to +\infty$ when $\delta \to +\infty$, we can easily conclude by using the continuity and the monotony of u on $[\rho, +\infty[$. □

1.9.5 — Taking the notations from 1.9.3 back, we suppose that α, β and γ are all three positive. Let

$$\varepsilon \ : \ \mathbb{R}_+^* \to \mathbb{R}, \ \delta \mapsto \frac{\alpha}{\delta^2} + \frac{\beta}{\delta} + \gamma\delta.$$

Functions $\operatorname{diam} \psi_t''$ and ε are obviously linked by the relation $\varepsilon_{|I_t} = \operatorname{diam} \psi_t''$. Thus to determine the minimizers of $\operatorname{diam} \psi_t''$, it could be useful to determine those of ε. For all $\delta \in \mathbb{R}_+^*$

$$\varepsilon'(\delta) \ = \ \frac{\gamma\delta^3 - \beta\delta - 2\alpha}{\delta^3}.$$

Therefore lemma 1.9.4 applied with $\lambda = \gamma$, $\mu = \beta$ and $\nu = 2\alpha$ implies that ε gets an unique minimizer $\delta^* \in \mathbb{R}^*$ satisfying

$$\gamma(\delta^*)^3 - \beta\delta^* - 2\alpha \ = \ 0.$$

As in 1.8.5, we can determine the exact expression of δ^*, but its expression is so complicated that we suggest to approximate it by using a numerical algorithm (e.g. Newton's method).

1.10 When ψ is just bounded in some discrete points

1.10.1 — In this chapter, we have supposed being able to bound ψ at every point of I, which is a strong hypothesis than those done in the introduction of this part I. Indeed in practice ψ will just be estimated in $t_1, \ldots, t_N$. Therefore when inequalities from (1.1) are available only for all $y \in \{t_1, \ldots, t_N\}$, if we want to bound a derivative of ψ in $t = t_j$ for such a $j \in \{1, \ldots, N\}$, we must precise how to use the previous models. Thus in this section we will assume that ψ is bounded in $t_1, \ldots, t_N$ and that $t = t_j$ for a such a $j \in \{1, \ldots, N\}$. In other words it means that ψ still satisfies (1.2) but that (1.1) is now restrained to

$$\forall\, i \in \{1, \ldots, N\}, \qquad \psi_-(t_i) \ \leqslant \ \psi(t_i) \ \leqslant \ \psi_+(t_i).$$

Then to bound ψ' or ψ'' as previously, we can only use:

- bounds from (1.3), (1.19) and (1.21) by choosing $\delta = t_j - t_{j-r}$, for all index r satisfying $1 \leqslant j-r < j$, *i.e.* $1 \leqslant r \leqslant j-1$ (in particular, we need to suppose $j \geqslant 2$ to do this),

- centred bounds from (1.10) and (1.15) by choosing $\delta_- = t_j - t_{j-r_-}$ and $\delta_+ = t_{j+r_+} - t_j$, for all indexes r_- and r_+ satisfying $1 \leqslant j-r_- < j < j+r_+ \leqslant N$, *i.e.* $1 \leqslant r_- \leqslant j-1$ and $1 \leqslant r_+ \leqslant N-j$ (in particular, we need to suppose $2 \leqslant j \leqslant N-1$ to do this),

- decentred bounds from (1.11) and (1.16) by choosing $\delta_1 = t_j - t_{j-r_1}$ and $\delta_2 = t_j - t_{j-r_2}$, for all indexes r_1 and r_2 satisfying $1 \leqslant j-r_2 < j-r_1 < j$, *i.e.* $1 \leqslant r_1 < r_2 \leqslant j-1$ (in particular, we need to suppose $j \geqslant 3$ to do this).

With these notations, our goal consists in choosing as well as possible:

- index $r \in \{1, \dots, j-1\}$ with bounds from (1.3), (1.19) and (1.21),

- indexes $r_- \in \{1, \dots, j-1\}$ and $r_+ \in \{1, \dots, N-j\}$ with centred bounds from (1.10) and (1.15),

- indexes $r_1 \in \{1, \dots, j-2\}$ and $r_2 \in \{r_1+1, \dots, j-1\}$ with decentred bounds from (1.11) and (1.16),

to minimize the distance between the computed bounds on $\psi'(t_j)$ or $\psi''(t_j)$.

1.10.2 — Theoretically the previous indexes r, r_- and r_+, r_1 and r_2 are obtained by minimization of the corresponding diameter function on a discrete subset. By example, when $p = 2$ with bounds from (1.3), we have to minimize the diameter function $\operatorname{diam} \psi'_{t_j}$ from 1.1.2 on $I_{t_j} \cap \{t_j - t_{j-s} \mid 1 \leqslant s \leqslant j-1\}\}$. And such a problem can be quite complicated and costly (in time) when j is high or when the value of differences $\psi_+(t_i) - \psi_-(t_i)$ depend on the index $i \in \{1, \dots, N\}$.
That is why in the rest of this section, we will suggest an alternative way to choose r, r_- and r_+, r_1 and r_2 when there exists $\mu > 0$ such that for all $i \in \{1, \dots, N\}$, differences $\psi_+(t_i) - \psi_-(t_i)$ are equal or not so far from μ. For the crossed bounds on $\psi'(t_j)$ (resp. $\psi''(t_j)$) from (1.19) (resp. (1.21)), it will also be necessary to assume that for all $i \in \{1, \dots, N\}$, differences $\psi''_+(t_i) - \psi''_-(t_i)$ (resp. $\psi'_+(t_i) - \psi'_-(t_i)$) are equal or sufficiently close to a same value $\mu'' > 0$ (resp. $\mu' > 0$).

1.10.3 — *How to choose r with bounds from* (1.3)*,* (1.19) *and* (1.21)*?*

We will explain it on the example of the bounds from (1.3), knowing that we can do the same by making reference to section 1.8 with bounds from (1.19) and section 1.9 with bounds from (1.21). So let us take the notations from section 1.1 back, and let us fix $j \in \{2, \dots, N\}$.

1) We first determine δ^* the minimizer of function ε from proposition 1.1.3 taken with

$$\alpha \;=\; 2\mu \quad \text{ and } \quad \beta \;=\; \frac{\psi^*_+ - \psi^*_-}{2}.$$

2) Since $\operatorname{diam} \psi'_t = \varepsilon_{|I_t}$, then we should ideally take $r \in \{1, \dots, j-1\}$ such that $t_j - t_{j-r} = \delta^*$. Unfortunately it will never exist such an index r in practice. But since for all $i \in \{1, \dots, N\}$, differences $\psi_+(t_i) - \psi_-(t_i)$ are equal or not so far from μ, the idea consists in choosing an index r such that $t_j - t_{j-r}$ is not so far from δ^*. Knowing that function ε is decreasing on $]0, \delta^*]$ and increasing on $[\delta^*, +\infty[$, we can proceed as follows:

   - If $t_j - t_1 \leqslant \delta^*$, we choose $r = j-1$.
   - If $t_j - t_1 > \delta^*$, we choose r the smallest index $s \in \{1, \dots, j-1\}$ such that $t_j - t_{j-s} > \delta^*$.

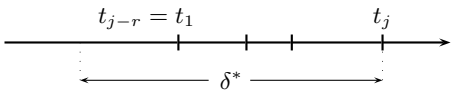

Figure 2 – Choice of index r when $t_j - t_1 \leqslant \delta^*$

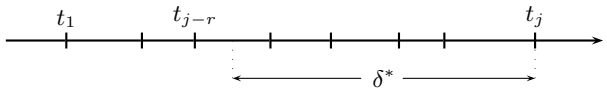

Figure 3 – Choice of index r when $t_j - t_1 > \delta^*$

1.10.4 — *How to choose r_- and r_+ with centred bounds from* (1.10) *and* (1.15)*?*

We will explain it on the example of centred bounds on $\psi'(t)$ from (1.10), knowing that the same method can be used with the centred bounds on $\psi''(t)$ from (1.15) by making reference to section 1.6. So let us take notations from section 1.3 back. In this case, indexes $r_- \in \{1, \ldots, j-1\}$ and $r_+ \in \{1, \ldots, N-j\}$ are chosen independently and similarly as the index r from paragraph 1.10.3.

1) We first determine δ^* the minimizer of function ε from proposition 1.3.5 taken with

$$\alpha \;=\; \mu \qquad \text{and} \qquad \beta \;=\; \frac{\psi_+^* - \psi_-^*}{6}.$$

2) Then r_- is selected as follows:

- If $t_j - t_1 \leqslant \delta^*$, we set $r_- = j - 1$.
- If $t_j - t_1 > \delta^*$, we choose r_- the smallest index $s \in \{1, \ldots, j-1\}$ such that $t_j - t_{j-s} > \delta^*$.

For index r_+:

- If $t_N - t_j \leqslant \delta^*$, we set $r_- = N - j$.
- If $t_N - t_j > \delta^*$, we choose r_- the smallest index $s \in \{1, \ldots, N-j\}$ such that $t_{j+s} - t_j > \delta^*$.

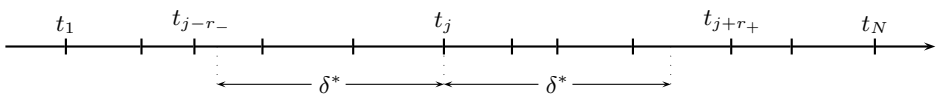

Figure 4 – Choice of indexes r_- and r_+ when $t_j - t_1 > \delta^*$ and $t_N - t_j > \delta^*$

1.10.5 — *How to choose r_1 and r_2 with decentred bounds from* (1.11) *and* (1.16)*?*

As previously we will limit our explanations by focusing on the decentred bounds on $\psi'(t)$ from (1.11), knowing that we can do it similarly with the decentred bounds on $\psi''(t)$ from (1.16) by refering to section 1.7.

1) We first determine (δ_1^*, δ_2^*) (an approximation of) the minimizer of function ε from 1.4.3 taken with

$$\alpha \;=\; 2\mu \qquad \text{and} \qquad \beta \;=\; \frac{\psi_+^* - \psi_-^*}{6}.$$

2) Here the choice of r_2 depends on the choice of r_1.

- If $t_j - t_2 \leqslant \delta_1^*$, we set down $r_1 = j-2$ and $r_2 = j-1$.
- If $t_j - t_2 > \delta_1^*$, then we first choose r_1 the smallest index $s \in \{1, \ldots, j-2\}$ such that $t_j - t_{j-s} > \delta_1^*$. After that:
  - If $t_j - t_1 \leqslant \delta_2^*$, we set $r_2 = j-1$.
  - If $t_j - t_1 > \delta_2^*$, we choose r_2 the smallest index $s \in \{r_1+1, \ldots, j-1\}$ such that $t_j - t_{j-s} > \delta_2^*$.

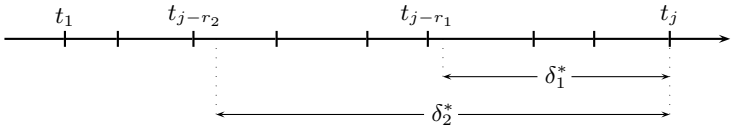

Figure 5 – Choice of indexes r_1 and r_2 when $t_j - t_2 > \delta_1^*$ and $t_j - t_1 > \delta_2^*$

1.10.6 — REMARK. When we suppose that differences $\psi_+(t_i) - \psi_-(t_i)$ are equal or sufficiently close to a same value μ, computations of δ^* (resp. δ^*, δ_1^* and δ_2^*) in the first time from 1.10.3 (resp. 1.10.4, 1.10.5) can be done *once and for all* before the bounds computations. In particular, it means that for a use in real time, it is sufficient to research the corresponding index r (resp. r_- and r_+, r_1 and r_2) and doing the bounds computations mentioned in a second time.

1.10.7 — *A possible improvement.*

The choice of the indexes r from 1.10.3, r_- and r_+ from 1.10.4, and r_1 and r_2 from 1.10.5 is quite arbitrary. That's why we suggest these possible improvements:

- For the index r, still on the example of bounds from (1.3), when $t_j - t_1 > \delta^*$ and $r \neq 1$, we can perform computations of bounds from (1.3) with $\delta = t_j - t_{j-r}$ and $\delta = t_j - t_{j-r+1}$, and by choosing $\psi'_-(t)$ the lower bound and $\psi'_+(t)$ the upper bound on $\psi'(t)$ by applying the selection principle as follows:

$$\left\{ \begin{array}{rcl} \psi'_-(t) & = & \max\left\{\psi'_-(t\,; t_j - t_{j-r}),\ \psi'_-(t\,; t_j - t_{j-r+1})\right\}, \\ \psi'_+(t) & = & \min\left\{\psi'_+(t\,; t_j - t_{j-r}),\ \psi'_+(t\,; t_j - t_{j-r+1})\right\}. \end{array} \right. \tag{1.22}$$

- For r_- and r_+, still on the example of centred bounds from (1.10), when $t_j - t_1 > \delta_1^*$ and $t_N - t_j > \delta^*$ with $r_- \neq 1$ and $r_+ \neq 1$, by setting

$$\Delta_- \ = \ \{t_j - t_{j-r_-},\ t_j - t_{j-r_-+1}\} \quad \text{and} \quad \Delta_+ \ = \ \{t_{j+r_+1} - t_j,\ t_{j-r_+} - t_j\},$$

  we can compute bounds from (1.10) for all couples (δ_-, δ_+) in $\Delta_- \times \Delta_+$. In the end, by using the selection principle, we can select $\psi'_-(t)$ the lower bound and $\psi'_+(t)$ the upper bound on $\psi'(t)$ given by:

$$\left\{ \begin{array}{rcl} \psi'_-(t) & = & \max\left\{\psi'_-(t\,; \delta_-, \delta_+) \,\middle|\, (\delta_-, \delta_+) \in \Delta_- \times \Delta_+\right\}, \\ \psi'_+(t) & = & \min\left\{\psi'_+(t\,; \delta_-, \delta_+) \,\middle|\, (\delta_-, \delta_+) \in \Delta_- \times \Delta_+\right\}. \end{array} \right. \tag{1.23}$$

- For r_1 and r_2, still on the example of decentred bounds from (1.11), when $t_N - t_2 > \delta_1^*$ and $t_N - t_1 > \delta_2^*$ with $r_1 \neq 1$ and $r_2 \neq r_1 + 1$, by setting

$$\Delta_1 \ = \ \{t_j - t_{j-r_1},\ t_j - t_{j-r_1+1}\} \quad \text{and} \quad \Delta_2 \ = \ \{t_j - t_{j-r_2},\ t_j - t_{j-r_2+1}\},$$

  we can compute bounds from (1.11) for all couples (δ_1, δ_2) in $\Delta_1 \times \Delta_2$. In the end by using the selection principle, we can select $\psi'_-(t)$ the lower bound and $\psi'_+(t)$ upper bound on $\psi'(t)$ given by:

$$\left\{ \begin{array}{rcl} \psi'_-(t) & = & \max\left\{\psi'_-(t\,; \delta_1, \delta_2) \,\middle|\, (\delta_1, \delta_2) \in \Delta_1 \times \Delta_2\right\}, \\ \psi'_+(t) & = & \min\left\{\psi'_+(t\,; \delta_1, \delta_2) \,\middle|\, (\delta_1, \delta_2) \in \Delta_1 \times \Delta_2\right\}. \end{array} \right. \tag{1.24}$$

REMARK. In practice ψ will correspond to a time-variable function. Moreover, N will be a very large integer (N larger than 10^3) and $t_{i+1} - t_i$ sufficiently small for all $i \in \{1, \ldots, N-1\}$. Therefore taking the notations from 1.10.3 (resp. 1.10.4, 1.10.5) back, even if j is close to 1 (resp. 1 or N, 1), we will have most of time $t_j - t_{j-r} > \delta^*$ (resp. $t_j - t_{j-r_-} > \delta^*$ and $t_{j+r_+} - t_j > \delta^*$, $t_j - t_{j-r_1} > \delta_1^*$ and $t_j - t_{j-r_2} > \delta_2^*$) with $r \neq 1$ (resp. $r_- \neq 1$ and $r_+ \neq 1$, $r_1 \neq 1$ and $r_2 \neq r_1 + 1$). Thus it will be possible to compute bounds on $\psi'(t)$ as in (1.22) (resp. (1.23), (1.24)).

1.11 Some remarks about out models

1.11.1 — We have proved that the functions ε associated to the various diameter functions reach their global minimum. Even if we have not found or mentioned for all of them the exact expression of their global minimizer, let us remind that it will be sufficient to approximate it in practice.

1.11.2 — Now let us talk about the advantages and disadvantages of the centred and decentred bounds presented in sections 1.3 and 1.4 for $\psi'(t)$ and in sections 1.6 and 1.7 for $\psi''(t)$. To simplify our analysis, we will suppose that differences $\psi_+(y) - \psi_-(y)$ are independent of $y \in I$ and we will only focus on bounds on $\psi'(t)$, knowing that similar phenomenons can be observed with $\psi''(t)$.
For various values of parameters $\psi_+(t) - \psi_-(t)$ and $\psi_+^* - \psi_-^*$ the following table shows:

– the value of δ^* the minimizer of function ε from proposition 1.3.5 (linked to the diameter function associated to bounds from (1.10)) and its evaluation $\varepsilon(\delta^*)$,

– a pair (δ_1^*, δ_2^*) that minimizes function ε from paragraph 1.4.3 (linked to the diameter function associated to bounds from (1.11)) and its associated evaluation $\varepsilon(\delta_1^*, \delta_2^*)$, obtained by using a MATLAB minimization algorithm.

Parameters		Results with ε from 1.3.5		Results with ε from 1.4.3		
$\psi_+(t) - \psi_-(t)$	$\psi_+^* - \psi_-^*$	δ^*	$\varepsilon(\delta^*)$	δ_1^*	δ_2^*	$\varepsilon(\delta_1^*, \delta_2^*)$
0.05	1	0.5313	0.1412	0.3760	1.7466	0.5084
0.05	2	0.4217	0.1778	0.2984	1.3863	0.6406
0.05	4	0.3347	0.2241	0.2368	1.1003	0.8071
0.1	1	0.6694	0.2241	0.4737	2.2006	0.8071
0.1	2	0.5313	0.2823	0.3760	1.7466	1.0169
0.1	4	0.4217	0.3557	0.2984	1.3863	1.2812
0.2	1	0.8434	0.3557	0.5968	2.7725	1.2812
0.2	2	0.6694	0.4481	0.4737	2.2006	1.6141
0.2	4	0.5313	0.5646	0.3760	1.7466	2.0337

Table 1 – Values of minimums and minimizers associated of functions ε from 1.3.5 and 1.4.3

We can observe that for a same choice of parameters, minimums of functions ε from 1.3.5 are about four times smaller than those of functions ε from 1.4.3, that means that the centred bounds are much more accurate than the decentred.
Nevertheless when ψ is a time-valued function with time expressed for instance in seconds, it means that we must wait around δ^* seconds to compute the more accurate centred bounds on $\psi'(t)$. And since δ^* is quite high, the use of centred bounds to perform a real time analysis seems not reasonable.

1.11.3 — In this chapter, we have seen various methods that enable to bound a same k^{th} derivative of a function $\psi : I \to \mathbb{R}$ in a point t. For each one of them, we have explained how to compute $\psi_-^{(k)}(t)$ a lower bound and $\psi_+^{(k)}(t)$ an upper bound on $\psi^{(k)}(t)$ as close as possible, by minimizing the associated diameter function. Even if for some of these methods, the global minimum of the diameter function will be lower than for others methods, *it does not mean* that the more accurate lower or upper bounds on $\psi^{(k)}(t)$ will be given by these first methods.
Consequently, it can be useful to compute bounds on $\psi^{(k)}(t)$ by using various techniques and by applying the selection principle to select the best bounds among those.

Conclusion

According to the problem exposed in introduction of this part I, we are now able to bounds the derivatives of the function f in times $t_1, \ldots, t_N$ by applying the previous models with $\psi = f$ or $\psi = f'$, and $p = 2$ or $p = 3$. Various examples of their use will be presented and analysed in detail in chapter 3. But before that, we will expose in chapter 2 a way that enables to establish coherence between the various (available) bounds on f, f' and f''.

Chapter 2

The forward-backward corrections

Introduction

By using the models from chapter 1 we are now able to bound the k^{th} derivatives of the function f, for all $k \in \{1, \ldots, d-1\}$. In this chapter, we will present how to establish coherence between the various available bounds on f and its derivatives in $t_1, \ldots, t_N$.
In order to do it, we fix $t \in I$ and $\varphi : I \to \mathbb{R}$ a(n unknown) function satisfying:

- For all $k \in \{0, \ldots, p-1\}$, we can determine (known) real numbers $\varphi_-^{(k)}(t)$ and $\varphi_+^{(k)}(t)$ such that

$$\varphi_-^{(k)}(t) \leqslant \varphi^{(k)}(t) \leqslant \varphi_+^{(k)}(t). \tag{2.1}$$

- There exists $p \in \mathbb{N}^*$ such that φ has a p^{th} bounded derivative, *i.e.* there exists (known) real constants φ_-^* and $\varphi_+^* \in \mathbb{R}$ such that for all $\xi \in I$

$$\varphi_-^* \leqslant \varphi^{(p)}(\xi) \leqslant \varphi_+^*. \tag{2.2}$$

In sections 2.1 and 2.2, we will see how to bound φ at each point $y \in I$ such that $y > t$ and $y < t$ respectively just by using data from (2.1) and (2.2), and how these bounds can be used to correct other bounds on φ in such a point y. From section 2.3, we will suppose the inequalities (2.1) satisfied for all $t \in \{t_1, \ldots, t_N\}$, and we will present in this same section two algorithms that perform the previous corrections on bounds on φ in $t_1, \ldots, t_N$. An analysis of these two algorithms will be led in section 2.4. In section 2.5 we will present an multi-order algorithm that establishes coherence between all bounds on φ and its derivatives in $t_1, \ldots, t_N$. To terminate this chapter, we will illustrate in section 2.6 the performances of the multi-order algorithm on a simple example.

2.1 Forward corrections

2.1.1 — For all $\delta > 0$ satisfying $t + \delta \in I$, by Taylor-Lagrange formula there exists $\xi \in]t, t+\delta[$ such that

$$\varphi(t+\delta) = \left(\sum_{k=0}^{p-1} \frac{\delta^k}{k!} \varphi^{(k)}(t) \right) + \frac{\delta^p}{p!} \varphi^{(p)}(\xi).$$

Therefore we deduce from (2.1) and (2.2):

$$\tau_-^{\mathrm{F}}(t+\delta) \leqslant \varphi(t+\delta) \leqslant \tau_+^{\mathrm{F}}(t+\delta) \quad \text{with} \quad \begin{cases} \tau_-^{\mathrm{F}}(t+\delta) = \left(\sum_{k=0}^{p-1} \frac{\delta^k}{k!}\varphi_-^{(k)}(t)\right) + \frac{\delta^p}{p!}\varphi_-^*, \\ \tau_+^{\mathrm{F}}(t+\delta) = \left(\sum_{k=0}^{p-1} \frac{\delta^k}{k!}\varphi_+^{(k)}(t)\right) + \frac{\delta^p}{p!}\varphi_+^*. \end{cases} \tag{2.3}$$

To make it clear, these formulas applied to each value of p in $\{1,2,3\}$ lead to

$$\begin{aligned} \tau_\pm^{\mathrm{F}}(t+\delta) &= \varphi_\pm(t) + \delta\varphi_\pm^* && \text{when } p=1, \\ \tau_\pm^{\mathrm{F}}(t+\delta) &= \varphi_\pm(t) + \delta\varphi_\pm'(t) + \frac{\delta^2}{2}\varphi_\pm^* && \text{when } p=2, \\ \tau_\pm^{\mathrm{F}}(t+\delta) &= \varphi_\pm(t) + \delta\varphi_\pm'(t) + \frac{\delta^2}{2}\varphi_\pm''(t) + \frac{\delta^3}{6}\varphi_\pm^* && \text{when } p=3. \end{aligned}$$

It means that if we are able to bound the p^{th} derivative of φ on overall I and its first $(p-1)^{\text{th}}$ derivatives just at *one point* t of I, then we can bound φ at *each point* of $D_t^{\mathrm{F}} = I \cap [t, +\infty[$. Nevertheless we can easily observe that the difference $\Delta\tau^{\mathrm{F}} : \delta \mapsto \tau_+^{\mathrm{F}}(t+\delta) - \tau_-^{\mathrm{F}}(t+\delta)$ is increasing with respect to δ on D_t^{F}, with $\Delta\tau^{\mathrm{F}}(\delta) \to +\infty$ when $\delta \to +\infty$ and $\sup I = +\infty$. Therefore the lower δ the higher the accuracy of lower and upper bounds $\tau_\pm^{\mathrm{F}}(t+\delta)$ of φ in $t+\delta$ will be.

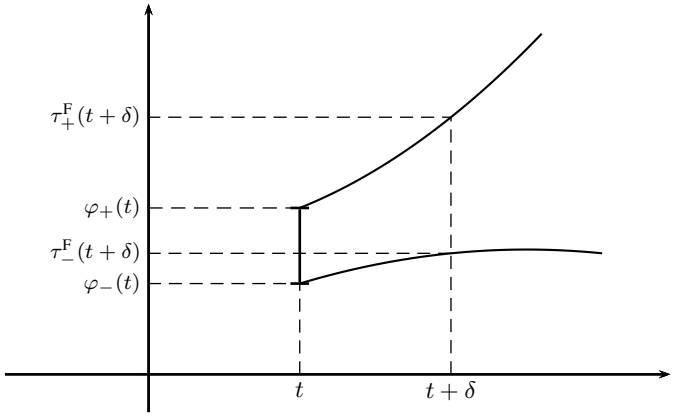

Figure 6 – Example of bounds on φ on D_t^{F} deduced from (2.3)

2.1.2 — We now suppose that for a fixed $\delta > 0$ such that $t+\delta \in I$, we have determined real numbers $\varphi_-(t+\delta)$ and $\varphi_+(t+\delta)$ such that

$$\varphi_-(t+\delta) \leqslant \varphi(t+\delta) \leqslant \varphi_+(t+\delta). \tag{2.4}$$

Then by applying the selection principle with $A = \varphi(t+\delta)$, we can use (2.3) to adjust the bounds from (2.4) as follows:

$$\varphi_-(t+\delta) \leftarrow \max\left\{\varphi_-(t+\delta),\ \tau_-^{\mathrm{F}}(t+\delta)\right\} \quad \text{and} \quad \varphi_+(t+\delta) \leftarrow \min\left\{\varphi_+(t+\delta),\ \tau_+^{\mathrm{F}}(t+\delta)\right\}. \tag{2.5}$$

These corrections are called *forward corrections* , hence the "F" on the bounds $\tau_\pm^{\mathrm{F}}(t+\delta)$. They are illustrated by figure 7 by considering the same situation as in figure 6.

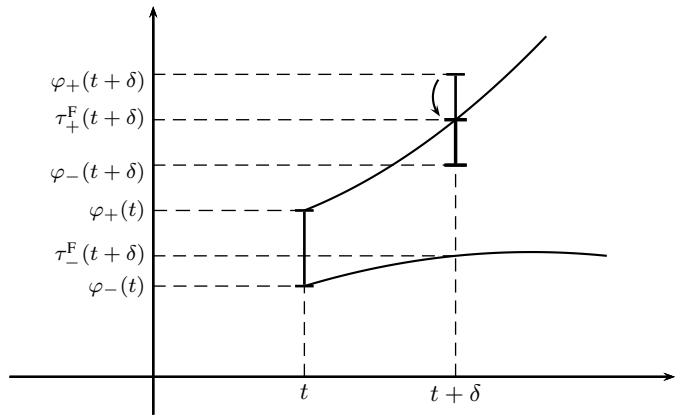

Figure 7 – Example of a forward correction (2.5) on an upper bound of φ in $t+\delta$

2.2 Backward corrections

2.2.1 — The backward corrections principle is analogous to the forward corrections one, but we have to deal with an additional difficulty due to signs considerations. Indeed for all $\delta > 0$ such that $t-\delta \in I$, Taylor-Lagrange formula insures the existence of $\xi \in]t-\delta, t[$ such that

$$\varphi(t-\delta) \;=\; \left(\sum_{k=0}^{p-1} \frac{(-1)^k\,\delta^k}{k!}\,\varphi^{(k)}(t)\right) + \frac{(-1)^p\,\delta^p}{p!}\,\varphi^{(p)}(\xi).$$

Therefore by (2.1) and (2.2) we obtain

$$\tau_-^{\mathrm{B}}(t-\delta) \;\leqslant\; \varphi(t-\delta) \;\leqslant\; \tau_+^{\mathrm{B}}(t-\delta) \tag{2.6}$$

with

$$\begin{cases} \tau_-^{\mathrm{B}}(t-\delta) \;=\; \left(\displaystyle\sum_{k=0}^{p-1} \frac{(-1)^k\delta^k}{k!}\varphi_{-\sigma\langle k\rangle}^{(k)}(t)\right) + \frac{(-1)^p\delta^p}{p!}\varphi_{-\sigma\langle p\rangle}^{*}, \\ \tau_+^{\mathrm{B}}(t-\delta) \;=\; \left(\displaystyle\sum_{k=0}^{p-1} \frac{(-1)^k\delta^k}{k!}\varphi_{\sigma\langle k\rangle}^{(k)}(t)\right) + \frac{(-1)^p\delta^p}{p!}\varphi_{\sigma\langle p\rangle}^{*}. \end{cases}$$

To clarify this, these formulas applied for each value of p in $\{1,2,3\}$ give

$$\begin{aligned} \tau_\pm^{\mathrm{B}}(t-\delta) \;&=\; \varphi_\pm(t) - \delta\varphi_\mp^* && \text{when } p=1, \\ \tau_\pm^{\mathrm{B}}(t-\delta) \;&=\; \varphi_\pm(t) - \delta\varphi_\mp'(t) + \frac{\delta^2}{2}\varphi_\pm^* && \text{when } p=2, \\ \tau_\pm^{\mathrm{B}}(t-\delta) \;&=\; \varphi_\pm(t) - \delta\varphi_\mp'(t) + \frac{\delta^2}{2}\varphi_\pm''(t) - \frac{\delta^3}{6}\varphi_\mp^* && \text{when } p=3. \end{aligned}$$

Now we are also able to bound φ on $D_t^{\mathrm{B}} = I \cap]-\infty, t]$. As in 2.1.1 we can observe that the lower δ the higher the accuracy of bounds $\tau_\pm^{\mathrm{B}}(t-\delta)$ on $\varphi(t-\delta)$ will be.

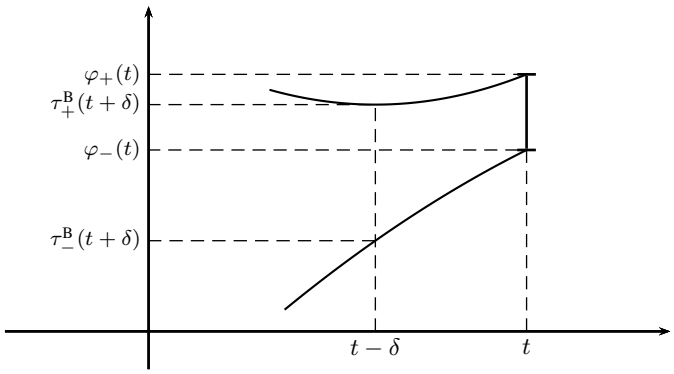

Figure 8 – Example of bounds on φ on D_t^{F} deduced from (2.6)

2.2.2 — We now suppose that for a fixed $\delta > 0$ such that $t - \delta \in I$, there exists real numbers $\varphi_-(t-\delta)$ and $\varphi_+(t-\delta)$ such that

$$\varphi_-(t-\delta) \;\leqslant\; \varphi(t-\delta) \;\leqslant\; \varphi_+(t-\delta). \tag{2.7}$$

Then as in 2.1.2 we can use (2.6) to adjust the bounds from (2.7) by doing

$$\varphi_-(t-\delta) \;\leftarrow\; \max\left\{\varphi_-(t-\delta),\, \tau_-^{\mathrm{B}}(t-\delta)\right\} \quad \text{and} \quad \varphi_+(t-\delta) \;\leftarrow\; \min\left\{\varphi_+(t-\delta),\, \tau_+^{\mathrm{B}}(t-\delta)\right\}. \tag{2.8}$$

These corrections are called *backward corrections* , hence the "B" on the bounds $\tau_\pm^{\mathrm{B}}(t-\delta)$. They are illustrated by the following figure:

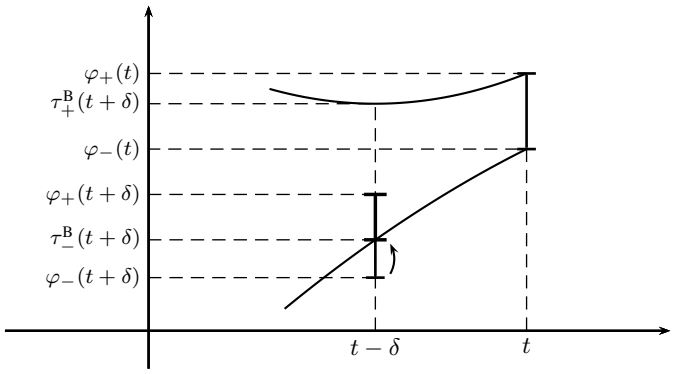

Figure 9 – Example of a backward correction (2.8) on a lower bound of φ in $t - \delta$

2.3 The forward-backward and backward-forward algorithms

2.3.1 — From now on, we will suppose the inequalities (2.1) satisfied for all $t \in \{t_1, \ldots, t_N\}$:

$$\forall\, (i,k) \in \{1,\ldots,N\} \times \{0,\ldots,p-1\}, \quad \varphi_-^{(k)}(t_i) \;\leqslant\; \varphi^{(k)}(t_i) \;\leqslant\; \varphi_+^{(k)}(t_i). \tag{2.9}$$

In this section we present various algorithms that enable corrections on bounds on φ in $t_1, \ldots, t_N$ by using the forward and backward corrections presented in the previous sections. In order to do it, for all $i \in \{1, \ldots, N\}$, we first store the lower and upper bounds on $\varphi, \ldots, \varphi^{(p-1)}$ in t_i in two vectors:

$$\Phi_-^i \;=\; \left(\varphi_-^{(k)}(t_i)\right)_{0\leqslant k\leqslant p-1} \qquad \text{and} \qquad \Phi_+^i \;=\; \left(\varphi_+^{(k)}(t_i)\right)_{0\leqslant k\leqslant p-1}.$$

Let us remark that by doing this, for all $(i,k) \in \{1,\ldots,N\} \times \{0,\ldots,p-1\}$, we have

$$\varphi_\pm^{(k)}(t_i) \;=\; \Phi_\pm^i(k+1).$$

2.3.2 — We first develop an algorithm that uses the forward corrections (2.5) to adapt the bounds on φ in $t_1, \ldots, t_N$. As explained in paragraph 2.1.1 the accuracy of bounds $\tau_\pm^{\mathrm{F}}(t+\delta)$ deduced from the Taylor-Lagrange formula is higher when δ is close to 0. Thus for i going from 2 to N, it seems natural to perform the forward corrections (2.5) with $t = t_i$ and $\delta = t_i - t_{i-1}$. It leads to the *forward algorithm* :

Forward

INPUT: $p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$.

FOR ALL i from 2 to N:

$$\Phi_-^i(1) \leftarrow \max\left\{\Phi_-^i(1),\ \left(\sum_{k=0}^{p-1}\frac{(t_i-t_{i-1})^k}{k!}\,\Phi_-^{i-1}(k+1)\right) + \frac{(t_i-t_{i-1})^p}{p!}\,\varphi_-^*\right\}$$

$$\Phi_+^i(1) \leftarrow \min\left\{\Phi_+^i(1),\ \left(\sum_{k=0}^{p-1}\frac{(t_i-t_{i-1})^k}{k!}\,\Phi_+^{i-1}(k+1)\right) + \frac{(t_i-t_{i-1})^p}{p!}\,\varphi_+^*\right\}$$

OUTPUT: $\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N$ with their updated values.

To better understanding, the following figure illustrates the principle of this algorithm:

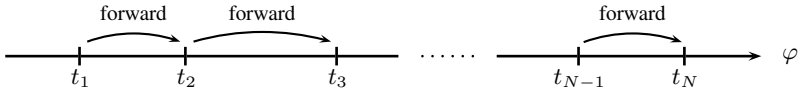

Figure 10 – Illustration of the forward algorithm

By the same way, for i going from $N-1$ to 1, we can perform the backward corrections (2.8) with $t = t_i$ and $\delta = t_{i+1} - t_i$ to adapt the bounds on φ in $t_1, \ldots, t_N$. It leads to the *backward algorithm* :

Backward

INPUT: $p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$.

FOR ALL i from $N-1$ to 1:

$$\Phi_-^i(1) \leftarrow \max\left\{\Phi_-^i(1),\ \left(\sum_{k=0}^{p-1}\frac{(t_{i+1}-t_i)^k}{k!}\,\Phi_{-\sigma\langle k\rangle}^{i-1}(k+1)\right) + \frac{(t_{i+1}-t_i)^p}{p!}\,\varphi_{-\sigma\langle p\rangle}^*\right\}$$

$$\Phi_+^i(1) \leftarrow \min\left\{\Phi_+^i(1),\ \left(\sum_{k=0}^{p-1}\frac{(t_{i+1}-t_i)^k}{k!}\,\Phi_{\sigma\langle k\rangle}^{i-1}(k+1)\right) + \frac{(t_{i+1}-t_i)^p}{p!}\,\varphi_{\sigma\langle p\rangle}^*\right\}$$

OUTPUT: $\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N$ with their updated values.

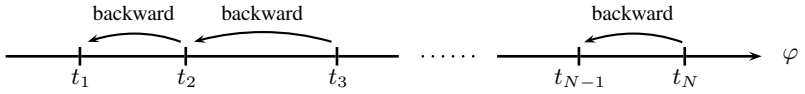

Figure 11 – Illustration of the `Backward` algorithm applied with φ in $t_1, \ldots, t_N$

2.3.3 — We can easily note that the forward (resp. backward) algorithm is *idempotent*, which means that nothing happens if we reapply the forward (resp. backward) algorithm to the output of the first call of the forward (resp. backward) algorithm. Consequently if we want to correct the bounds on φ in $t_1, \ldots, t_N$ by using these algorithms, we suggest to successively apply the forward and backward algorithms until a bound on φ in $t_1, t_2, \ldots,$ or t_N is adjusted. That is what the *forward-backward algorithm* does:

`Forward-backward`

INPUT: $p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$.

$v_-^1 = 0, v_+^1 = 0, \ldots, v_-^N = 0, v_+^N = 0$

WHILE there exists $i \in \{1, \ldots, N\}$ s.t. $\Phi_-^i(1) \neq v_-^i$ or $\Phi_+^i(1) \neq v_+^i$:

FOR ALL i from 1 to N:

$v_-^i = \Phi_-^i(1), v_+^i = \Phi_+^i(1)$

$\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Forward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$

$\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Backward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$

OUTPUT: $\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N$ with their updated values.

In the same way we can implement the *backward-forward algorithm* that has the same structure as the forward-backward algorithm, but in which the backward algorithm is applied before the forward one:

`Backward-forward`

INPUT: $p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$.

$v_-^1 = 0, v_+^1 = 0, \ldots, v_-^N = 0, v_+^N = 0$

WHILE there exists $i \in \{1, \ldots, N\}$ s.t. $\Phi_-^i(1) \neq v_-^i$ or $\Phi_+^i(1) \neq v_+^i$:

FOR ALL i from 1 to N:

$v_-^i = \Phi_-^i(1), v_+^i = \Phi_+^i(1)$

$\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Backward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$

$\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Forward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$

OUTPUT: $\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N$ with their updated values.

2.3.4 — Considering these algorithms, it is legitimate to ask the following questions:

1) Do the forward-backward and backward-forward algorithm converge in a finite number of iterations? And if so, can this number be estimated or determined?

2) Are there cases in which one of these algorithms gives a more accurate output than the other one?

The following section is intended to answer these two questions.

2.4 Analysis of the forward-backward and backward-forward algorithms

2.4.1 — Let us start with some basic observations and by introducing some specific notations.

- The forward and backward algorithms presented in 2.3.2 can only modify $\Phi^i_-(1)$ and $\Phi^i_+(1)$ the values of the lower and upper bounds on φ in t_i, for all $i \in \{1,\dots,N\}$. The other components of Φ^i_- and Φ^i_+ are associated to bounds on $\varphi', \dots, \varphi^{(p-1)}$ and just required to adjust those on φ, which means that for all $k \in \{1,\dots,p-1\}$ we will always have $\Phi^i_-(k+1) = \varphi^{(k)}_-(t_i)$ and $\Phi^i_+(k+1) = \varphi^{(k)}_+(t_i)$.
- It is also useful to remind that the opposite of every lower (resp. upper) bound on φ is an upper (resp. lower) bound on $-\varphi$. Therefore

$$\begin{aligned}&\texttt{Forward}\left(p,t_1,\dots,t_N,\Phi^1_-,\Phi^1_+,\dots,\Phi^N_-,\Phi^N_+,\varphi^*_-,\varphi^*_+\right)\\ &\qquad = -\texttt{Forward}\left(p,t_1,\dots,t_N,-\Phi^1_+,-\Phi^1_-,\dots,-\Phi^N_+,-\Phi^N_-,-\varphi^*_+,-\varphi^*_-\right),\end{aligned}$$

$$\begin{aligned}&\texttt{Backward}\left(p,t_1,\dots,t_N,\Phi^1_-,\Phi^1_+,\dots,\Phi^N_-,\Phi^N_+,\varphi^*_-,\varphi^*_+\right)\\ &\qquad = -\texttt{Backward}\left(p,t_1,\dots,t_N,-\Phi^1_+,-\Phi^1_-,\dots,-\Phi^N_+,-\Phi^N_-,-\varphi^*_+,-\varphi^*_-\right),\end{aligned}$$

  and we can note that it is sufficient to study how the upper bounds on φ are modified by our algorithms.
- For all $i \in \{2,\dots,N\}$ we will denote by δ_i the difference $t_i - t_{i-1}$.
- Finally for all $i \in \{1,\dots,N\}$, since it is sufficient to study how the upper bounds on φ are modified by our algorithm, we will denote by $\varphi^0(t_i)$ the initial value of $\Phi^i_+(1)$, the bound on φ in t_i given by (2.9). Then from paragraph 2.4.3, for all $n \in \mathbb{N}$, we will denote by $\varphi^{\mathrm{F}}_n(t_i)$ (resp. $\varphi^{\mathrm{B}}_n(t_i)$) the upper bound on $\varphi(t_i)$ obtained after the n^{th} forward (resp. backward) call in the forward-backward algorithm, and by $\phi^{\mathrm{B}}_n(t_i)$ (resp. $\phi^{\mathrm{F}}_n(t_i)$) the upper bound on $\varphi(t_i)$ obtained after the n^{th} backward (resp. forward) call in the backward-forward algorithm, with the convention $\varphi^{\mathrm{F}}_0(t_i) = \varphi^{\mathrm{B}}_0(t_i) = \phi^{\mathrm{F}}_0(t_i) = \phi^{\mathrm{B}}_0(t_i) = \varphi^0(t_i)$. Therefore when $n \geqslant 1$,

$$\varphi^{\mathrm{F}}_n(t_1) \;=\; \varphi^{\mathrm{B}}_{n-1}(t_1), \quad \phi^{\mathrm{F}}_n(t_1) = \phi^{\mathrm{B}}_n(t_1), \quad \varphi^{\mathrm{B}}_n(t_N) \;=\; \varphi^{\mathrm{F}}_n(t_N), \quad \phi^{\mathrm{B}}_n(t_N) \;=\; \phi^{\mathrm{F}}_{n-1}(t_N),$$

  if $i \geqslant 2$,

$$\varphi^{\mathrm{F}}_n(t_i) \;=\; \min\left\{\varphi^{\mathrm{B}}_{n-1}(t_i)\,,\; \varphi^{\mathrm{F}}_n(t_{i-1}) + \left(\sum_{k=1}^{p-1}\frac{\delta_i^k}{k!}\,\varphi^{(k)}_+(t_{i-1})\right) + \frac{\delta_i^p}{p!}\,\varphi^*_+\right\},$$

$$\phi^{\mathrm{F}}_n(t_i) \;=\; \min\left\{\phi^{\mathrm{B}}_n(t_i)\,,\; \phi^{\mathrm{F}}_n(t_{i-1}) + \left(\sum_{k=1}^{p-1}\frac{\delta_i^k}{k!}\,\varphi^{(k)}_+(t_{i-1})\right) + \frac{\delta_i^p}{p!}\,\varphi^*_+\right\},$$

  and if $i \leqslant N-1$,

$$\varphi^{\mathrm{B}}_n(t_i) \;=\; \min\left\{\varphi^{\mathrm{F}}_n(t_i)\,,\; \varphi^{\mathrm{B}}_n(t_{i+1}) + \left(\sum_{k=1}^{p-1}\frac{(-1)^k\delta_i^k}{k!}\,\varphi^{(k)}_{\sigma\langle k\rangle}(t_{i+1})\right) + \frac{(-1)^p\delta_i^p}{p!}\,\varphi^*_{\sigma\langle p\rangle}\right\},$$

$$\phi^{\mathrm{B}}_n(t_i) \;=\; \min\left\{\phi^{\mathrm{F}}_{n-1}(t_i)\,,\; \phi^{\mathrm{B}}_n(t_{i+1}) + \left(\sum_{k=1}^{p-1}\frac{(-1)^k\delta_i^k}{k!}\,\varphi^{(k)}_{\sigma\langle k\rangle}(t_{i+1})\right) + \frac{(-1)^p\delta_i^p}{p!}\,\varphi^*_{\sigma\langle p\rangle}\right\}.$$

  In particular, we obviously have

$$\begin{array}{ccccccccccccc}\varphi^0(t_i) & \geqslant & \varphi^{\mathrm{F}}_1(t_i) & \geqslant & \varphi^{\mathrm{B}}_1(t_i) & \geqslant & \varphi^{\mathrm{F}}_2(t_i) & \geqslant & \varphi^{\mathrm{B}}_2(t_i) & \geqslant & \cdots & \geqslant & \varphi(t_i),\\ \varphi^0(t_i) & \geqslant & \phi^{\mathrm{B}}_1(t_i) & \geqslant & \phi^{\mathrm{F}}_1(t_i) & \geqslant & \phi^{\mathrm{B}}_2(t_i) & \geqslant & \phi^{\mathrm{F}}_2(t_i) & \geqslant & \cdots & \geqslant & \varphi(t_i).\end{array}$$

2.4.2 — It is reasonable to believe that the forward-backward and backward-forward algorithms have a relatively similar behaviour. An argument which supports this conjecture is that applying the forward (resp. backward) algorithm (to any input data) is like applying the backward (resp. forward) algorithm (to other input data). To clarify and demonstrate it, we introduce the function

$$\psi \ : \ -I \to \mathbb{R}, \ t \mapsto \varphi(-t)$$

that morally reverses the time when φ is a time-variable function. Clearly, ψ has the same regularity that φ and satisfies, for all $t \in I$ and $k \in \{0, \dots, p\}$,

$$\psi^{(k)}(-t) \ = \ (-1)^k \varphi(t).$$

Therefore for all $(i,k) \in \{1, \dots, N\} \times \{0, \dots, p-1\}$, we have

$$\psi_-^{(k)}(-t_i) \ \leqslant \ \psi^{(k)}(-t_i) \ \leqslant \ \psi_+^{(k)}(-t_i) \quad \text{with} \quad \left\{ \begin{array}{l} \psi_-^{(k)}(-t_i) \ = \ (-1)^k \varphi_{-\sigma\langle k \rangle}^{(k)}(t_i), \\ \psi_+^{(k)}(-t_i) \ = \ (-1)^k \varphi_{\sigma\langle k \rangle}^{(k)}(t_i), \end{array} \right. \tag{2.10}$$

and for all $\xi \in I$, we have

$$\psi_-^* \ \leqslant \ \psi^{(p)}(-\xi) \ \leqslant \ \psi_+^* \quad \text{with} \quad \left\{ \begin{array}{l} \psi_-^* \ = \ (-1)^k \varphi_{-\sigma\langle k \rangle}^*, \\ \psi_+^* \ = \ (-1)^k \varphi_{\sigma\langle k \rangle}^*. \end{array} \right. \tag{2.11}$$

For all $i \in \{1, \dots, N\}$, let us now introduce:

$$\Psi_-^i \ = \ \Big(\psi_-^{(k)}(-t_i) \Big)_{0 \leqslant k \leqslant p-1} \quad \text{and} \quad \Psi_+^i \ = \ \Big(\psi_+^{(k)}(-t_i) \Big)_{0 \leqslant k \leqslant p-1}.$$

For the needs of this paragraph, we will denote here by:

– $\varphi_-^{\mathrm{F}}(t_i)$ (resp. $\varphi_+^{\mathrm{F}}(t_i)$) the value of $\Phi_-^i(1)$ (resp. $\Phi_+^i(1)$), the lower (resp. upper) bound on $\varphi(t_i)$ after the call of the forward algorithm with p, $t_1, \dots, t_N$, $\Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N$ in input,

– $\psi_-^{\mathrm{B}}(t_i)$ (resp. $\psi_+^{\mathrm{B}}(t_i)$) the value of $\Psi_-^i(1)$ (resp. $\Psi_+^i(1)$), the lower (resp. upper) bound on $\psi(-t_i)$ after the call of the backward algorithm with p, $-t_N, \dots, -t_1$, $\Psi_-^N, \Psi_+^N, \dots, \Psi_-^1, \Psi_+^1$ in input,

– $\varphi_+^0(t_i)$ (resp. $\psi_+^0(-t_i)$) the value of $\Phi_+^i(1)$ (resp. $\Psi_+^i(1)$) before the application of the forward (resp. backward) algorithm with the input data mentioned below.

Let us prove that $\psi_+^{\mathrm{B}}(-t_i) = \varphi_+^{\mathrm{F}}(t_i)$ for all $i \in \{1, \dots, N\}$.

- By the structure of the forward and backward algorithms we have respectively $\varphi_+^{\mathrm{F}}(t_1) = \varphi_+^0(t_1)$ and $\psi_+^{\mathrm{B}}(-t_1) = \psi_+^0(-t_1)$, hence $\varphi^{\mathrm{F}}(t_1) = \psi^{\mathrm{B}}(-t_1)$ since $\psi_+^0(-t_1) = \varphi_+^0(t_1)$ according to (2.10).
- If $i \in \{1, \dots, N-1\}$ satisfies $\varphi^{\mathrm{F}}(t_{i-1}) = \psi^{\mathrm{B}}(-t_{i-1})$, since the inequalities (2.10) and (2.11) imply $\varphi^0(t_i) = \psi^0(-t_i)$, $\varphi_+^{(k)}(t_{i-1}) = (-1)^k \psi_{\sigma\langle k \rangle}^{(k)}(t_{i-1})$ for all $k \in \{1, \dots, p\}$, and $\varphi_+^* = (-1)^p \psi_{\sigma\langle p \rangle}^*$, then we get

$$\begin{aligned} \varphi_+^{\mathrm{F}}(t_i) \ &= \ \min \left\{ \varphi_+^0(t_i), \ \varphi_+^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!} \, \varphi_+^{(k)}(t_{i-1}) \right) + \frac{\delta_i^p}{p!} \, \varphi_+^* \right\} \\ &= \ \min \left\{ \psi_+^0(-t_i), \ \psi_+^{\mathrm{F}}(-t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_i^k}{k!} \, \psi_{\sigma\langle k \rangle}^{(k)}(t_{i-1}) \right) + \frac{\delta_i^p}{p!} \, \psi_{\sigma\langle k \rangle}^* \right\} \ = \ \psi_+^{\mathrm{B}}(-t_i). \end{aligned}$$

Thus we have proved by induction that $\psi_+^{\mathrm{B}}(-t_i) = \varphi_+^{\mathrm{F}}(t_i)$ for all $i \in \{1, \dots, N\}$. According to 2.4.1 it implies that $\psi_-^{\mathrm{B}}(-t_i) = \varphi_-^{\mathrm{F}}(t_i)$ for all $i \in \{1, \dots, N\}$. And since an observation done in this same paragraph insures that the other components of vectors Φ_-^1, Φ_+^1, …, Φ_-^N, Φ_+^N (resp. Ψ_-^1, Ψ_+^1, …, Ψ_-^N, Ψ_+^N) are unmodified by the forward (resp. backward) algorithm, we have obtained the point (*i*) of the following proposition, its point (*ii*) being obtained by a similar way.

Proposition. *For all $i \in \{1, \dots, N\}$ let Ψ_-^i and Ψ_+^i be given as previously.*

(*i*) *The modifications made by the forward algorithm that takes p, t_1, …, t_N, Φ_-^1, Φ_+^1, …, Φ_-^N, Φ_+^N in input can be obtained by an application of the backward algorithm with p, $-t_N$, …, $-t_1$, Ψ_-^N, Ψ_+^N, …, Ψ_-^1, Ψ_+^1 in input.*

(*ii*) *The modifications made by the backward algorithm that takes p, t_1, …, t_N, Φ_-^1, Φ_+^1, …, Φ_-^N, Φ_+^N in input can be obtained by an application of the forward algorithm with p, $-t_N$, …, $-t_1$, Ψ_-^N, Ψ_+^N, …, Ψ_-^1, Ψ_+^1 in input.*

In a nutshell, for all $U = \left(U_-^i,\, U_+^i\right)_{1 \leqslant i \leqslant N}$ in $\left(\mathbb{R}^p \times \mathbb{R}^p\right)^N$, by setting $U^{\mathrm{R}} = \left(U_-^{N-i},\, U_+^{N-i}\right)_{1 \leqslant i \leqslant N}$, the previous assertions can be summarized as follows:

$$\begin{aligned}
&\texttt{Forward}\left(p, t_1, \dots, t_N, \Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right) \\
&\qquad = \left[\texttt{Backward}\left(p, -t_N, \dots, -t_1, \Psi_-^N, \Psi_+^N, \dots, \Psi_-^1, \Psi_+^1, \psi_-^*, \psi_+^*\right)\right]^{\mathrm{R}}, \\
&\texttt{Backward}\left(p, t_1, \dots, t_N, \Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right) \\
&\qquad = \left[\texttt{Forward}\left(p, -t_N, \dots, -t_1, \Psi_-^N, \Psi_+^N, \dots, \Psi_-^1, \Psi_+^1, \psi_-^*, \psi_+^*\right)\right]^{\mathrm{R}}.
\end{aligned}$$

2.4.3 — The answers to the questions asked in paragraph 2.3.4 will be based on the following result:

Lemma. *For all $i \in \{2, \dots, N\}$ the quantity*

$$C_i \;=\; \left[\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!} \left(\varphi_+^{(k)}(t_{i-1}) + (-1)^k \varphi_{\sigma\langle k\rangle}^{(k)}(t_i)\right)\right] + \frac{\delta_i^p}{p!} \left(\varphi_+^* + (-1)^p \varphi_{\sigma\langle p\rangle}^*\right)$$

is non-negative.

Proof.— Let us suppose that there exists $i \in \{2, \dots, N\}$ such that $C_i < 0$, and let us fix $n \in \mathbb{N}^*$. As reminded in 2.4.1, we have $\varphi_{n+1}^{\mathrm{F}}(t_{i-1}) \leqslant \varphi_n^{\mathrm{B}}(t_{i-1})$, and so

$$\begin{aligned}
\varphi_{n+1}^{\mathrm{F}}(t_i) \;&\leqslant\; \varphi_{n+1}^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_{i-1})\right) + \frac{\delta_i^p}{p!}\, \varphi_+^* \\
&\leqslant\; \varphi_n^{\mathrm{B}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_{i-1})\right) + \frac{\delta_i^p}{p!}\, \varphi_+^*. \qquad (2.12)
\end{aligned}$$

By the same way, $\varphi_n^{\mathrm{B}}(t_i) \leqslant \varphi_n^{\mathrm{F}}(t_i)$, and so

$$\begin{aligned}
\varphi_n^{\mathrm{B}}(t_{i-1}) \;&\leqslant\; \varphi_n^{\mathrm{B}}(t_i) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_i^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_i)\right) + \frac{(-1)^p \delta_i^p}{p!}\, \varphi_{\sigma\langle p\rangle}^* \\
&\leqslant\; \varphi_n^{\mathrm{F}}(t_i) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_i^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_i)\right) + \frac{(-1)^p \delta_i^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*. \qquad (2.13)
\end{aligned}$$

Thanks to (2.12) and (2.13), we obtain $\varphi_{n+1}^{\mathrm{F}}(t_i) \leqslant \varphi_n^{\mathrm{F}}(t_i) + C_i$, which implies that for all $n \in \mathbb{N}^*$,

$$\varphi(t_i) \;\leqslant\; \varphi_{n+1}^{\mathrm{F}}(t_i) \;\leqslant\; \varphi_1^{\mathrm{F}}(t_i) + nC_i. \tag{2.14}$$

But since C_i is negative, we have $\varphi_1^{\mathrm{F}}(t_i)+nC_i \to -\infty$ when $n \to +\infty$, hence a contradiction! Therefore C_i is necessarily non-negative for all $i \in \{2, \dots, N\}$. □

REMARK. It is important to give a graphical interpretation of the previous lemma. For this purpose, let us fix $i \in \{2, \dots, N\}$ and let us introduce the functions

$$\tau_{i-1}^{\mathrm{F}} \;:\; \mathbb{R} \times \mathbb{R} \to \mathbb{R},\; (\delta, y) \mapsto y + \left(\sum_{k=1}^{p-1} \frac{\delta^k}{k!}\, \varphi_+^{(k)}(t_{i-1}) \right) + \frac{\delta^p}{p!}\, \varphi_+^*,$$

$$\tau_i^{\mathrm{B}} \;:\; \mathbb{R} \times \mathbb{R} \to \mathbb{R},\; (\delta, y) \mapsto y + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta^k}{k!}\, \varphi_{\sigma\langle k \rangle}^{(k)}(t_i) \right) + \frac{(-1)^p \delta^p}{p!}\, \varphi_{\sigma\langle p \rangle}^*.$$

From an algebraic point of view, for all $y \in \mathbb{R}$, we can easily show that

$$C_i \;=\; \tau_i^{\mathrm{B}}\big(\delta_i,\, \tau_{i-1}^{\mathrm{F}}(\delta_i, y)\big) - y.$$

Therefore from a graphical point of view, the sign of C_i determines the relative position between $\mathcal{C}_{i-1}^{\mathrm{F}}(y)$, the graph of $\tau_{i-1}^{\mathrm{F}}(\,\cdot\,, y)$, and $\mathcal{C}_i^{\mathrm{B}}(z)$, the graph of $\tau_i^{\mathrm{B}}(\,\cdot\,, z)$ where $z = \tau_{i-1}^{\mathrm{F}}(\delta_i, y)$.

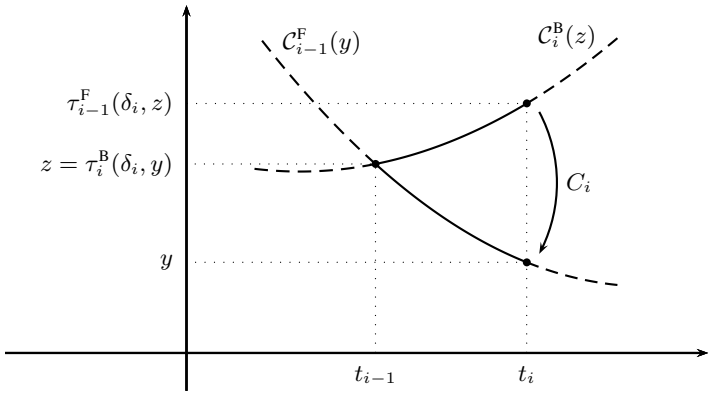

Figure 12 – Graphical representation of the quantity C_i

And since for all $n \in \mathbb{N}^*$, we have

$$\begin{aligned} \varphi_{n+1}^{\mathrm{F}}(t_i) \;&=\; \min\left\{\varphi_n^{\mathrm{B}}(t_i)\,,\; \tau_{i-1}^{\mathrm{F}}\big(\delta_i, \varphi_{n+1}^{\mathrm{F}}(t_{i-1})\big)\right\}, \\ \varphi_n^{\mathrm{B}}(t_{i-1}) \;&=\; \min\left\{\varphi_n^{\mathrm{F}}(t_{i-1})\,,\; \tau_i^{\mathrm{B}}\big(\delta_i, \varphi_n^{\mathrm{B}}(t_i)\big)\right\}, \end{aligned}$$

with

$$\varphi_{n+1}^{\mathrm{F}}(t_{i-1}) \;\leqslant\; \varphi_n^{\mathrm{B}}(t_{i-1}) \qquad \text{and} \qquad \varphi_n^{\mathrm{B}}(t_i) \;\leqslant\; \varphi_n^{\mathrm{B}}(t_i)\,,$$

we can note that quantity C_i must be non-negative, otherwise a graphical analyse would lead step by step to the inequality (2.14) obtained in the proof of the lemma. Figure 13 illustrates this phenomenon.

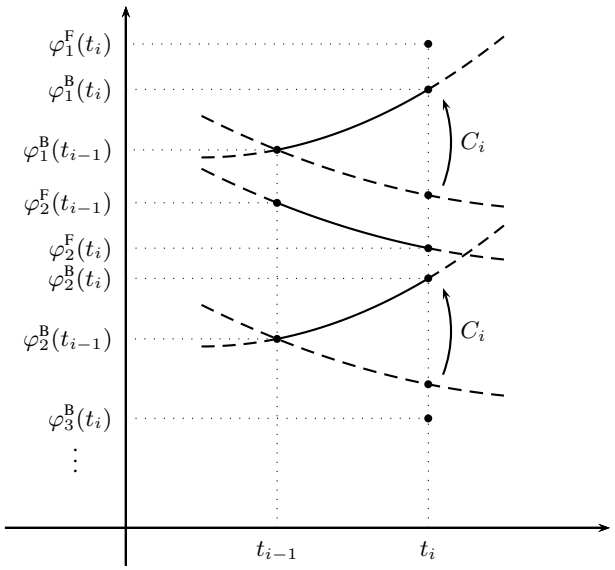

Figure 13 – Graphical illustration of the contradiction obtained in the proof of the lemma when $C_i < 0$

2.4.4 — Theorem. *For all $i \in \{1, \dots, N\}$ we have:*

(*i*) $\varphi_2^{\mathrm{B}}(t_i) = \varphi_2^{\mathrm{F}}(t_i) = \varphi_1^{\mathrm{B}}(t_i)$,

(*ii*) $\phi_2^{\mathrm{F}}(t_i) = \phi_2^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$.

Proof.— According to the proposition 2.4.2, it is sufficient to establish the point (*i*).

1) We start by proving that for all $i \in \{1, \dots, N\}$,

$$\varphi_2^{\mathrm{F}}(t_i) \;=\; \varphi_1^{\mathrm{B}}(t_i)\,. \tag{2.15}$$

By contradiction, let us suppose that there exists $i \in \{1, \dots, N\}$ such that $\varphi_2^{\mathrm{F}}(t_i) < \varphi_1^{\mathrm{B}}(t_i)$, and let us define:

$$j \;=\; \min\left\{ i \in \{1, \dots, N\} \;\middle|\; \varphi_2^{\mathrm{F}}(t_i) < \varphi_1^{\mathrm{B}}(t_i) \right\}.$$

According to the structure of the forward algorithm, we have $j \geqslant 2$, and by definition of j, we have $\varphi_2^{\mathrm{F}}(t_{j-1}) = \varphi_1^{\mathrm{B}}(t_{j-1})$ and $\varphi_2^{\mathrm{F}}(t_j) < \varphi_1^{\mathrm{B}}(t_j)$. Thus

$$\begin{aligned} \varphi_2^{\mathrm{F}}(t_j) \;&=\; \varphi_2^{\mathrm{F}}(t_{j-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_j^k}{k!}\, \varphi_+^{(k)}(t_{j-1}) \right) + \frac{\delta_j^p}{p!}\, \varphi_+^* \\ &=\; \varphi_1^{\mathrm{B}}(t_{j-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_j^k}{k!}\, \varphi_+^{(k)}(t_{j-1}) \right) + \frac{\delta_j^p}{p!}\, \varphi_+^*. \end{aligned} \tag{2.16}$$

In particular if $\varphi_1^{\mathrm{B}}(t_{j-1}) = \varphi_1^{\mathrm{F}}(t_{j-1})$, then

$$\varphi_1^{\mathrm{F}}(t_j) \;\leqslant\; \varphi_1^{\mathrm{F}}(t_{j-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_j^k}{k!}\, \varphi_+^{(k)}(t_{j-1}) \right) + \frac{\delta_j^p}{p!}\, \varphi_+^* \;=\; \varphi_2^{\mathrm{F}}(t_j)\,,$$

and we end up with $\varphi_2^{\mathrm{F}}(t_j) < \varphi_1^{\mathrm{B}}(t_j) \leqslant \varphi_1^{\mathrm{F}}(t_j) \leqslant \varphi_2^{\mathrm{F}}(t_j)$, which is obviously impossible. Therefore $\varphi_1^{\mathrm{B}}(t_{j-1}) < \varphi_1^{\mathrm{F}}(t_{j-1})$, and

$$\varphi_1^{\mathrm{B}}(t_{j-1}) \;=\; \varphi_1^{\mathrm{B}}(t_j) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_j^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_{j-1}) \right) + \frac{\delta_j^p}{p!}\, \varphi_{\sigma\langle p\rangle}^{*}. \tag{2.17}$$

With the notations from the lemma 2.4.3, the relations (2.16) and (2.17) finally imply

$$C_j \;=\; \varphi_2^{\mathrm{F}}(t_j) - \varphi_1^{\mathrm{B}}(t_j) \;<\; 0,$$

which is in contradiction with this same lemma. Consequently we have proved that the equality (2.15) is available for all $i \in \{1, \ldots, N\}$.

2) Now let us prove that for all $i \in \{1, \ldots, N\}$,

$$\varphi_2^{\mathrm{B}}(t_i) \;=\; \varphi_2^{\mathrm{F}}(t_i)\,. \tag{2.18}$$

By the structure of the backward algorithm we have $\varphi_2^{\mathrm{B}}(t_N) = \varphi_2^{\mathrm{F}}(t_N)$. Now let $i \in \{2, \ldots, N\}$ satisfying (2.18), that implies $\varphi_2^{\mathrm{B}}(t_i) = \varphi_1^{\mathrm{B}}(t_i)$ according to (2.15). Then

$$\varphi_2^{\mathrm{B}}(t_{i-1}) \;=\; \min \left\{ \varphi_2^{\mathrm{F}}(t_{i-1})\,,\; \varphi_1^{\mathrm{B}}(t_i) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_i^k}{k!} \varphi_{\sigma\langle k\rangle}^{(k)}(t_i) \right) + \frac{(-1)^p \delta_i^p}{p!} \varphi_{\sigma\langle p\rangle}^{*} \right\}.$$

Since (2.15) also implies $\varphi_2^{\mathrm{F}}(t_{i-1}) = \varphi_1^{\mathrm{B}}(t_{i-1})$, we finally obtain:

$$\varphi_2^{\mathrm{F}}(t_{i-1}) \;=\; \varphi_1^{\mathrm{B}}(t_{i-1}) \;\leqslant\; \varphi_1^{\mathrm{B}}(t_i) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_i^k}{k!} \varphi_{\sigma\langle k\rangle}^{(k)}(t_i) \right) + \frac{(-1)^p \delta_i^p}{p!} \varphi_{\sigma\langle p\rangle}^{*},$$

which shows that we also have $\varphi_2^{\mathrm{B}}(t_{i-1}) = \varphi_2^{\mathrm{F}}(t_{i-1})$. In the end we have proved by induction the availability of equality (2.18) for all $i \in \{1, \ldots, N\}$. □

REMARK. Let us signal that the result established in the step 2) of this proof can be immediately obtained by using the one from step 1), since by starting from the output of the first forward call and by applying the forward-backward algorithm on $\psi : t \mapsto \varphi(-t)$, proposition 2.4.2 states that the second forward on ψ corresponds to the second backward on φ.

2.4.5 — According to the theorem 2.4.4, we are now able to answer the question 1) from 2.3.4: the forward-backward and backward-forward algorithms precisely converge in *one iteration*. In other words, it means that these algorithms can be rewritten:

`Forward-backward`
INPUT: $p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$. $\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Forward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$ $\left(\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Backward` $\left(p, t_1, \ldots, t_N, \Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$ OUTPUT: $\Phi_-^1, \Phi_+^1, \ldots, \Phi_-^N, \Phi_+^N$ with their updated values.

Backward-forward
INPUT: $p, t_1, \dots, t_N, \Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*$. $\left(\Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Backward` $\left(p, t_1, \dots, t_N, \Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$ $\left(\Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N\right) \leftarrow$ `Forward` $\left(p, t_1, \dots, t_N, \Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N, \varphi_-^*, \varphi_+^*\right)$ OUTPUT: $\Phi_-^1, \Phi_+^1, \dots, \Phi_-^N, \Phi_+^N$ with their updated values.

2.4.6 — Let us now focus on the question 2) from 2.3.4. To answer it, we need to prove some other preliminary results. Here is the first one:

Proposition. *For all $i \in \{1, \dots, N\}$ we have $\phi_1^{\mathrm{F}}(t_i) \leqslant \varphi_1^{\mathrm{F}}(t_i)$ and $\varphi_1^{\mathrm{B}}(t_i) \leqslant \phi_1^{\mathrm{F}}(t_i)$.*

Proof.— Thanks to the proposition 2.4.2 it is sufficient to prove that $\phi_1^{\mathrm{F}}(t_i) \leqslant \varphi_1^{\mathrm{F}}(t_i)$ for all $i \in \{1, \dots, N\}$.

- Giving the structure of the forward-backward algorithm we have $\phi_1^{\mathrm{F}}(t_1) \leqslant \varphi^0(t_1) = \varphi_1^{\mathrm{F}}(t_1)$.
- Let $i \in \{2, \dots, N\}$ such that $\phi_1^{\mathrm{F}}(t_{i-1}) \leqslant \varphi_1^{\mathrm{F}}(t_{i-1})$. As a reminder,

$$\varphi_1^{\mathrm{F}}(t_i) \;=\; \min\left\{\varphi^0(t_i),\; \varphi_1^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_i)\right) + \frac{\delta_i^p}{p!}\, \varphi_+^*\right\}.$$

If $\varphi_1^{\mathrm{F}}(t_i) = \varphi^0(t_i)$, then we immediately have $\phi_1^{\mathrm{F}}(t_i) \leqslant \varphi^0(t_i) = \varphi_1^{\mathrm{F}}(t_i)$. So let us suppose that

$$\varphi_1^{\mathrm{F}}(t_i) \;=\; \varphi_1^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_i)\right) + \frac{\delta_i^p}{p!}\, \varphi_+^*.$$

In this case, our initial hypothesis $\phi_1^{\mathrm{F}}(t_{i-1}) \leqslant \varphi_1^{\mathrm{F}}(t_{i-1})$ implies

$$\begin{aligned}\varphi_1^{\mathrm{F}}(t_i) \;&\geqslant\; \phi_1^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_i)\right) + \frac{\delta_i^p}{p!}\, \varphi_+^* \\ &\geqslant\; \min\left\{\phi_1^{\mathrm{B}}(t_i)\,,\; \phi_1^{\mathrm{F}}(t_{i-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_i^k}{k!}\, \varphi_+^{(k)}(t_i)\right) + \frac{\delta_i^p}{p!}\, \varphi_+^*\right\} \;=\; \phi_1^{\mathrm{F}}(t_i)\,.\end{aligned}$$

To sum up, we have proved the expected result by induction. □

2.4.7 — Lemma. *For all index $j \in \{2, \dots, N\}$ such that $\varphi_1^{\mathrm{B}}(t_{j-1}) < \varphi_1^{\mathrm{F}}(t_{j-1})$ and $\varphi_1^{\mathrm{B}}(t_j) = \varphi_1^{\mathrm{F}}(t_j)$, we have $\varphi_1^{\mathrm{F}}(t_j) = \varphi^0(t_j)$.*

Proof.— Let j be an index from $\{2, \dots, N\}$ such that $\varphi_1^{\mathrm{B}}(t_{j-1}) < \varphi_1^{\mathrm{F}}(t_{j-1})$ and $\varphi_1^{\mathrm{B}}(t_j) = \varphi_1^{\mathrm{F}}(t_j)$. Thus

$$\begin{aligned}\varphi_1^{\mathrm{B}}(t_{j-1}) \;&=\; \varphi_1^{\mathrm{B}}(t_j) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_j^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_j)\right) + \frac{\delta_j^p}{p!}\, \varphi_{\sigma\langle p\rangle}^* \\ &=\; \varphi_1^{\mathrm{F}}(t_j) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_j^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_j)\right) + \frac{\delta_j^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*.\end{aligned}$$

Let us suppose that $\varphi_1^{\mathrm{F}}(t_j) < \varphi^0(t_j)$. Then

$$\varphi_1^{\mathrm{F}}(t_j) \;=\; \varphi_1^{\mathrm{F}}(t_{j-1}) + \left(\sum_{k=1}^{p-1} \frac{\delta_j^k}{k!}\, \varphi_+^{(k)}(t_j)\right) + \frac{\delta_j^p}{p!}\, \varphi_+^*.$$

Finally, with the notation from the lemma 2.4.3, we obtain

$$C_j \;=\; \varphi_1^{\mathrm{B}}(t_j) - \varphi_1^{\mathrm{F}}(t_j) \;<\; 0,$$

which is in contradiction with this same result. Therefore $\varphi_1^{\mathrm{F}}(t_j) = \varphi^0(t_j)$. □

2.4.8 — Proposition. *Let $i \in \{1, \ldots, N-1\}$.*

(*i*) *For all $j \in \{i+1, \ldots, N\}$, we have*

$$\varphi_1^{\mathrm{B}}(t_i) \;\leqslant\; \varphi_1^{\mathrm{B}}(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l)\right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*\right], \tag{2.19}$$

$$\phi_1^{\mathrm{B}}(t_i) \;\leqslant\; \phi_1^{\mathrm{B}}(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l)\right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*\right]. \tag{2.20}$$

(*ii*) *We assume that $\varphi_1^{\mathrm{B}}(t_i) < \varphi_1^{\mathrm{F}}(t_i)$. Let Γ_i be the set of indexes $l \in \{i+1, \ldots, N\}$ such that $\varphi_1^{\mathrm{B}}(t_l) = \varphi_1^{\mathrm{F}}(t_l)$. Then $\Gamma_i \neq \varnothing$, and for $j = \min \Gamma_i$, we have*

$$\varphi_1^{\mathrm{B}}(t_i) \;=\; \varphi^0(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l)\right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*\right]. \tag{2.21}$$

Proof.— (*i*) We will only prove the availability of (2.19), (2.20) being obtained similarly. We proceed by induction on $j \in \{i+1, \ldots, N\}$, case $j = i+1$ being obviously true by the definition of $\varphi_1^{\mathrm{B}}(t_i)$. So let $j \in \{i+1, \ldots, N-1\}$ be an index such that

$$\varphi_1^{\mathrm{B}}(t_i) \;\leqslant\; \varphi_1^{\mathrm{B}}(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l)\right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*\right].$$

Then by the definition of $\varphi_1^{\mathrm{B}}(t_j)$, we have

$$\varphi_1^{\mathrm{B}}(t_j) \;\leqslant\; \varphi_1^{\mathrm{B}}(t_{j+1}) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_{j+1}^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_{j+1})\right) + \frac{(-1)^p \delta_{j+1}^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*,$$

which immediately implies

$$\varphi_1^{\mathrm{B}}(t_i) \;\leqslant\; \varphi_1^{\mathrm{B}}(t_{j+1}) + \sum_{l=i+1}^{j+1} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l)\right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*\right].$$

(*ii*) By the structure of the backward algorithm, we have $N \in \Gamma_i$. Therefore $\Gamma_i \neq \varnothing$, and $j = \min \Gamma_i$ is well-defined. By the definition of j, we have $\varphi_1^{\mathrm{B}}(t_l) < \varphi_1^{\mathrm{F}}(t_l)$ for all $l \in \{i, \ldots, j-1\}$, hence

$$\varphi_1^{\mathrm{B}}(t_l) \;=\; \varphi_1^{\mathrm{B}}(t_{l+1}) + \left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_{l+1}^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_{l+1})\right) + \frac{(-1)^p \delta_{l+1}^p}{p!}\, \varphi_{\sigma\langle p\rangle}^*.$$

Thus

$$\varphi_1^{\mathrm{B}}(t_i) \;=\; \varphi_1^{\mathrm{B}}(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l) \right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^{*} \right]. \tag{2.22}$$

And since we also have $\varphi_1^{\mathrm{B}}(t_{j-1}) < \varphi_1^{\mathrm{F}}(t_{j-1})$ and $\varphi_1^{\mathrm{B}}(t_j) = \varphi_1^{\mathrm{F}}(t_j)$, then the lemma 2.4.7 implies that $\varphi_1^{\mathrm{F}}(t_j) = \varphi^0(t_j)$. Consequently, we have $\varphi_1^{\mathrm{B}}(t_j) = \varphi^0(t_j)$, and so the equality (2.21) can be deduced from (2.22). □

2.4.9 — Theorem. *For all $i \in \{1, \ldots, N\}$, we have $\varphi_1^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$.*

Proof.— Let $i \in \{1, \ldots, N\}$.

- Let us suppose that $\varphi_1^{\mathrm{B}}(t_i) < \varphi_1^{\mathrm{F}}(t_i)$. Regarding the structure of the backward algorithm, it necessarily implies $i \in \{1, \ldots, N-1\}$, and so we can introduce j as at the point (*ii*) of the proposition 2.4.8, which satisfies

$$\varphi_1^{\mathrm{B}}(t_i) \;=\; \varphi^0(t_j) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l) \right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^{*} \right].$$

  Now since $\phi_1^{\mathrm{B}}(t_j) \leqslant \varphi^0(t_j)$, the point (*i*) of the proposition 2.4.8 implies

$$\phi_1^{\mathrm{B}}(t_i) \;\leqslant\; \varphi^0(t_i) + \sum_{l=i+1}^{j} \left[\left(\sum_{k=1}^{p-1} \frac{(-1)^k \delta_l^k}{k!}\, \varphi_{\sigma\langle k\rangle}^{(k)}(t_l) \right) + \frac{(-1)^p \delta_l^p}{p!}\, \varphi_{\sigma\langle p\rangle}^{*} \right] \;=\; \varphi_1^{\mathrm{B}}(t_i)\,.$$

  By the proposition 2.4.6, since we also have $\varphi_1^{\mathrm{B}}(t_i) \leqslant \phi_1^{\mathrm{B}}(t_i)$, we can conclude that $\varphi_1^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$.

- If $\phi_1^{\mathrm{F}}(t_i) \leqslant \phi_1^{\mathrm{B}}(t_i)$, then by the proposition 2.4.2, we can go back to the previous case in order to conclude that $\varphi_1^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$.

- Let us finally suppose that $\varphi_1^{\mathrm{B}}(t_i) = \varphi_1^{\mathrm{F}}(t_i)$ and $\phi_1^{\mathrm{F}}(t_i) = \phi_1^{\mathrm{B}}(t_i)$. Then proposition 2.4.6 leads to

$$\phi_1^{\mathrm{F}}(t_i) \;\leqslant\; \varphi_1^{\mathrm{F}}(t_i) \;=\; \varphi_1^{\mathrm{B}}(t_i) \;\leqslant\; \phi_1^{\mathrm{B}}(t_i) \;=\; \phi_1^{\mathrm{F}}(t_i)\,,$$

  hence $\varphi_1^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$. □

2.4.10 — Using the previous results we can now provide the answer to the question 2) from 2.3.4, which is given by the following result:

Corollary. *The forward-backward and backward-forward algorithms are equivalent, that means, given the same input arguments, their output arguments are precisely the same.*

Proof.— According to 2.4.1 and their rewritings proposed in paragraph 2.4.5, the forward-backward and backward-forward algorithms are equivalent if and only if $\varphi_1^{\mathrm{B}}(t_i) = \phi_1^{\mathrm{F}}(t_i)$ for all $i \in \{1, \ldots, N\}$. Therefore we can conclude by applying theorem 2.4.9. □

2.4.11 — In a nutshell, to correct bounds on φ at t_1, …, t_N from (2.9), we can indifferently use the forward-backward or backward-forward algorithm in their final forms from 2.4.5, instead of their initial versions presented in paragraph 2.3.3. It is also important to note that we can explicitly determine the number of operations done by these rewritten algorithms (which is the same for both of them thanks to theorem 2.4.9) since it only depends on p, the derivative order of function φ, and on N, the number of data.

2.5 A forward-backward multi-order algorithm

2.5.1 — In the previous sections, we have presented and studied algorithms that perform forward and backward corrections to adjust the bounds on φ at $t_1, \ldots, t_n$ from (2.9). We have more precisely proved in section 2.4 that the forward-backward and backward-forward algorithms from the paragraph 2.4.5 enable these in a totally equivalent way. Now let us remark the two following facts:

– For all $k \in \{1, \ldots, p-1\}$ we can also use the forward-backward or backard-forward algorithms to correct bounds on $\varphi^{(k)}$ from (2.9) by replacing p by $p-k$, φ by $\varphi^{(k)}$ and, for all $i \in \{1, \ldots, N\}$, vectors Φ^i_- and Φ^i_+ by $\Phi^i_-[k+1, p]$ and $\Phi^i_+[k+1, p]$ in the previous paragraphs.

– For all $k \in \{0, \ldots, p-1\}$ the higher the accuracy on bounds on $\varphi^{(k+1)}$, …, $\varphi^{(p-1)}$ the higher the accuracy on bounds on $\varphi^{(k)}$ from the forward-backward or backward-forward algorithms will be.

Thus we can implement the *forward-backward multi-order* algorithm that establishes coherence between the various bounds on φ and its derivatives from (2.9) as follows:

`Forward-backward multi-order`

INPUT: $p, t_1, \ldots, t_N, \Phi^1_-, \Phi^1_+, \ldots, \Phi^N_-, \Phi^N_+, \varphi^*_-, \varphi^*_+$.

FOR ALL k from $p-1$ to 0:

$(\Phi^1_\pm[k{+}1\!:\!p], \ldots, \Phi^N_\pm[k{+}1\!:\!p]) \leftarrow$ `Forward-backward` $(p-k, t_1, \ldots, t_N, \Phi^1_\pm[k{+}1\!:\!p], \ldots, \Phi^N_\pm[k{+}1\!:\!p], \varphi^*_\pm)$

or equivalently

$(\Phi^1_\pm[k{+}1\!:\!p], \ldots, \Phi^N_\pm[k{+}1\!:\!p]) \leftarrow$ `Backward-forward` $(p-k, t_1, \ldots, t_N, \Phi^1_\pm[k{+}1\!:\!p], \ldots, \Phi^N_\pm[k{+}1\!:\!p], \varphi^*_\pm)$

OUTPUT: $\Phi^1_-, \Phi^1_+, \ldots, \Phi^N_-, \Phi^N_+$ with their updated values.

REMARK. In this algorithm we have written $\Phi^1_\pm[k{+}1\!:\!p]$ instead of $\Phi^1_-[k{+}1\!:\!p], \Phi^1_+[k{+}1\!:\!p]$, $\Phi^N_\pm[k{+}1\!:\!p]$ instead of $\Phi^N_-[k{+}1\!:\!p], \Phi^N_+[k{+}1\!:\!p]$, and $\varphi^*_\pm$ instead of φ^*_-, φ^*_+ in order to save space.

2.5.2 — The equivalence between the forward-backward and backward-forward algorithms can be used to parallelize some computations of the forward-backward multi-order algorithm. To do this it is sufficient to use the forward-backward algorithm in the for-loop when k is an even number and the backward-forward algorithm when k is an odd number. For instance when φ is three times differentiable, *i.e.* when $p = 3$, this morally consists in performing successively:

1) forward on $\varphi''(t_2), \ldots, \varphi''(t_N)$,

2) backward on $\varphi''(t_{N-1}), \ldots, \varphi''(t_1)$,

3) backward on $\varphi'(t_{N-1}), \ldots, \varphi'(t_1)$,

4) forward on $\varphi'(t_2), \ldots, \varphi'(t_N)$,

5) forward on $\varphi(t_2), \ldots, \varphi(t_N)$,

6) backward on $\varphi(t_{N-1}), \ldots, \varphi(t_1)$.

Now let us highlight that for all i going from $N-1$ to 1:

– As soon as bounds on $\varphi''(t_i)$ are adjusted by the backward algorithm in step 2), they will no longer be modified by the forward-backward multi-order algorithm.

– Adjustments on $\varphi'(t_i)$ done by the backward algorithm from step 3) only require bounds on $\varphi''(t_{i+1})$.

Therefore, computations from steps 2) and 3) can be parallelized by performing in the same time, for i going from $N-1$ to 1, the backward corrections on $\varphi''(t_i)$ and $\varphi'(t_i)$. In the same way computations from steps 4) and 5) can be parallelized by performing at the same time, for i going from 2 to N, the forward corrections on $\varphi'(t_i)$ and $\varphi(t_i)$.

2.5.3 — To finish this section, let us indicate how to bound φ, ..., $\varphi^{(p-1)}$ on overall I just by using their bounds at $t_1, \ldots, t_N$ given by the forward-backward multi-order algorithm. In which follows, each bound obtained in output of this algorithm will be called a *corrected bound*.
Here again, it is sufficient to detail for the function φ. So let us fix $k \in \{0, \ldots, p-1\}$ and $t \in I$.

- If $t < t_1$, it is sufficient to take bounds deduced from (2.8) and the corrected bounds on φ in $t = t_1$.
- If $t > t_N$, it is sufficient to take bounds deduced from (2.5) and the corrected bounds on φ in $t = t_N$.
- Let us suppose that there exists $i \in \{2, \ldots, N-1\}$ such that $t \in]t_i, t_{i+1}[$. We get a lower bound $\tau_-^{\mathrm{F}}(t)$ and an upper bound $\tau_+^{\mathrm{F}}(t)$ on $\varphi(t)$ by using (2.5) and the corrected bounds on φ in $t = t_i$, and a lower bound $\tau_-^{\mathrm{B}}(t)$ and an upper bound $\tau_+^{\mathrm{B}}(t)$ on $\varphi(t)$ by using (2.8) and the corrected bounds on φ in $t = t_{i+1}$. Therefore by using the selection principle we get $\varphi_-(t) \leqslant \varphi(t) \leqslant \varphi_+(t)$ with $\varphi_-(t) = \max\left\{\tau_-^{\mathrm{F}}(t), \tau_-^{\mathrm{B}}(t)\right\}$ and $\varphi_+(t) = \min\left\{\tau_+^{\mathrm{F}}(t), \tau_+^{\mathrm{B}}(t)\right\}$.

2.6 A basic example

2.6.1 — In this section we present some numerical simulations that illustrate the general performances of the forward-backward multi-order algorithm from paragraph 2.5.1 on a simple example. We suppose here that $I = \mathbb{R}$ and that function φ is given by:

$$\varphi \ : \ I \to \mathbb{R}, \ t \mapsto \cos(t) - t^2.$$

Function φ is clearly two times differentiable on $\mathbb{R}$ with $\varphi''(t) = -\cos(t) - 2$ for all $t \in \mathbb{R}$. Therefore the inequalities from (2.2) are satisfied with $p = 2$, $\varphi_-^* = -3$ and $\varphi_+^* = -1$.
We then suppose that $N = 41$ and, for all $i \in \{1, \ldots, N\}$,

$$t_i \ = \ \frac{i - 21}{40}.$$

In doing so,

$$t_1 = -2 \ < \ t_2 = -1.9 \ < \ t_3 = -1.8 \ < \ \cdots \ < \ t_{N-1} = 1.9 \ < \ t_N = 2.$$

We finally generate bounds $\varphi_-(t_i)$ and $\varphi_+(t_i)$ at $\varphi(t_i)$ and $\varphi'_-(t_i)$ and $\varphi'_+(t_i)$ at $\varphi'(t_i)$ from (2.9) by defining, for all $i \in \{1, \ldots, N\}$,

$$\varphi_-(t_i) \ = \ \varphi(t_i) - U_i \quad \text{and} \quad \varphi_+(t_i) \ = \ \varphi(t_i) + (1 - U_i),$$

$$\varphi'_-(t_i) \ = \ \varphi'(t_i) - U'_i \quad \text{and} \quad \varphi'_+(t_i) \ = \ \varphi'(t_i) + (1 - U'_i),$$

where U_i and U'_i are two uniform continuous distributions on $[0, 1]$ depending on index i.

2.6.2 — We have done ten simulations on data generated as in 2.6.1. For each one of them we have indicated in table 2 ρ (resp. ρ') corresponding to the average percentage of reduction of the distance between the bounds on φ (resp. φ') given in input of the forward-backward multi-order algorithm and those obtained in output of this algorithm. We have also indicate τ the computation time (expressed in milliseconds) required for each one of these simulations.

	ρ	ρ'	τ
1^{st}	69.1%	56.8%	$1.65\,\mathrm{ms}$
2^{nd}	66.6%	66.8%	$2.14\,\mathrm{ms}$
3^{rd}	80.0%	64.2%	$1.88\,\mathrm{ms}$
4^{th}	71.6%	68.8%	$2.49\,\mathrm{ms}$
5^{th}	72.2%	55.0%	$2.22\,\mathrm{ms}$
6^{th}	74.6%	60.8%	$1.66\,\mathrm{ms}$
7^{th}	78.4%	55.9%	$2.26\,\mathrm{ms}$
8^{th}	76.9%	64.0%	$2.09\,\mathrm{ms}$
9^{th}	73.4%	63.6%	$3.11\,\mathrm{ms}$
10^{th}	76.7%	55.0%	$1.66\,\mathrm{ms}$

Table 2 – Performances of the forward-backward multi-order algorithm on ten simulations

We can observe that our algorithm significantly improve the accuracy of bounds on φ and φ', and that its computations are executed in a very short time. So even if the hazard on our data would be in practice less important than here, these encouraging results highlight the efficiency of the forward and backward corrections presented in this chapter.

REMARK. Even if it is not mentioned, by the way of these simulations we can check:

1) the validity of statements made in paragraph 2.4.5, *i.e.* that the forward-backward (resp. backward-forward) algorithms from 2.3.3 and 2.4.5 are indeed equivalent. In doing so we can also observe that their rewritings are (obviously) significantly faster their initial versions.

2) the veracity of corollary 2.4.10, which says that the forward-backward and backward-forward algorithms are equivalent.

Conclusion

In this chapter we have presented various models that establish coherence between bounds on a function and its derivative, and we have suggested and implemented an efficient and optimal algorithm which enables this. According to the study of chapter 1 it means that we are now able to solve the problem presented in introduction. We will do this on a concrete example (and more sophisticated and realistic that the one presented in section 2.6) in the following chapter.

Chapter 3

Numerical simulations

Introduction

In this section, we present various simulations that illustrate possible uses of the models exposed in chapters 1 and 2. Since our study was in particular done for the needs of the RATP, these simulations are inspired by situations taken from the Parisian underground, even if they are totally fictional: how to estimate the position, speed and acceleration of a train online?
In section 3.1, we specify the various hypotheses done for our simulations and we precise the problem that will be considered. In section 3.2, we suggest a treatment that will be used to solve this problem. We will perform in section 3.3 a preliminary simulation to show how to initialize our treatment and analyse these first results. Then we will suggest some improvements or adaptations that will be exposed and commented in section 3.4. Finally in section 3.5 will be presented an adaptation of our treatment usable to lead *a posteriori* analyses. In doing so, we will be able to compare in detail the benefits and disadvantages of the use of centred and decentred bounds from chapter 1.

3.1 Hypothesis and objectives

3.1.1 — *Hypothesis.*
In this section:

- The distances will be expressed in meters (m) and the times in seconds (s). Sometimes speeds will be expressed in kilometres by hour $(\mathrm{km/h})$ for the understanding, as they often are in the daily life.
- $f : I \to \mathbb{R}$ will represent the distance crossed by a train, which will correspond to its position since the time t_1 supposed to be 0. By doing a simplification of the physical models (e.g. train not deformable), it can be assimilated to the distance crossed by a punctual mobile (for instance corresponding to the front of the train) on a line. We will suppose that $I = \mathbb{R}_+$ and that f is three times differentiable. Therefore as explained in introduction, if $t \in \mathbb{R}_+$, $f(t)$ corresponds to the *position* of the train at the time t, $f'(t)$ its *speed*, $f''(t)$ its *acceleration* and $f'''(t)$ a quantity called its *jerk*.
- We will generate 1800 data. Thus we set $N = 1800$ and, for all $i \in \{2, \dots, N\}$, we define

$$t_i \;=\; (i-1)\cdot 0.05 + \lambda_i \cdot 0.01,$$

with λ_i a continuous uniform distribution on $[-1, 1]$ (depending on index i). As previously,

$$t_1 \;<\; t_2 \;<\; \cdots \;<\; t_N,$$

and times $t_1, t_2, \ldots, t_N$ are normally not uniformly distributed.

- We suppose the railway road divided in sections of $0.1\,\mathrm{m}$ and that on each time t_i we are able to determine in which section the (front of the) train is. Therefore, for all $i \in \{0, \ldots, N\}$, there exists an unique $k_i \in \mathbb{N}$ such that $k_i \cdot 0.1 \leqslant f(t_i) \leqslant (k_i + 1) \cdot 0.1$. Thus assumption ($A_1$) from the introduction of this part I is obtained by setting:

$$f_-(t_i) \;=\; k_i \cdot 0.1 \qquad \text{and} \qquad f_+(t_i) \;=\; (k_i + 1) \cdot 0.1. \tag{3.1}$$

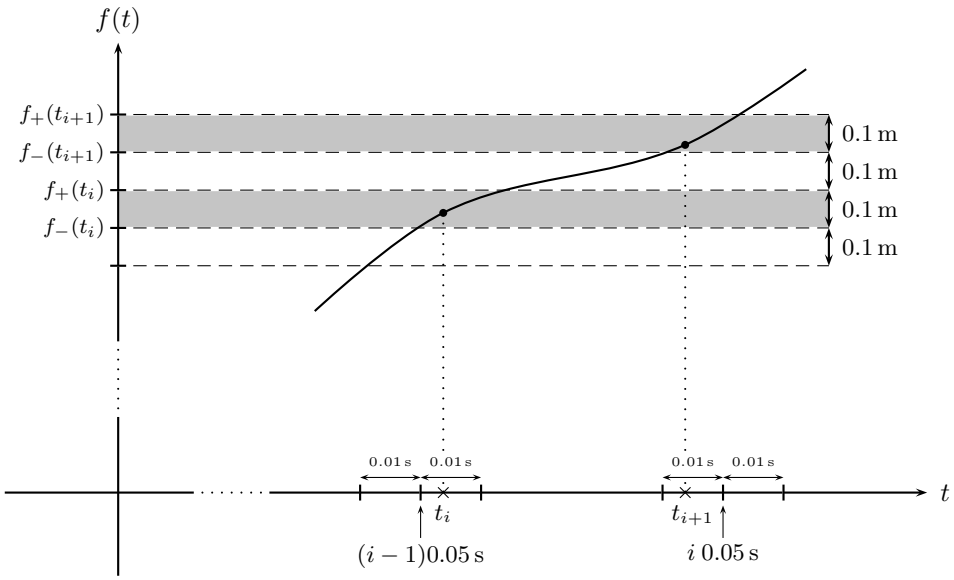

Figure 14 – Illustration of the generated times t_i and lower $f_-(t_i)$ and upper $f_+(t_i)$ bounds on $f(t_i)$

- We suppose that f'' and f''' are bounded. More precisely, for all $\theta \in \mathbb{R}_+$, we assume that

$$\begin{array}{rcccl} -1.3 & \leqslant & f''(\theta) & \leqslant & 1.2, \\ -1 & \leqslant & f'''(\theta) & \leqslant & 1. \end{array} \tag{3.2}$$

  In doing so assumption (A_2) from the introduction of this part I is satisfied with $d = 2$ and $d = 3$.

- Each second $s \in \mathbb{N}^*$, we suppose that we receive an new sample of measurements (obtained since the previous second) including the times t_i and their associated lower $f_-(t_i)$ and upper $f_+(t_i)$ bounds on $f(t_i)$, for all index i such that $t_i \in [s-1, s[$.

- All the simulations presented in this chapter were done by using *the same input data*. It means that differences of performances can only be explained by modifications of our various treatments.

3.1.2 — *Objectives.*

Let us remind that by the way of our simulations, we want to see how to use our models in order to bound the position, speed and acceleration of a train in real time. Then according to the problem exposed in introduction of this part I, our goal will be the following. Each second s, by using data from paragraph 3.1.1, we want to:

- determine quantities $f_-^{(k)}(t_i)$ and $f_+^{(k)}(t_i)$ such that $f_-^{(k)}(t_i) \leqslant f^{(k)}(t_i) \leqslant f_+^{(k)}(t_i)$ with difference $f_+^{(k)}(t_i) - f_-^{(k)}(t_i)$ as low as possible, for all $k \in \{1, 2\}$ and all index i such that $t_i \in [s-1, s[$,
- establish coherence between all the available bounds on f, f' and f'' in times t_i such that $t_i < s$, in order to make them as accurate as possible.

3.1.3 — REMARK. Let us remind that in practice, f, f' and f'' are (obviously) unknown functions. But in order to generate our data and verify the validity of our models, we have generate their exact trajectories. We have first constructed f''' as a piecewise continuous affine function. Then f'', f' and f are piecewise polynomial functions obtained by successive integrations. For the clarity and not overwrite this chapter of figures, the graphical representations of f''', f'', f' and f are presented in its appendix.

3.2 Suggested treatment

3.2.1 — In order to solve the problem exposed in the previous section, we suggest an iterative treatment that will be divided in two steps. For each second $s \in \mathbb{N}^*$, this treatment will consist in:

STEP 1. For all index j such that $t_j \in [s-1, s[$:

1.1. Bounds computation on $f'(t_j)$ by applying (1.11) with the available data on $\psi = f$ and $\psi''' = f'''$.

1.2. Bounds computation on $f'(t_j)$ by applying (1.3) with the available data on $\psi = f$ and $\psi'' = f''$.

1.3. Selection principle on bounds on $f'(t_j)$ obtained in steps 1.1 and 1.2.

1.4. Bounds computation on $f''(t_j)$ by applying (1.16) with the available data on $\psi = f$ and $\psi''' = f'''$.

1.5. Bounds computation on $f''(t_j)$ by applying (1.3) with the available data on $\psi = f'$ (from step 1.3) and $\psi'' = f'''$.

1.6. Selection principle on bounds on $f''(t_j)$ from steps 1.4 and 1.5 and by using the uniform bounds from (3.2).

1.7. Bounds computation on $f'(t_j)$ by applying (1.19) with the available data on $\psi = f$, $\psi'' = f''$ (from step 1.6) and $\psi''' = f'''$.

1.8. Selection principle on bounds on $f'(t_j)$ from steps 1.3 and 1.7.

1.9. Bounds computation on $f''(t_j)$ by applying (1.21) with the available data on $\psi = f$, $\psi'' = f''$ (from step 1.8) and $\psi''' = f'''$.

1.10. Selection principle on bounds on $f''(t_j)$ from steps 1.6 and 1.9.

STEP 2. Correction of data from the five latest samples (or s latest when $s \leqslant 4$) by using the forward-backward multi-order algorithm with $p = 3$, $\varphi = f$ and so $\varphi^{(p)} = f'''$.

The use of this iterative treatment in real time is illustrated in figure 15.

3.2.2 — *About the step 1.*

We need to explain how bounds on f' or f'' from steps 1.1, 1.2, 1.4, 1.5, 1.7 and 1.9 are computed. In order to do this, we refer to section 1.10. As explained in paragraph 1.10.1, we need to choose:

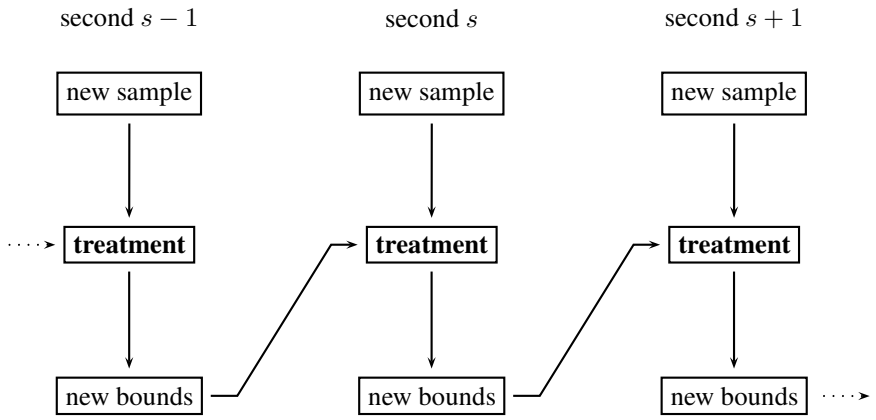

Figure 15 – Illustration of the iterative treatment presented in 3.2.1

- indexes r_1 and r_2 that enable to apply (1.11) (resp. (1.16)) with $\delta_1 = t_j - t_{j-r_1}$ and $\delta_2 = t_j - t_{j-r_2}$ at step 1.1 (resp. 1.4),
- an index r that enables to apply (1.3) (resp. (1.3), (1.19), (1.21)) with $\delta = t_j - t_{j-r}$ at step 1.2 (resp. 1.5, 1.7, 1.9).

An analysis of our treatment shows that differences $f_+(t_i) - f_-(t_i)$, $f'_+(t_i) - f'_-(t_i)$ and $f''_+(t_i) - f''_-(t_i)$ will be reduced over time and will depend on $i \in \{1, \ldots, N\}$. Thus as suggested in 1.10.3 or 1.10.5, we will suppose that differences $f_+(t_i) - f_-(t_i)$, $f'_+(t_i) - f'_-(t_i)$ and $f''_+(t_i) - f''_-(t_i)$ are respectively close to a same value $\mu > 0$, $\mu' > 0$ and $\mu'' > 0$ in order to determine:

- r_1 and r_2 by referring to 1.10.5 at steps 1.1 and 1.4,
- r by referring to 1.10.3 at steps 1.2, 1.5, 1.7 and 1.9.

3.2.3 — *About the step 2.*

Let us explain why we have not limited the use of the forward-backward multi-order algorithm to the sample associated to the second s in the step 2. In fact, when we will compute bounds of f' and f'' in the step 1 of our treatment, bounds computed for the sample associated to a second s will required bounds on f or f' from the previous samples (e.g. associated to seconds $s-1$ or $s-2$). And since the higher the accuracy of bounds from the past, the higher the accuracy of new bounds will be, it explains our choice.

3.3 A first simulation

3.3.1 — For this first simulation, we have supposed that the distance between bounds on $f(t_i)$ is not so far from $\mu = 0.1\,\mathrm{m}$ for all $i \in \{1, \ldots, N\}$, as are the initial data from (3.1). To compute bounds on f' or f'' from steps 1.5, 1.7 and 1.9 we have supposed that the differences between the required bounds on f' (resp. f'') are not so far from the minimum of the function ε from 1.4.3 (resp. 1.7.3) taken with:

$$\alpha = 2\mu = 0,2 \text{ and } \beta = \frac{1-(-1)}{6} = \frac{1}{3} \quad \text{(resp. } \alpha = 4\mu = 0,4 \text{ and } \beta = \frac{2[1-(-1)]}{6} = \frac{2}{3}\text{).}$$

In doing so, bounds on f' or f'' are computed as explained in 3.2.2. Taking the notations from this paragraph back, the various indexes r_1, r_2 or r computed to do this are obtained by minimizing a certain function ε (see paragraphs 1.10.3 or 1.10.5). For each step, the following tables indicate the minimizer and minimum value of these considering functions ε:

step	minimizer (δ_1^*, δ_2^*)		minimum value
	δ_1^*	δ_2^*	$\varepsilon(\delta_1^*, \delta_2^*)$
1.1	0.3760	1.7466	1.0169
1.4	0.3760	1.7466	2.3288

Table 3 – Minimizers and minimums of functions ε considered at steps 1.1 and 1.4

step	minimizer δ^*	minimum $\varepsilon(\delta^*)$
1.2	0.4	1
1.5	1.4142	2.8284
1.7	0.3777	1.0114
1.9	1.9052	2.4301

Table 4 – Minimizers and minimums of functions ε considered at steps 1.2, 1.5, 1.7 and 1.9

REMARK. The attentive reader will have noted that the function considered at step 1.2 is the same as the one represented in figure 1.

3.3.2 — *Output data analyses.*

In table 5 we have indicated the average distance between the various bounds on f, f' or f'' obtained on each step of our treatment, and their associated minimal and maximal values.

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.2642	1.9822/2.3381
	bounds from step 1.5	2.8550	2.4203/3.1767
	bounds from step 1.6	1.9646	1.2745/2.3380
	bounds from step 1.9	2.2345	1.7741/2.4310
	bounds from step 1.10	1.9193	1.2745/2.3374
	final bounds (step 2)	1.8129	1.2745/2.3009
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.9773	0.7230/1.0250
	bounds from step 1.2	0.9737	0.7899/1.0121
	bounds from step 1.3	0.8288	0.3890/1.0091
	bounds from step 1.7	0.9174	0.5990/1.0250
	bounds from step 1.8	0.8263	0.3890/1.0073
	final bounds (step 2)	0.6875	0.3890/1.0002
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0694	0.0116/0.1000

Table 5 – General performances of this preliminary simulation

We can first observe that in step 1, using various methods to bound a same derivative and the selection principle to adjust them is really effective. For instance with f', the selection principle from step 1.3

improves the accuracy of bounds obtained at steps 1.1 or 1.2 of approximatively 15%. Nevertheless by analysing the output data from steps 1.8 and 1.10, the improvements from steps 1.7 and 1.9 are relatively limited. This phenomenon will be studied in detail in paragraph 3.3.5.
About the use of the `FB_algorithm` algorithm from step 2, we can note that it improves in average of:

– 6% the accuracy of bounds on f'' from step 1.10,

– 16% the accuracy of bounds on f' from step 1.8,

– more than 30% the accuracy of the initial bounds on f from inequalities 3.1.

Moreover that represents:

– an improvement of 20% in comparison to the first bounds on f'' from step 1.4,

– an improvement of 30% in comparison to the first bounds on f' computed at steps 1.1 or 1.2.

Thus as expected, that illustrates the efficiency of the forward-backward multi-order algorithm presented in chapter 2. However we can observe that some bounds on f are not corrected by our treatment (since the maximal error after step 2 is equal to 0.1). In fact this phenomenon is just observed at the end of the simulation, when the train is stopped.

3.3.3 — *Graphical representations.*
We have produced various graphical representations on f and f' to illustrate the performances of our treatment. At first we have illustrated the effect of our treatment on bounds on f:

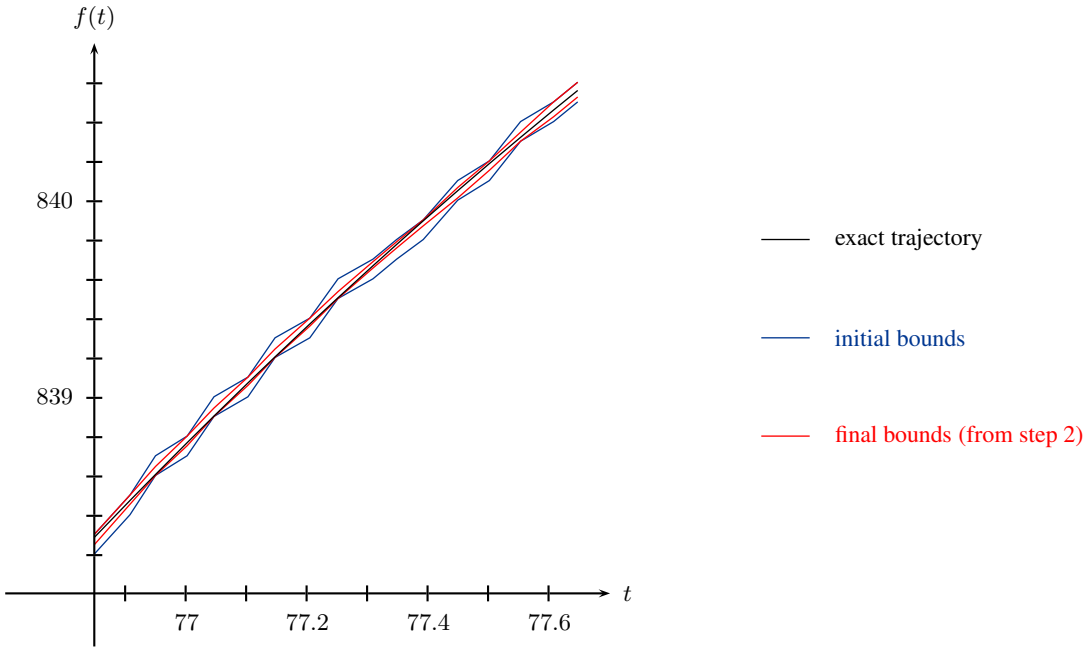

Figure 16 – Bounds on f obtained with our treatment

We can observe that the oscillations of initial bounds on f from (3.1), due to the measurement principle

exposed in paragraph 3.1.1 and illustrated by the figure 14, are significantly smoothed by the forward-backward multi-order algorithm (used at step 2).

The same phenomenon can be observed with f' and f'', but will limit our graphic analysis to outputs on f' because they are more impressive than those on f''. As shown in figure 17, we can note that the oscillations on the various bounds on f' computed at step 1 are smoothed by the `FB_optimal` algorithm. Let us signal that these oscillations are probably due to the use of initial bounds on f from the current sample (not yet corrected).
As previously, we can also observe that bounds on f' from steps 1.1 and 1.7 are practically the same, and especially for the lower bounds. That is probably due to the fact that bounds on f' are computed by using the same past data at these two steps by doing which is explained at paragraph 3.3.1: with other initial values of $f_+(t_i) - f_-(t_i)$ ($i \in \{1, \dots N\}$) or uniform bounds on f'' or f''', it generally does not happen. In other words, it just means that we had bad luck in our special case...

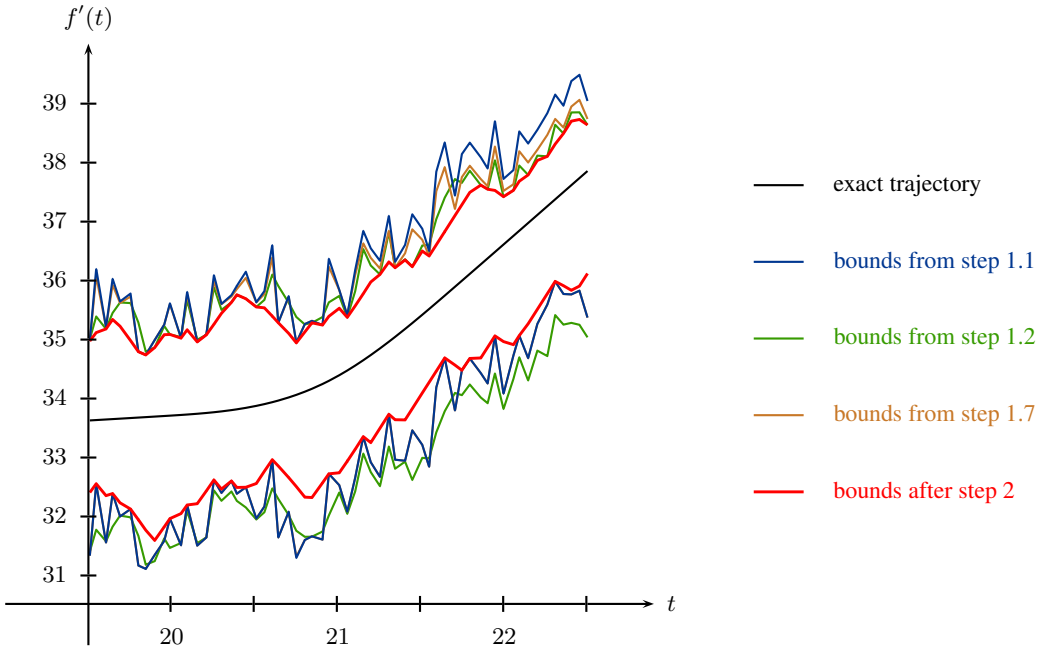

Figure 17 – Bounds on f' (in $\mathrm{km/h}$) obtained with our treatment

3.3.4 — *Computation time.*
Each sample is handled in less than $0.001\,\mathrm{s}$ in comparison to the second available to perform our various computations. That means our treatment is totally adapted for an use in real time, and that it can be improved (by doing more or quite complicated computations) without any problem. Some improvements of this first treatment that require more computations will be presented and analysed in section 3.4.

3.3.5 — *About the usefulness of the various computations from step 1.*
As observed in paragraphs 3.3.2 and 3.3.3, some bounds computations on f' or f'' done in step 1 seem useless. To end the study of this preliminary simulation, we present here the outputs of various adaptations of our treatment that omit some of bounds computations on f'.

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.2684	1.9896/2.3381
	bounds from step 1.5	2.9708	2.4207/3.1905
	bounds from step 1.6	1.9670	1.2745/2.3380
	bounds from step 1.9	2.2576	1.7739/2.4375
	bounds from step 1.10	1.9175	1.2745/2.3352
	bounds from step 2	1.8148	1.2745/2.3042
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	—	— / —
	bounds from step 1.2	0.9755	0.7965/1.0121
	bounds from step 1.3	—	— / —
	bounds from step 1.7	0.9635	0.6819/1.1202
	bounds from step 1.8	0.8478	0.4031/1.0103
	bounds from step 2	0.7511	0.4031/1.0005
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0712	0.0143/0.1000

Table 6 – Performances of our treatment without steps 1.1 and so 1.3

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.2668	1.9850/2.3381
	bounds from step 1.5	2.9439	2.4324/3.2139
	bounds from step 1.6	1.9660	1.2745/2.3380
	bounds from step 1.9	2.2519	1.7731/2.4511
	bounds from step 1.10	1.9204	1.2745/2.3376
	bounds from step 2	1.8142	1.2745/2.3011
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.9788	0.7235/1.0250
	bounds from step 1.2	—	— / —
	bounds from step 1.3	—	— / —
	bounds from step 1.7	0.9623	0.6769/1.1202
	bounds from step 1.8	0.8425	0.3890/1.0234
	bounds from step 2	0.7132	0.3890/1.0160
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0706	0.0116/0.1000

Table 7 – Performances of our treatment without steps 1.2 and so 1.3

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.2646	1.9822/2.3381
	bounds from step 1.5	2.8570	2.4203/3.1767
	bounds from step 1.6	1.9647	1.2745/2.3380
	bounds from step 1.9	2.2374	1.7742/2.4360
	bounds from step 1.10	1.9195	1.2745/2.3374
	bounds from step 2	1.8130	1.2745/2.3009
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.9775	0.7230/1.0250
	bounds from step 1.2	0.9738	0.7899/1.0121
	bounds from step 1.3	0.8290	0.3890/1.0091
	bounds from step 1.7	—	— / —
	bounds from step 1.8	—	— / —
	bounds from step 2	0.6901	0.3890/1.0002
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0696	0.0116/0.1000

Table 8 – Performances of our treatment without steps 1.7 and so 1.8

In comparison to the outputs of table 5, the analysis of tables 6 to 8 shown that doing all the computations from step 1 seems unnecessary. In particular and as highlighted in the previous paragraphs, computations from steps 1.7 and 1.8 improve almost anything. However the only analysis of tables 6 to 8 proved that computing bounds on f' in two different ways is useful, according to the outputs of the various selection principle from steps 1.3 or 1.8.
In summary, that means that all of these formulas could be useful, but it is maybe not necessary to use all of them to improve significantly the accuracy of the outputs. In our context, computations from steps 1.7 and 1.8 seem dispensable. We can also observe analogous phenomenons with bounds computations on f'': in this case, computations from steps 1.5 and 1.6 are clearly useless.

3.4 Some improvements or adaptations

3.4.1 — We can try to adapt the treatment done in section 3.3 just by taking into account its inputs and outputs. Let us remind that for a given index $j \in \{1, \ldots, N\}$, bounds on $f'(t_j)$ or $f''(t_j)$ from step 1 are computed by using bounds on f, f' or f'' in t_i, with $i \leqslant j$. Here we suggest to take into account if t_i belongs to the current sample or not. If so, that means that the bounds on f, f' or f'' in t_i come from a previous substep of step 1. If not, these bounds have been already corrected by our treatment and come from the step 2 of our treatment applied on the last sample.

- At steps 1.1, 1.2, 1.4, 1.7 and 1.9, data on f from current and previous samples may be used. According to table 5 the distance between bounds on f from the current sample are distant of $0.1\,\mathrm{m}$ when those from the previous sample are distant in average of $0.07\,\mathrm{m}$. Thus we can suppose that bounds on f used on these steps are close to the average of these values, *i.e.* $0.085\,\mathrm{m}$.

- Similarly at step 1.5 bounds on f' from current and previous samples may be used. Those from the current sample are obtained in step 1.3 and distant in average of $0.8288\,\mathrm{m/s} \simeq 0.83\,\mathrm{m}$, while those from the last sample are distant in average of $0.6875\,\mathrm{m/s} \simeq 0.69\,\mathrm{m}$. Thus we can suppose that bounds on f' used at this step are close to the average of these two values, *i.e.* $0.76\,\mathrm{m/s}$.

- At step 1.7 (resp. 1.9) bounds on $f''(t_i)$ (resp. $f'(t_i)$), and so from the current sample, are also used. Since these bounds come from step 1.6 (resp. 1.8), we can suppose that the distance between these bounds is not so far from $1.9646\,\mathrm{m/s^2} \simeq 0.96\,\mathrm{m/s^2}$ (resp. $0.9174\,\mathrm{m/s} \simeq 0.92\,\mathrm{m/s}$).

In doing so bounds from step 1 are computed as explained in 3.2.2. In comparison to tables 3 and 4 from paragraph 3.3.1, the following tables indicate the minimizers and minimum values of the various functions ε considered to compute these various bounds on f' and f''.

step	minimizer (δ_1^*, δ_2^*)		minimum value
	δ_1^*	δ_2^*	$\varepsilon(\delta_1^*, \delta_2^*)$
1.1	0.3561	1.6545	0.9124
1.4	0.3561	1.6545	2.2060

Table 9 – New minimizers and minimums of functions ε considered at steps 1.1 and 1.4

step	minimizer δ^*	minimum $\varepsilon(\delta^*)$
1.2	0.3688	0.9220
1.5	1.1747	2.3495
1.7	0.3721	0.8677
1.9	1.6973	2.1451

Table 10 – New minimizers and minimums of functions ε considered at steps 1.2, 1.5, 1.7 and 1.9

The performances associated to this adapted simulation are given by the following table:

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.2595	1.9439/2.3427
	bounds from step 1.5	2.7642	2.2347/3.1493
	bounds from step 1.6	1.9562	1.2342/2.3406
	bounds from step 1.9	2.2318	1.7125/2.4476
	bounds from step 1.10	1.9065	1.2342/2.3406
	bounds from step 2	1.7886	1.2342/2.3079
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.9761	0.7099/1.0256
	bounds from step 1.2	0.9710	0.7675/1.0033
	bounds from step 1.3	0.8389	0.3908/1.0009
	bounds from step 1.7	0.9155	0.5905/1.0250
	bounds from step 1.8	0.8352	0.3908/1.0009
	bounds from step 2	0.6815	0.3908/0.9970
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0691	0.0094/0.1000

Table 11 – General performances of this adapted simulation

In the end, we can note that there are almost no improvements in comparison to the preliminary simulation done in section 3.3. Consequently it seems difficult to improve the general performances of our treatment just by doing this kind of adjustments. It is maybe possible to do better by adjusting the previous parameters in real time, but we have not tried to do it for now.

3.4.2 — An other idea consists in trying to improve the previous treatments by doing what it is indicated in paragraph 1.10.7. With this aim in mind, we compute the required index r, r_1 and r_2 as in paragraph 3.4.1 (since our output data are a little bit better). Here are the associated output performances:

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	2.1399	1.8490/2.3377
	bounds from step 1.5	2.2395	1.9269/2.3427
	bounds from step 1.6	1.8830	1.1634/2.3377
	bounds from step 1.9	2.1235	1.6468/2.4049
	bounds from step 1.10	1.8310	1.1634/2.3295
	bounds from step 2	1.7138	1.1634/2.1893
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.8859	0.6552/1.0219
	bounds from step 1.2	0.9099	0.7463/1.0012
	bounds from step 1.3	0.7599	0.3569/1.0004
	bounds from step 1.7	0.8423	0.5779/1.0219
	bounds from step 1.8	0.7512	0.3569/1.0004
	bounds from step 2	0.6095	0.2912/0.9287
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0662	0.0094/0.1000

Table 12 – Performances of the simulation adapted as suggested in 1.10.7

In comparison to the simulation performed in 3.4.1, we can note that:

– the final bounds on f'' (from step 2) are improved of a little more than 4% in average,

– the final bounds on f' are improved of a little more than than 10% in average,

– the final bounds on f are improved of a little more than 4%.

This constitutes more significant improvements, in particular for the bounds on f'. And since here the computation time is almost unchanged in comparison to the previous simulations, we can say that it constitutes an interesting improvement.

3.4.3 — Observing that bounds on f, f' and f'' are significantly more accurate after the use of our treatment, we have thought that reapplying our treatment to each sample could improve their accuracy. In order to do this, we have first treat each sample as in 3.4.2, and a second time by taking in account the improvements provide by the first treatment in computations from the step 1. For the second treatment, we have thus supposed bounds on f close to $0.066\,\mathrm{m}$, those on f' close to $0.61\,\mathrm{m/s}$ and those on f'' close to $1.71\,\mathrm{m/s^2}$, which leads to the following choice of parameters:

step	minimizer (δ_1^*, δ_2^*)		minimum value
	δ_1^*	δ_2^*	$\varepsilon(\delta_1^*, \delta_2^*)$
1.1	0.3273	1.5207	0.7708
1.4	0.3273	1.5207	2.0276

Table 13 – Minimizers and minimums of functions ε for steps 1.1 and 1.4 of the second treatment

step	minimizer δ^*	minimum $\varepsilon(\delta^*)$
1.2	0.3250	0.8124
1.5	1.1045	2.2091
1.7	0.3484	0.7172
1.9	1.5320	1.9302

Table 14 – Minimizers and minimums of functions ε for steps 1.2, 1.5, 1.7 and 1.9 of the second treatment

Unfortunately it does not improve almost anything, as shown by the following table:

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	1.9530	1.5251/2.3685
	bounds from step 1.5	2.0099	1.6194/2.3685
	bounds from step 1.6	1.6749	1.1313/2.1732
	bounds from step 1.9	1.8834	1.4379/2.3252
	bounds from step 1.10	1.6682	1.1313/2.1732
	bounds from step 2	1.7102	1.1634/2.1893
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	0.7231	0.3734/1.0329
	bounds from step 1.2	0.7758	0.4900/1.0182
	bounds from step 1.3	0.5843	0.2877/0.9085
	bounds from step 1.7	0.6781	0.3149/0.9735
	bounds from step 1.8	0.5836	0.2877/0.9085
	bounds from step 2	0.6074	0.2912/0.9286
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0661	0.0094/0.1000

Table 15 – Performances when we apply twice the treatment used in 3.4.2

In the end, various application of our treatment to a same sample seems inefficient. To conclude, we thus suggest to use it as explained in paragraph 3.4.2.

3.5 An alternative use

3.5.1 — As we have seen in section 1.11, even if the centred bounds seem more precise than the decentered, their use to perform real-time analyses seems inappropriate. That is why the treatment suggested in paragraph 3.2.1 does not use them. To highlight and illustrate the differences between these centred and decentred bounds, we now suppose that we want to solve *a posteriori* the problem exposed in section 3.1. Thus the use of the centred formulas appears now reasonable.
In order to use them, we will modify the suggested treatment from section 3.2 by using:

– the centred bounds from (1.10) instead of the decentred (1.11) in step 1.1,

– the centred bounds from (1.15) instead of the decentred from (1.16) in step 1.4.

Then bounds on f' or f'' from step 1 are computed by referring to paragraphs 1.10.3 and 1.10.4, in the same way as in section 3.3. By taking notations from these paragraphs back, the various indexes r, r_- and r_+ used to compute bounds on f' and f'' are obtained by supposing bounds f distant of $0.1\,\mathrm{m}$, and those on f' (resp. f'') of the minimum of function ε from 1.3.5 (resp. 1.6.5) taken with

$$\alpha = \mu = 0.1 \quad \text{and} \quad \beta = \frac{1-(-1)}{6} = \frac{1}{3} \quad (\text{resp. } \alpha = 4\alpha = 0.4 \quad \text{and} \quad \beta = \frac{1-(-1)}{3} = \frac{2}{3}).$$

The following tables show the values of the various functions ε considered to computes indexes r_-, r_+ and r in question:

step	minimizer δ^*	minimum $\varepsilon(\delta^*, \delta^*)$
1.1	0.6694	0.2241
1.4	1.0627	1.0627

Table 16 – Minimizer and minimum of functions ε considered at steps 1.1 and 1.4

step	minimizer δ^*	minimum $\varepsilon(\delta^*)$
1.2	0.4	1
1.5	1.1045	2.2091
1.7	0.4839	0.7483
1.9	1.2713	1.4476

Table 17 – Minimizer and minimum of functions ε considered at steps 1.2, 1.5, 1.7 and 1.9

3.5.2 — *Output data analyses.*

The following table presents the average distance between bounds on f, f' and f'' obtained at each step of our treatment:

		average	min/max
error on f'' (in $\mathrm{m/s^2}$)	bounds from step 1.4	1.0315	0.0000/7.8986
	bounds from step 1.5	1.8855	1.5591/1.9972
	bounds from step 1.6	1.0065	0.0000/1.9972
	bounds from step 1.9	1.4600	1.1012/1.8612
	bounds from step 1.10	1.0048	0.0000/1.6689
	bounds from step 2	0.9641	0.0000/1.1541
error on f' (in $\mathrm{m/s}$)	bounds from step 1.1	1.0363	0.0000/3.6597
	bounds from step 1.2	3.4341	2.8137/3.6437
	bounds from step 1.3	0.9759	0.0000/1.7720
	bounds from step 1.7	2.4700	1.7470/3.5433
	bounds from step 1.8	0.9759	0.0000/1.7720
	bounds from step 2	0.9173	0.0000/1.1264
error on f (in m)	initial bounds	0.1000	0.1000/0.1000
	final bounds (step 2)	0.0504	0.0049/0.1000

Table 18 – Performances of our *a posteriori* treatment (using centred bounds instead of the decentred)

In comparison to the results obtained in table 5 with the preliminary simulation using the decentred bounds on f' and f'', we can note that:

– the average distance between bounds on f'' from step 1.5 is divided by more than 2, and those from step 2 are all the more accurate,

– the average distance between bounds on f' from step 1.1 is approximatively divided by 4, and those from step 2 by 3,

– the average distance between the initial bounds on f is in the end reduced of 50%, against 30% when the decentred bounds were used.

Thus as expected, the centred bounds produce more accurate bounds than the decentred. But as confirmed by analysing the values of δ^* given by the table 16 (which correspond to the waiting time to get the input data used in computations of the centred bounds), their use in real time seems totally unreasonable.

3.5.3 — *Computation time.*

The computation time is almost unchanged in comparison to the one obtained witrh the preliminary simulation using the decentred formula.

3.5.4 — *Graphical representation.*

The improvements due to the use of the centred bounds noted in the paragraph 3.5.2 can be observed graphically. In comparison to the output on f' represented in figure 17, the following picture shows an important improvement of the accuracy of the bounds from steps 1.1 and 2:

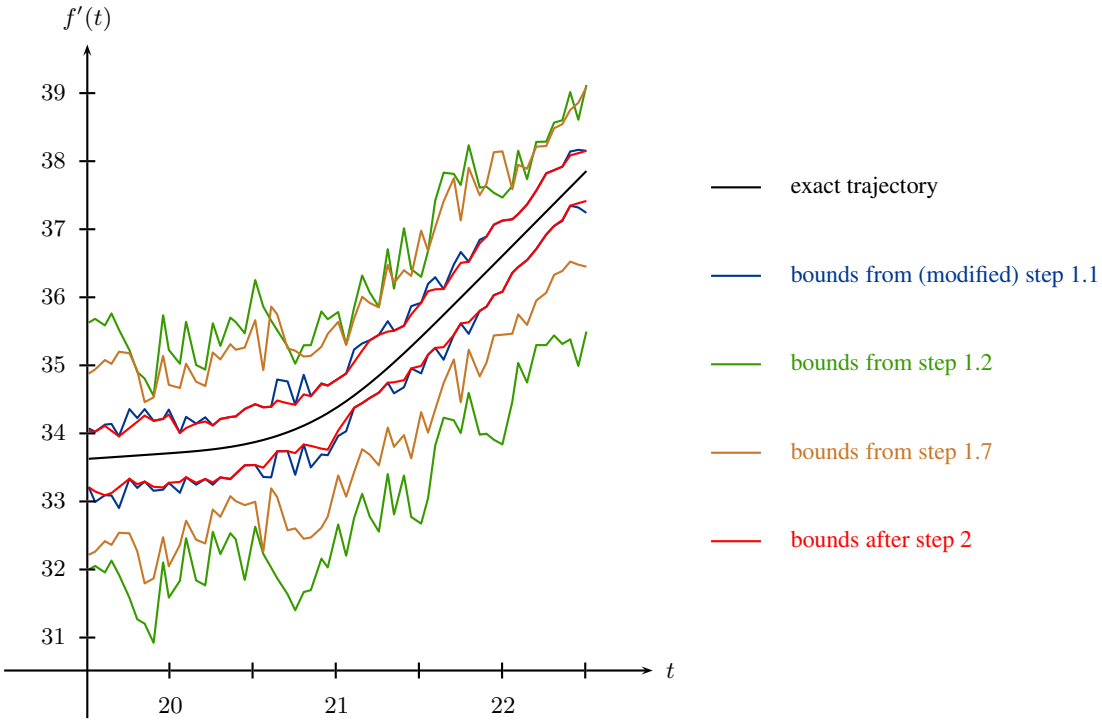

Figure 18 – Bounds on f' (in $\mathrm{km/h}$) with our *a posteriori* treatment (using centred bounds)

Even if it is more difficult to observe with f (due to a scale problem), we can see in figure 19 that bounds on f obtained at step 2 are closer than those obtained in figure 16. We can for instance look it on the neighbourhood of $t = 77.4\,\mathrm{s}$: the lower bounds on f are significantly higher than the initial in comparison to those obtained with the real-time treatment from section 3.3 (where one of them was manifestly not corrected).

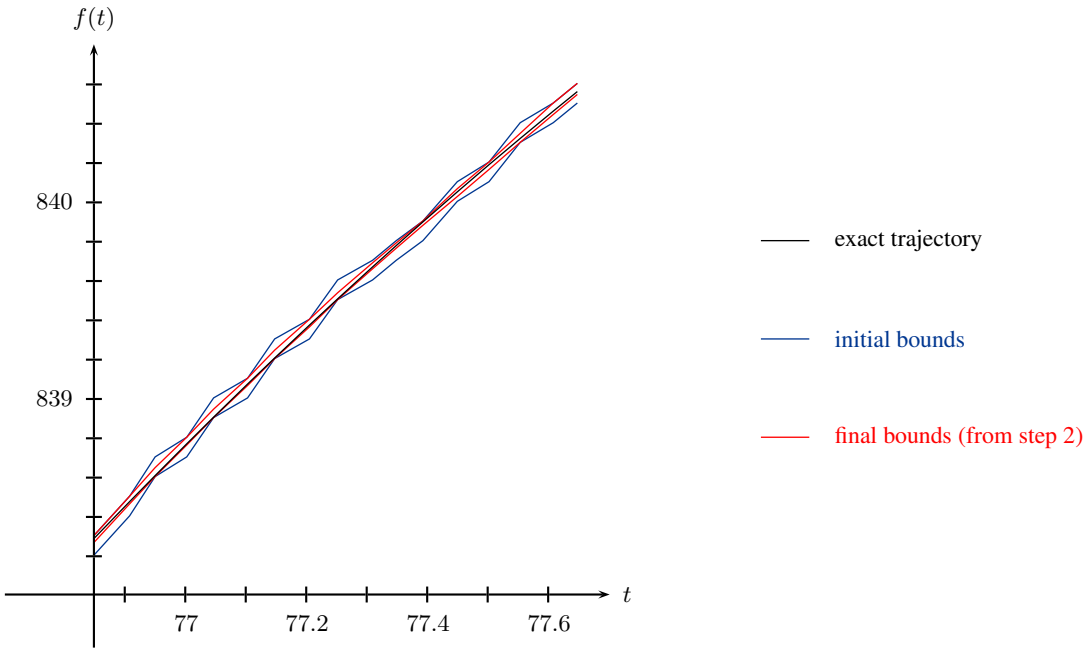

Figure 19 – Bounds on f obtained with our treatment

3.5.5 — REMARK. As for the real-time treatment presented in section 3.3, it is then possible to improve this *a posteriori* treatment by doing analogous improvements than those presented in section 3.4. But knowing that we will obtain similar result by doing them, we will not present them in this document.

Conclusion

On a sophisticated example inspired of the Parisian underground, this chapter has shown how the theories from chapter 1 and 2 can be used to solve the problem presented in introduction of this part I. We have developed and improved two treatments that enable to bound a function and its derivative in real time for the first one, and *a posteriori* for the other.
By doing these simulations, we have also highlighted the efficiency of our various models. Now it could be interesting to see how to improve it. Let us give some possible ways to do this:

- As suggested in 3.4.1, by choosing the best data to use in step 1 in real time, and not arbitrary as we have done in this chapter. However in doing so, care should be taken to ensure that the calculations are not too costly in time.
- When we are estimating the position of a mobile (here a train), by taking into account the fact that the mobile is accelerating or decelerating, when such informations are available.
- By coupling our treatments with some other models, etc.

There is probably a lot of other possible improvements as those mentioned here, but these will depend before anything else of the nature of the input data. Some other improvements may also be suggested by an experience feedback.

Appendix

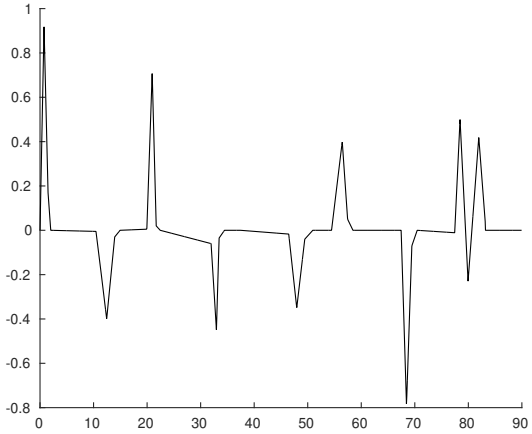

Figure 20 – Exact trajectory of f''' (in $\mathrm{m/s^3}$)

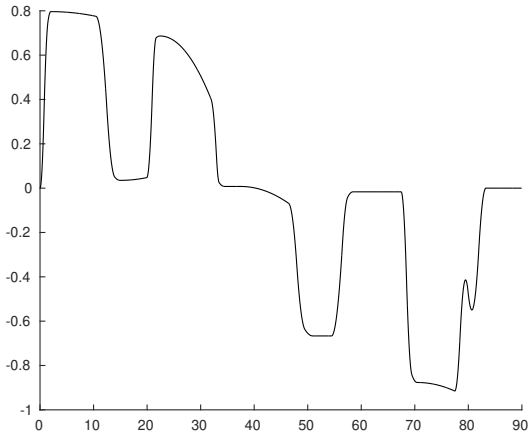

Figure 21 – Exact trajectory of f'' (in $\mathrm{m/s^2}$)

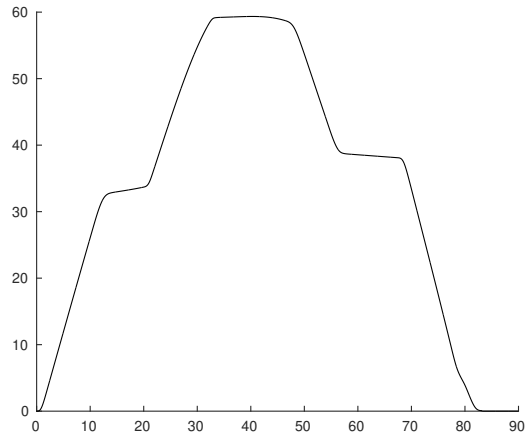

Figure 22 – Exact trajectory of f' (in $\mathrm{km/h}$)

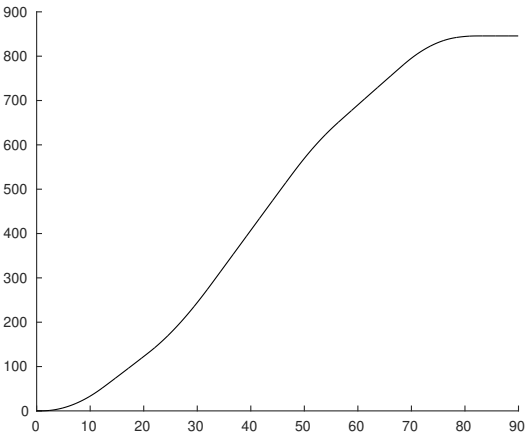

Figure 23 – Exact trajectory of f (in m)

Chapter 4

Some possible extensions

Introduction

This chapter gathers various results or ideas that could be used to generalize or improve the models or algorithms presented in the previous chapters, but not really studied in detail, tested in practice or presented in an optimal form. In section 4.1 we extend our works from chapter 1 by presenting various techniques that enable to bound a function by using (other) bounds on it and uniforms bounds on one of its derivatives. We introduce a new kind of forward-backward corrections in section 4.2. Finally we suggest some possible adaptations of our models to functions defined on (a part of) $\mathbb{R}^n$ to $\mathbb{R}$ in section 4.3.

4.1 How to bound a function by one of its derivative?

4.1.1 — In chapter 1, we have presented various techniques that enable to bound the derivatives of a function ψ in a point t just by using uniform bounds on one of its p^{th} derivatives. So as in this chapter, let us fix $t \in I$ and $\psi : I \to \mathbb{R}$ a function satisfying inequalities (1.1) and (1.2), that means:

- for all $y \in I$, we can determine (known) quantities $\psi_-(y)$ and $\psi_+(y)$ such that:

$$\psi_-(y) \ \leqslant \ \psi(y) \ \leqslant \ \psi_+(y).$$

- there exists $p \in \mathbb{N}^*$ such that ψ admits a p^{th} bounded derivative, i.e. there exists (known) real constants ψ_-^* and ψ_+^* such that for all $\xi \in I$:

$$\psi_-^* \ \leqslant \ \psi^{(p)}(\xi) \ \leqslant \ \psi_+^*.$$

In this section, we will how to bound ψ in t by leading similar computations, and here again just by considering cases $p = 2$ or $p = 3$.

4.1.2 — Let us start with the case $p = 2$. As we did for $\psi'(t)$ or $\psi''(t)$ in sections 1.2 or 1.5, we need to find here an expression of $\psi(t)$ that only depends on ψ and ψ''. To do this, we start from the following Taylor expansions, available for all δ_1 and δ_2 in $\mathbb{R}^*$ such that $t + \delta_1 \in I$ and $t + \delta_2 \in I$,

$$\psi(t + \delta_1) \ = \ \psi(t) + \delta_1 \psi'(t) + \frac{\delta_1^2}{2} \psi''(\xi_1), \tag{4.1}$$

$$\psi(t + \delta_2) \ = \ \psi(t) + \delta_2 \psi'(t) + \frac{\delta_2^2}{2} \psi''(\xi_2), \tag{4.2}$$

where ξ_j is strictly between t and $t+\delta_j$ for all $j \in \{1,2\}$. For all $\alpha_1, \alpha_2 \in \mathbb{R}$, by doing $\alpha_1(4.1)+\alpha_2(4.2)$ and by ordering $\alpha_1+\alpha_2 \neq 0$ and $\alpha_1\delta_1+\alpha_2\delta_2 = 0$, we obtain

$$\psi(t) \;=\; \frac{\alpha_1\psi(t+\delta_1)+\alpha_2\psi(t+\delta_2)}{\alpha_1+\alpha_2} - \frac{\alpha_1\delta_1^2\psi''(\xi_1)}{2(\alpha_1+\alpha_2)} - \frac{\alpha_2\delta_2^2\psi''(\xi_2)}{2(\alpha_1+\alpha_2)}.$$

And since we can multiply α_1 and α_2 by a same non-null quantity without changing this expression of $\psi(t)$, we can fix $\alpha_1 = \delta_2$, and so $\alpha_2 = -\delta_1$ (since $\alpha_1\delta_1+\alpha_2\delta_2 = 0$). To guarantee $\alpha_1+\alpha_2 \neq 0$, we this need to impose $\delta_1 \neq \delta_2$. In the end for all $\delta_1, \delta_2 \in \mathbb{R}^*$ such that $t+\delta_1 \in I$, $t+\delta_2 \in I$ and $\delta_1 \neq \delta_2$,

$$\psi(t) \;=\; \frac{\delta_2\psi(t+\delta_1)-\delta_1\psi(t+\delta_2)}{\delta_2-\delta_1} + \frac{\delta_1\delta_2^2\psi''(\xi_2)}{2(\delta_2-\delta_1)} - \frac{\delta_2\delta_1^2\psi''(\xi_1)}{2(\delta_2-\delta_1)}. \tag{4.3}$$

4.1.3 — *Centred bounds on ψ when $p = 2$.*

Let us suppose that δ_1 and δ_2 have opposite signs. At the risk of exchanging their roles, we can suppose $\delta_1 > 0$ and $\delta_2 < 0$. Then by setting $\delta_+ = \delta_1$, $\delta_- = -\delta_2$, $\xi_+ = \xi_1$ and $\xi_- = \xi_2$, formula (4.3) can be rewritten:

$$\psi(t) \;=\; \frac{\delta_-\psi(t+\delta_+)+\delta_+\psi(t-\delta_-)}{\delta_-+\delta_+} - \frac{\delta_-\delta_+^2\psi''(\xi_+)}{2(\delta_-+\delta_+)} - \frac{\delta_+\delta_-^2\psi''(\xi_-)}{2(\delta_-+\delta_+)}. \tag{4.4}$$

Let us define

$$P_t \;=\; \{(\delta_-,\delta_+) \in \mathbb{R}_+^* \mid t--\delta_- \in I, t+\delta_+ \in I\}.$$

Thus for all $(\delta_-,\delta_+) \in P_t$, by setting

$$\begin{cases} \psi_-(t\,;\delta_-,\delta_+) \;=\; \dfrac{\delta_-\psi_-(t+\delta_+)+\delta_+\psi_-(t-\delta_-)}{\delta_-+\delta_+} - \dfrac{\delta_-\delta_+\psi_+^*}{2}, \\ \psi_+(t\,;\delta_-,\delta_+) \;=\; \dfrac{\delta_-\psi_+(t+\delta_+)+\delta_+\psi_+(t-\delta_-)}{\delta_-+\delta_+} - \dfrac{\delta_-\delta_+\psi_-^*}{2}, \end{cases}$$

we deduce of (1.1), (1.2) and (4.4) that

$$\psi_-(t\,;\delta_-,\delta_+) \;\leqslant\; \psi(t) \;\leqslant\; \psi_+(t\,;\delta_-,\delta_+).$$

And as in chapter 1 the values of δ_- and δ_+ that minimize the distance between bounds $\psi_-(t\,;\delta_-,\delta_+)$ and $\psi_+(t\,;\delta_-,\delta_+)$ on $\psi(t)$ can be found by studying the diameter function

$$\operatorname{diam}\psi_t \;:\; P_t \to \mathbb{R}_+,\; (\delta_-,\delta_+) \mapsto \psi_+(t\,;\delta_-,\delta_+) - \psi_-(t\,;\delta_-,\delta_+).$$

4.1.4 — *Decentred bounds on ψ when $p = 2$.*

We now suppose that δ_1 and δ_2 have the same sign, and more precisely that $\delta_1 < 0$ and $\delta_2 < 0$ (the only interesting case for our applications). Then by doing our usual transformations, for all $\delta_1, \delta_2 \in \mathbb{R}_+^*$ such that $t-\delta_1 \in I$, $t-\delta_2 \in I$ and $\delta_1 < \delta_2$, formula (4.3) can be rewritten:

$$\psi(t) \;=\; \frac{\delta_2\psi(t-\delta_1)-\delta_1\psi(t-\delta_2)}{\delta_2-\delta_1} + \frac{\delta_1\delta_2^2\psi''(\xi_2)}{2(\delta_2-\delta_1)} - \frac{\delta_2\delta_1^2\psi''(\xi_1)}{2(\delta_2-\delta_1)}. \tag{4.5}$$

Let us define

$$P_t \;=\; \{(\delta_1,\delta_2) \in \mathbb{R}_+^* \mid t-\delta_1 \in I, t-\delta_2 \in I, \delta_1 < \delta_2\}.$$

Thus for all $(\delta_1,\delta_2) \in P_t$, by setting

$$\begin{cases} \psi_-(t\,;\delta_1,\delta_2) \;=\; \dfrac{\delta_2\psi_-(t-\delta_1)-\delta_1\psi_+(t-\delta_2)}{\delta_2-\delta_1} + \dfrac{\delta_1\delta_2^2\psi_-^*}{2(\delta_2-\delta_1)} - \dfrac{\delta_2\delta_1^2\psi_+^*}{2(\delta_2-\delta_1)}, \\ \psi_+(t\,;\delta_1,\delta_2) \;=\; \dfrac{\delta_2\psi_+(t-\delta_1)-\delta_1\psi_-(t-\delta_2)}{\delta_2-\delta_1} + \dfrac{\delta_1\delta_2^2\psi_+^*}{2(\delta_2-\delta_1)} - \dfrac{\delta_2\delta_1^2\psi_-^*}{2(\delta_2-\delta_1)}, \end{cases}$$

we deduce of (1.1), (1.2) and (4.5) that

$$\psi_-(t\,;\delta_1,\delta_2) \;\leqslant\; \psi(t) \;\leqslant\; \psi_+(t\,;\delta_1,\delta_2).$$

Here again the values of δ_1 and δ_2 that minimize the distance between these bounds on $\psi(t)$ can be obtained by studying the diameter function

$$\operatorname{diam}\psi_t \;:\; P_t \to \mathbb{R}_+,\; (\delta_1,\delta_2) \mapsto \psi_+(t\,;\delta_1,\delta_2) - \psi_-(t\,;\delta_1,\delta_2).$$

4.1.5 — *Decentred bounds on ψ when $p = 3$.* We now suppose that $p = 3$. To obtain an expression of $\psi(t)$ that only depends on various evaluations of ψ and ψ''', we need to use three Taylor expansions of the third order to eliminate the contributions in ψ' and ψ''. That is why we will only compute decentred bounds on $\psi(t)$. So let δ_1, δ_2 and δ_3 be three real numbers satisfying

$$t-\delta_1 \;\in\; I, \quad t-\delta_2 \;\in\; I, \quad t-\delta_3 \;\in\; I \quad \text{and} \quad 0 \;<\; \delta_1 \;<\; \delta_2 \;<\; \delta_3. \tag{4.6}$$

By the Taylor-Lagrange formula, there exists $(\xi_1,\xi_2,\xi_3) \in]t-\delta_1,t[\,\times\,]t-\delta_2,t[\,\times\,]t-\delta_3,t[$ such that

$$\psi(t-\delta_1) \;=\; \psi(t) - \delta_1\psi'(t) + \frac{\delta_1^2}{2}\,\psi''(t) - \frac{\delta_1^3}{6}\,\psi'''(t_1), \tag{4.7}$$

$$\psi(t-\delta_2) \;=\; \psi(t) - \delta_2\psi'(t) + \frac{\delta_2^2}{2}\,\psi''(t) - \frac{\delta_2^3}{6}\,\psi'''(t_2), \tag{4.8}$$

$$\psi(t-\delta_3) \;=\; \psi(t) - \delta_3\psi'(t) + \frac{\delta_3^2}{2}\,\psi''(t) - \frac{\delta_3^3}{6}\,\psi'''(t_3). \tag{4.9}$$

For all $\alpha_1,\alpha_2,\alpha_3 \in \mathbb{R}$, by ordering

$$\begin{aligned} \alpha_1+\alpha_2+\alpha_3 \;&\neq\; 0, \\ \alpha_1\delta_1+\alpha_2\delta_2+\alpha_3\delta_3 \;=\; \alpha_1\delta_1^2+\alpha_2\delta_2^2+\alpha_3\delta_3^2 \;&=\; 0, \end{aligned} \tag{4.10}$$

and by doing α_1(4.7) + α_2(4.8) + α_3(4.9), we can obtain the following expression of $\psi(t)$:

$$\psi(t) \;=\; \frac{\alpha_1\psi(t-\delta_1)+\alpha_2\psi(t-\delta_2)+\alpha_3\psi(t-\delta_3)}{\alpha_1+\alpha_2+\alpha_3} + \frac{\alpha_1\delta_1^3\psi'''(\xi_1)+\alpha_2\delta_2^3\psi'''(\xi_2)+\alpha_3\delta_3^3\psi'''(\xi_3)}{6(\alpha_1+\alpha_2+\alpha_3)}. \tag{4.11}$$

Since we can multiply α_1, α_2 and α_3 by a same non-null quantity without changing this expression of $\psi(t)$, we can first fix $\alpha_1 = 1$. In doing so,

$$(4.10) \iff \left\{ \begin{array}{rcl} \alpha_2\delta_2+\alpha_3\delta_3 &=& -\delta_1 \\ \alpha_2\delta_2^2+\alpha_3\delta_3^2 &=& -\delta_1^2 \end{array} \right. \iff \left\{ \begin{array}{rcl} \alpha_2 &=& \dfrac{\delta_1(\delta_1-\delta_3)}{\delta_2(\delta_3-\delta_2)}, \\ \alpha_3 &=& \dfrac{\delta_1(\delta_2-\delta_1)}{\delta_3(\delta_3-\delta_2)}. \end{array} \right.$$

Now by multiplying α_1, α_2 and α_3 by $\delta_2\delta_3(\delta_3-\delta_2)$, which is a non-null quantity since thanks to (4.6), we can finally suppose that

$$\alpha_1 \;=\; \delta_2\delta_3(\delta_3-\delta_2), \quad \alpha_2 \;=\; -\delta_1\delta_3(\delta_3-\delta_1) \quad \text{and} \quad \alpha_3 \;=\; \delta_1\delta_2(\delta_2-\delta_1).$$

By using (4.6) we can easily observe that

$$\alpha_1 \;>\; 0, \quad \alpha_2 \;<\; 0, \quad \alpha_3 \;>\; 0 \quad \text{and} \quad \alpha_1+\alpha_2+\alpha_3 \;=\; (\delta_2-\delta_1)(\delta_3-\delta_2)(\delta_3-\delta_1) \;>\; 0. \tag{4.12}$$

Consequently, by setting

$$\left\{\begin{array}{rcl}\psi_-(t\,;\delta_1,\delta_2,\delta_3) &=& \dfrac{\alpha_1\psi_-(t-\delta_1)+\alpha_2\psi_+(t-\delta_2)+\alpha_3\psi_-(t-\delta_3)}{\alpha_1+\alpha_2+\alpha_3}+\dfrac{\alpha_1\psi_-^*+\alpha_2\psi_+^*+\alpha_3\psi_-^*}{6(\alpha_1+\alpha_2+\alpha_3)},\\ \psi_+(t\,;\delta_1,\delta_2,\delta_3) &=& \dfrac{\alpha_1\psi_+(t-\delta_1)+\alpha_2\psi_-(t-\delta_2)+\alpha_3\psi_+(t-\delta_3)}{\alpha_1+\alpha_2+\alpha_3}+\dfrac{\alpha_1\psi_+^*+\alpha_2\psi_-^*+\alpha_3\psi_+^*}{6(\alpha_1+\alpha_2+\alpha_3)},\end{array}\right.$$

we can deduce of (4.11) and (4.12) that

$$\psi_-(t\,;\delta_1,\delta_2,\delta_3)\ \leqslant\ \psi(t)\ \leqslant\ \psi_+(t\,;\delta_1,\delta_2,\delta_3).$$

Here again the triplets $(\delta_1,\delta_2,\delta_3)$ satisfying (4.6) that minimize the distance between these bounds on $\psi(t)$ can be found by minimizing the corresponding diameter function (where α_1, α_2 and α_3 must obviously be seen as functions of δ_1, δ_2 and δ_3).

4.1.6 — On the concrete example of the problem solved in chapter 3, these models can for instance be used in the suggested treatment from 3.2.1 to improve the accuracy of the initial bounds on f. By doing it before the step 1.1, it would logically improve the accuracy of bound on f' and f'' computed after that.

4.2 Other forward-backward corrections

4.2.1 — The models presented in chapter 2 enable to determine admissible values on $\varphi^{(k)}(t)$, the k^{th} derivative in a point t of a function $\varphi : I \to \mathbb{R}$ admitting a p^{th} bounded derivative, by having some informations on $\varphi^{(k)}$, …, $\varphi^{(p)}$. The models presented in this section are intended to refine that by considering admissible values for couples of the form $\big(\varphi^{(k)}(t),\varphi^{(k+1)}(t)\big)$, with $k\in\{0,\dots,p-1\}$, by knowing bounds on couple of the form $\varphi^{(k)}(t)\pm\delta_\pm\varphi^{(k+1)}(t)$, for given $\delta_-,\delta_+>0$. Then the idea consists in using this new bounds on $\varphi^{(k)}(t)\pm\delta_\pm^{(k+1)}(t)$ to adjust those on $\varphi^{(k)}(t)$ and $\varphi^{(k+1)}(t)$.

EXAMPLE. When $p=2$, let us suppose that we can determine quantities $\varphi_\pm(t)$ and $\varphi'_\pm(t)$ such that

$$\begin{array}{ccccc}\varphi_-(t) &\leqslant& \varphi(t) &\leqslant& \varphi_+(t),\\ \varphi'_-(t) &\leqslant& \varphi'(t) &\leqslant& \varphi'_+(t),\end{array}$$

but also, for given δ_- and δ_+ in $\mathbb{R}_+^*$, quantities $\varphi_\pm(t-\delta_-)$ and $\varphi_\pm(t+\delta_+)$ such that

$$\begin{array}{ccccc}\varphi_-(t-\delta_-) &\leqslant& \varphi(t-\delta_-) &\leqslant& \varphi_+(t-\delta_-),\\ \varphi'_-(t+\delta_+) &\leqslant& \varphi'(t+\delta_+) &\leqslant& \varphi'_+(t+\delta_+).\end{array}$$

Then by using the Taylor-Lagrange formula, we can bound $\varphi(t)+\delta_+\varphi'(t)$ and $\varphi(t)-\delta_-\varphi'(t)$ as follows:

$$\begin{array}{ccccc}\varphi_-(t+\delta_+)-\dfrac{\delta_+^2\varphi_+^*}{2} &\leqslant& \varphi(t)+\delta_+\varphi'(t) &\leqslant& \varphi_+(t+\delta_+)-\dfrac{\delta_+^2\varphi_-^*}{2},\\ \varphi_-(t-\delta_-)-\dfrac{\delta_-^2\varphi_+^*}{2} &\leqslant& \varphi(t)-\delta_-\varphi'(t) &\leqslant& \varphi_+(t-\delta_-)-\dfrac{\delta_-^2\varphi_-^*}{2}.\end{array}$$

4.2.2 — The study leads in this section will use the following terminology:

Definition. *We call* rectangle *of* $\mathbb{R}^2$ *each Cartesian product of the form* $I\times J$*, where* I *and* J *are two closed and bounded intervals of* $\mathbb{R}$.

4.2.3 — By setting $x = \varphi^{(k)}(t)$ and $y = \varphi^{(k+1)}(t)$ in 4.2.1, the corrections suggested in this section requires the resolution of a same geometrical problem. Given eight real numbers x^-, x^+, y^-, y^+, m^-, m^+, μ^-, μ^+ and two positive numbers δ_- and δ_+, it consists in determining the smallest rectangle of $\mathbb{R}^2$ that contains all the couples $(x, y) \in \mathbb{R}^2$ satisfying the following constraints:

$$x \leqslant x^+, \quad (1) \qquad x + \delta_+ y \leqslant m^+, \quad (5)$$
$$x \geqslant x^-, \quad (2) \qquad x + \delta_+ y \geqslant m^-, \quad (6)$$
$$y \leqslant y^+, \quad (3) \qquad x - \delta_- y \leqslant \mu^+, \quad (7)$$
$$y \geqslant y^-, \quad (4) \qquad x - \delta_- y \geqslant \mu^-. \quad (8)$$

For instance on the example from 4.2.1, we will have

$$x^\pm = \varphi_\pm(t), \quad y^\pm = \varphi'_\pm(t), \quad m^\pm = \varphi_\pm(t + \delta_+) - \frac{\delta_+^2 \varphi_\mp^*}{2} \quad \text{and} \quad \mu^\pm = \varphi_\pm(t - \delta_-) - \frac{\delta_-^2 \varphi_\mp^*}{2}.$$

For all $i \in \{1, \dots, 8\}$, let E_i be the set of couples $(x, y) \in \mathbb{R}^2$ satisfying the constraint (i), and

$$E = \bigcap_{i=1}^{8} E_i$$

the set of couples $(x, y) \in \mathbb{R}^2$ satisfying the constraints (1) to (8). In other word, we are looking for the smallest rectangle of $\mathbb{R}^2$ that contains E.
We can first observe this problem is well-posed. Indeed, let $\mathcal{P}$ be the set of rectangles of $\mathbb{R}^2$ that contains E. Since $E \subset [x^-, x^+] \times [y^-, y^+]$, then we have $\mathcal{P} \neq \varnothing$. By setting

$$P^* = \bigcap_{P \in \mathcal{P}} P,$$

it is clear that P^* is also a rectangle of $\mathbb{R}^2$, that $E \subset P^*$, and that P^* is included in any rectangle of $\mathbb{R}^2$ that contains E. Therefore P^* is the unique solution of our problem.
When $P^* \neq \varnothing$, which is equivalent to say that $E \neq \varnothing$, there exists unique intervals I^* and J^* such that $P^* = I^* \times J^*$. And since in practice x will be associated to $\psi^{(k)}(t)$ and y to $\psi^{(k+1)}(t)$, we will be able to conclude that $\psi^{(k)}(t) \in I^*$ and that $\psi^{(k+1)}(t) \in J^*$.
In the following paragraphs, we will first expose a characterisation of the fact that P^* is non empty. Then under the assumption $E \neq \varnothing$, we will explicit the bounds of the intervals I^* and J^* that define P^*.

4.2.4 — The resolution of the problem exposed in 4.2.3 will we be lead by geometrical considerations.

- For all $i \in \{1, \dots, 8\}$, the set E_i described by the constraint (i) corresponds to a half-space of $\mathbb{R}^2$. In the following, we will denote by l_i the line that defines the boundary of E_i. These lines have for equations:

$$l_1 : x = x^+, \qquad l_3 : y = y^+, \qquad l_5 : y = \frac{m^+ - x}{\delta_+}, \qquad l_7 : y = \frac{x - \mu^+}{\delta_-},$$
$$l_2 : x = x^-, \qquad l_4 : y = y^-, \qquad l_6 : y = \frac{m^- - x}{\delta_+}, \qquad l_8 : y = \frac{x - \mu^-}{\delta_-}.$$

- For all indexes i and j such that $1 \leqslant i < j \leqslant 8$ and lines l_i and l_j are secant, we will denote by (x_{ij}, y_{ij}) the unique intersection point between l_i and l_j.
- We will more precisely set $A = (x_{23}, y_{23})$, $B = (x_{13}, y_{13})$, $C = (x_{14}, y_{14})$, $D = (x_{24}, y_{24})$, and $A' = (x_{68}, y_{68})$, $B' = (x_{58}, y_{58})$, $C' = (x_{57}, y_{57})$, $D' = (x_{67}, y_{67})$. Thus when rectangle $E_1 \cap \cdots \cap E_4$ (resp. parallelogram $E_5 \cap \cdots \cap E_8$) is non-empty, then points A, B, C, D (resp. A', B', C', D') are its vertexes.

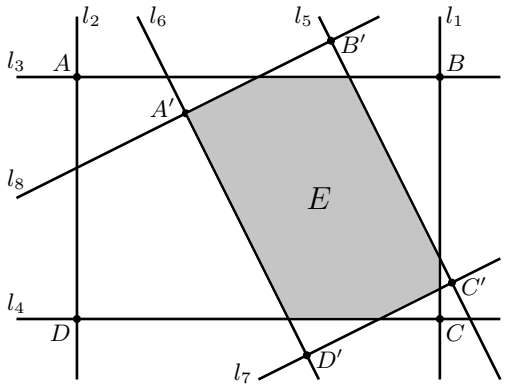

Figure 24 – Geometrical representation of E with lines $l_1, \ldots, l_8$ when $E \neq \varnothing$

4.2.5 — Definition. *Let $k \in \{1, \ldots, 8\}$ and $i_1, \ldots, i_k$ be indexes such that $1 \leqslant i_1 < \cdots < i_k \leqslant 8$. We will say that constraints $(i_1), \ldots, (i_k)$ are* compatible *if the set $E_{i_1} \cap \cdots \cap E_{i_k}$ of points from $\mathbb{R}^2$ satisfying these constraints is non-empty.*

4.2.6 — To characterize the fact that E is non-empty, which means that constraints (1) to (8) are satisfied, we can start by finding criteria that insure that some of these constraints are compatible.

Proposition. *We have the twelve following compatibility criterion:*

- *Constraints* (1) *and* (2) *are compatible if and only if l_1 is on the right of l_2, that means:*

$$x^- \leqslant x^+. \tag{R$_1$}$$

- *Constraints* (3) *et* (4) *are compatible if and only if l_3 is above l_4, that means*

$$y^- \leqslant y^+. \tag{R$_2$}$$

- *Constraints* (5) *et* (6) *are compatible if and only if l_5 is above l_6, that means*

$$m^- \leqslant m^+. \tag{R$_3$}$$

- *Constraints* (7) *et* (8) *are compatible if and only if l_7 is below l_8, that means*

$$\mu^- \leqslant \mu^+. \tag{R$_4$}$$

- *Constraints* (1), (6) *and* (8) *are compatible if and only if A' is on the right of l_1, that means*

$$x_{68} = \frac{\delta_- m^- + \delta_+ \mu^-}{\delta_- + \delta_+} \leqslant x^+. \tag{R$_5$}$$

- *Constraints* (2), (5) *and* (7) *are compatible if and only if C' is on the right of l_2, that means*

$$x_{57} = \frac{\delta_- m^+ + \delta_+ \mu^+}{\delta_- + \delta_+} \geqslant x^-. \tag{R$_6$}$$

- *Constraints* (3), (6) *and* (7) *are compatible if and only if* D' *is below* l_3*, that means*

$$y_{67} \;=\; \frac{m^- - \mu^+}{\delta_- + \delta_+} \;\leqslant\; y^+. \tag{R$_7$}$$

- *Constraints* (4), (5) *and* (8) *are compatible if and only if* B' *is above* l_4*, that means*

$$y_{58} \;=\; \frac{m^+ - \mu^-}{\delta_- + \delta_+} \;\geqslant\; y^-. \tag{R$_8$}$$

- *Constraints* (2), (4) *and* (5) *are compatible if and only if* D *is below* l_5*, that means*

$$y_{25} \;=\; \frac{m^+ - x^-}{\delta_+} \;\geqslant\; y^-. \tag{R$_9$}$$

- *Constraints* (1), (3) *and* (6) *are compatible if and only if* B *is above* l_6*, that means*

$$y_{16} \;=\; \frac{m^- - x^+}{\delta_+} \;\leqslant\; y^+. \tag{R$_{10}$}$$

- *Constraint* (2), (3) *and* (7) *are compatible if and only if* A *is above* l_7*, that means*

$$y_{27} \;=\; \frac{x^- - \mu^+}{\delta_-} \;\leqslant\; y^+. \tag{R$_{11}$}$$

- *Constraints* (1), (4) *and* (8) *are compatible if and only if* C *is below* l_8*, that means*

$$y_{18} \;=\; \frac{x^+ - \mu^-}{\delta_-} \;\geqslant\; y^-. \tag{R$_{12}$}$$

Proof.— Criteria associated to relations (R$_1$) to (R$_4$) are obvious. Let us establish criterion associated to relation (R$_5$), criteria associated to (R$_6$) à (R$_{12}$) being obtained similarly.

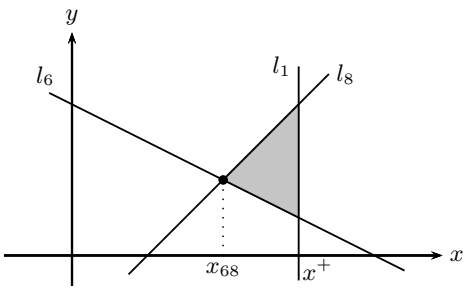

Figure 25 – Graphical illustration of compatibility between constraints (6), (8) and (1)

By the definition of (x_{68}, y_{68}), we can first note that :

$$\frac{m^- - x_{68}}{\delta_+} \;=\; y_{68} \;=\; \frac{x_{68} - \mu^-}{\delta_-},$$

hence :

$$x_{68} \;=\; \frac{\delta_- m^- + \delta_+ \mu^-}{\delta_- + \delta_+}.$$

Now let us suppose that (R_5) is satisfied, i.e. $x_{68} \leqslant x^+$. Then (x_{68}, y_{68}) satisfies (1) by hypothesis, and also (6) et (8) by definition. Reciprocally, let us suppose that there exists $(x, y) \in \mathbb{R}^2$ satisfying (1), (6) and (8). Since (6) and (8) are satisfied, by doing $\delta_-(6) + \delta_+(8)$, we obtain $(\delta_- + \delta_+)x \geqslant \delta_- m^- + \delta_+ \mu^-$, *i.e.*

$$x \;\geqslant\; \frac{\delta_- m^- + \delta_+ \mu^+}{\delta_- + \delta_+} \;=\; x_{68}.$$

And since (1) is also satisfied, we finally obtain $x^+ \geqslant x_{68}$, which precisely corresponds to (R_5). □

4.2.7 — Theorem. *The following conditions are equivalent:*

(*i*) *E is non-empty.*

(*ii*) *Relations* (R_1) *to* (R_{12}) *are satisfied.*

Proof.— Implication (*i*) ⇒ (*ii*) is clear since for all $k \in \{1, \ldots, 8\}$ and indexes $i_1, \ldots, i_k$ such that $1 \leqslant i_1 < \cdots < i_k \leqslant 8$, we have $E_{i_1} \cap \cdots \cap E_{i_k} \neq \varnothing$ as soon as $E = E_1 \cap \cdots \cap E_8 \neq \varnothing$. Reciprocally, let us suppose that (*ii*) is satisfied, and let us prove that (*i*) is also satisfied. In order to do this, we need to find a point $M \in \mathbb{R}^2$ that verifies each of constraints (1) to (8).
We will reason according to the position of point $A' = (x_{68}, y_{68})$ in $\mathbb{R}^2$. As we have seen in the proof of proposition 4.2.6, by using the equations of l_6 and l_8,

$$x_{68} \;=\; \frac{\delta_- m^- + \delta_+ \mu^-}{\delta_- + \delta_+},$$

which implies

$$y_{68} \;=\; \frac{m^- - \mu^-}{\delta_- + \delta_+}.$$

1) Since (R_5) is satisfied, proposition 4.2.6 says that A' must be on the left of l_1. Let Λ_1 be the straight line of equation

$$y \;=\; y^- + \frac{x - x^+}{\delta_-},$$

i.e. the parallel line to l_7 and l_8 passing through point C. We can prove that relation (R_{12}) is also equivalent to say that A' is above Λ_1. Indeed saying that A' is above de Λ_1 means that

$$\frac{m^- - \mu^-}{\delta_- + \delta_+} \;\geqslant\; \frac{1}{\delta_-}\left(\frac{\delta_- m^- + \delta_+ \mu^-}{\delta_- + \delta_+} - x^+\right) + y^-, \quad \text{i.e.} \quad \frac{x^+ - \mu^-}{\delta_-} \;\geqslant\; y^-.$$

which precisely corresponds to relation (R_{12}). By the same way we can prove that relation (R_{10}) means that A' is below Λ_2, the straight line of equation

$$y \;=\; y^+ + \frac{x^+ - x}{\delta_+},$$

which is the parallel to l_5 and l_6 passing through B. Finally, by defining

$$\mathcal{Z} \;=\; \left\{(x, y) \in \mathbb{R}^2 \;\middle|\; x \leqslant x^+,\; y \geqslant \frac{x - x^-}{\delta_-} + y^-,\; y \leqslant \frac{x^+ - x}{\delta_+} + y^+\right\},$$

we have shown that A' must belong to $\mathcal{Z}$.
Obviously when A' belongs to E, it is sufficient to set $M = A'$ to obtain the expected result. Thus in the following, we will suppose that $A' \notin E$. First we have $A' \in E_1$ by the definition of $\mathcal{Z}$. Next we have $\{A'\} = d_6 \cap d_8$ by the definition of A', hence $A' \in E_6 \cap E_8$, and so $A' \in E_5 \cap E_7$ since conditions (R_3) and (R_4) are satisfied. Consequently hypothesis $A' \notin E$ can be rewritten:

$$A' \;\in\; (\mathcal{Z} \smallsetminus E_2) \cup (\mathcal{Z} \smallsetminus E_3) \cup (\mathcal{Z} \smallsetminus E_4).$$

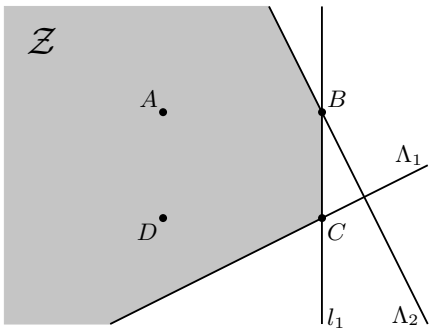

Figure 26 – Representation of $\mathcal{Z}$, the admissible region to point A' in $\mathbb{R}^2$

2) Let us suppose that $A' \in R_1$, with

$$R_1 \;=\; \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; y < y^-,\; y \geqslant \frac{x - x^+}{\delta_-} + y^-,\; y \leqslant \frac{x - x^-}{\delta_-} + y^- \right\}.$$

In practice, R_1 corresponds to the region of $\mathbb{R}^2$ constituted of the points strictly below l_4 the straight line of equation $y = y^-$, above Λ_1, the straight line of equation $y = y^- + (x - x^+)/\delta_-$, and below Λ_3, the straight line of equation $y = y^- + (x - x^-)/\delta_-$ (the parallel to Λ_1 passing through D).
By definition, relation (R_9) means that B' is above l_4. Let M be (x_{48}, y_{48}) the intersection point between l_4, the straight line passing through C and D, and l_8, the straight line passing through A' and B'. Let us show that $M \in E$.

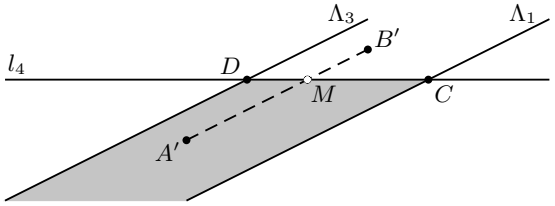

Figure 27 – Point M solution when $A' \in R_1$

- Since $M \in l_4$, we have $y_{48} = y^-$, and since $M \in l_8$, we deduce that $x_{48} = \delta_- y^- + \mu^-$ by using the equation of l_8 given in 4.2.4.
- According to (R_{12}), we have $x^+ - \mu^- \geqslant \delta_- y^-$, hence $x_{48} = \delta_- y^- + \mu^- \leqslant x^+$. Therefore $M \in E_1$.
- Since $A' = (x_{68}, y_{68})$ is supposed to be below Λ_3, then

$$0 \;\leqslant\; \left(y^- + \frac{x_{68} - x^-}{\delta_-} \right) - y_{68} \;=\; y^- + \frac{\mu^- - x^-}{\delta_-},$$

and so $x_{48} = \delta_- y^- + \mu^- \geqslant x^-$. Thus $M \in E_2$.
- By definition, $M \in d_4 \subset E_4$, and we have in particular $M \in E_3$ since (R_2) is satisfied.
- According to (R_8) we have $m^+ - \mu^- - (\delta_- + \delta_+) y^- \geqslant 0$. Therefore

$$\frac{m^+ - x_{48}}{\delta_+} - y_{48} \;=\; \frac{m^+ - \mu^- - (\delta_- + \delta_+) y^-}{\delta_+} \;\geqslant\; 0,$$

which proves that M is below l_5, i.e. $M \in E_5$.

- Since $A' \notin E_2$, we have
$$y^- - \frac{m^- - \mu^-}{\delta_- + \delta_+} > 0,$$
hence $(\delta_- + \delta_+)y^- + \mu^- - m^- > 0$. Therefore
$$y_{48} - \frac{m^- - x_{48}}{\delta_+} = \frac{(\delta_- + \delta_+)y^- + \mu^- - m^-}{\delta_+} > 0,$$
which proves that M is (strictly) above l_6, and so belong to E_6.
- By definition, $M \in d_8 \subset E_8$, and we have in particular $M \in E_7$ since (R_4) is satisfied.

Consequently under the hypotheses (*ii*) and $A' \in R_1$, we have $M \in E$.

3) Let us suppose that $A' \in R_2$, with
$$R_2 = \left\{ (x,y) \in \mathbb{R}^2 \;\middle|\; x < x^-,\ y \geqslant \frac{x - x^-}{\delta_-} + y^-,\ y \leqslant \frac{x - x^-}{\delta_-} + y^+ \right\}.$$
In practice, R_2 corresponds to the region of $\mathbb{R}^2$ constituted of the points stricly on the left of l_2, the straight line of equation $x = x^-$, above Λ_3, the straight line of equation $y = y^- + (x - x^-)/\delta_-$, and below Λ_4, the straight line of equation $y = y^+ + (x - x^-)/\delta_-$.
If $B' \in E_2$, we can show by using (R_1) and (R_4) that $M = (x_{28}, y_{28})$, the intersection point between l_2 (the straight line passing through A and D) and l_8 (the straight line passing through A' and B'), belongs to E. But if $B' \notin E_2$, we can show by using (R_3) and (R_9) that $M = (x_{25}, y_{25}) \in E$.

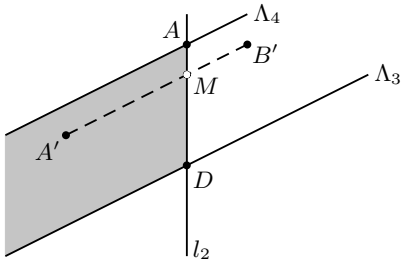

Figure 28 – Point M solution $A' \in R_2$ and $B' \in E_2$

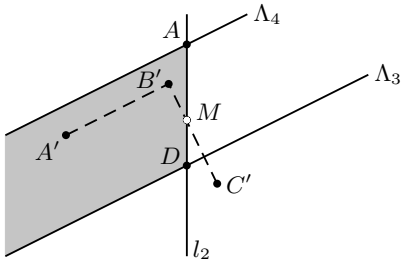

Figure 29 – Point M solution when $A' \in R_2$ and $B' \notin E_2$

4) Let us suppose that $A' \in R_3$, with

$$R_3 \;=\; \left\{(x,y)\in\mathbb{R}^2 \;\middle|\; y \geqslant \frac{x-x^-}{\delta_-}+y^-,\; y\leqslant \frac{x^- - x}{\delta_+}+y^+\right\}.$$

In practice, R_3 corresponds to the region of $\mathbb{R}^2$ constituted of the points above Λ_4, the straight line of equation $y = y^- + (x-x^-)/\delta_-$, and below Λ_5, the straight line of equation $y = y^+ + (x^- - x)/\delta_+$. Here again we need to distinguish two cases. If B' is below Λ_5, knowing that (R_3), (R_6) and (R_9) are satisfied, we can prove that $M = (x_{25}, y_{25})$, the intersection point between l_2 and l_5, belongs to E. But if B' is above Λ_5, knowing that (R_1), (R_2) and (R_{11}) are satisfied, we can prove that $A = (x^-, y^+)$ belongs to E.

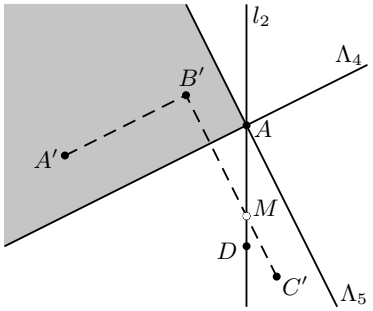

Figure 30 – Point M solution when $A' \in R_3$ and B' is below Λ_4

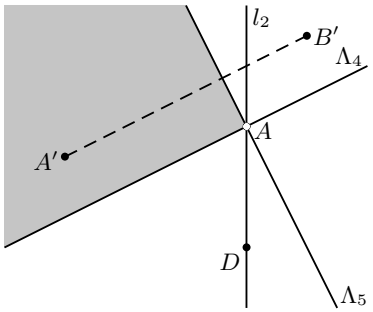

Figure 31 – A is solution when $A' \in R_3$ and B' is above Λ_4

5) Let us end by considering that $A' \in R_4$, with

$$R_4 \;=\; \left\{(x,y)\in\mathbb{R}^2 \;\middle|\; y > y^+,\; y \geqslant \frac{x^- - x}{\delta_+}+y^+,\; y\leqslant \frac{x^+ - x}{\delta_+}+y^+\right\}.$$

In practice, R_4 corresponds to the region of $\mathbb{R}^2$ constituted by the points strictly above l_3, the straight line of equation $y = y^+$, above Λ_4, the straight line of equation $y = y^+ + (x^- - x)/\delta_+$, and below Λ_2, the straight line of equation $y = y^+ + (x^+ - x)/\delta_+$.
In this case, knowing that (R_2), (R_3), (R_7) and (R_{10}) are satisfied, we can prove that $M = (x_{36}, y_{36})$, the intersection point between l_3 and l_6, belongs to E.

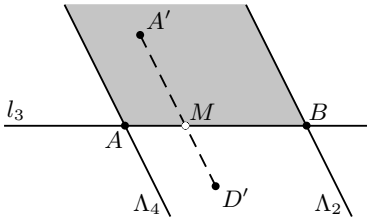

Figure 32 – Point M solution when $A' \in R_4$

In the end, having

$$(\mathcal{Z} \setminus E_2) \cup (\mathcal{Z} \setminus E_3) \cup (\mathcal{Z} \setminus E_4) \;=\; R_1 \cup R_2 \cup R_3 \cup R_4,$$

we deduce of points 1) to 5) that implication $(ii) \Rightarrow (i)$ is also true. □

4.2.8 — Let us assume that E is non-empty. According to theorem 4.2.7, that means that relations (R_1) to (R_{12}) are satisfied. From now on, we will make explicit the bounds of the intervals that define P^*, the smallest rectangle of $\mathbb{R}^2$ that contains E. So let us fix $(x, y) \in E$.

- By doing (5) $-\ \delta_+$(4), we can note that

  $$x \;\leqslant\; x_{45} \;=\; m^+ - \delta_+ y^-.$$

  Similarly, by doing (7) $+\ \delta_-$(3), we can note that

  $$x \;\leqslant\; x_{37} \;=\; \mu^+ + \delta_- y^+.$$

  Finally, the linear combination δ_-(5) $+\ \delta_-$(7) leads to $(\delta_- + \delta_+)x \leqslant \delta_- m^+ + \delta_+ \mu^+$, from which we deduce

  $$x \;\leqslant\; x_{57} \;=\; \frac{\delta_- m^+ + \delta_+ \mu^+}{\delta_- + \delta_+}.$$

  And since $x \leqslant x^+$ according to (1), by setting

  $$x_+^* \;=\; \min\left(x^+, x_{37}, x_{45}, x_{57}\right),$$

  we have obtained $x \leqslant x_+^*$.

- By the same way, by setting

  $$x_-^* \;=\; \max\left(x^-, x_{36}, x_{48}, x_{68}\right),$$

  where

  $$x_{36} \;=\; m^- - \delta_+ y^+, \quad x_{48} \;=\; \delta_- y^- + \mu^- \quad \text{and} \quad x_{68} \;=\; \frac{\delta_- m^- + \delta_+ \mu^-}{\delta_- + \delta_+},$$

  we can prove that $x \geqslant x_-^*$.

- By doing (5) $-\ \delta_+$(2), we obtain

  $$y \;\leqslant\; y_{25} \;=\; \frac{m^+ - x^-}{\delta_+}.$$

  Similarly, relation (8) $-\ \delta_-$(1) leads to

  $$y \;\leqslant\; y_{18} \;=\; \frac{x^+ - \mu^-}{\delta_-}.$$

Finally the difference (5) − (8) provides $(\delta_- + \delta_+)y \leqslant m^+ - \mu^-$, hence

$$y \;\leqslant\; y_{58} \;=\; \frac{m^+ - \mu^-}{\delta_- + \delta_+}.$$

And since $y \leqslant y^+$ according to (3), by setting

$$y_+^* \;=\; \min\left(y^+, y_{18}, y_{25}, y_{58}\right),$$

we can conclude that $y \leqslant y_+^*$.

- By the same way, if

$$y_-^* \;=\; \max\left(y^-, y_{16}, y_{27}, y_{67}\right),$$

where

$$y_{16} \;=\; \frac{m^- - x^+}{\delta_+}, \quad y_{27} \;=\; \frac{x^- - \mu^+}{\delta_-} \quad \text{and} \quad y_{67} \;=\; \frac{m^- - \mu^+}{\delta_- + \delta_+},$$

we can prove that $y \geqslant y_-^*$.

In the end, we have established the following result:

Lemma. *If E is non-empty, then it is contained in the rectangle $\Pi = [x_-^*, x_+^*] \times [y_-^*, y_+^*]$.*

4.2.9 — Let us take the notations of x_-^*, x_+^*, y_-^*, y_+^* and Π introduced in 4.2.8 back.

Theorem. *If E is non-empty, then P^* the smallest rectangle from $\mathbb{R}^2$ that contains E is equal to Π.*

Proof.— Let us assume that E is non-empty. Then by theorem 4.2.7, conditions (R_1) to (R_{12}) are satisfied, and according to lemma 4.2.8, we have already proved that Π is a rectangle of $\mathbb{R}^2$ that contains E, *i.e.* that $P^* \subset \Pi$. To establish the theorem, it is now sufficient to determine:

– for all $x \in \{x_-^*, x_+^*\}$, an element $y \in \mathbb{R}$ such that $(x, y) \in E$, and,

– for all $y \in \{y_-^*, y_+^*\}$, an element $x \in \mathbb{R}$ such that $(x, y) \in E$.

We will only do it for $x = x_+^*$, the other cases being obtained similarly.

1) Let us assume that $x_+^* = x_{37}$, and let us show that $M = (x_{37}, y_{37}) = (\delta_- y^+ + \mu^+,\ y^+)$ belongs to E.

- Since $x_+^* = x_{37}$, then we have $x_{37} = x_+^* \leqslant x^+$ by the definition of x_+^*, and so $M \in E_1$.
- Since (R_{11}) is equivalent to $x^- \leqslant \delta_- y^+ + \mu^+ = x_{37}$, we deduce that $M \in E_2$.
- We have $M \in l_3 \subset E_3$ by definition. In particular we also have $M \in E_4$ according to (R_2).
- Since $x_+^* = x_{37}$, we have $x_{37} \leqslant x_{57}$. And since

$$x_{37} \;\leqslant\; x_{57} \;=\; \frac{\delta_- m^+ + \delta_+ \mu^+}{\delta_- + \delta_+} \quad\Longleftrightarrow\quad x_{37} \;\leqslant\; m^+ - \delta_+ y^+ \quad\Longleftrightarrow\quad y_{37} \;=\; y^+ \;\leqslant\; \frac{m^+ - x_{37}}{\delta_+},$$

  we can note that M is below l_5, that means $M \in E_5$.
- We can easily show that condition (R_7) precisely means that M is above l_6, hence $M \in E_6$.
- We have $M \in d_7 \subset E_7$. In particular, we also have $M \in E_8$ according to (R_4).

Consequently, we have already proved that $M \in E$.

2) If $x_+^* = x_{45}$, by using the relations (R_2), (R_3), (R_8), (R_9) and inequalities $x_{45} \leqslant x^+$ and $x_{45} \leqslant x_{57}$, that are implied by the hypothesis $x_+^* = x_{45}$, we can demonstrate as in point 1) that

$$(x_{45}, y_{45}) \;=\; (m^+ - \delta_+ y^-,\; y^-) \;\in E.$$

3) If $x_+^* = x_{57}$, by using conditions (R_2), (R_6), (R_{11}) and inequalities $x_{57} \leqslant x^+$, $x_{57} \leqslant x_{37}$ and $x_{37} \leqslant x_{45}$, that are implied by the hypothesis $x_+^* = x_{57}$, then we can prove as in point 1) that

$$A' \;=\; (x_{57}, y_{57}) \;=\; \left(\frac{\delta_- m^+ + \delta_+ \mu^+}{\delta_- + \delta_+}, \frac{m^+ - \mu^+}{\delta_- + \delta_+} \right) \;\in\; E.$$

4) To conclude, we need to suppose that $x_+^* = x^+$. This case will be treated a little differently from previous cases. Let us define

$$\left\{ \begin{array}{rcl} \beta^+ & = & \min(y^+, y_{15}, y_{18}), \\ \beta^- & = & \max(y^-, y_{16}, y_{17}), \end{array} \right.$$

where

$$y_{15} \;=\; \frac{m^+ - x^+}{\delta_+}, \quad y_{16} \;=\; \frac{m^- - x^+}{\delta_+}, \quad y_{17} \;=\; \frac{x^+ - \mu^+}{\delta_-} \quad \text{and} \quad y_{18} \;=\; \frac{x^+ - \mu^-}{\delta_-}.$$

- We first prove that $\beta^- \leqslant \beta^+$.
  - We have $y^- \leqslant y^+$ according to (R_2), but also $y_{16} \leqslant y^+$ according to (R_{10}). Finally, since $x_+^* = x^+$, we deduce that $x^+ \leqslant x_{37} = \delta_- y^+ + \mu^+$, which is equivalent to $y_{17} \leqslant y^+$.
  - Since $x_+^* = x^+$, we obtain $x^+ \leqslant x_{45}$, and so $y^- \leqslant y_{15}$. Inequality $y_{16} \leqslant y_{15}$ immediately results from (R_2). Finally since $x_+^* = x^+$, we also have $x^+ \leqslant x_{57}$, which is equivalent to $y_{17} \leqslant y_{15}$.
  - Condition (R_{12}) implies that $y^- \leqslant y_{18}$, and (R_2) implies that $y_{17} \leqslant y_{18}$. Finally by using (R_5), we can demonstrate that $y_{17} \leqslant y_{18}$.

  For all $u \in \{y^-, y_{16}, y_{17}\}$ et $v \in \{y^+, y_{15}, y_{18}\}$, we have proved that $u \leqslant v$. Hence $\beta^- \leqslant \beta^+$.
- Now let us fix $y \in [\beta^-, \beta^+]$, and let us consider $M = (x^+, y)$. We have $M \in l_1 \subset E_1$, and also $M \in E_2$ since (R_1) is satisfied. In addition we have $M \in E_3 \cap E_5 \cap E_8$ (resp. $M \in E_4 \cap E_6 \cap E_7$) according to inequality $y \leqslant \beta^+$ (resp. $y \geqslant \beta^-$). Therefore $M \in E$.

In any case, we have been able to determine $y \in \mathbb{R}$ such that $(x_+^*, y) \in E$. Hence the result. □

4.2.10 — Let us summarize the study leads in the previous paragraphs. According to theorem 4.2.7, if conditions (R_1) to (R_{12}) are satisfied, then E the set of couples (x, y) from $\mathbb{R}^2$ satisfying the constraints (1) to (8) from 4.2.3 is non-empty. When it happens, theorem 4.2.9 insures that bounds x^- and x^+ on x and y^- and y^+ on y can be corrected by doing:

$$x^- \;\leftarrow\; x_-^*, \quad x^+ \;\leftarrow\; x_+^*, \quad y^- \;\leftarrow\; y_-^* \quad \text{and} \quad y^+ \;\leftarrow\; y_+^*,$$

quantities $x_\pm^*$ and $y_\pm^*$ having been defined in 4.2.8.
When it is possible to do it with $x = \psi^{(k)}(t)$ and $y = \psi^{(k+1)}(t)$, then we obtain new corrections on bounds on $\psi^{(k)}(t)$ and $\psi^{(k+1)}(t)$. These corrections are still associated to a *forward-backward* process since as seen on the example 4.2.1, constrains of type (1) to (8) on $x = \psi^{(k)}(t)$ and $y = \psi^{(k+1)}(t)$ are in practice obtained by applying the Taylor-Lagrange formula on ψ at points lower and upper that t.
Now an interesting work would consist in implementing numerically these new forward-backward corrections, in order to evaluate their efficiency and comparing it with the one obtained with our classical forward-backward corrections from chapter 2. It is also important to signal that if the classical forward-backward corrections are efficient when they are used with sufficiently close data, nothing says that it will be the same with this new kind corrections. Unfortunately for a lack of time, we have essentially lead this theoretical study, but we have not yet taken the time to do enough numerical simulations.

4.3 In higher dimension

4.3.1 — In this section, we fix $n \in \mathbb{N} \setminus \{0,1\}$, and we will see how it is possible to generalize in a very elementary way the models and algorithms presented in the previous chapters to functions defined on I and valued in $\mathbb{R}^n$. It is well-known that there is no general hypotheses that makes Taylor-Lagrange formula true to functions valued in $\mathbb{R}^n$. For instance, if $g : \mathbb{R} \to \mathbb{R}^2,\ x \mapsto (\cos x, \sin x)$, then for all $\xi \in]0, 2\pi[$ we have

$$g(2\pi) - g(0) \ = \ (0,0) \ \neq \ 2\pi(-\sin\xi, \cos\xi) \ = \ (2\pi - 0)g'(\xi).$$

However if $g : I \to \mathbb{R}^n$ has a m^{th} derivative, $x \in I$ and $h \in \mathbb{R}^*$ satisfies $x+h \in I$, for all $j \in \{1, \ldots, n\}$, it is still possible to apply the Taylor-Lagrange formula to $g_j : I \to \mathbb{R}$, hence the existence of ξ_j strictly between x and $x + h$ such that

$$g_j(x+h) \ = \ \left(\sum_{k=0}^{m-1} \frac{h^k}{k!}\, g_j^{(k)}(x) \right) + \frac{h^m}{m!}\, g_j^{(m)}(\xi_j).$$

Therefore theories from the previous chapters can be extended to functions from I to $\mathbb{R}^n$ by working component by component. Thus to solve the problem presented in introduction of this part with a function $f : I \to \mathbb{R}^n$ satisfying for instance analogous hypotheses than those done in chapter 3, that means:

- for all $i \in \{1, \ldots, N\}$, $f(t_i)$ is bounded,
- f'' and f''' are uniformly bounded,

we can concretely apply the treatment presented in sections 3.2 or 3.5 on each component of f. Let us highlight that in doing this, the various components of f can be handled independently, hence a possible parallelization of the various computations.

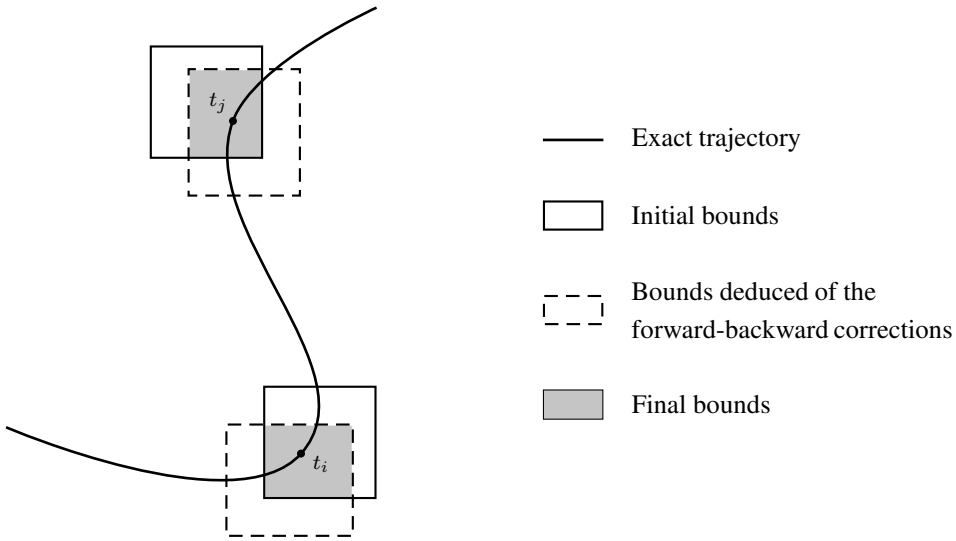

Figure 33 – Example of forward-backward corrections when $n = 2$

4.3.2 — EXAMPLE. We have tested the effect of the extension suggested in 4.3.1 on a very simple example. Denoting by f_2 the function f considered in chapter 3 for our one-dimensional simulations, we have adapted them to solve a problem in dimension $n = 2$ on the function f given by:

$$f \ : \ \mathbb{R} \to \mathbb{R}^2, \ t \mapsto (t, f_2(t)).$$

By generating analogous input data on $f_1 : t \mapsto t$ and using precisely the same as those used in chapter 3 for f_2, we have noted similar performances in comparison to those observed one the one-dimensional case: equivalent computation time and improvements on each component of f. In particular we have obtained exactly the same performances on f_2, that illustrates the independence of the treatment done component by component.

4.3.3 — The extension in higher dimension presented in this section suggests to equip $\mathbb{R}^n$ of the norm

$$\|\cdot\|_\infty \ : \ \mathbb{R}^n \to \mathbb{R}_+, \ x \mapsto \max\{|x_j| \mid 1 \leqslant j \leqslant n\},$$

which is always possible since all the norms on $\mathbb{R}^n$ are equivalent. If the initial bounds on f or $f^{(d)}$ are expressed in an other norm of $\mathbb{R}^n$, it can require a pejoration of our initial data, which could be nevertheless make up by our treatment. An additional work but not lead until now (since we were only considering a unidimensional problem) could consist in extending our models by taking the initial norm (associated to the input data) on $\mathbb{R}^n$.

REMARK. More generally, for functions valued in $\mathbb{R}^n$, our goal will consist in reducing the certified domains (in the sense of inclusion) that bounds f and its derivatives in $t_1, \ldots, t_N$. Whatever the shape of these domains, we can thus seek to reduce the diameter of each one of these domains. When $n = 1$ in the theory presented above, these domains are intervals and we have tried to reduced their diameter: it justifies the terminology of *diameter functions* adopted in chapter 1.

Conclusion

There are probably many other possible generalizations or extensions to the works presented in chapters 1 to 3. We have only presented here those that we had already considered, or that we thought were the most essential. If the ones presented in sections 4.1 and 4.2 seem now relatively easy to test in practice, the generalization of our models in higher dimension mentioned in section 4.3 could be much more difficult, both theoretically and numerically.

Conclusion et perspectives

L'étude menée dans cette partie nous a permis de voir comment, à partir de bornes en certains points sur une application (inconnue en pratique) $f : I \to \mathbb{R}^n$, où $n \in \mathbb{N}^*$, et de bornes uniformes sur au moins l'une de ses dérivées d-ième, il était possible de borner les applications f', ..., $f^{(d-1)}$ d'une part, et mettre en cohérence différentes bornes sur f, f', ..., $f^{(d-1)}$ d'autre part. En ce sens, notre travail se révèle particulièrement adapté aux situations où l'on ne parvient pas à modéliser mathématiquement, que ce soit de façon totale ou partielle, le phénomène ou le système étudié. Dans notre cas, nous souhaitions estimer les position, vitesse et accélération d'un train sans avoir une connaissance suffisante de ses caractéristiques techniques (mécanismes d'accélération ou de freinage) ; nous pouvions seulement nous fier à des bornes ponctuelles sur sa position, et des bornes uniformes sur son accélération et son jerk.
Rappelons que notre approche, à notre connaissance novatrice en la matière, est totalement déterministe. À cet égard, elle se distingue des méthodes probabilistes (e.g. filtrage de Kalman), souvent inadaptées lorsque qu'un très haut niveau de sureté est requis, comme c'était le cas pour les applications que nous considérions. Par ailleurs, nos modèles et algorithmes peuvent être sans crainte couplés à d'autres existant sur le sujet, sans que cela risque d'impacter le niveau de fiabilité, et peuvent en outre être utilisés pour détecter certaines incohérences dans les jeux de données traités.
En combinant les différents algorithmes que nous avons conçus, nous avons résolu une problématique inspiré du monde ferroviaire. Nous avons à cet effet illustré l'efficacité de nos modèles, et montré qu'ils permettent de réaliser tant des analyses en temps réel qu'*a posteriori*.

Si l'on peut déjà se satisfaire de ces premiers résultats encourageants, de nombreuses pistes permettant d'améliorer, de généraliser ou d'appliquer notre travail nous semblent intéressantes à considérer. Donnons-en quelques unes parmi bien d'autres :

- Au chapitre 1, nous avons proposé différentes façons de borner les dérivées k-ièmes d'une fonction ψ dérivable $p = 2$ ou $p = 3$ fois, pour $0 < k < p$. Une suite logique consisterait à mener ce travail pour des entiers k et p quelconques, le cas $k = 0$ ayant été évoqué à la section 4.1.

- Toujours au chapitre 1, en comparant les avantages et incovénients des formules centrées et décentrées, nous avons montré qu'il existe une concession entre leur précision et leur possible utilisation en temps réel. Une autre perspective consisterait à trouver récupérer des bornes aussi précises que celles déduites des formules centrées, mais qui puissent être raisonnablement utilisées en temps réel. Nous avions à cet égard tenté de nous inspirer de ce que permettent les schémas compacts de S. K. LELE dans [13] pour la résolution d'équations différentielles, mais leur adaptation à notre étude ne semble pas si aisée.

- Une extension des corrections forward-backward proposées au chapitre 2 a été considérée à la section 4.2. Il pourrait être intéressant d'étudier en détail l'efficacité de ces nouvelles corrections, en particulier par rapport à celles du chapitre 2, ou bien encore de trouver d'autres modèles similaires permettant de mettre en cohérences différentes bornes sur une fonction et ses dérivées.

- Nos modèles ont été originellement développées pour des fonctions à valeurs réelles. À la section 4.3, nous avons expliqué comment les appliquer très simplement à des applications à valeurs dans $\mathbb{R}^n$, pour

un entier n arbitrairement grand, en équipant $\mathbb{R}^n$ de la norme infinie. Il nous semblerait désormais important de trouver des méthodes qui prennent en compte la topologie initiale des données sur $\mathbb{R}^n$ à traiter, et plus généralement d'étendre tout cela à des applications de $\mathbb{R}^p$ dans $\mathbb{R}^q$, pour p et q choisis quelconques dans $\mathbb{N}^*$.

- Nous évoquions enfin le possible couplage de nos modèles à d'autres. Signalons à cet égard une possible application à certaines techniques de machine learning, qui essaient de retrouver les équations régissant un système dynamique spatio-temporel de la forme :

  $$\partial_t u \;=\; F(u, \partial_x u, \ldots, x, \theta),$$

  Ici, u désigne une fonction, l'inconnue du système, dépendant d'une variable temporelle t et d'une variable d'espace x connu ou partiellement, F une fonctionnelle non nécessairement connue, et θ un paramètre. Pour parvenir à leurs fins, ils se basent sur diverses observations de u, et éventuellement $\partial_t u$, $\partial_x u$, etc. à disposition. Pour de plus amples informations, nous renvoyons le lecteur à l'article [2] d'I. AYED, E. de BÉZENAC, A. BRAJARD et P. GALLINARI.
  À partir de l'information disponible sur u, et l'une de ses dérivées dans le meilleur des cas, ou, le cas échéant, sur certains processus liés à la dynamique du système, nos modèles permettraient sans doute d'obtenir plus d'information sur les dérivées de u, puis raffiner toute les informations sur u et ses dérivées ainsi récupérées. Ainsi, les équations obtenues par ces techniques de machine learning expliqueraient d'autant mieux le système ou phénomène étudié.

Bien entendu, il existe sans doute bien d'autres applications, adaptations ou améliorations de nos travaux ; nous en proposerons d'ailleurs de nouvelles à l'issue de l'étude menée dans le partie II de ce me manuscrit. Mais insistons sur le fait qu'en pratique, celles-ci dépendront avant tout du système ou phénomène étudié, et nous seront suggérées par la nature des données à disposition, ou encore par un retour d'expérience.

II

Theorical and numerical study of a shape from shading problem

Introduction

Notations et conventions spécifiques

La plupart des notations ou conventions spécifiques à cette étude seront spécifiées tout au long du texte, ou pour les plus traditionnelles d'entre elles, à la fin de ce document. Nous ne préciserons donc ici que les plus essentielles d'entre elles.
Sauf mention explicite du contraire, étant donné $d \in \mathbb{N}^*$:

- Nous noterons abusivement 0 le vecteur nul de $\mathbb{R}^d$.
- La famille $(e_1, \dots, e_d)$ désignera la base canonique de $\mathbb{R}^d$.
- Si $x \in \mathbb{R}^n$ et $j \in \{1, \dots, d\}$, nous noterons généralement x_j ou $x(j)$ la j-ième composante de x, de sorte que $x = (x_1, \dots, x_d)$ ou $x = \big(x(1), \dots, x(d)\big)$. Dans quelques rares cas, nous préférerons noter celle-ci x_{j-1} ou $x(j-1)$, de sorte que $x = (x_0, \dots, x_{d-1})$ ou $x = \big(x(0), \dots, x(d-1)\big)$.
  Bien entendu, nous nous efforcerons de chasser toute ambiguïté à cet égard durant cette étude.
- Si $x, y \in \mathbb{R}^n$, nous noterons $x{\cdot}y = x_1y_1 + \cdots + x_ny_n$ le produit scalaire usuel de x et y, et $|x| = \sqrt{x{\cdot}x}$ la norme euclidienne de x.
- $\mathbb{R}^d$ sera supposé muni de sa structure usuel d'espace euclidien, et, au besoin, toute partie de $\mathbb{R}^d$ sera vue comme un sous-espace métrique de $\mathbb{R}^d$. Pour toute partie A de $\mathbb{R}^d$, nous noterons alors $\overline{A}$ l'adhérence de A dans $\mathbb{R}^d$, et ∂A sa frontière dans $\mathbb{R}^d$.
- Nous noterons respectivement $\mathrm{B}(0,1)$ et $\mathrm{B}'(0,1)$ les boules unités ouverte et fermée de $\mathbb{R}^d$.

De plus, si X est une partie de $\mathbb{R}^d$:

- Nous noterons respectivement $\mathcal{C}(X)$ (resp. $\mathcal{C}^1(X)$, $\mathcal{C}^\infty(X)$) l'ensemble des fonctions continues (resp. différentiables à différentielles continues, indéfiniment différentiables) de X dans $\mathbb{R}$.
- Nous noterons $\mathcal{B}(X)$ l'ensemble des fonctions continues et bornées de X dans $\mathbb{R}$. Pour u et v deux fonctions de $\mathcal{B}(X)$, nous écrirons en outre $u \leqslant v$ dès que $u(x) \leqslant v(x)$ pour tout $x \in X$.

Enfin, toutes les simulations présentées dans cette partie ont été réalisées sous `python`, et ce même si nombre de sorties graphiques ont été reproduites, pour des raisons esthétiques, sous LaTeX.

Contexte de l'étude, organisation et contributions

Dans cette partie, on s'intéresse à la résolution d'un problème de *shape from shading* (abrégé en s.f.s. dans ce qui suit), qui consiste à reconstituer la surface représentée sur une image en noir et blanc, par la seule connaissances des différentes nuances de gris et, bien entendu, de la position exacte de certains

points de la surface. Une première modélisation mathématique de ce problème à partir d'une équation différentielle non linéaire a été proposée par B. K. HORN dans [9] en 1975. Depuis, de nombreux auteurs ont considéré ce problème et diverses approches ont été proposées.
Notre étude est essentiellement basée sur les travaux initialement menés par P.-L. LIONS, E. ROUY et A. TOURIN, par exemple dans [16] ou [19], et plus récemment repris par E. PRADOS et O. FAUGERAS, par exemple dans [17] ou [18]. Dans ce cas, la solution du problème de s.f.s. sera associée à la solution de viscosité d'une équation de Hamilton-Jacobi stationnaire du premier ordre, et nous tenterons de fait d'en déterminer une approximation. En l'état, nos contributions consistent en un bref état de l'art théorique et numérique de cette approche du problème. Nous proposons également une formulation explicite d'un schéma d'approximation considéré comme implicite, et nous effectuons, code `python` à l'appui, une optimisation de certains algorithmes utilisés jusqu'alors. De façon plus précise :

- La littérature au sujet des solutions de viscosité étant généralement très spécialisée, nous avons souhaité, au chapitre 5, rappeler les définitions et résultats théoriques nécessaires à l'étude du problème s.f.s., en proposant parfois même des preuves directes de résultats établis dans des cadres plus généraux ailleurs (voir par exemple le résultat d'unicité de la section 5.3). Pour la compréhension, nous mettons cette théorie en œuvre sur des exemple simples, et nous montrons comment l'appliquer à notre problème.

- Les chapitres suivants sont consacrés à la résolution numérique du problème de s.f.s. Nous y reprenons pour l'essentiel un schéma d'approximation, qualifié d'implicite par E. PRADOS et O. FAUGERAS dans [18], et nous montrons comment le rendre explicite, dans un cadre unidimensionnel au chapitre 6, puis dans un cadre bi-dimensionnel au chapitre 7. Dans les deux cas, nous proposons un algorithme permettant une résolution effective, et manifestement bien plus rapide, du problème de s.f.s. Nous présentons enfin diverses simulations numériques illustrant les performances de notre algorithme. Dans un soucis de reproductibilité, l'intégralité du code `python` permettant de reproduire nos simulations est mis en libre accès (à ce jour, nos fichiers sont livrés en l'état, mais un notebook `jupyter` est en préparation).
  Si le problème de s.f.s. et l'équation associée sont traditionnellement étudiés en dimension 2, la seule que l'on considère d'un point de vue pratique, l'étude spécifique du cas unidimensionnel permet de fixer les idées et de mieux comprendre comment celles-ci se généralisent en dimension supérieure. C'est pourquoi il nous a semblé utile de consacrer un chapitre entier à l'étude de ce problème en dimension 1.

Bien entendu, et même si ne nous les évoquerons pas ici, il existe bien d'autres manières d'aborder et de résoudre un tel problème. À cet égard, signalons par exemple les méthodes dites de *fast marching*, présentées par J. A. SETHIAN dans [20], puis appliquées au problème de s.f.s. par et R. KIMMEL et J. A. SETHIAN dans [12]. Ou bien encore les travaux de B. WU, W. C. LIU, A. GRUMPE et C. WÔHLER qui, dans [21], considèrent des modèles prenant en compte la variation de certains paramètres physiques (*e.g.* l'albedo).

Formulation du problème

Comme nous l'avons précédemment expliqué, notre objectif consiste à reconstituer la surface représentée sur une image en noir et blanc, par la seule connaissances des différentes nuances de gris et de la position exacte de certains points de la surface, et ce dans un cadre tant unidimensionnel que bidimensionnel. En ce sens, si l'image correspond à une traditionnelle photographie en noir et blanc en dimension 2, nous pourrons l'assimiler à une "bande grisée" en dimension 1.

D'un point de vue mathématique, fixant $n \in \{1, 2\}$:

- Nous supposerons que la surface $\mathcal{S}$ observée sur l'image peut être décrite comme suit :

$$\mathcal{S} \;=\; \big\{\big(x, u(x)\big) \;\big|\; x \in \overline{\Omega}\big\}\,,$$

  où Ω désigne un ouvert connexe et borné de $\mathbb{R}^n$ (physiquement assimilé à la région photographiée) et u une fonction bornée de $\overline{\Omega}$ dans $\mathbb{R}$ (correspondant physiquement à la l'altitude).

- L'espace sera supposé éclairé par une unique source lumineuse ponctuelle et située au dessus de $\mathcal{S}$. Ainsi, à tout point $\big(x, u(x)\big)$ de $\mathcal{S}$, où $x \in \overline{\Omega}$, il est d'associer un unique vecteur $L(x)$ *unitaire* et *orienté vers la source lumineuse*.

- L'intensité lumineuse en un point $x \in \overline{\Omega}$ sera modélisée par une quantité $I(x)$ comprise entre 0 et 1, d'où une fonction $I : \overline{\Omega} \to [0, 1]$. Très concrètement, I mesurera le niveau de gris sur l'image, le noir correspondant à une intensité nulle (zone d'ombre) et le blanc à une intensité égale à 1.

Sous ces conditions, et en utilisant un modèle physique usuel (scène lambertienne et d'albédo constant égal à 1), l'*équation de la brillance* (*brightness equation*) stipule que pour tout $x \in \overline{\Omega}$, l'intensité lumineuse $I(x)$ de $\mathcal{S}$ au point $\big(x, u(x)\big)$ est décrite par la relation :

$$I(x) \;=\; \max\left(0,\; \frac{n(x)\cdot L(x)}{|n(x)|}\right) \;=\; \begin{cases} \cos\alpha(x) & \text{si } \dfrac{-\pi}{2} \leqslant \alpha(x) \leqslant \dfrac{\pi}{2}, \\ 0 & \text{sinon,} \end{cases}$$

$n(x) = \big(n_1(x), n_2(x), n_3(x)\big)$ désignant un vecteur normal à $\mathcal{S}$ au point $\big(x, u(x)\big)$ tel que $n_3(x) \geqslant 0$, et $\alpha(x)$ l'angle entre les vecteurs $L(x)$ et $n(x)$.

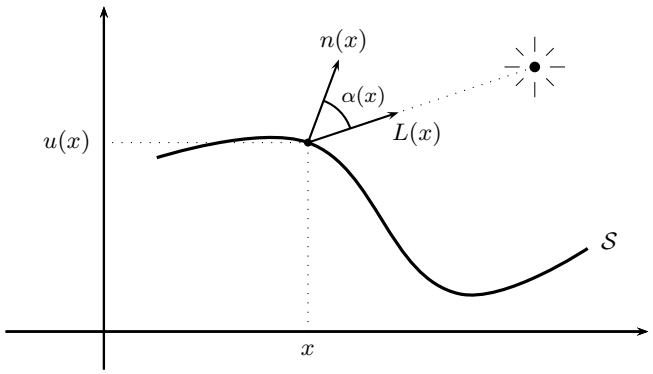

Figure 34 – Représentation de $\mathcal{S}$, $L(x)$ et $n(x)$ en dimension 1

En pratique, nos scènes seront éclairées par le soleil. Dans ce cas, il est raisonnable de supposer la source lumineuse située à l'infini, et donc qu'il existe un unique triplet $(l, l_3) \in \mathbb{R}^n \times \mathbb{R}$, avec $l_3 > 0$, tel que $L(x) = (l, l_3)$ pour tout $x \in \overline{\Omega}$. De fait, sachant que pour tout $x \in \overline{\Omega}$, un vecteur normal à $\mathcal{S}$ en $\big(x, u(x)\big)$ est donné par $(-\nabla u(x), 1)$, la précédente équation se réécrit :

$$I(x) \;=\; \max\left(0,\; \frac{-\nabla u(x)\cdot l + l_3}{\sqrt{1 + |\nabla u(x)|^2}}\right). \tag{II.1}$$

Pour reconstituer la surface $\mathcal{S}$ recherchée, tout revient donc à déterminer la fonction u qui la décrit. Au regard de (II.1), connaissant $I : \overline{\Omega} \to [0,1]$, il est possible de récupérer de l'information sur u à partir de l'équation :

$$I(x) \;=\; \frac{-\nabla u(x)\cdot l + l_3}{\sqrt{1+|\nabla u(x)|^2}}. \tag{II.2}$$

Ce faisant, si u satisfait à (II.2), elle satisfait en particulier à (II.1), mais le profil de u obtenu par résolution de (II.2) ne sera pas forcément fidèle à la réalité. Signalons qu'il s'agit là d'une limitation uniquement dû au modèle physique : en effet, lorsque I s'annule, (II.1) peut-être satisfaite pour différents profils de u.
Dans cette seconde partie de manuscrit, la surface $\mathcal{S}$ sera donc reconstruite par résolution de l'équation simplifiée (II.2). Comme nous le verrons dès le chapitre 5, il s'agit d'une équation de Hamilton-Jacobi du premier ordre, mal posée en l'état : non unicité flagrante, existence de solutions généralement non classiques, etc. Nous y montrerons en particulier comment le formalisme des solutions de viscosité permet, en partie, de palier ces difficultés.

Chapter 5

Generalities about Hamilton-Jacobi equations

Introduction

In this chapter, we remind various results about the Hamilton-Jacobi equations. To begin with, a definition and first examples of these (ill-posed) equations are given in section 5.1. Then we introduce in section 5.2 the notion of *viscosity solutions* in the continuous case. In sections 5.3 and 5.4, we state uniqueness and existence results, by giving a direct proof of this first one, and we apply all this to the s.f.s. problem in section 5.5. We shortly explain how to extend this to the discontinuous case in section 5.6. Finally in section 5.7, we will introduce theoritically the approximation schemes that will be used in the following chapters in order to solve numerically the s.f.s. problem in one-dimension as well as in two-dimension.

5.1 Definitions and first observations

5.1.1 — From a formal point of view, a first order *Hamilton-Jacobi equation* is an equation of unknown $u : \Omega \to \mathbb{R}$ that can be written:

$$\forall\, x \in \Omega, \qquad H(x, u(x), \nabla u(x)) \;=\; 0, \tag{5.1}$$

where Ω is an open subset of $\mathbb{R}^n$ (with $n \in \mathbb{N}^*$) and $H : \Omega \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is a continuous function called a *Hamiltonian*. Let us remark that if a function u satisfies (5.1), then for all $\lambda \in \mathbb{R}^*$, it will also satisfy $\lambda H(x, u(x), \nabla u(x)) = 0$ for all $x \in \Omega$. As we will see in section 5.2, this obvious observation will have importance consequences when we will deal with viscosity solutions.
A special case of Hamilton-Jacobi equations of the first order is given by those which only depend on x and $\nabla u(x)$, *i.e.* that can be written:

$$\forall\, x \in \Omega, \qquad H(x, \nabla u(x)) \;=\; 0, \tag{5.2}$$

where Ω is still an open subset of $\mathbb{R}^n$ (with $n \in \mathbb{N}^*$) and $H : \Omega \times \mathbb{R}^n \to \mathbb{R}$ a continuous function also called a *Hamiltonian*.

5.1.2 — EXAMPLE. The *eikonal equation*

$$\forall\, x \in \Omega, \qquad |\nabla u(x)| \;=\; 1 \tag{5.3}$$

is a Hamilton-Jacobi equation of the form (5.2), and functions

$$H \;:\; \Omega \times \mathbb{R}^n \to \mathbb{R},\ (x,p) \mapsto |p| - 1 \qquad \text{and} \qquad -H \;:\; \Omega \times \mathbb{R}^n \to \mathbb{R},\ (x,p) \mapsto 1 - |p|$$

are two Hamiltonians associated to (5.3).

5.1.3 — In connection with the shape from shading problem we are interested in, by fixing $n \in \{1, 2\}$, we can note that the brightness equation (II.2) presented in introduction of this part II is a Hamilton-Jacobi of the form (5.2), with for Hamiltonian

$$H \,:\, \Omega \times \mathbb{R}^n \to \mathbb{R},\ x \mapsto I(x)\sqrt{1 + |\nabla u(x)|^2} + \nabla u(x) \cdot l - l_3.$$

REMARK. In the particular case where the unit light vector $(l, l_3) \in \mathbb{R}^n \times \mathbb{R}$ is defined by

$$l \,=\, 0 \quad \text{and} \quad l_3 \,=\, 1,$$

and the intensity satisfies, for all $x \in \Omega$,

$$I(x) \,=\, \frac{\sqrt{2}}{2},$$

we can easily prove that the brightness equation is equivalent to the eikonal equation (5.3).

5.1.4 — Let us formulate some observations about equations of the form (5.2).

- *Boundary conditions.*
  We can note that equation (5.2) is ill-posed. Indeed since this equation only depends on x and $\nabla u(x)$, if u is a solution of (5.2) then for all $\lambda \in \mathbb{R}$ function $u + \lambda$ is also solution of (5.2). To avoid this difficulty, we can fix for instance the value of u in at least one point of $\overline{\Omega}$.
  In control theory, by assuming that Ω is convex, the function

  $$d \,:\, \Omega \to \mathbb{R},\ x \mapsto \inf_{y \in \partial\Omega} \big(|y - x|\big)$$

  that measures the distance of a point from Ω to its boundary $\partial\Omega$ turns out to be a natural weak solution of the eikonal equation (5.3), that satisfies the boundary condition:

  $$\forall\, x \in \partial\Omega, \qquad d(x) \,=\, 0.$$

  More generally, let $\varphi : \partial\Omega \to \mathbb{R}$ be a continuous function and

  $$v \,:\, \Omega \to \mathbb{R},\ x \mapsto \inf_{y \in \partial\Omega} \big(\varphi(y) + |y - x|\big).$$

  While $\varphi(x) \leqslant \varphi(y) + |y - x|$ for all $x, y \in \partial\Omega$, function v turns out to be a weak solution of the eikonal equation (5.3) that satisfies the boundary condition

  $$\forall\, x \in \partial\Omega, \qquad v(x) \,=\, \varphi(x).$$

  That is why equation (5.2) is generally coupled to Dirichlet boundary condition

  $$\forall\, x \in \partial\Omega, \qquad u(x) \,=\, \varphi(x), \tag{5.4}$$

  for a given continuous function $\varphi : \partial\Omega \to \mathbb{R}$.
  But in the case of our shape from shading problem, prescribing Dirichlet boundary conditions seems unreasonable since it physically consists in knowing the altitude of the shape in the boundary of the black and white picture (which is associated to the domain Ω).

- *Existence, uniqueness and regularity of the solutions.*
  Now let us suppose $n = 1$, $\Omega =]0, 1[$, and focus on the problem still given by the eikonal equation

  $$\begin{cases} |u'(x)| \,=\, 1 & \text{if} \quad x \in]0, 1[, \\ u(x) \,=\, 0 & \text{if} \quad x \in \{0, 1\}. \end{cases} \tag{5.5}$$

Rolle theorem shows that all solution of (5.5) is never differentiable on overall $]0,1[$. Therefore this problem does not admit a classical solution, and we will have to look for solutions in a weaker sense.
Now for all $k \in \mathbb{N}$, let us consider the 2^{-k}-periodic function u_k from $[0,1]$ to $\mathbb{R}$ defined, for all $x \in \left[0, 2^{-k}\right]$, by

$$u_k(x) \;=\; \frac{1}{2^{k+1}} - \left| \frac{1}{2^{k+1}} - x \right|.$$

It is clear that u_k and $-u_k$ are continuous functions on $[0,1]$ and differentiable on $]0,1[$ except at a finite number of points. Thus these functions constitutes acceptable weak solutions of problem (5.5), and this problem still admits and infinity of solutions.

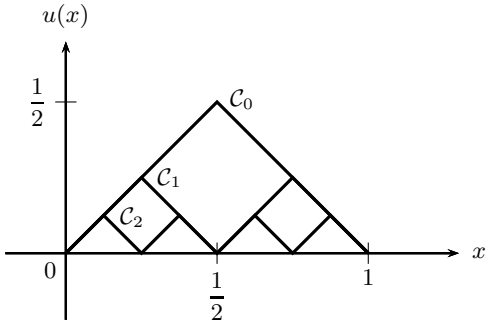

Figure 35 – Representative curves $\mathcal{C}_0$, $\mathcal{C}_1$, $\mathcal{C}_2$ of functions u_0, u_1, u_2 satisfying (5.5)

The previous considerations show that we can at most hope to find continuous solutions to Hamilton-Jacobi equations of the form (5.2). In addition since it seems necessary and natural to couple this kind of equations to Dirichlet boundary conditions of the form (5.4), this is clearly not sufficient.
That is why we need to introduce a class of weak solution that makes problems of the form (5.2)-(5.4) well-posed. The *viscosity solutions* presented in the following section are intended to do it.

5.2 Continuous viscosity solutions

5.2.1 — For the rest of this chapter, we fix $n \in \mathbb{N}^*$ and we denote by Ω an open and bounded subset of $\mathbb{R}^n$. Until section 5.6, we will also denote by H a continuous application defined on $\Omega \times \mathbb{R}^n$. In this section 5.2, we will more generally consider the equation of unknown $u : \Omega \to \mathbb{R}$:

$$\forall\, x \in \Omega, \qquad H(x, \nabla u(x)) \;=\; f(x), \tag{5.6}$$

where f is a continuous function from Ω to $\mathbb{R}$.

Definition. *Let $u \in \mathcal{C}(\Omega)$.*

(*i*) *We say that u is a* viscosity subsolution *of equation* (5.6)*, or that u satisfies $H(x, \nabla u(x)) \leqslant f(x)$* in a viscosity sense*, if for all $\phi \in \mathcal{C}^1(\Omega)$ and all x_0 local maximum of $(u - \phi)$ we have*

$$H(x_0, \nabla \phi(x_0)) \;\leqslant\; f(x_0).$$

(*ii*) *We say that u is a* viscosity supersolution *of equation* (5.6)*, or that u satisfies $H(x, \nabla u(x)) \geqslant f(x)$* in a viscosity sense*, if for all $\phi \in \mathcal{C}^1(\Omega)$ and all x_0 local minimum of $(u - \phi)$ we have*

$$H(x_0, \nabla \phi(x_0)) \;\geqslant\; f(x_0).$$

(*iii*) *We say that u is a* viscosity solution *of equation* (5.6)*, or that u satisfies $H(x, \nabla u(x)) = f(x)$* in a viscosity sense*, if u is both a subsolution and a supersolution of this equation.*

5.2.2 — EXAMPLE. An important example of viscosity solutions is given by the classical solutions. Indeed if $u \in \mathcal{C}^1(\Omega)$ is a classical solution of (5.6), then we have $H(x, \nabla u(x)) = f(x)$ for all $x \in \Omega$. And since for all $\phi \in \mathcal{C}^1(\Omega)$ and all x_0 local extremum of $(u - \phi)$ we have $\nabla u(x_0) = \nabla \phi(x_0)$, then $H(x_0, \nabla \phi(x_0)) = 0$ and u is therefore a viscosity solution of (5.6).

5.2.3 — It is important to note that definitions from 5.2.1 depend on the choice of H. This will be confirmed on a concrete example in 5.2.7. For now we can prove the following result, which shows how the definitions from 5.2.1 behave by switching to the opposite.

Proposition. *Let $u \in \mathcal{C}(\Omega)$ and let us denote by v the function $-u$.*

(*i*) *Function u satisfies $H(x, \nabla u(x)) \leqslant f(x)$ in a viscosity sense if and only if function v satisfies $-H(x, -\nabla v(x)) \geqslant -f(x)$ in a viscosity sense.*

(*ii*) *Function u satisfies $H(x, \nabla u(x)) \geqslant f(x)$ in a viscosity sense if and only if function v satisfies $-H(x, -\nabla v(x)) \leqslant -f(x)$ in a viscosity sense.*

(*iii*) *Function u satisfies $H(x, \nabla u(x)) = f(x)$ in a viscosity sense if and only if function v satisfies $-H(x, -\nabla v(x)) = -f(x)$ in a viscosity sense.*

Proof.— It is sufficient to prove (*i*), (*ii*) being obtained by a totally similar way and (*iii*) being an immediate consequence of the previous points by definition 5.2.1. So let $\phi \in \mathcal{C}(\Omega)$ and $x_0 \in \Omega$.

- Let us suppose that u satisfies $H(x, \nabla u(x)) \leqslant f(x)$ in a viscosity sense. If x_0 is a local maximum of $v - \phi = -u - \phi$, then x_0 is a local minimum of $-(-u - \phi) = u - (-\phi)$. Therefore

$$H(x_0, -\nabla \phi(x_0)) \leqslant -f(x_0), \quad \text{hence} \quad -H(x_0, -\nabla \phi(x_0)) \geqslant f(x_0),$$

  and the arbitrariness on ϕ and x_0 shows that $-v$ satisfies $H(x, -\nabla v(x)) \geqslant -f(x)$ in a viscosity sense.

- Reciprocally let us suppose that v satisfies $-H(x, -\nabla v(x)) \geqslant -f(x)$ in a viscosity sense. If x_0 is a local minimum of $u - \phi$, then x_0 is a local maximum of $-(u - \phi) = v - (-\phi)$. Therefore

$$-H(x_0, -\nabla(-\phi)(x_0)) \geqslant f(x_0), \quad \text{hence} \quad H(x_0, \nabla \phi(x_0)) \leqslant -f(x_0),$$

  and the arbitrariness on ϕ and x_0 shows that u satisfies $H(x, \nabla u(x)) \leqslant f(x)$ in a viscosity sense. □

5.2.4 — Until section 5.6, unless otherwise indicated, we will only talk about *subsolution*, *supersolution* or *solution* to mean *viscosity subsolution*, *viscosity supersolution* or *viscosity solution*. Let us highlight that all of our definitions deal with continuous functions. We will shortly explain how to extend them to discontinuous functions in section 5.6.
Moreover points (*i*), (*ii*) and (*iii*) of definition 5.2.1 will be extended to continuous functions from $\overline{\Omega}$ to $\mathbb{R}$ as long as their restrictions to Ω satisfy the corresponding conditions. By doing so, we will say that a function $u \in \mathcal{C}\left(\overline{\Omega}\right)$ is a *solution* (resp. *subsolution*, *supersolution*) of the Dirichlet problem (5.2)-(5.4)

$$\begin{cases} H(x, \nabla u(x)) = 0 & \text{if} \quad x \in \Omega \\ u(x) = \varphi(x) & \text{if} \quad x \in \partial\Omega \end{cases}$$

if u is a solution (resp. subsolution, supersolution) of (5.2) and if it satisfies (5.4).

5.2.5 — As it stands, it seems difficult to show that a function is a solution of viscosity of equation (5.6) just by applying definition 5.2.1. To make this easier, given a function $u \in \mathcal{C}(\Omega)$ and $x \in \Omega$, we introduce the *superdifferential* of u in x, which is the closed and convex subset of $\mathbb{R}^n$ given by:

$$\mathrm{D}^+u(x) \;=\; \left\{ p \in \mathbb{R}^n \;\middle|\; \limsup_{y \to x} \frac{u(y) - u(x) - p\cdot(y-x)}{|y-x|} \leqslant 0 \right\},$$

and the *subdifferential* of u in x, which is the closed and convex subset of $\mathbb{R}^n$ given by:

$$\mathrm{D}^-u(x) \;=\; \left\{ p \in \mathbb{R}^n \;\middle|\; \liminf_{y \to x} \frac{u(y) - u(x) - p\cdot(y-x)}{|y-x|} \geqslant 0 \right\}.$$

We can graphically observe that a point $p \in \mathbb{R}^n$ is in $\mathrm{D}^+u(x)$ (resp. $\mathrm{D}^-u(x)$) if and only if the hyperplane $y \mapsto u(x) + (p \mid y - x)$ is above (resp. under) the graph of u in a neighbourhood of x.
We can also verify that $\mathrm{D}^\pm(-u)(x) = -\mathrm{D}^\mp u(x)$ and $\mathrm{D}^\pm(u+\lambda)(x) = \mathrm{D}^\pm u(x)$ for all $\lambda \in \mathbb{R}$.

EXAMPLE. If $n = 1$, $\Omega = \mathbb{R}$ and $u : \mathbb{R} \to \mathbb{R},\ x \mapsto |x|$, we have

$$\mathrm{D}^+u(x) \;=\; \begin{cases} \{1\} & \text{if} \quad x > 0, \\ \{-1\} & \text{if} \quad x < 0, \\ \varnothing & \text{if} \quad x = 0, \end{cases} \qquad \text{and} \qquad \mathrm{D}^-u(x) \;=\; \begin{cases} \{1\} & \text{if} \quad x > 0, \\ \{-1\} & \text{if} \quad x < 0, \\ [-1,1] & \text{if} \quad x = 0. \end{cases}$$

5.2.6 — The following result establishes a link between the superdifferential and subdifferential introduced in the previous paragraph and the viscosity solutions. Its proof can be for instance found in [6].

Theorem. *Let $u \in \mathcal{C}(\Omega)$.*

(i) u is a subsolution of (5.6) *if and only if $H(x,p) \leqslant f(x)$ for all $(x,p) \in \Omega \times \mathrm{D}^+u(x)$.*

(ii) u is a supersolution of (5.6) *if and only if $H(x,p) \geqslant f(x)$ for all $(x,p) \in \Omega \times \mathrm{D}^-u(x)$.*

5.2.7 — EXAMPLE. We suppose here that $n = 1$, $\Omega =]0,1[$, and

$$H \;:\;]0,1[\,\times \mathbb{R} \to \mathbb{R},\ (x,p) \mapsto |p| - 1.$$

By taking the notations of functions u_k from paragraph 5.1.4 back and by using the various results from 5.2.5, we can verify that

$$\mathrm{D}^+u_0(x) \;=\; \begin{cases} \{1\} & \text{if} \quad x < \dfrac{1}{2}, \\ \{-1\} & \text{if} \quad x > \dfrac{1}{2}, \\ [-1,1] & \text{if} \quad x = \dfrac{1}{2}, \end{cases} \qquad \text{and} \qquad \mathrm{D}^-u_0(x) \;=\; \begin{cases} \{1\} & \text{if} \quad x < \dfrac{1}{2}, \\ \{-1\} & \text{if} \quad x > \dfrac{1}{2}, \\ \varnothing & \text{if} \quad x = \dfrac{1}{2}. \end{cases}$$

Therefore theorem 5.2.6 easily implies that u_0 is a viscosity solution of the problem

$$\begin{cases} H(x,u'(x)) \;=\; 0 & \text{if} \quad x \in]0,1[, \\ u(x) \;=\; 0 & \text{if} \quad x \in \{0,1\}, \end{cases} \tag{5.7}$$

and that $-u_0$ is not a viscosity solution of this problem. Now we can also deduce from proposition 5.2.3 that $-u_0$ is a viscosity solution of

$$\begin{cases} -H(x,u'(x)) \;=\; 0 & \text{if} \quad x \in]0,1[, \\ u(x) \;=\; 0 & \text{if} \quad x \in \{0,1\}, \end{cases} \tag{5.8}$$

but not u_0. Similarly, we can show that when $k \geqslant 1$, functions u_k or $-u_k$ are not viscosity solutions to either of these two problems.

5.3 Uniqueness result

5.3.1 — In this section, we will thus see how it is possible to insure, in a viscosity sense, the uniqueness of a solution to problem (5.2)-(5.4)

$$\begin{cases} H(x, \nabla u(x)) \ = \ 0 & \text{if} \quad x \in \Omega, \\ u(x) \ = \ \varphi(x) & \text{if} \quad x \in \partial\Omega. \end{cases}$$

As said in 5.2.4, let us remind that we will talk about *solution* instead of *viscosity solution*. Let us start by introducing some vocabulary and reminding some general results of real analysis.

- We call *modulus* each function $m : [0, +\infty[\to [0, +\infty[$ that is non-decreasing, continuous in 0 and that satisfies $m(0) = 0$.
- If (X, d) is a compact metric space and $f : X \to \mathbb{R}$ a continuous application, then:
  - f is bounded and reaches its bounds.
  - f is uniformly continuous and admits a *modulus of continuity*, that can be considered as modulus m_f such that for all $x, y \in X$:
    $$|f(x) - f(y)| \ \leqslant \ m_f(d(x, y)).$$
    Indeed such a function is given by
    $$m_f \ : \ [0, +\infty[\to [0, +\infty[, \ t \mapsto \sup\{|f(x) - f(y)| \mid x, y \in X, \ |x - y| \leqslant t\}.$$

5.3.2 — The uniqueness result is based on a *maximum principle* . More precisely, we say that equation (5.2) satisfies a *maximum principle* if for all functions $u, v \in \mathcal{C}\left(\overline{\Omega}\right)$ such that u and v are respectively subsolution and supersolution of (5.2), we have:

$$\big[\forall\, x \in \partial\Omega, \quad u(x) \ \leqslant \ v(x)\big] \implies \big[\forall\, x \in \Omega, \quad u(x) \ \leqslant \ v(x)\big].$$

It is then clear that if (5.2) satisfies a maximum principle, the problem (5.2)-(5.4) has at most one solution. Indeed, let be $u, v \in \mathcal{C}\left(\overline{\Omega}\right)$ two solutions of it. Then:

– u and v are both subsolutions and supersolutions of (5.2),

– $u(x) = v(x)$ for all $x \in \partial\Omega$ by (5.4).

In particular, having $u(x) \leqslant v(x)$ for all $x \in \partial\Omega$ with u subsolution and v supersolution of (5.2), that satisfies a maximum principle, we get $u(x) \leqslant v(x)$ for all $x \in \Omega$. But as u and v play symmetrical roles, we also have $v(x) \leqslant u(x)$ for all $x \in \Omega$, that implies $u = v$ on overall Ω.
We can show that equation (5.2) satisfies a maximal principle under the following assumptions:

(A_1) There exists a modulus m such that, for all $x, y \in \Omega$ and $p \in \mathbb{R}^n$:

$$H(x, p) - H(y, p) \ \leqslant \ m\big(|x - y|(1 + |p|)\big).$$

(A_2) For all $x \in \Omega$, the function $H(x, \cdot\,) : \mathbb{R}^n \to \mathbb{R}, \ p \mapsto H(x, p)$ is convex.

(A_3) There exists $\alpha < 0$ and $\psi \in \mathcal{C}\left(\overline{\Omega}\right)$ with $\psi_{|\Omega} \in \mathcal{C}^1(\Omega)$ such that, for all $x \in \Omega$:

$$H(x, \nabla\psi(x)) \ \leqslant \ \alpha.$$

In the rest of this section, we will give a direct proof of this by referring to the work of H. ISHII in [10], M. G. CRANDALL, L. C. EVANS and P.-L. LIONS in [7] and G. BARLES in [4].

5.3.3 — The classical way to prove the uniqueness result consists in performing a *split into two variables*. Given $u, v \in \mathcal{C}\left(\overline{\Omega}\right)$ and $\varepsilon > 0$, the split we will consider will be defined as follows:

$$\Phi_\varepsilon \,:\, \overline{\Omega} \times \overline{\Omega} \to \mathbb{R},\ (x,y) \mapsto u(x) - v(y) - \frac{|x-y|^2}{\varepsilon^2}.$$

Since u, v are continuous and $\overline{\Omega}$ is compact (because closed and bounded in $\mathbb{R}^n$), functions u, v, $(u-v)$ and Φ_ε are bounded and reach they bounds. Thus there exists:

– $R \geqslant 0$ such that $|u(x)| \leqslant R$ and $|v(x)| \leqslant R$ for all $x \in \overline{\Omega}$,

– $M \in \mathbb{R}$ such that $M = \sup\left\{u(x) - v(x) \;\middle|\; x \in \overline{\Omega}\right\}$,

– $(x_\varepsilon, y_\varepsilon) \in \overline{\Omega} \times \overline{\Omega}$ such that the real $M_\varepsilon = \Phi(x_\varepsilon, y_\varepsilon)$ satisfies $M_\varepsilon \geqslant \Phi_\varepsilon(x,y)$ for all $(x,y) \in \overline{\Omega} \times \overline{\Omega}$.

Conserving the previous notations, we can establish the following result:

Lemma. *We assume that $u(x) \leqslant v(x)$ for all $x \in \partial\Omega$ and that $M > 0$. Then:*

(*i*) *$M_\varepsilon \to M$ when $\varepsilon \to 0$.*

(*ii*) *$\dfrac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2} \to 0$ when $\varepsilon \to 0$.*

(*iii*) *There exists $\eta > 0$ such that $x_\varepsilon, y_\varepsilon \in \Omega$ for all $\varepsilon \in]0, \eta]$.*

Proof.— Let $\varepsilon > 0$.

(*i*) By the definition of M_ε, we clearly have

$$M \;\leqslant\; M_\varepsilon, \tag{5.9}$$

since $M_\varepsilon = \Phi_\varepsilon(x_\varepsilon, y_\varepsilon) \geqslant \Phi_\varepsilon(x,x) = u(x) - v(x)$ for all $x \in \overline{\Omega}$. In particular, by (5.9) and the definition of R,

$$M \;\leqslant\; M_\varepsilon \;=\; u(x_\varepsilon) - v(x_\varepsilon) - \frac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2} \;\leqslant\; 2R - \frac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2},$$

that implies, since $M > 0$,

$$|x_\varepsilon - y_\varepsilon|^2 \;\leqslant\; 2R\varepsilon^2. \tag{5.10}$$

Otherwise as $\overline{\Omega}$ is compact and v continuous, if m_v is a modulus of continuity of v,

$$M_\varepsilon \;\leqslant\; u(x_\varepsilon) - v(y_\varepsilon) \;=\; u(x_\varepsilon) - v(x_\varepsilon) + v(x_\varepsilon) - v(y_\varepsilon) \;\leqslant\; M + m_v(|x_\varepsilon - y_\varepsilon|).$$

Then by (5.9) and (5.10), we obtain

$$0 \;\leqslant\; M_\varepsilon - M \;\leqslant\; [u(x_\varepsilon) - v(y_\varepsilon)] - M \;\leqslant\; m_v\left(\varepsilon\sqrt{2R}\right). \tag{5.11}$$

Therefore when ε goes to 0, we deduce from (5.11) that M_ε goes to M.

(*ii*) When ε goes to 0, we also deduce from (5.11) that $u(x_\varepsilon) - v(y_\varepsilon)$ goes to M. Thus since by (*i*),

$$u(x_\varepsilon) - v(y_\varepsilon) - \frac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2} \;=\; M_\varepsilon \;\xrightarrow[\varepsilon\to 0]{}\; M,$$

it necessarily implies that $\dfrac{|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2}$ goes to 0 when ε goes to 0.

(*iii*) By contradiction, let us suppose that for all $\eta > 0$, there exists $\varepsilon \in]0,\eta]$ such that $x_\varepsilon \in \partial\Omega$. Then we can easily construct a sequence $(\varepsilon(n))_{n\in\mathbb{N}}$ of non-negative numbers that converges to 0 and satisfies $x_{\varepsilon(n)} \in \partial\Omega$ for all $n \in \mathbb{N}$.
If $n \in \mathbb{N}$, since $u\big(x_{\varepsilon(n)}\big) \leqslant v\big(x_{\varepsilon(n)}\big)$ by hypothesis, we deduce from (5.10) and (*ii*) that:

$$\begin{aligned} M_{\varepsilon(n)} \;&\leqslant\; u\big(x_{\varepsilon(n)}\big) - v\big(y_{\varepsilon(n)}\big) - \frac{\big|x_{\varepsilon(n)} - y_{\varepsilon(n)}\big|^2}{[\varepsilon(n)]^2} \\ &\leqslant\; m_v\Big(\varepsilon(n)\sqrt{2R}\Big) - \frac{\big|x_{\varepsilon(n)} - y_{\varepsilon(n)}\big|^2}{[\varepsilon(n)]^2} \;\xrightarrow[n\to+\infty]{}\; 0. \end{aligned}$$

Therefore, we have $M \leqslant 0$ by (5.9) whereas $M > 0$ by hypothesis, which is impossible!
Consequently there exists $\eta_1 > 0$ such that $x_\varepsilon \in \Omega$ as soon as $\varepsilon \in]0,\eta_1]$, and we establish in the same way the existence of $\eta_2 > 0$ such that $y_\varepsilon \in]0,\eta_2]$ as soon as $\varepsilon \in]0,\eta_2]$. Then we conclude that (*iii*) is true by taking $\eta = \min(\eta_1,\eta_2)$. □

5.3.4 — Lemma. *Let $\alpha < 0$, $f \in \mathcal{C}(\Omega)$ a function such that $f(x) \leqslant \alpha$ for all $x \in \Omega$, and $u, v \in \mathcal{C}\big(\overline{\Omega}\big)$ functions that satisfy in a viscosity sense*

$$H(x,\nabla u(x)) \;\leqslant\; f(x) \quad \text{and} \quad H(x,\nabla v(x)) \;\geqslant\; 0.$$

Under the assumption (A_1)*, if $u(x) \leqslant v(x)$ for all $x \in \partial\Omega$, then $u(x) \leqslant v(x)$ for all $x \in \Omega$.*

Proof.— Let us take the notations of M and, for all $\varepsilon > 0$, of x_ε and y_ε introduced in 5.3.3 back. Our goal is to prove that $M \leqslant 0$; by contradiction, let us suppose that $M > 0$. By (*iii*) of lemma 5.3.3, we can fix $\eta > 0$ such that $x_\varepsilon, y_\varepsilon \in \Omega$ when $\varepsilon \in]0,\eta]$. Then for all $\varepsilon \in]0,\eta]$:

– x_ε is a local maximum of $x \mapsto u(x) - \left[v(y_\varepsilon) + \dfrac{|x - y_\varepsilon|^2}{\varepsilon^2}\right]$, so by hypothesis,

$$f(x_\varepsilon) \;\geqslant\; H\left(x_\varepsilon,\; \frac{2|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2}\right). \tag{5.12}$$

– y_ε is a local minimum of $y \mapsto v(y) - \left[u(x_\varepsilon) - \dfrac{|x_\varepsilon - y|^2}{\varepsilon^2}\right]$, so by hypothesis,

$$0 \;\leqslant\; H\left(y_\varepsilon,\; \frac{2|x_\varepsilon - y_\varepsilon|}{\varepsilon^2}\right). \tag{5.13}$$

Therefore by (5.12), (5.13) and hypothesis on f, for all $\varepsilon \in]0,\eta]$, we have

$$\left|H\left(x_\varepsilon,\; \frac{2|x_\varepsilon - y_\varepsilon|^2}{\varepsilon^2}\right) - H\left(y_\varepsilon,\; \frac{2|x_\varepsilon - y_\varepsilon|}{\varepsilon^2}\right)\right| \;\geqslant\; -f(x_\varepsilon) \;\geqslant\; -\alpha \;>\; 0. \tag{5.14}$$

In the same time, since by assumption (A_1) and (ii) of lemma 5.3.3,

$$\left|H\left(x_\varepsilon,\ \frac{2|x_\varepsilon-y_\varepsilon|^2}{\varepsilon^2}\right)-H\left(y_\varepsilon,\ \frac{2|x_\varepsilon-y_\varepsilon|^2}{\varepsilon^2}\right)\right| \leqslant m\left(|x_\varepsilon-y_\varepsilon|+\frac{|x_\varepsilon-y_\varepsilon|^2}{\varepsilon^2}\right) \xrightarrow[\varepsilon\to 0]{} 0,$$

we get a contradiction with (5.14). It means that we cannot have $M>0$, hence the result. □

5.3.5 — Theorem. (*Maximum principle*)
Under the assumptions (A_1), (A_2) *and* (A_3), *equation* (5.2) *satisfies a maximum principle.*

Proof.— Let $u\in\mathcal{C}\left(\overline{\Omega}\right)$ and $v\in\mathcal{C}\left(\overline{\Omega}\right)$ be respectively subsolution and supersolution of (5.2), with

$$\forall\, x\in\partial\Omega,\qquad u(x)\ \leqslant\ v(x). \tag{5.15}$$

Let us first note that since ψ and v are continuous on the compact $\overline{\Omega}$, at the risk of replacing ψ by $\psi-C$ with C a sufficiently large constant, we can always suppose that

$$\forall\, x\in\partial\Omega,\qquad \psi(x)\ \leqslant\ v(x). \tag{5.16}$$

We now fix $\theta\in\,]0,1[$ and we introduce the function

$$u_\theta\ :\ \overline{\Omega}\to\mathbb{R},\ x\mapsto\theta u(x)+(1-\theta)\psi(x).$$

- By (5.15) and (5.16), for all $x\in\partial\Omega$ we have

$$u_\theta(x)\ =\ \theta u(x)+(1-\theta)\psi(x)\ \leqslant\ \theta v(x)+(1-\theta)v(x)\ =\ v(x).$$

- Otherwise we can prove by easy computations that for all $x\in\Omega$:

$$\mathrm{D}^+u_\theta(x)\ =\ \{\theta p+(1-\theta)\nabla\psi(x)\mid p\in\mathrm{D}^+u(x)\}.$$

  Therefore if $x\in\Omega$ and $q\in\mathrm{D}^+u_\theta(x)$, writing $q=\theta p+(1-\theta)\nabla\psi(x)$ for such a $p\in\mathrm{D}^+u(x)$, we successively deduce from (A_2) and (A_3) that

$$H(x,q)\ =\ H(x,\ \theta p+(1-\theta)\nabla\psi(x))\ \leqslant\ \theta H(x,p)+(1-\theta)H(x,\nabla\psi(x))\ \leqslant\ (1-\theta)\alpha\ \leqslant\ \alpha,$$

  which shows that u_θ satisfies $H(x,\nabla u_\theta(x))\leqslant\alpha$ in the viscosity sense by theorem 5.2.6.

In the end, lemma 5.3.4 applied to the constant function $f:x\mapsto\alpha$ assures that $u_\theta(x)\leqslant v(x)$ for all $x\in\Omega$. But since θ was arbitrarily chosen in $]0,1[$, the result follows as θ goes to 1. □

5.3.6 — According to 5.3.2 and 5.3.5, we can now say loud and clear:

Corollary. (*Uniqueness result*)
Under the assumptions (A_1), (A_2) *and* (A_3), *the Dirichlet problem* (5.2)-(5.4) *has at most one solution.*

5.3.7 — EXAMPLE. As in example 5.2.7, we suppose here that $n=1$, $\Omega=\,]0,1[$, and

$$H\ :\ ]0,1[\times\mathbb{R},\ (x,p)\mapsto|p|-1.$$

It clearly satisfies assumptions (A_1) and (A_2), but also (A_3) for any constant function ψ on $]0,1[$. Thus corollary 5.3.6 shows that function

$$u_0\ :\ [0,1]\to\mathbb{R},\ x\mapsto\frac{1}{2}-\left|\frac{1}{2}-x\right|$$

is the unique viscosity solution of problem (5.7). By proposition 5.2.3, we can also note that $-u_0$ is the unique viscosity solution of problem (5.8).

5.4 Existence result

5.4.1 — The existence result uses the control theory, which is out the scope of this document. Thus we will only remind here the corresponding result and illustrating it on a concrete example. In order to do this, we need to introduce some notations. So let us suppose that (A_2) is satisfied, *i.e.* $H(x,\,\cdot\,)$ convex for all $x \in \Omega$, and that H can be extended as a continuous function on $\overline{\Omega} \times \mathbb{R}^n$, still noted H.

- For all $x \in \overline{\Omega}$, the *Legendre transform* of $H(x,\,\cdot\,)$ is the function

$$\mathcal{L}H(x,\,\cdot\,) \,:\, \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\},\; q \mapsto \sup\{p\cdot q - H(x,p) \mid p \in \mathbb{R}^n\}.$$

- For all $T > 0$ and $x, y \in \overline{\Omega}$, we denote by $\Gamma_{x,y}(T)$ the set of paths $\gamma : [0,T] \to \mathbb{R}^n$ such that $\gamma(0) = x$, $\gamma(T) = y$, $\gamma(]0,T[) \subset \overline{\Omega}$ and $\gamma' \in \mathrm{L}^\infty(]0,T[,\mathbb{R}^n)$.

- By using the previous notations, for all $x, y \in \overline{\Omega}$, we finally set

$$L(x,y) \;=\; \inf\left\{ \int_0^T \mathcal{L}H\big(\gamma(s),\,-\gamma'(s)\big)\,\mathrm{d}s \;\middle|\; T>0,\; \gamma \in \Gamma_{x,y}(T) \right\}.$$

We will see that an existence result can be obtained under the following assumptions:

(A_4) Ω is smooth and connected.

(A_5) $\inf\{H(x,p) \mid p \in \mathbb{R}^n\} \leqslant 0$ for all $x \in \Omega$.

(A_6) For all $x \in \Omega$, function $H(x,\,\cdot\,)$ is coercive, *i.e.* $H(x,p) \to +\infty$ when $|p| \to +\infty$.

(A_7) H can be extended as a continuous function on $\overline{\Omega} \times \mathbb{R}^n$, still noted H.

(A_8) For all $x, y \in \partial\Omega$, function φ satisfies the following compatibility condition:

$$\varphi(x) - \varphi(y) \;\leqslant\; L(x,y).$$

5.4.2 — Here is reminded the existence result. As mentioned by E. PRADOS and O. FAUGERAS in [18], it was described by P.-L. LIONS in [14] (see its chapter 5, and more precisely its theorem 5.3).

Theorem. (*Existence result*)

(*i*) *Under the assumptions* (A_2) *and* (A_4) *to* (A_8)*, function*

$$u \,:\, \overline{\Omega} \to \mathbb{R},\; x \mapsto \inf_{y \in \partial\Omega} \big(\varphi(y) + L(x,y)\big)$$

is a continuous viscosity solution to the Dirichlet problem (5.2)-(5.4)*.*

(*ii*) *More precisely, under the assumptions* (A_2) *and* (A_4) *to* (A_7)*, the function* u *above is a viscosity solution to the Dirichlet problem* (5.2)-(5.4) *if and only if assumption* (A_8) *is satisfied.*

5.4.3 — EXAMPLE. As in examples 5.2.7 or 5.3.7, let us suppose that $n = 1$, $\Omega =]0,1[$, and

$$H \,:\,]0,1[\to \mathbb{R},\; (x,p) \mapsto |p| - 1.$$

It is clear that assumptions (A_2) and (A_4) to (A_7) are satisfied. On the other hand, by doing easy calculations, we can verify that for all $x \in]0,1[$, the Legendre transform $\mathcal{L}H(x,\,\cdot\,)$ of $H(x,\,\cdot\,)$ verifies

$$\mathcal{L}H(x,q) \;=\; \left\{ \begin{array}{lll} 1 & \text{if} & |q| \;\leqslant\; 1, \\ +\infty & \text{if} & |q| \;>\; 1. \end{array} \right.$$

Therefore, for all $x, y \in \partial\Omega = \{0,1\}$,

$$\begin{aligned} L(x,y) \;&=\; \inf \left\{ \int_0^T \mathcal{L}H\big(\gamma(s),\,-\gamma'(s)\big)\,\mathrm{d}s \;\middle|\; T>0,\; \gamma \in \Gamma_{x,y}(T) \right\} \\ &=\; \inf \left\{ \int_0^T \mathcal{L}H\big(\gamma(s),\,-\gamma'(s)\big)\,\mathrm{d}s \;\middle|\; T>0,\; \gamma \in \Gamma_{x,y}(T),\; \|\gamma'\|_\infty \leqslant 1 \right\} \\ &=\; \inf \left\{ T \mid T>0,\; \gamma \in \Gamma_{x,y}(T),\; \|\gamma'\|_\infty \leqslant 1 \right\} \\ &=\; |y-x|. \end{aligned}$$

Consequently assumption (A_8) can be rewritten

$$\left\{ \begin{array}{l} \varphi(0) - \varphi(1) \;\leqslant\; L(0,1) \;=\; 1, \\ \varphi(1) - \varphi(0) \;\leqslant\; L(1,0) \;=\; 1, \end{array} \right. \qquad \textit{i.e.} \qquad |\varphi(1) - \varphi(0)| \;\leqslant\; 1.$$

According to theorem 5.4.2, this shows that the equation

$$\forall\, x \in]0,1[, \qquad H(x,u'(x)) = 0$$

has a continuous viscosity solution $u : [0,1] \to \mathbb{R}$ if and only if $|u(0) - u(1)| \leqslant 1$.

5.5 Application to the shape from shading problem

5.5.1 — In sections 5.3 and 5.4, given $n \in \mathbb{N}^*$, Ω an open subset of $\mathbb{R}^n$, $H : \Omega \times \mathbb{R}^n \to \mathbb{R}$ a Hamiltonian and $\varphi : \partial\Omega \to \mathbb{R}$ a continuous function, we have presented various assumptions on H and φ that insure the existence and uniqueness of a viscosity solution to the problem

$$\left\{ \begin{array}{lll} H(x,\nabla u(x)) \;=\; 0 & \text{if} & x \in \Omega, \\ u(x) \;=\; \varphi(x) & \text{if} & x \in \partial\Omega. \end{array} \right.$$

As explained in the introduction of this part II, in practice n will be equal to 1 or 2, Ω a rectangular domain, and H given, for all $(x,p) \in \Omega \times \mathbb{R}^n$, by

$$H(x,p) \;=\; I(x)\sqrt{1+|p|^2} + l\cdot p - l_3. \tag{5.17}$$

As a reminder, I is a (known) function from $\overline{\Omega}$ to $[0,1]$ corresponding to the brightness intensity, and $L = (l, l_3) \in \mathbb{R}^n \times \mathbb{R}_+^*$ is associated to the unit light vector. For this choice of H, problem (5.2)-(5.4) will be called the *shape from shading problem* (*s.f.s. problem* in abbreviated form).
In this section, we will present various assumptions on the brightness intensity I and the unit light vector L that insure the existence and uniqueness of a solution to this s.f.s. problem. For the existence, this will be possible by referring to theorem 5.4.2, and by referring to corollary 5.3.6 for the uniqueness.

5.5.2 — In order to determine assumptions that insure the existence and uniqueness to the s.f.s. problem, we will start by proving some preliminary results. Here is the first one:

Proposition. *For all $x \in \Omega$, function $H(x, \cdot\,)$ is convex.*

Proof.— Let us fix $x \in \Omega$, and let us denote more simply by H_x the function $H(x, \cdot\,)$. Whatever the value of n, it is clear that this function is indefinitely differentiable on $\mathbb{R}^n$.

1) We first treat the case $n = 1$. For all $p \in \mathbb{R}$, we have

$$H_x'(p) \;=\; \frac{I(x)p}{\sqrt{1+p^2}} + l \qquad \text{and so} \qquad H_x''(p) \;=\; \frac{1}{(1+p^2)^{3/2}} \;\geqslant\; 0.$$

Hence the convexity of H_x.

2) We now assume that $n = 2$. Then if $p \in \mathbb{R}^2$, for all $j \in \{1, 2\}$ we have

$$\partial_j H_x(p) \;=\; I(x)\,\frac{p_j}{\sqrt{1+|p|^2}} + l_j,$$

and so

$$\partial_j^2 H_x(p) \;=\; I(x)\,\frac{1+p_j^2+|p|^2}{(1+|p|^2)^{3/2}} \qquad \text{and} \qquad \partial_1(\partial_2 H_x(p)) \;=\; I(x)\,\frac{p_1 p_2}{(1+|p|^2)^{3/2}}.$$

Therefore M the Hessian of H_x in p is the matrix given by

$$M \;=\; \frac{I(x)}{(1+|p|^2)^{3/2}} \begin{pmatrix} 1+p_1^2+|p|^2 & p_1 p_2 \\ p_1 p_2 & 1+p_2^2+|p|^2 \end{pmatrix}.$$

We can thus verify that $\operatorname{tr} M \geqslant 0$ and $\det M \geqslant 0$, which implies that all the eigenvalues of M are non-negative. Therefore M is non-negative, and H_x is a convex function. □

5.5.3 — Proposition. *For all $x \in \Omega$ we have*

$$\inf\{H(x,p) \mid p \in \mathbb{R}^2\} \;=\; \begin{cases} \sqrt{I^2(x) - |l|^2} - l_3 & \text{if } I(x) \geqslant |l|, \\ -\infty & \text{if } I(x) < |l|. \end{cases}$$

Proof.— Let us fix $x \in \Omega$ and let us take the notation of H_x from the proof of proposition 5.5.2 back.

- We first suppose that $I(x) > |l|$. Whatever the value of n in $\{1, 2\}$, we can prove by doing classical computations that

$$p^* \;=\; \frac{-l}{\sqrt{I^2(x) - |l|^2}}$$

is the unique critical point of H_x. And since H_x is convex in accordance with proposition 5.5.2, we deduce that it corresponds to its global minimizer. Hence

$$\inf\{H_x(p) \mid p \in \mathbb{R}\} \;=\; H_x(p^*) \;=\; \sqrt{I^2(x) - |l|^2} - l_3.$$

- We now suppose that $I(x) = |l|$. Then for all $p \in \mathbb{R}^2 \setminus \{0\}$,

$$H_x(p) \;=\; |l|\sqrt{1+|p|^2} + l\cdot p - l_3 \;=\; |l||p|\sqrt{1+\frac{1}{|p|^2}} - l\cdot(-p) - l_3.$$

According to the Cauchy-Schwarz inequality, we have $l \cdot p \geqslant -|l||p|$, hence

$$H_x(p) \;\geqslant\; |l||p|\sqrt{1+\frac{1}{|p|^2}} - |l||p| - l_3 \;=\; |l||p| \underbrace{\left(\sqrt{1+\frac{1}{|p|^2}} - 1\right)}_{\geqslant 1} - l_3 \;\geqslant\; -l_3.$$

And since $H_x(0) = -l_3$, we can conclude that:

$$\inf\{H_x(p) \mid p \in \mathbb{R}^n\} \;=\; -l_3 \;=\; \sqrt{I^2(x) - |l|^2} - l_3.$$

- Let us finally suppose that $I(x) < |l|$. Since $I(x) \geqslant 0$, it implies that $|l| > 0$. Now let us fix $t > 0$ and let us define $p_t = t|l|$. Then,

$$H_x(p_t) \;=\; I(x)\sqrt{1+t^2|l|^2} - t|l|^2 - l_3 \;=\; t|l| \left[I(x)\sqrt{1+\frac{1}{t^2|l|^2}} - |l| \right] - l_3. \tag{5.18}$$

And since $I(x) < |l|$, we easily obtain $H_x(p_t) \to -\infty$ when $t \to +\infty$, hence

$$\inf\{H_x(p) \mid p \in \mathbb{R}\} \;=\; -\infty. \qquad \square$$

5.5.4 — Proposition. *Let $x \in \Omega$. The following conditions are equivalent:*

(*i*) $I(x) > |l|$.

(*ii*) *Function $H(x, \cdot\,)$ is coercive,* i.e. $H(x,p) \to +\infty$ *when* $|p| \to +\infty$.

Proof.— (*i*) $\Rightarrow$ (*ii*) : Let us assume that $I(x) > |l|$, and let us fix $p \in \mathbb{R}^2 \smallsetminus \{0\}$. Since

$$H(x,p) \;=\; I(x)\sqrt{1+|p|^2} + l\cdot p - l_3 \;=\; I(x)|p|\sqrt{1+\frac{1}{|p|^2}} + l\cdot p - l_3,$$

then by the Cauchy-Schwarz inequality, we get

$$H(x,p) \;\geqslant\; I(x)|p|\sqrt{1+\frac{1}{|p|^2}} - |l||p| - l_3 \;>\; |l|\,|p| \underbrace{\left(\sqrt{1+\frac{1}{|p|^2}} - 1\right)}_{\geqslant 1} - l_3.$$

Therefore, we clearly obtain $H(x,p) \to +\infty$ when $|p| \to +\infty$. Hence (*i*) $\Rightarrow$ (*ii*).

(*ii*) $\Rightarrow$ (*i*) : Let us suppose that (*i*) is not satisfied, *i.e.* $I(x) \leqslant |l|$. Then as in the proof of proposition 5.5.3, let us fix $t > 0$ and let us set $p_t = t|l|$. According to (5.18), we have

$$H(x,p_t) \;=\; t|l| \left[I(x)\sqrt{1+\frac{1}{t^2|l|^2}} - |l| \right] - l_3.$$

Therefore if $I(x) < |l|$, then for t large enough, we get

$$I(x)\sqrt{1+\frac{1}{t^2|l|^2}} - |l| \;<\; 0,$$

hence $H(x,p_t) \to -\infty$ when $t \to +\infty$. On the other hand, when $I(x) = |l|$, (5.18) implies

$$H(x,p_t) \;=\; t|l|^2 \underbrace{\left(\sqrt{1+\frac{1}{|p|^2}}-1\right)}_{\geqslant 1} - l_3,$$

hence $H(x,p_t) \to -l_3$ when $t \to +\infty$. In any case, (*ii*) is not satisfied. □

5.5.5 — We will start by showing how to insure the existence of a viscosity solution to the s.f.s. problem. According to theorem 5.4.2, we have to see when assumptions (A_2) and (A_4) to (A_8) are satisfied. We can first note that the validity of (A_8) depends on the choice of function φ, and must be considered in practice on a case-by-case basis. To insure that the other assumptions are satisfied, we will see that it is sufficient to suppose:

(A_9) I is continuous.

(A_{10}) $I(x) > |l|$ for all $x \in \Omega$.

Theorem. (*Existence of a solution to the s.f.s. problem*)

(*i*) *If* (A_9) *is satisfied, then* (A_7) *is satisfied.*

(*ii*) *In addition if* (A_8) *and* (A_{10}) *are also satisfied, then the s.f.s. problem has a solution.*

Proof.— Point (*i*) is obvious. According to theorem 5.4.2, to obtain (*ii*), we need to prove that assumptions (A_2) and (A_4) to (A_6) are satisfied, (A_7) being true thanks to (*i*) and (A_8) being supposed true.

- (A_2) is true thanks to proposition 5.5.2.
- Since Ω is supposed to be rectangular, then assumption (A_4) is satisfied.
- If (A_{10}) is satisfied, then since $I(x) \leqslant 1 = |l|^2 + l_3$,

$$\inf\{H(x,p) \mid p \in \mathbb{R}^n\} \;\leqslant\; \sqrt{I^2(x)-|l|^2} - l_3^2 \;\leqslant\; \sqrt{l_3^2} - l_3 \;=\; l_3 - l_3 \;=\; 0,$$

  and (A_5) is also satisfied.
- (A_6) is also true as soon as (A_{10}) thanks to proposition 5.5.4. □

5.5.6 — Now let us consider the uniqueness of solutions to the s.f.s. problem. We will show that the uniqueness to this problem can be obtained with this additional hypothesis:

(A_{11}) There exists $\alpha < 1$ such that $I(x) \leqslant \alpha$ for all $x \in \Omega$.

Theorem. (*Uniqueness of a solution to the s.f.s. problem*)
Under the assumptions (A_9) *and* (A_{11})*, the s.f.s. problem has at most one solution.*

Proof.— According to corollary 5.3.6, it is sufficient to prove that assumptions (A_1) to (A_3) are satisfied.

- For all $x, y \in \Omega$ and $p \in \mathbb{R}^n$, we have

$$H(x,p) - H(y,p) \;\leqslant\; |H(x,p) - H(y,p)| \;=\; |I(x) - I(y)|\sqrt{1+|p|^2}.$$

  And since for all $t \in \mathbb{R}_+$, we have $1 + t^2 \leqslant (1+t)^2$, then

$$H(x,p) - H(y,p) \;\leqslant\; |I(x) - I(y)|\sqrt{(1+|p|)^2} \;=\; |I(x) - I(y)|(1+|p|).$$

  Thus if (A_9) is satisfied, it is now clear that (A_1) is also satisfied.

- (A_2) is still satisfied thanks to proposition 5.5.2.
- Let us introduce

$$\psi \ : \ \overline{\Omega} \to \mathbb{R}, \ x \mapsto -\frac{x \cdot l}{l_3}.$$

It is clear that $\psi \in \mathcal{C}\left(\overline{\Omega}\right)$ and that $\psi_{|\Omega} \in \mathcal{C}^1(\Omega)$ with, for all $x \in \Omega$,

$$\nabla\psi(x) \ = \ -\frac{l}{l_3}.$$

But $L = (l, l_3)$ is supposed to be an unit vector, hence $|l|^2 + l_3^2 = 1$. Therefore if $x \in \Omega$, we have

$$H\big(x, \nabla\psi(x)\big) \ = \ I(x)\sqrt{1 + \frac{|l|^2}{l_3^2}} - \frac{|l|^2}{l_3} - l_3 \ = \ I(x)\sqrt{\frac{|l|^2 + l_3^2}{l_3^2}} - \frac{|l|^2 + l_3^2}{l_3} \ = \ \frac{I(x) - 1}{l_3}.$$

Now we can easily observe that if (A_{11}), then (A_3) is also satisfied. □

5.5.7 — *When there exists points $x \in \Omega$ such that $I(x) = 1$.*

According to 5.5.6, if there exists points $x \in \Omega$ such that $I(x) = 1$, then we can not insure the uniqueness of a solution to the s.f.s. problem. To start, let us illustrate it on a concrete example. In order to do this, let us suppose $n = 1$, $\Omega =]-1, 1[$, and let us introduce

$$u \ : \ [-1, 1] \to \mathbb{R}, \ x \mapsto \sqrt{1 - x^2}.$$

By supposing $I(x) = u(x)$ for all $x \in]-1, 1[$ and $\varphi(-1) = \varphi(1) = 0$, we clearly have $I(0) = u(0) = 1$, and thanks to theorem 5.2.6, we can prove that u and its opposite $-u$ are two different viscosity solutions of the s.f.s. problem, as shown by the following figure:

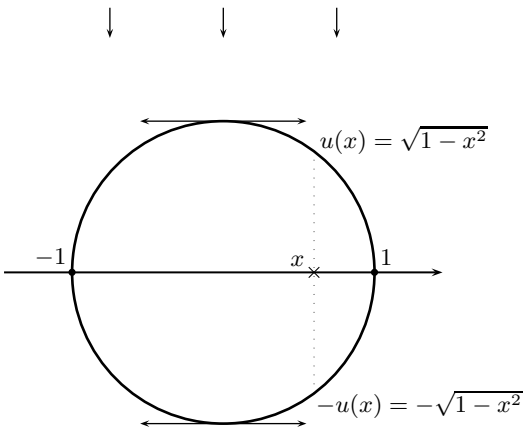

Figure 36 – Two different viscosity solutions of a same s.f.s. problem when I reaches 1

This very simple example confirms that we can generally loss the uniqueness of a solution to the s.f.s. problem as soon as the brightness intensity at some point of Ω is equal to 1.
In [18], E. PRADOS and O. FAUGERAS have extended the result established by P.-L. LIONS, E. ROUY

and A. TOURIN in [16] by proving that if $I^{-1}(\{1\})$, the set of points $x \in \Omega$ such that $I(x) = 1$, is equal to a finite union of disjoint connected compact sets, then by setting

$$\Omega' \ = \ \Omega \setminus I^{-1}(\{1\}),$$

the uniqueness of a solution to the s.f.s. problem can be obtained again by supposing that φ can be extended as a continuous function from $\partial\Omega'$ to $\mathbb{R}$ and by solving

$$\left\{ \begin{array}{lll} H(x, \nabla u(x)) \ = \ 0 & \text{if} & x \in \Omega', \\ u(x) \ = \ \varphi(x) & \text{if} & x \in \partial\Omega', \end{array} \right.$$

instead of the initial s.f.s. problem.

5.6 About the discontinuous viscosity solutions

5.6.1 — The aim of this section is not to explain how to generalize the previous theories to discontinuous functions, but just to mention that it is possible to do it. Since we have not studied this in detail, we will only present general ideas that enable it by giving some references that talk about it with a more accuracy (and probably rigorousness) level and by mentioning its possible application to the s.f.s. problem.
As explained by G. Barles in [4], the first point is that contrary to the classical point of view adopted with partial differential equations, we will here consider equations posed on closed subset, *i.e.* of the form:

$$\forall\, x \in \overline{\Omega}, \qquad G(x, u(x), \nabla u(x)) \ = \ 0, \tag{5.19}$$

for such a locally bounded function G defined on $\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^n$.
By doing this, if H is now a continuous function defined on $\overline{\Omega} \times \mathbb{R}^n$, the Dirichlet problem (5.2)-(5.4) can be seen as an equation of the form (5.19) for G defined, for all $x \in \overline{\Omega}$, $u \in \mathbb{R}$ and $p \in \mathbb{R}^n$, by

$$G(x, u, p) \ = \ \left\{ \begin{array}{ll} H(x, p) & \text{if} \quad x \in \Omega, \\ u - \varphi(x) & \text{if} \quad x \in \partial\Omega. \end{array} \right. \tag{5.20}$$

5.6.2 — Notation. Given $d \in \mathbb{N}^*$ and f a locally bounded function defined on U an open subset of $\mathbb{R}^d$, we will denote by f^* its upper semi-continuous (u.s.c. in abbreviated form) envelope and by f_* its lower semi-continuous (l.s.c. in abbreviated form) envelope. For all $x \in U$, let us remind:

$$f^*(x) \ = \ \limsup_{y \to x} f(y) \qquad \text{and} \qquad f_*(x) \ = \ \liminf_{y \to x} f(y).$$

In the following, if the function f depends on several variables, f^* and f_* will always correspond to the u.s.c. and l.s.c. envelopes through all its variables.

EXAMPLE. With the definition of G from relation (5.20), for all $x \in \overline{\Omega}$, $u \in \mathbb{R}$ and $p \in \mathbb{R}^n$, we have

$$\left\{ \begin{array}{lll} G^*(x, u, p) \ = \ G_*(x, u, p) \ = \ H(x, \nabla u(x)) & \text{if} & x \in \Omega, \\ G^*(x, u, p) \ = \ \min\{H(x, p),\ u - \varphi(x)\} & \text{if} & x \in \partial\Omega, \\ G_*(x, u, p) \ = \ \max\{H(x, p),\ u - \varphi(x)\} & \text{if} & x \in \partial\Omega. \end{array} \right.$$

5.6.3 — In comparison to the definition obtained in 5.2.1, here is given its generalization that enable to deal with the discontinuous case:

Definition. *A locally bounded and u.s.c. function $u : \overline{\Omega} \to \mathbb{R}$ is called a* viscosity subsolution *of* (5.19) *if for all $\phi \in \mathcal{C}^1(\overline{\Omega})$ and all x_0 local maximum of $(u - \phi)$, we have*

$$G_*(x_0, u(x_0), \nabla u(x_0)) \ = \ 0.$$

A locally bounded and l.s.c. function $v : \overline{\Omega} \to \mathbb{R}$ is called a viscosity supersolution *of* (5.19) *if for all $\phi \in \mathcal{C}^1(\overline{\Omega})$ and all x_0 local minimum of $(u - \phi)$, we have*

$$G^*(x_0, u(x_0), \nabla u(x_0)) \ = \ 0.$$

Finally, a function from $\overline{\Omega}$ to $\mathbb{R}$ is said to be a viscosity solution *of* (5.19) *if it is both a subsolution and a supersolution of this equation.*

5.6.4 — With this extension, we can now find viscosity solutions to equations or problems that do not admit a continuous viscosity solution. That is for instance the case with the Dirichlet problem associated to the eikonal equation

$$\left\{ \begin{array}{rcll} H(x, u'(x)) & = & 0 & \text{if} \quad x \in]0,1[\\ u(x) & = & 2x & \text{if} \quad x \in \{0,1\} \end{array} \right. \qquad \text{with} \qquad H \ : \ [0,1] \to \mathbb{R}, \ (x,p) \mapsto |p| - 1.$$

Thanks to the mean value theorem, this problem does not have any continuous viscosity solution. By using the generalized definitions from 5.6.3, we can now prove that the function

$$u \ : \ [0,1] \to \mathbb{R}, \ x \mapsto \left\{ \begin{array}{ll} x & \text{if} \quad x \in [0,1[, \\ 2 & \text{if} \quad x = 1. \end{array} \right.$$

is a discontinuous viscosity solution of this problem.
As for the continuous case, there exists existence and uniqueness taking into account viscosity solutions as defined in 5.6.3. As observed by E. PRADOS and O. FAUGERAS in [18], such an uniqueness result is for instance given by G. BARLES in [4] (corollary 4.1). For an existence result, we can refer to the theorem V.4.13 given by M. BARDI and I. CAPUZZO-DOLCETTA in [3]. They more specifically consider Hamilton-Jacobi Bellman equations, *i.e.* equations of the form:

$$\lambda u(x) + \sup\{-g(x,a)\cdot\nabla u(x) - c(x,a) \mid a \in A\} \ = \ 0, \tag{5.21}$$

where $\lambda \in \mathbb{R}$, A is a topological space, $g : \Omega \times A \to \mathbb{R}^n$ and $c : \Omega \times A \to \mathbb{R}$.

5.6.5 — In relation to 5.6.4, we can terminate this section by showing how the Hamiltonian associated to the s.f.s. problem, which is given by (5.17), can be rewritten in order to obtain an Hamilton-Jacobi Bellman equation of the form (5.21). In order to do this, we can first establish the following result:

Lemma. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable and convex function. Then for all $p \in \mathbb{R}^n$, we have*

$$f(p) \ = \ \sup\{f(q) - \nabla f(q)\cdot(q-p) \mid q \in \mathbb{R}^n\}.$$

Proof.— Let us fix $p \in \mathbb{R}^n$. Since $f(p) = f(q) - \nabla f(q)\cdot(q-p)$, we clearly have

$$f(p) \ \leqslant \ \sup\{f(q) - \nabla f(q)\cdot(q-p) \mid q \in \mathbb{R}^n\}.$$

Thus to prove our result, we now need to establish:

$$\forall\, q \in \mathbb{R}^n, \qquad f(p) \ \geqslant \ f(q) - \nabla f(q)\cdot(q-p). \tag{5.22}$$

- We start by supposing $n = 1$. The result being obvious when $q = p$, we will suppose that $q \neq p$. Given $t \in]0, 1[$, let us introduce $z_t = tp + (1-t)q$. Then we have

$$q - p \;=\; \frac{1}{t}\,(q - z_t),$$

and, by using the convexity of f,

$$f(q) - f(z_t) \;\geqslant\; \frac{1}{t}\,[f(q) - f(p)].$$

Thus if $q - p > 0$, we obtain

$$\frac{f(q)-f(p)}{q-p} \;=\; \frac{1/t}{1/t}\cdot\frac{f(q)-f(p)}{q-p} \;\leqslant\; \frac{f(q)-f(z_t)}{q-z_t} \;\xrightarrow[t\to+\infty]{}\; f'(q),$$

whereas when $q - p < 0$,

$$\frac{f(q)-f(p)}{q-p} \;=\; \frac{1/t}{1/t}\cdot\frac{f(q)-f(p)}{q-p} \;\geqslant\; \frac{f(q)-f(z_t)}{q-z_t} \;\xrightarrow[t\to+\infty]{}\; f'(q),$$

Whatever the sign of $q - p$ is, we can now easily claim that (5.22) is true.

- Let us suppose that $n \geqslant 2$. By fixing $q \in \mathbb{R}^2$ and by introducing

$$\gamma \;:\; \mathbb{R} \to \mathbb{R}^n,\; t \mapsto tp + (1-t)q,$$

since $f \circ \gamma : \mathbb{R} \to \mathbb{R}$ is convex and differentiable, we can deduce of the previous case that

$$f(p) \;=\; (f\circ\gamma)(1) \;\geqslant\; (f\circ\gamma)(0) - (f\circ\gamma)'(0)\cdot(0-1) \;=\; f(q) - \nabla f(q)\cdot(q-p). \qquad \square$$

5.6.6 — Let us fix $(x, p) \in \Omega \times \mathbb{R}^n$. By applying lemma 5.6.5 with $f = H(x,\,\cdot\,)$, we find

$$H(x,p) \;=\; \sup\left\{H(x,q) - \nabla_q H(x,q)\cdot(q-p) \mid q \in \mathbb{R}^n\right\},$$

where $\nabla_q H(x,q)$ corresponds to the gradient of $H(x,\,\cdot\,)$ in q. After calculations, it leads to

$$H(x,p) \;=\; \sup\left\{\frac{I(x)}{\sqrt{1+|q|^2}} + \left(\frac{I(x)\,q}{\sqrt{1+|q|^2}} + l\right)\cdot p - l_3 \;\middle|\; q \in \mathbb{R}^n\right\},$$

Knowing that the application

$$\varphi \;:\; \mathbb{R}^2 \to \mathrm{B}(0,1)\,,\; q \mapsto \frac{q}{\sqrt{1+|q|^2}}$$

is bijective, by setting $b = \varphi(q)$, we obtain

$$H(x,p) \;=\; \sup\left\{I(x)\sqrt{1-|b|^2} + \big(I(x)b + l\big)\cdot p - l_3 \;\middle|\; b \in \mathrm{B}(0,1)\right\}.$$

And since $\overline{\mathrm{B}(0,1)} = \mathrm{B}'(0,1)$, we have finally proved the following result, which effectively shows that the s.f.s. problem can be expressed with an Hamilton-Jacobi Bellman equation of the form (5.21):

Proposition. *For all $(x, p) \in \Omega \times \mathbb{R}^n$, we have*

$$H(x,p) \;=\; \sup\left\{I(x)\sqrt{1-|b|^2} + \big(I(x)b + l\big)\cdot p - l_3 \;\middle|\; b \in \mathrm{B}'(0,1)\right\}.$$

5.7 Approximation schemes

5.7.1 — To end this chapter, we will present the notion of *approximation scheme*, that can be used in order to compute a numerical approximation of the viscosity solution u of a Hamilton-Jacobi equation. As usual, given a mesh $\mathcal{M}$ of $\overline{\Omega}$, we will determine an approximation U of u at each node of $\mathcal{M}$. This approximation U will be computed iteratively:

1) We start from U^0 a first approximation of u at each node of $\mathcal{M}$.

2) Given $n \in \mathbb{N}$ and U^n an approximation of u at each node of $\mathcal{M}$, U^{n+1} is computed by mimicking the problem we want to solve, *i.e.* for all $x \in \mathcal{M}$,

$$H(x, \nabla u(x)) \ = \ 0.$$

Since the exact value of $\nabla u(x)$ will be unknown in practice, we will discretize it by using the values of U^n instead of those of u in $\mathcal{M}$ (that we are trying to compute).

3) Computations from step 2) will be done as soon as $\left\|U^{n+1} - U^n\right\|_\infty > \varepsilon$, for a given accuracy threshold $\varepsilon > 0$. Then, for the first index n such that $\left\|U^{n+1} - U^n\right\|_\infty \leqslant \varepsilon$, we will set $U = U^{n+1}$.

This approach motivates the introduction of *approximation schemes*, as done by G. BARLES and P. E. SOUGANIDIS in [5]. From a formal point of view, such a scheme is a locally bounded function

$$S \ : \ \left(\mathbb{R}_+^*\right)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}, \ (h, x, t, u) \mapsto S(h, x, t, u).$$

For all $h \in \left(\mathbb{R}_+^*\right)^n$, $x \in \Omega$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, $S(h, x, u(x), u)$ will morally corresponds to an approximation of $H(x, \nabla u(x))$. Thus given $h \in \left(\mathbb{R}_+^*\right)^n$, a function $u_h \in \mathcal{B}\big(\overline{\Omega}\big)$ will be called a *solution* (resp. a *subsolution*) of $S(h, \cdot, \cdot, \cdot)$ if, for all $x \in \overline{\Omega}$,

$$S\big(h, x, u_h(x), u_h\big) \ = \ 0 \quad \text{(resp. } S\big(h, x, u_h(x), u_h\big) \ \leqslant \ 0\text{)}.$$

In practice:

- h will correspond to the mesh size of $\mathcal{M}$,
- a solution u_h of $S(h, \cdot, \cdot, \cdot)$ will correspond to a numerical approximation of the considered Hamilton-Jacobi equation,
- and the values of such a solution u_h will be determined by solving equations of the form: given h and v, for all $x \in \mathcal{M}$, find $t \in \mathbb{R}$ such that

$$S(h, x, t, v) \ = \ 0.$$

In the sequel, we will see how to insure the existence of solutions to an given approximation scheme, and how to insure their convergence to the viscosity solution of the considered Hamilton-Jacobi equation.

REMARK. Since in practice, we are interested in computing approximation of the viscosity solution of an Hamilton-Jacobi equation on a given mesh $\mathcal{M}$ of $\overline{\Omega}$, the definition of approximation schemes could be limited to functions of the form

$$S \ : \ \left(\mathbb{R}_+^*\right)^n \times \mathcal{M} \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$$

where the integer N corresponds to the number of nodes of the mesh $\mathcal{M}$. In fact, the proof of the convergence of a solution of an approximation scheme to the viscosity solution of the Hamilton-Jacobi equation seems simpler by considering a continuous point of view, *i.e.* approximation schemes defined on $\left(\mathbb{R}_+^*\right)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big)$ instead of $\left(\mathbb{R}_+^*\right)^n \times \mathcal{M} \times \mathbb{R} \times \mathbb{R}^N$.

5.7.2 — First, we have to insure the existence of a solution to a given approximation scheme, a property called *stability*, which is traditionally defined as follows:

Definition. *An approximation scheme* $S : \big(\mathbb{R}_+^*\big)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}$ *is said to be* stable *if for all* $h \in (\mathbb{R})^n$*, it admits a solution* $u_h \in \mathcal{B}\big(\overline{\Omega}\big)$ *that can bounded independently of* h*.*

5.7.3 — An important notion, that will insure the stability of an approximation scheme as well as the convergence of its solutions to the viscosity solution of the considered Hamilton-Jacobi equation is the *monotonicity* of the scheme.

Definition. *An approximation scheme* $S : \big(\mathbb{R}_+^*\big)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}$ *is said to be* monotonous *if for all* $h \in (\mathbb{R}^n)^n$*,* $x \in \overline{\Omega}$ *and* $t \in \mathbb{R}$*, the function* $S(h, x, t, \cdot\,)$ *is non-increasing,* i.e. *if for all* $u, v \in \mathcal{B}\big(\overline{\Omega}\big)$*, we have* $S(h, x, t, u) \geqslant S(h, x, t, v)$ *as soon as* $u \leqslant v$*.*

5.7.4 — A stability result is for instance given by E. PRADOS and O. FAUGERAS in [18] (theorem 8). Given an approximation scheme $S : \big(\mathbb{R}_+^*\big)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}$, it uses on the following assumptions:

- (A_{12}) S is monotonous (in the sense of the definition 5.7.3).
- (A_{13}) For all $h \in \big(\mathbb{R}_+^*\big)^n$, $x \in \overline{\Omega}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, the function $S(h, x, \cdot\,, u)$ is non-decreasing and has a positive limit in the neighbourhood of $+\infty$.
- (A_{14}) There exists $d \in \mathbb{N}^*$ such that for all $h \in \big(\mathbb{R}_+^*\big)^n$ and $x \in \overline{\Omega}$, there exists $\Xi_{h,x} \subset \overline{\Omega}^d$ and $\sigma_{h,x} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that $S(h, x, t, u) = \sigma_{h,x}\big(t, (u(\xi))_{\xi \in \Xi_{h,x}}\big)$ for all $t \in \mathbb{R}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$.
- (A_{15}) For all $h \in \big(\mathbb{R}_+^*\big)^n$, the scheme S admits a subsolution.
- (A_{16}) All the subsolutions of S are upper bounded independently with respect to the first variable of S.

Theorem. (Stability result)
If the assumptions (A_{12}) *to* (A_{16}) *are satisfied, then* S *is stable.*

REMARK. The general idea of the proof of this result is given by E. PRADOS and O. FAUGERAS in [18]: given $h \in \mathbb{R}_+^*$ and u_0 a subsolution of $S(h, \cdot\,, \cdot\,, \cdot\,)$, they construct by induction a sequence $(u_n)_{n \in \mathbb{N}}$ of subsolutions of S. Then by considering u the limit of this sequence (which is well-defined), they prove that it is a solution of $S(h, \cdot\,, \cdot\,, \cdot\,)$. However, they have not justified why u_n, for all $n \in \mathbb{N}$, and u are in $\mathcal{B}\big(\overline{\Omega}\big)$, what should be done before talking about subsolution or solution…

5.7.5 — It seems important to highlight that if the notion of stability is essential to insure that an approximation scheme is not without solution, it does not depend on the considered Hamilton-Jacobi equation. Thus in order to obtain a convergence result, it will be necessary to link the scheme with the equation. The notion of *consistency* will enable this. As used by G. BARLES and P. E. SOUGANIDIS in [5], it deals with equation of the form (5.19). So let us take the notation of G from 5.6.1 back.

Definition. *An approximation scheme* $S : \big(\mathbb{R}_+^*\big)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}$ *is said to be* consistent *if, for all* $x \in \overline{\Omega}$ *and* $\phi \in \mathcal{C}^\infty\big(\overline{\Omega}\big)$*, we have*

$$\limsup_{\substack{h \to 0 \\ y \to x \\ \xi \to 0}} \frac{S(h, y,\ \phi(y) + \xi,\ \phi + \xi)}{h} \ \leqslant\ G^*\big(x, \psi(x), \nabla\phi(x)\big)$$

and

$$\liminf_{\substack{h \to 0 \\ y \to x \\ \xi \to 0}} \frac{S(h, y,\ \phi(y) + \xi,\ \phi + \xi)}{h} \ \geqslant\ G_*\big(x, \psi(x), \nabla\phi(x)\big).$$

5.7.6 — Given an approximation scheme $S : \left(\mathbb{R}_+^*\right)^n \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\left(\overline{\Omega}\right) \to \mathbb{R}$, it is possible to obtain the following convergence result, proved by G. BARLES and P. E. SOUGANIDIS in [5]. It requires the three following additional assumptions:

(A_{17}) S is stable.

(A_{18}) S is consistent.

(A_{19}) Equation (5.19) satisfies a *strong uniqueness property*, *i.e.* if u is an u.s.c. solution of (5.19) and v a l.s.c. solution of (5.19), then $u(x) \leqslant v(x)$ for all $x \in \overline{\Omega}$.

Theorem. (*Convergence result*)
Under the assumptions (A_{12}) *and* (A_{17}) *to* (A_{19})*, when* h *goes to* 0*, the solution* u_h *of* $S(h, \cdot\,, \cdot\,, \cdot\,)$ *converges locally uniformly to the unique discontinuous solution of equation* (5.19)*.*

Conclusion

The notion of viscosity solution turns out to be a way to insure the existence and uniqueness of a solution to Hamilton-Jacobi equations. As we have seen in introduction, our shape from shading problem is describes by a Hamilton-Jacobi equation. To solve it, we now need to see when it admits a viscosity solution, which constitutes one of the aims of the following chapter.

Chapter 6

Numerical resolution of the one-dimensional shape from shading problem

Introduction

By using the notion of viscosity solutions, we have seen in chapter 5 how to insure the existence and uniqueness of the shape from shading problem. In this new chapter, we intend to solve it numerically in the one-dimensional case. Even if it is not the most interesting case in practice, its study can be useful for a better understanding of the bi-dimensional case.
Thus here, we will assume that $n = 1$ and, without loss of generality, $\Omega =]0, \alpha[$ for some $\alpha > 0$. Then we are interesting in computing an approximation of the viscosity solution of the s.f.s. problem:

$$\begin{cases} H\big(x, u'(x)\big) \;=\; 0 & \text{if} \quad x \in]0, \alpha[, \\ u(x) \;=\; \varphi(x) & \text{if} \quad x \in \{0, \alpha\}, \end{cases} \tag{6.1}$$

where H is given, for all $(x, p) \in \Omega \times \mathbb{R}$, by

$$H(x, p) \;=\; I(x)\sqrt{1 + p^2} + lp - l_3.$$

Let us recall that I is a function from $\overline{\Omega}$ to $[0, 1]$ corresponding to the brightness intensity, and that $(l, l_3) \in \mathbb{R}^2$ is an unit vector such that $l_3 > 0$ and that is oriented to the light source. Moreover, as we have proved in 5.6.6, or as we can check in this simple case, by introducing

$$\Upsilon \;:\; [-1, 1] \times \big(\overline{\Omega} \times \mathbb{R}\big) \to \mathbb{R},\; (b\,; x, p) \mapsto I(x)\sqrt{1 - b^2} + \big(I(x)b + l\big)p - l_3,$$

for all $(x, p) \in \Omega \times \mathbb{R}$, we also have

$$H(x, p) \;=\; \sup_{b \in [-1,1]} \Upsilon(b\,; x, p). \tag{6.2}$$

We have seen in paragraph 5.5.7 that when there exists points $x \in \Omega$ such that $I(x) = 1$, we can loose the uniqueness of a solution to problem (6.1). As suggested in this same paragraph, when the set of points from Ω where I equals 1 is a finite union of disjoint connected compact sets, it is possible to avoid this difficulty by solving

$$\begin{cases} H\big(x, u'(x)\big) \;=\; 0 & \text{if} \quad x \in \Omega', \\ u(x) \;=\; \varphi(x) & \text{if} \quad x \in \partial\Omega'. \end{cases}$$

instead of (6.1), with H defined as above, $\Omega' = \{x \in \Omega \mid I(x) < 1\}$ and φ extended as a continuous function from $\partial\Omega'$ to $\mathbb{R}$. So even if we will not do it here in order to simplify our talk, everything we will do in this chapter can be extended to this particular case by replacing Ω by Ω' and $\partial\Omega$ by $\partial\Omega'$.

In section 6.1, we will construct an approximation scheme $S : \mathbb{R}_+^* \times [-1,1] \times \mathbb{R} \times \mathcal{B}([-1,1]) \to \mathbb{R}$ that we will be used to compute numerical approximation of the viscosity solution of the s.f.s. problem (6.1). An explicit expression of this approximation scheme will be determined in section 6.2. Then, we will explain in section 6.3 how to determine an explicit solution $t \in \mathbb{R}$ of the equation $S(h,x,t,u) = 0$, for fixed $h \in \mathbb{R}$, $x \in \overline{\Omega}$ and $u \in \mathcal{B}(\overline{\Omega})$, that will be necessary in order to solve numerically the s.f.s. problem (6.1). In section 6.4, we will expose a numerical algorithm that will enable to solve numerically the s.f.s. problem (6.1), and we will finally analyse its performances by presenting various numerical simulations in section 6.5.

6.1 An approximation scheme associated to the s.f.s. problem

6.1.1 — In this section, we will show how to obtain an approximation scheme associated to equation (6.1). So let us fix $h \in \mathbb{R}_+^*$ and let us define

$$\Omega_h \;=\; \big\{x \in \Omega \;\big|\; \forall\, s \in \{-1,1\},\; x + sh \in \overline{\Omega}\big\} \qquad \text{and} \qquad \partial\Omega_h \;=\; \overline{\Omega} \smallsetminus \Omega_h.$$

These definitions obviously imply that $\overline{\Omega} = \Omega_h \cup \partial\Omega_h$. Let us indicate that the notation of $\partial\Omega_h$ is abusive since it does not rigorously correspond to the boundary of Ω_h. However $\partial\Omega_h$ can be morally seen as a boundary of Ω_h that has h to diameter, as shown by the following figure:

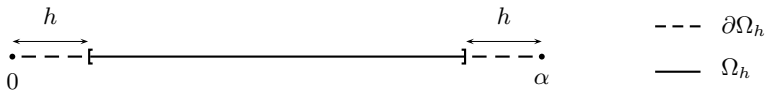

Figure 37 – Representation of the domains Ω_h and $\partial\Omega_h$

Now let explain how to define our approximation scheme S. So let us also fix $x \in \overline{\Omega}$, $t \in \mathbb{R}$ and $u \in \mathcal{B}(\overline{\Omega})$. As we have seen in section 5.7, an important assumption that will insure the stability of S as well as the convergence of its solutions to the viscosity solutions of (6.1) is its monotonicity.

- Let us suppose that $x \in \Omega_h$. In order to insure the monotonicity of our scheme, given $b \in [-1,1]$, we introduce the differential quotient

$$q(b\,;h,x,t,u) \;=\; s\,\frac{t - u(x-sh)}{h} \qquad \text{where} \qquad \left\{\begin{array}{lll} s = 1 & \text{if} & I(x)b + l \geqslant 0, \\ s = -1 & \text{if} & I(x)b + l < 0. \end{array}\right.$$

  By doing this, we always have

$$\big(I(x)b + l\big)s \;\geqslant\; 0.$$

  Consequently, for fixed b, h and x, the quantity

$$\Upsilon\big(b\,;x,\; q(b\,;h,x,t,u)\big) \;=\; I(x)\sqrt{1-b^2} + \big(I(x)b + l\big)s \cdot \frac{t - u(x-sh)}{h} - l_3$$

  is non-decreasing with respect to t and non-increasing with respect to u. Therefore we can set

$$\begin{aligned} S(h,x,t,u) \;&=\; \sup_{b\in[-1,1]} \Upsilon\big(b\,;x,\; q(b\,;h,x,t,u)\big) \\ &=\; \sup_{b\in[-1,1]} \left\{ I(x)\sqrt{1-b^2} + \big(I(x)b + l\big)s \cdot \frac{t - u(x-sh)}{h} - l_3 \right\}. \end{aligned} \tag{6.3}$$

  Let us observe that since $q(b\,;h,x,t,u)$ morally corresponds to an approximation of $u'(x)$, then thanks to (6.2), $S(h,x,t,u)$ corresponds to an approximation of $H\big(x,u'(x)\big)$.

- Since our s.f.s. problem is given with Dirichlet boundary conditions, we will suppose being able to extend φ on $\partial\Omega_h$ in order to insure the continuity of $S(h,x,\cdot,\cdot)$, which can be easily in one-dimension. Then if $x \in \partial\Omega_h$, we can set

$$S(h,x,t,u) \;=\; t - \varphi(x). \tag{6.4}$$

Finally, we have constructed an approximation scheme

$$S \;:\; \mathbb{R}_+^* \times [-1,1] \times \mathbb{R} \times \mathcal{B}([-1,1]) \to \mathbb{R},\; (h,x,t,u) \mapsto S(h,x,t,u)$$

that is non-increasing with respect to its last variable ($u \in \mathcal{B}([0,1])$) and non-decreasing with respect to its third variable ($t \in \mathbb{R}$). It essentially corresponds to the "implicit decentred scheme" used by E. PRADOS and O. FAUGERAS in [18]. In fact, we will prove in the following sections that despite its present expression, we can obtain explicit expressions of its values and solutions.

6.1.2 — The following result, proved by E. PRADOS and O. FAUGERAS in [18] in the bi-dimensional case, will be useful to prove the stability of the previous approximation scheme S as well as to initialize our numerical algorithms.

Proposition. *For all $h \in \mathbb{R}_+^*$, the function*

$$u_0 \;:\; [-1,1] \to \mathbb{R},\; x \mapsto \frac{-lx}{l_3} + C,$$

where $C \in \mathbb{R}$ is chosen so that $u_0(x) \leqslant \min_{\partial\Omega_h} \varphi$ for all $x \in \overline{\Omega}$, is a subsolution of $S(h,\,\cdot\,,\,\cdot\,,\,\cdot\,)$.

6.1.3 — Now let us explain why the approximation scheme S constructed in 6.1.1 is stable. According to theorem 5.7.4, it is sufficient to explain when assumptions (A_{12}) to (A_{16}) are satisfied.

- The scheme S was constructed in order to insure that it is non-increasing with respect to its fourth variable ($u \in \mathcal{B}([-1,1])$). Thus ($A_{12}$) is always satisfied.
- Similarly, S was constructed in order to insure that it is non-decreasing with respect to its third variable ($t \in \mathbb{R}$). In addition, for fixed $h \in \mathbb{R}_+^*$, $x \in \overline{\Omega}$, $u \in \mathcal{B}\big(\overline{\Omega}\big)$ and $t \to +\infty$:
  - Thanks to (6.4), if $x \in \partial\Omega_h$, then we clearly have $S(h,x,t,u) = t - \varphi(x) \to +\infty$.
  - Thanks to (6.3), if $x \in \Omega_h$, except when $I(x) = l = 0$, there exists $b \in [-1,1]$ such that $I(x)b + l > 0$ (indeed, $I(x) \in [0,1]$ and (l,l_3) is an unit vector such that $l_3 > 0$), hence $S(h,x,t,u) \to +\infty$.

  This shows that (A_{13}) is satisfied as soon as we do not have $l = I(x) = 0$.
- Let us fix $h \in \mathbb{R}_+^*$ and $x \in \overline{\Omega}$.
  - If $x \in \partial\Omega_h$, for all $t \in \mathbb{R}$ and $U \in \mathbb{R}^2$, let us set

$$\sigma_{h,x}(t,U) \;=\; t - \varphi(x).$$

  - Now let us suppose that $x \in \Omega_h$, and let us fix $t \in \mathbb{R}$ and $U = (U_-,U_+) \in \mathbb{R}^2$. By setting

$$\sigma_{h,x}(t,U) \;=\; \sup_{b\in[-1,1]} \left\{ I(x)\sqrt{1-b^2} + \big(I(x)b + l\big)s \cdot \frac{t-\alpha}{h} - l_3 \right\},$$

    where

$$\begin{cases} s = 1, & \alpha = U_-, \quad \text{if} \quad I(x)b + l \geqslant 0, \\ s = -1, & \alpha = U_+, \quad \text{if} \quad I(x)b + l < 0, \end{cases}$$

    we can easily observe that

$$S(h,x,t,u) \;=\; \sigma_{h,x}\Big(t, \big(u(x-h), u(x+h)\big)\Big).$$

By extending φ to $\partial\Omega_h$ in order to insure the continuity of $S(h,x,\,\cdot\,,\,\cdot\,)$, we can prove by refering to lemma 4 of [18] that

$$\sigma_{h,x}\ :\ \mathbb{R}\times\mathbb{R}^2\to\mathbb{R},\ (t,U)\mapsto\sigma_{h,x}(t,U)$$

is continuous, and so that (A_{14}) holds for $d=2$ and $\Xi_{h,x}=\{x-h,x+h\}$.

- According to proposition 6.1.2, assumption (A_{15}) is satisfied.
- Thanks to proposition 6 of [18], assumption (A_{16}) is also satisfied.

6.1.4 — Now let us terminate this section by indicating when the solutions of the approximation scheme S converge to the viscosity solution of the s.f.s. problem (6.1). In order to do this, we refer to theorem 5.7.6, which says that it is sufficient to satisfy assumptions (A_{12}) and (A_{17}) to (A_{19}).

- We have seen that previously that S was constructed in order to satisfy assumption (A_{12}).
- The stability property (A_{17}) was studied in the previous paragraph (6.1.3).
- Conditions that insure that the consistency assumption (A_{18}) and the strong uniqueness property from (A_{19}) are satisfied and are given by E. PRADOS and O. FAUGERAS in [18]. We will not expose them here and we invite the reader that want to know more detail about it to refer this report.

6.2 An explicit expression of the approximation scheme

6.2.1 — As constructed in 6.1.1, given $h\in\mathbb{R}_+^*$, $x\in\Omega_h$, $t\in\mathbb{R}$ and $u\in\mathcal{B}(\overline{\Omega})$, $S(h,x,t,u)$ just corresponds to the supremum of a real function:

$$S(h,x,t,u)\ =\ \sup_{b\in[-1,1]}\left\{I(x)\sqrt{1-b^2}+\big(I(x)b+l\big)s\cdot\frac{t-u(x-sh)}{h}-l_3\right\},$$

where the value of $s\in\{-1,1\}$ depends on the sign of $I(x)b+l$:

$$\left\{\begin{array}{ll} s=1 & \text{if}\quad I(x)b+l\geqslant 0,\\ s=-1 & \text{if}\quad I(x)b+l<0,\end{array}\right.$$

which does not make calculation of its value easy as soon as $I(x)\neq 0$. In this section, for such fixed quantities h, x, t and u, we propose to determine an explicit expression of $S(h,x,t,u)$.

6.2.2 — To achieve our objectives, we will use the following result, that can be proved by doing very classical calculations.

Lemma. *For all $\delta,\alpha>0$, $w,c\in\mathbb{R}$, the function*

$$\psi\ :\ \big[-\sqrt{\alpha},\sqrt{\alpha}\big]\to\mathbb{R},\ b\mapsto\delta\sqrt{\alpha-b^2}+wb+c$$

is concave. It has an unique critical point b^* on $]-\sqrt{\alpha},\sqrt{\alpha}[$, which corresponds to its global maximiser and which satisfies

$$b^*\ =\ \frac{w\sqrt{\alpha}}{\sqrt{\delta^2+w^2}}\qquad\textit{and}\qquad\psi(b^*)\ =\ c+\sqrt{\alpha(\delta^2+w^2)}.$$

6.2.3 — In this paragraph, we will fix $h \in \mathbb{R}_+^*$, $x \in \Omega_h$ such that $I(x) > 0$, $t \in \mathbb{R}$ and $u \in \mathcal{B}(\overline{\Omega})$. Since now, h, x, t and u are fixed once for all, for all $b \in [-1,1]$, we will abusively denote by $\Psi(b)$ the quantity $\Upsilon\big(b\,;x,\ q(b\,;h,x,t,u)\big)$. Hence a function $\Psi : [-1,1] \to \mathbb{R}$.

1) Let us introduce

$$R_1 \ = \ \{b \in [-1,1] \mid I(x)b + l > 0\} \qquad \text{and} \qquad R_{-1} \ = \ \{b \in [-1,1] \mid I(x)b + l < 0\}.$$

By doing this,

$$S(h,x,t,u) \ = \ \sup_{[-1,1]} \Psi \ = \ \max\left\{ \sup_{R_1} \Psi,\ \sup_{R_{-1}} \Psi \right\}.$$

2) For all $s \in \{-1,1\}$, let $\psi_s : [-1,1] \to \mathbb{R}$ be the function defined, for all $b \in [-1,1]$, by

$$\begin{aligned} \psi_s(b) \ &= \ I(x)\sqrt{1-b^2} \ + \ \big(I(x)b + l\big)s\,\frac{t-u(x-sh)}{h} \ - \ l_3 \\ &= \ \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1}_{\alpha_s} - b^2} \ + \ \underbrace{sI(x)\,\frac{t-u(x-sh)}{h}}_{w_s}\,b \ + \ \underbrace{sl\,\frac{t-u(x-sh)}{h} - l_3}_{c_s}. \end{aligned}$$

Functions Ψ and ψ_s are obviously linked by the relation $\Psi_{|R_s} = \psi_{s\,|R_s}$. Moreover, thanks to lemma 6.2.2, the function ψ_s is concave and has an unique critical point b_s^* on $[-1,1]$ corresponding to its global maximizer, and

$$\psi_s(b_s^*) \ = \ c_s + \sqrt{\alpha_s(\delta_s^2 + w_s^2)} \ = \ sl\,\frac{t-u(x-sh)}{h} - l_3 + I(x)\sqrt{1 + \frac{[t-u(x-sh)]^2}{h^2}}. \quad (6.5)$$

3) Let us suppose that $|l| \leqslant I(x)$, and let us fix $s \in \{-1,1\}$.

- If Ψ has no critical point on R_s, then it reaches its maximum on the boundary of R_s, *i.e.* at s or $-l/I(x)$. But it can not be reached at s since $\Psi_{|R_s} = \psi_{s\,|R_s}$ and ψ_s reaches its maximum on $]-1,1[$, according to lemma 6.2.2. Hence

$$\sup_{R_s} \Psi \ = \ \Psi\left(\frac{-l}{I(x)}\right) \ = \ \sqrt{I^2(x) - l^2} - l_3. \quad (6.6)$$

- If Ψ has a critical point on R_s, then it is also a critical point of ψ_s. Thus by referring to point 2), we deduce that $b_s^* \in R_s$, and that b_s^* also corresponds to the global maximizer of Ψ on R_s, hence

$$\sup_{R_s} \Psi \ = \ \psi_s(b_s^*) \ = \ sl\,\frac{t-u(x-sh)}{h} - l_3 + I(x)\sqrt{1 + \frac{[t-u(x-sh)]^2}{h^2}}.$$

4) Let us suppose that $|l| > I(x)$. If $l > I(x)$, then $R_1 = [-1,1]$, so by referring to point 2),

$$\sup_{[-1,1]} \Psi \ = \ \psi_1(b_1^*) \ = \ l\,\frac{t-u(x-h)}{h} - l_3 + I(x)\sqrt{1 + \frac{[t-u(x-h)]^2}{h^2}}.$$

But if $l < -I(x)$, then $R_{-1} = [-1,1]$, and so thanks to point 2),

$$\sup_{[-1,1]} \Psi \ = \ \psi_{-1}\big(b_{-1}^*\big) \ = \ -l\,\frac{t-u(x+h)}{h} - l_3 + I(x)\sqrt{1 + \frac{[t-u(x+h)]^2}{h^2}}.$$

According to points 1) to 4), we have established the following result:

Theorem. *Let us fix $h \in \mathbb{R}_+^*$, $x \in \Omega_h$ such that $I(x) > 0$, $t \in \mathbb{R}$ and $u \in \mathcal{B}(\overline{\Omega})$.*

- *If $|l| \leqslant I(x)$, then by setting*

$$a_s(h,x,t,u) \;=\; \left\{ \begin{array}{ll} \psi_s(b_s^*) & \textit{if} \quad b_s^* \in R_s, \\ \Psi\big(-l/I(x)\big) & \textit{if} \quad b_s^* \notin R_s, \end{array} \right.$$

  we have

$$S(h,x,t,u) \;=\; \max\{a_s(h,x,t,u) \mid s \in \{-1,1\}\}.$$

- *If $|l| > I(x)$, then we have*

$$S(h,x,t,u) \;=\; \left\{ \begin{array}{ll} \psi_1(b_1^*) & \textit{if} \quad l > I(x), \\ \psi_{-1}(b_{-1}^*) & \textit{if} \quad l < -I(x). \end{array} \right.$$

As a reminder, the exact values of $\psi_s(b_s^*)$ and $\Psi\big(-l/I(x)\big)$ are respectively given in (6.5) *and* (6.6).

6.3 Explicit solutions of the approximation scheme

6.3.1 — In paragraph 6.1.3, we have exposed various conditions that insure that our approximation scheme S is stable, *i.e.* that for all $h \in \mathbb{R}_+^*$, there exists a function $u_h \in \mathcal{B}(\overline{\Omega})$ such that for all $x \in \overline{\Omega}$,

$$S(h,x,u_h(x),u_h) \;=\; 0.$$

In this section, given $h \in \mathbb{R}_+^*$ and $u \in \mathcal{B}(\overline{\Omega})$ such that for all $x \in \overline{\Omega}$, the equation

$$S(h,x,t,u) \;=\; 0 \tag{6.7}$$

has a solution $t \in \mathbb{R}$, we will determine the explicit expression of such a solution t. Since when $x \in \partial\Omega_h$, we have $S(h,x,t,u) = t - \varphi(x)$, it is clear that $t = \varphi(x)$ is the only solution in this case. As in the previous section, the non-obvious case consists in considering this equation when $x \in \Omega_h$. In order to do this, we will thus fix $h \in \mathbb{R}_+^*$, $x \in \Omega_h$ and $u \in \mathcal{B}(\overline{\Omega})$.

REMARK. Let us note that for an arbitrary choice of $u \in \mathcal{B}(\overline{\Omega})$, equation (6.7) can note have a solution. But as soon as S satisfies assumptions (A_{12}) to (A_{14}), we can for instance verify that for all subsolution u of $S(h,\,\cdot\,,\,\cdot\,,\,\cdot\,)$ and all $x \in \overline{\Omega}$, equation (6.7) has a solution $t \in \mathbb{R}$.

6.3.2 — To begin with, let us suppose that $I(x) = 0$. In this particular case, by using the definition of S from 6.1.1, equation $S(h,x,t,u) = 0$ can be immediately rewritten

$$sl\,\frac{t - u(x - sh)}{h} \;=\; 0 \qquad \text{where} \qquad \left\{ \begin{array}{lll} s = 1 & \text{if} & l \geqslant 0, \\ s = -1 & \text{if} & l < 0. \end{array} \right.$$

- If $l = 0$, we thus obtain $S(h,x,t,u) = 0$ for all $t \in \mathbb{R}$.
- If $l \neq 0$, then it is clear that its unique solution is

$$t \;=\; u(x - sh) + \frac{shl_3}{l}.$$

6.3.3 — From now, we will focus on the non-obvious case $I(x) > 0$. In paragraph 6.2.3, we were looking for an explicit value of $S(h,x,t,u)$, for fixed $h \in \mathbb{R}_+^*$, $x \in \Omega_h$ such that $I(x) > 0$, $t \in \mathbb{R}$ and $u \in \mathcal{B}(\overline{\Omega})$. Here, for such fixed quantities h, x and u, we are looking for the solutions $t \in \mathbb{R}$ of the equation (6.7), *i.e.* $S(h,x,t,u) = 0$.
Thus, in the rest of this section, by the same principle as in paragraph 6.2.3, since h, x and u will be fixed once for all, in accordance with the value of $t \in \mathbb{R}$, we will abusively denote by:

– $\Psi(\cdot\,;t)$ the function from $[-1,1]$ to $\mathbb{R}$ defined, for all $b\in[-1,1]$, by $\Psi(b;t)=\Upsilon\big(b;x,q(b;h,x,t,u)\big)$,

– for all $s\in\{-1,1\}$, by $\psi_s(\,\cdot\,;t)$ the function ψ_s, and by $b_s^*(t)$ its global maximiser.

Since the notations of R_1 and R_{-1} do not depend on t, we will reuse these notations. By doing this, for all $t\in\mathbb{R}$ and $s\in\{-1,1\}$, the quantity $a_s(h,x,t,u)$ used in theorem 6.2.3 to give the explicit expression of $S(h,x,t,u)$ can be rewritten:

$$a_s(h,x,t,u)\;=\;\left\{\begin{array}{ll}\psi_s\big(b_s^*(t)\,;t\big) & \text{if}\quad b_s^*(t)\in R_s,\\ \Psi\big(-l/I(x)\,;t\big) & \text{if}\quad b_s^*(t)\notin R_s.\end{array}\right.$$

6.3.4 — Thanks to the expression of $S(h,x,t,u)$ obtained in 6.2.3, it seems interesting to determine, for all $s\in\{-1,1\}$, the elements that can belong to

$$\Theta_s\;=\;\big\{t\in\mathbb{R}\;\big|\;\psi_s\big(b_s^*(t)\,;t\big)=0,\;b_s^*(t)\in R_s\big\}\,.$$

In order to do this, given $s\in\{-1,1\}$ and $t\in\mathbb{R}$, we will more simply set

$$q_s(t)\;=\;s\,\frac{t-u(x-sh)}{h}.$$

By doing this, we obviously have, for all $b\in[-1,1]$,

$$\psi_s\big(b\,;t\big)\;=\;I(x)\sqrt{1-b^2}+\big(I(x)b+l\big)q_s(t)-l_3.$$

Proposition. *Let us suppose that $I(x)>0$, and let us fix $s\in\{-1,1\}$ and $t\in\Theta_s$. Then*

$$Aq_s^2(t)+Bq_s(t)+C\;=\;0\qquad\textit{with}\qquad\left\{\begin{array}{l}A\;=\;I^2(x)-l^2,\\ B\;=\;2ll_3,\\ C\;=\;I^2(x)-l_3^2.\end{array}\right.\tag{6.8}$$

Consequently:

(*i*) *If $A=0$, then $B\neq0$ and*

$$t\;=\;u(x-sh)-\frac{shC}{B}.$$

(*ii*) *If $A\neq0$, then we have $\Delta=B^2-4AC\geqslant0$ and existence of $\sigma\in\{-1,1\}$ such that*

$$t\;=\;u(x-sh)+sh\,\frac{-B+\sigma\sqrt{\Delta}}{2A}.$$

Proof.— By referring to point 2) from 6.2.3, we have

$$\psi_s\big(b_s^*(t)\,;t\big)\;=\;I(x)\sqrt{1+q_s^2(t)}+slq_s(t)-l_3.$$

Thus we have

$$\begin{array}{rcl}\psi_s\big(b_s^*(t)\,;t\big)\;=\;0 & \Longrightarrow & I(x)\sqrt{1+q_s^2(t)}\;=\;l_3-lq_s(t)\\ & \Longrightarrow & I^2(x)[1+q_s^2(t)]\;=\;l_3^2-2ll_3+l^2q_s^2(t)\\ & \Longrightarrow & \underbrace{\big(I^2(x)-l^2\big)}_{A}q_s^2(t)+\underbrace{2ll_3}_{B}q_s(t)+\underbrace{\big(I^2(x)-l_3^2\big)}_{C}\;=\;0,\end{array}$$

which proves the validity of relation (6.8). Now let us prove points (*i*) and (*ii*).

(*i*) If $A = 0$, then $|l| = I(x)$ by the definition of A, and so $l \neq 0$ since we have supposed $I(x) > 0$. In addition, (6.8) can be rewritten

$$0 \;=\; Bq_s(t) + C \;=\; B\,\frac{t - u(x - sh)}{h} + C.$$

And since l is non-zero, then B is also non-zero, so we obtain the expected result.

(*ii*) If $A \neq 0$, then since (l, l_3) is an unit vector, we have

$$\Delta \;=\; (2ll_3)^2 - 4[I^2(x) - l^2][I^2(x) - l_3^2] \;=\; 4I^2(x)[l^2 + l_3^2 - I^2(x)] \;=\; 4I^2(x)[1 - I^2(x)] \;\geqslant\; 0.$$

Thus according to relation (6.8), we deduce that there exists $\sigma \in \{-1, 1\}$ such that

$$q_s(t) \;=\; \frac{-B + \sigma\sqrt{\Delta}}{2A},$$

and we can easily conclude by using the definition of $q_s(t)$. □

6.3.5 — The following proposition gives a characterization of the existence of a solution to equation (6.7), that will be useful in order to explicit the value of such a solution.

Proposition. *Let us fix $x \in \Omega_h$ such that $0 < I(x) < 1$. The following conditions are equivalent:*

(*i*) *There exists $t \in \mathbb{R}$ such that $S(h, x, t, u) = 0$.*

(*ii*) *There exists $s \in \{-1, 1\}$ such that $\Theta_s \neq \varnothing$.*

If these conditions are satisfied, then $\tau = \min(\Theta_{-1} \cup \Theta_1)$ is well-defined and satisfies $S(h, x, \tau, u) = 0$.

Proof.— (*i*) $\Rightarrow$ (*ii*) : According to theorem 6.2.3, the implication is clear as soon as $|l| > I(x)$. If $|l| \leqslant I(x)$, since $I(x) < 1 = l^2 + l_3$, we then have

$$\Psi\left(\frac{-l}{I(x)}\,;t\right) \;=\; \sqrt{I^2(x) - l^2} - l_3 \;<\; \sqrt{l_3^3} - l_3 \;=\; 0.$$

Consequently, thanks to theorem 6.2.3, we also obtain the expected implication.

(*ii*) $\Rightarrow$ (*i*) : Let us suppose that (*ii*) is satisfied. In this case, thanks to proposition 6.3.4, the sets Θ_1 and Θ_{-1} are finite, and so the quantity $\tau = \min(\Theta_{-1} \cup \Theta_1)$ is well-defined. We will prove that it satisfies $S(h, x, \tau, u) = 0$, which will immediately imply (*i*).

- By the definition of τ, let us fix $s \in \{-1, 1\}$ such that $\tau \in \Theta_s$. Then $b_s^*(\tau) \in R_s$, so according to theorem 6.2.3, we get

$$0 \;=\; \psi_s\big(b_s^*(\tau)\,;\tau\big) \;\leqslant\; \sup_{b\in[-1,1]} \Psi(b\,;\tau) \;=\; S(h, x, \tau, u). \tag{6.9}$$

- Now let us fix $s \in \{-1, 1\}$ and $t_s \in \Theta_s$. By the definition of τ, we thus have $\tau \leqslant t_s$. But since for all $b \in [-1, 1]$, function $\Psi(b\,;\,\cdot\,)$ is non-decreasing by construction, it implies

$$\sup_{b\in R_s} \Psi(b\,;\tau) \;\leqslant\; \sup_{b\in R_s} \Psi(b\,;t_s).$$

And since $t_s \in \Theta_s$, therefore thanks to point 2) from 6.2.3,

$$b_s^*(t_s) \in R_s \qquad \text{and} \qquad \sup_{b\in R_s} \Psi(b\,;t_s) \;=\; \psi_s\big(b_s^*(t_s)\,;t_s\big) \;=\; 0.$$

Consequently, for all $s \in \{-1, 1\}$, we have obtained

$$\sup_{b \in R_s} \Psi(b\,;\tau) \;\leqslant\; 0,$$

so thanks to point 1) from 6.2.3,

$$S(h, x, \tau, u) \;=\; \left\{ \sup_{b \in R_{-1}} \Psi(b\,;t_s),\; \sup_{b \in R_1} \Psi(b\,;t_s) \right\} \;\leqslant\; 0 \tag{6.10}$$

According to (6.9) and (6.10), we have already proved that $S(h, x, \tau, u) = 0$. □

REMARK. Let us observe that since (l, l_3) is an unit vector, the assumption $I(x) < 1$ is always satisfied as soon as $|l| > I(x)$.

6.3.6 — Let us terminate this section by explaining how to determine an explicit solution of the equation (6.7) in the non-obvious case $I(x) > 0$. In order to do this, let us suppose that $0 < I(x) < 1$ and that (6.7) has a solution. Then thanks to proposition 6.3.5, we have

$$S(h, x, \tau, u) \qquad \text{with} \qquad \tau \;=\; \min(\Theta_{-1} \cup \Theta_1).$$

To conclude, we now just need to determine the explicit expression of τ. In order to do this, it is sufficient to determine the elements that belong to Θ_{-1} and Θ_1, which can be done by referring to proposition 6.3.4. So let us reuse the notations from this proposition.

- If $A = 0$, then $I^2(x) = l^2$, and so $I(x) = |l|$. In this case, there exists an unique $s \in \{-1, 1\}$ such that $R_s \neq \varnothing$, and for this s, we have $R_s =]-1, 1[\smallsetminus \{-s\}$. Let us consider

$$t_s \;=\; u(x - sh) - \frac{shC}{B}.$$

Thanks to lemma 6.2.2, it is clear that

$$b_s^*(t_s) \;=\; \frac{q_s(t)}{\sqrt{1 + q_s^2(t)}} \;\in\;]-1, 1[\;\subset\; R_s.$$

Otherwise, by referring to relation (6.5) and by using the definitions of t_s, B and C, we get

$$\psi_s\big(b_s^*(t_s)\,;t_s\big) \;=\; -\frac{lC}{B} - l_3 + |l|\sqrt{1 + \frac{C^2}{B^2}} \;=\; \underbrace{-\frac{l(l^2 + l_3^2)}{2ll_3}}_{-\frac{(l^2+l_3^2)}{2l_3}} + \underbrace{\sqrt{\frac{l^2(l^2 + l_3^2)^2}{4l^2 l_3^2}}}_{\frac{(l^2+l_3^2)}{2l_3}} \;=\; 0.$$

Consequently, this proves that

$$\Theta_s \;=\; \{t_s\} \;=\; \left\{ u(x - sh) - \frac{shC}{B} \right\}.$$

Therefore, in this case, we have $\tau = t_s$, and

$$S(h, x, t_s, u) \;=\; 0.$$

- If $A \neq 0$, then for all $s, \sigma \in \{-1, 1\}$, by setting

$$t_{s,\sigma} \;=\; u(x - sh) + sh\,\frac{-B + \sigma\sqrt{\Delta}}{2A},$$

we can observe that

$$\Theta_s \;=\; \left\{ t_{s,\sigma} \;\middle|\; \sigma \in \{-1,1\},\; \psi_s\big(b_s^*(t_{s,\sigma})\,; t_{s,\sigma}\big) = 0,\; b_s^*(t_{s,\sigma}) \in R_s \right\}.$$

REMARK. Let us formulate the following observations:

- When $A \neq 0$, we have not found for which values of s and σ we have $\tau = t_{s,\sigma}$, *i.e.* $S(h, x, t_{s,\sigma}, u) = 0$.
- If $I(x) = 1$, proposition 6.3.5 becomes generally false, and so that we can not apply the method describe above in order to determine the value of a t satisfying $S(h, x, t, u) = 0$. In fact, in this case, the expression of $S(h, x, t, u)$ given by theorem 6.2.3 can be independent of t, and it could be impossible to determine the expression of such a t.
  From a theoretical point of view, we have explained in paragraph 5.5.7 that if there exists $x \in \Omega$ such that $I(x) = 1$, then the s.f.s. problem can have several viscosity solutions, but that we can recover the uniqueness of a viscosity solution by solving this problem on $\Omega' = \{x \in \Omega \mid I(x) < 1\}$ instead of Ω. Here, we have highlighted that it is also necessary to do it from a numerical point of view.

6.4 Associated algorithm

6.4.1 — Definition. *A* mesh *of* $\overline{\Omega} = [0, \alpha]$ *(resp.* $\Omega =]0, \alpha[$*) is a subset of the form:*

$$\begin{aligned} \overline{\Omega}_N \;&=\; \left\{ ih \;\middle|\; i \in \{0, \ldots, N\} \right\} \\ \textit{(resp. } \Omega_N \;&=\; \left\{ ih \;\middle|\; i \in \{1, \ldots, N-1\} \right\}\textit{,} \\ \partial\Omega_N \;&=\; \overline{\Omega}_N \setminus \Omega_N\textit{),} \end{aligned}$$

where N *is an integer larger than* 2 *and* h *is defined by*

$$h \;=\; \frac{\alpha}{N}. \tag{6.11}$$

With these notations, each element of the mesh is called a node*, and* h *is called the* mesh size*.*

6.4.2 — Let us formulate some remarks about the previous definition and let us specify some notations. So let N be an integer larger than 2.

- We can first observe that $\overline{\Omega}_N \;=\; \Omega_N \cup \partial\Omega_N$.
- Then it is important to note that meshes $\overline{\Omega}_N$ and Ω_N are fully determined by the integer N: that is why we have chosen to make it appear as an index. Therefore in this sense, $\overline{\Omega}_N$ does not correspond to the closure of Ω_N and $\partial\Omega_N$ does not correspond to its boundary.
- It is also important to observe that the mesh-size h is fully determined by N. In order to simplify the notations, we will not signal this dependence (but we will keep it in mind).
  In addition, for this choice of h, let us remark that Ω_N precisely corresponds to the nodes of $\overline{\Omega}_N$ that belong to the set Ω_h introduced in 6.1.1.

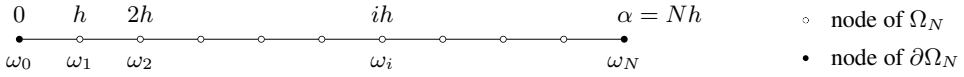

Figure 38 – Representation of the meshes $\overline{\Omega}_N$ and Ω_N of $\overline{\Omega}$ and Ω

- In what follows, given $N \geqslant 2$, h will automatically corresponds to the mesh-size of $\overline{\Omega}_N$, *i.e.* h will be automatically defined by the relation (6.11). In addition, given $i \in \{0, \dots, N\}$, we will denote by ω_i the node ih.

The following figure, that can be compared to figure 37, illustrates this all:

6.4.3 — Given $N \geqslant 2$ and h defined by the relation (6.11), we want to compute an approximation u_h of the viscosity solution u of the s.f.s. problem (6.1) at each node of $\overline{\Omega}_N$. This approximation will be computed as suggested by the proof of the stability result 5.7.4:

– We start with $u_0 \in \mathcal{B}\big(\overline{\Omega}\big)$ a subsolution of $S(h, \cdot\,, \cdot\,, \cdot\,)$.

– Given $n \in \mathbb{N}$ and u_n a subsolution of $S(h, \cdot\,, \cdot\,, \cdot\,)$, u_{n+1} is chosen such that for all $x \in \overline{\Omega}$,

$$S(h, x, u_{n+1}(x), u_n) \;=\; 0.$$

– In the end, we take u_h the limit of the sequence $(u_n)_{n\in\mathbb{N}}$.

Let us remind that in section 6.1, we have presented various conditions that insure that such a sequence $(u_n)_{n\in\mathbb{N}}$ is already well-defined and convergent, and that its limit u_h converges to u when h goes to 0, *i.e.* N is sufficiently large. Thus from now on, given $v \in \mathcal{B}\big(\overline{\Omega}\big)$, our goal will consist in solving, for all $x \in \overline{\Omega}_N$, an equation of the form $S(h, x, t, v) = 0$ with respect to $t \in \mathbb{R}$, which can be done as explained in paragraph 6.3.6.
Now let us observe that if $v \in \mathcal{B}\big(\overline{\Omega}\big)$, then for all $x \in \overline{\Omega}_N$, the value of $S(h, x, t, v)$ only depends on the values of h, x, t and those taken by v in other nodes from $\overline{\Omega}_N$. That is why in the sequel, by using the notation of $\sigma_{h,x}$ from 6.1.3, we will deal with the function $S_N : \overline{\Omega}_N \times \mathbb{R} \times \mathbb{R}^{N+1} \to \mathbb{R}$ defined, for all $x \in \overline{\Omega}_N$, $t \in \mathbb{R}$ and $V = \big(V(i)\big)_{0\leqslant i\leqslant N} \in \mathbb{R}^{N+1}$, by:

– $S_N(x, t, V) = \sigma_{h,x}\Big(t, \big(V(i-1), V(i+1)\big)\Big)$ if $x \in \Omega_N$ and $i \in \{1, \dots, N-1\}$ satisfies $x = \omega_i$,

– $S_N(x, t, V) = t - \varphi(x)$ if $x \in \partial\Omega_N$.

By doing this, V will be called a *subsolution* of S_N if $S_N(\omega_i, V(i), V) \leqslant 0$ for all $i \in \{0, \dots, N\}$.

6.4.4 — *Suggested algorithm.*

Given $N \geqslant 2$, h defined as in (6.11), $\varepsilon > 0$ an accuracy threshold and $n_{\max} \in \mathbb{N}^*$, let us explain how to obtain $U_{\text{app}} \in \mathbb{R}^{N+1}$ an approximation of u at each node of $\overline{\Omega}_N$. Just by knowing the values of I at each nodes of Ω_N, those of the components of the unit vector (l, l_3) oriented to the light source and those of the function φ at each node of $\partial\Omega_N$, U_{app} will be computed as follows:

1) We start with $U_0 = \big(U_0(i)\big)_{0\leqslant i\leqslant N}$ a subsolution of S_N satisfying $U_0(i) = \varphi(\omega_i)$ for index i such that $\omega_i \in \partial\Omega_N$.

2) Given $n \in \mathbb{N}$ and $U_n = \big(U_n(i)\big)_{0\leqslant i\leqslant N}$ a subsoltion of S_N, $U_{n+1} = \big(U_{n+1}(i)\big)_{0\leqslant i\leqslant N}$ is chosen such that, for all $i \in \{0, \dots, N\}$,

$$S_N(\omega_i, U_{n+1}(i), U_n) \;=\; 0.$$

In practice, $U_{n+1}(i)$ can be explicitly computed as explained in paragraph 6.3.6.

3) Computations from 2) are done as long as $\|U_{n+1} - U_n\|_\infty \geqslant \varepsilon$ and $n \leqslant n_{\max}$. If there exists $n \leqslant n_{\max}$ such that $\|U_{n+1} - U_n\|_\infty < \varepsilon$, we set $U_{\text{app}} = U_{n+1}$ for the smallest of these integer n. If this is not the case, we set $U_{\text{app}} = U_{n_{\max}}$.

Finally, for all $i \in \{0, \dots, N\}$, $U_{\text{app}}(i)$ corresponds to an approximation of the function u_h from 6.4.3 at the node ω_i. As a reminder, u_h converges to u the viscosity solution of (6.1) as soon as h goes to 0.

REMARK. Let us formulate some observations about the previous algorithm:

- At step 2), it is important to note that for all $i \in \{0, \dots, N\}$, the value of $U_{n+1}(i)$ depends on the values of U_n, but not those of U_{n+1}. Thus, computations done at this step can be performed independently, hence a possible parallelization.
- Still at step 2), starting with a subsolution U_n of S_N, for all $i \in \{0, \dots, N\}$, $U_{n+1}(i)$ is chosen such that $S_N(\omega_i, U_{n+1}(i), U_n) = 0$. Since S_N is non-increasing with respect to its last variable, then we have $S_N(h, U_{n+1}(i), U_{n+1}) \leqslant 0$ for all $i \in \{0, \dots, N\}$, and so U_{n+1} is also a subsolution of S_N.
- As explain in introduction, if there exists a point of Ω such that $I(x) = 1$, then we can adapt our algorithm by replacing Ω_N by $\Omega'_N = \left\{x \in \overline{\Omega}_N \;\middle|\; x \in \Omega'\right\}$ and $\partial\Omega_N$ by $\partial\Omega'_N = \overline{\Omega}_N \setminus \Omega'_N$.

6.5 Numerical simulations

6.5.1 — In this section, we will present various simulations done in order to solve the s.f.s. problem (6.1) by a direct use of the algorithm from paragraph 6.4.4, or after various adaptations of it. Thus we will take the notations from this paragraph back, and unless otherwise indicated, for all the simulations presented in this section:

- In step 1), by setting:

$$\mu = \min_{x \in \partial\Omega_N} \varphi(x) \quad \text{and} \quad C = \begin{cases} \mu & \text{if} \quad l \geqslant 0, \\ \mu + \dfrac{Nhl}{l_3} & \text{if} \quad l < 0, \end{cases}$$

  then we have $C \leqslant \max_{\partial\Omega_N} \varphi$. Thus thanks to proposition 6.1.2, the subsolution U_0 was chosen such that, for all index i satisfying $\omega_i \in \Omega_N$,

$$U_0(i) = C - \frac{ilh}{l_3}.$$

- Computations from step 2) will be done by referring to paragraph 6.3.6, and by using a vectorization.
- The accuracy threshold ε from 3) will be always equal to 10^{-3}.
- The *number of iterations* of the algorithm from 6.4.4 will correspond to the first integer n such that $\|U_{n+1} - U_n\|_\infty < \varepsilon$. We will denote it by ι, and we will then say that the algorithm from 6.4.4 *converges* in ι iterations.
  Concretely, ι corresponds to the number of times the computations from step 2) are performed.
- The vector

$$U_{\text{app}} = \big(U_{\text{app}}(i)\big)_{0 \leqslant i \leqslant N} \in \mathbb{R}^{N+1}$$

  will always corresponds to the vector obtained in output of our algorithm. In addition, given $u : \Omega \to \mathbb{R}$ a viscosity solution of (6.1),

$$U = \big(U(i)\big)_{0 \leqslant i \leqslant N} \in \mathbb{R}^{N+1}$$

will always corresponds to the vector such that $U(i) = u(\omega_i)$ for all $i \in \{0, \dots, N\}$. With these notations, we will consider the following errors:

$$\eta_\infty = \|U_{\text{app}} - U\|_\infty \quad \text{and} \quad \eta_{\text{ave}} = \frac{1}{N+1} \sum_{i=0}^{N} |U_{\text{app}}(i) - U(i)|.$$

- In accordance with the remark from 6.4.4, if there exists points $x \in \Omega$ such that $I(x) = 1$, we have in practice supposed that $\Omega' = \{x \in \Omega \mid I(x) < 0.99\}$ for numerical reasons (round of errors). We will then denote by n_{fix} the number of points from Ω_N where I is larger than 0.99.
- We will denote by I_{min} and I_{max} the minimal and maximal values of I on Ω_N.
- Finally, we will denote by τ the computation time of the considered simulation, from the first computations done in step 1) to the last computations done in step 2).

6.5.2 — For our first simulations, we can solve numerically the s.f.s. problem (6.1) with $\Omega =]0,1[$, $(l, l_3) = (0,1)$ and $I(x) = \sqrt{2}/2$ for all $x \in]0,1[$. As observed in 5.1.3, this leads to the eikonal problem largely studied in chapter 5:

$$\begin{cases} |u'(x)| = 1 & \text{if} \quad x \in]0,1[, \\ u(x) = \varphi(x) & \text{if} \quad x \in \{0,1\}. \end{cases}$$

Then for $N = 20$, we have applied the algorithm presented in 6.4.4 with the following choices of φ:

- $\varphi(0) = \varphi(1) = 0$. In this case, thanks to 5.2.7 and 5.3.7, we know that

$$u : [0,1] \to \mathbb{R},\ x \mapsto \frac{1}{2} - \left|\frac{1}{2} - x\right|$$

is the only continuous viscosity solution of the s.f.s. problem. Numerically, our algorithm converges in $\iota = 11$ iterations, with η_∞ and η_{ave} smaller than 10^{-15}, *i.e.* $\eta_\infty = \eta_{\text{ave}} = 0$ and so $U_{\text{app}}(i) = u(\omega_i)$ for all $i \in \{0, \dots, 21\}$ with machine precision.

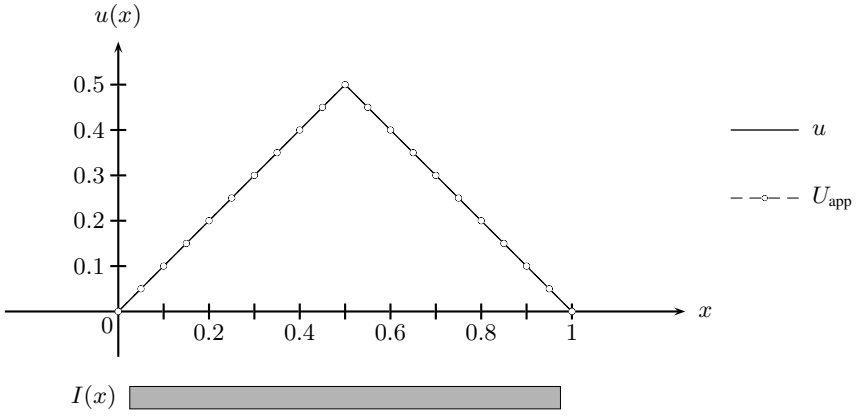

Figure 39 – Graphical representations of u and U_{app} when $\varphi(0) = \varphi(1) = 0$

- $\varphi(0) = 0$ and $\varphi(1) = 2$. We have observed in 5.6.4 that the s.f.s. problem does not admit a continuous viscosity solution, but that

$$u \,:\, [0,1] \to \mathbb{R},\ x \mapsto \begin{cases} x & \text{if} \quad x \in [0,1[, \\ 2 & \text{if} \quad x = 1. \end{cases}$$

is a discontinuous viscosity solution of it. Numerically, our algorithm converges in $\iota = 20$ iterations, with η_∞ and η_{ave} smaller than 10^{-15}, *i.e.* $\eta_\infty = \eta_{\text{ave}} = 0$ and so $U_{\text{app}}(i) = u(\omega_i)$ for all $i \in \{0, \ldots, 21\}$ with machine precision.

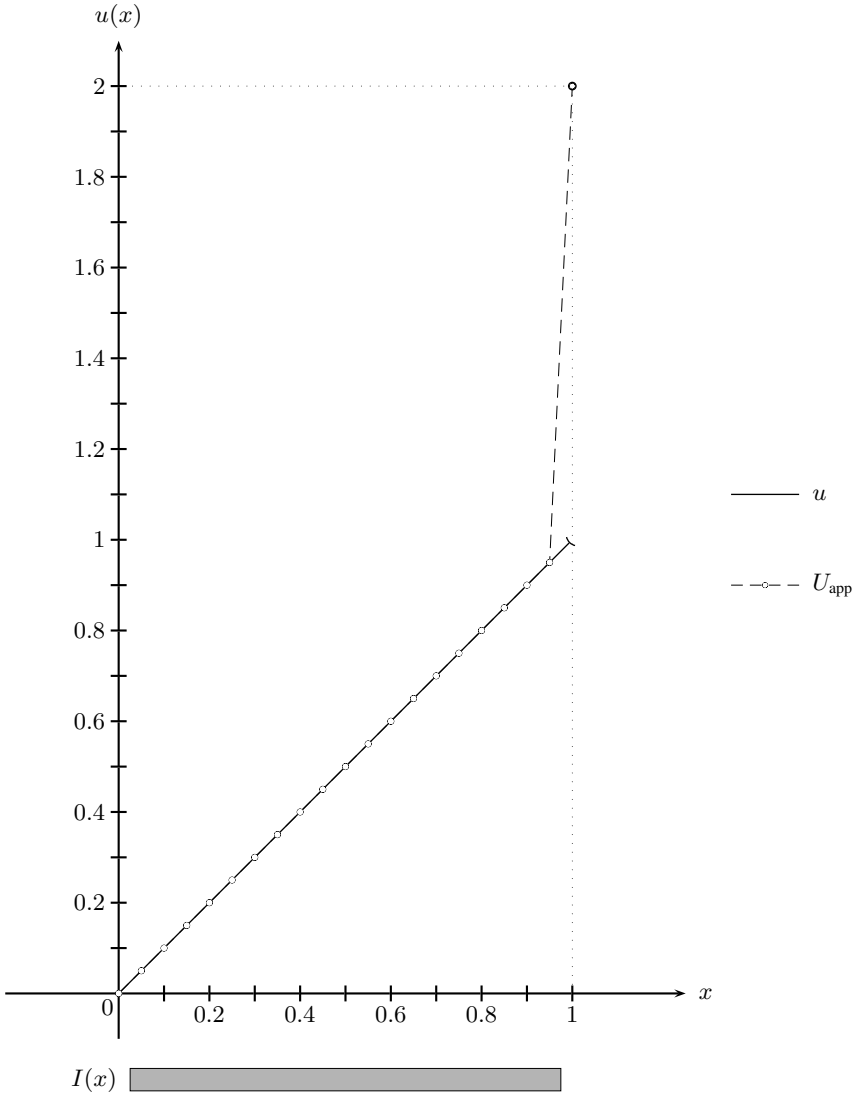

Figure 40 – Graphical representations of u and U_{app} when $\varphi(0) = 0$ and $\varphi(1) = 2$

These first simulations show that our algorithm enables to obtain approximations of a continuous viscosity solution of a s.f.s. problem as well as a discontinuous. We can also note that for these particular choices

of I and (l, l_3), we have always recovered the exact value of the viscosity solution of the s.f.s. problem: that is probably because the exact solutions are piecewise linear, so that in this case, discretizations of u' in our approximation scheme S are exact.
Let us also signal that these two simulations were done almost instantaneously: in less than $4\,\mathrm{ms}$.

6.5.3 — Let us consider a more sophisticated example, by trying to recover the following shape:

$$u \,:\, [0,1] \to \mathbb{R},\ x \mapsto \frac{1-x^2}{4}.$$

So here, $\Omega =]0,1[$, and thanks to the brightness equation (II.1) presented in the introduction of this part II, I is therefore defined by:

$$I \,:\,]0,1[\to \mathbb{R},\ x \mapsto \frac{lx + 2l_3}{\sqrt{4+x^2}}.$$

In this case, according to theorem 5.2.6, we can indeed prove that u is a viscosity solution of the s.f.s. problem. Thus by fixing $N = 20$ and supposing that $\varphi(x) = u(x)$ for all $x \in \partial\overline{\Omega}_N$, for various choices of (l, l_3), the table 19 presented the corresponding values of η_∞, η_{ave}, ι, but also the number n_{fix} of nodes of Ω_N where the intensity I was larger than 0.99 and the computation time τ (expressed in milliseconds).

l	l_3	I_{min}	I_{max}	η_∞	η_{ave}	ι	n_{fix}	τ
$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	0.52	0.82	$1.19{\cdot}10^{-2}$	$5.65{\cdot}10^{-3}$	20	0	$4.90\,\mathrm{ms}$
$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	0.72	0.94	$1.19{\cdot}10^{-2}$	$5.65{\cdot}10^{-3}$	26	0	$6.82\,\mathrm{ms}$
$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	0.88	1	$1.00{\cdot}10^{-2}$	$4.05{\cdot}10^{-3}$	53	3	$11.8\,\mathrm{ms}$
0	1	0.90	1	$8.75{\cdot}10^{-3}$	$3.12{\cdot}10^{-3}$	28	5	$7.29\,\mathrm{ms}$
$\frac{-1}{2}$	$\frac{\sqrt{3}}{2}$	0.57	0.85	$1.19{\cdot}10^{-2}$	$5.65{\cdot}10^{-3}$	21	0	$4.51\,\mathrm{ms}$

Table 19 – Values of I_{min}, I_{max}, η_∞, η_{ave}, ι, n_{fix} and τ for various choices of (l, l_3)

In comparison to the simulations from the previous paragraph, we now obtain approximations that are effectively different to the viscosity solution u of the s.f.s. problem, but really close to it, as shown by the errors η_∞ and η_{ave}. Even if the number of iterations ι seems increasing as soon as the values of I are close to 1, we finally obtain good approximations of u practically instantaneously.

In order to illustrate our simulations, we have represented I, U_{app} and u for various choices of l and l_3 at figures 41 and 42: $l = -1/2$ at figure 41, $l = 1/2$ at figure 42, and $l_3 = \sqrt{3}/2$ for both of them. It is interesting to observe that U_{app} have been computed by starting from the exact value of u in $x = 0$ in the first case, and from its exact value in $x = 1$ in the other one. In fact, the convergence over iterations of U_{app} to the viscosity solution u seems to follow the light source direction. Unfortunately, we are not able to explain analytically this phenomenon, which can be observed in the other cases, for other simulations, and more generally also in two-dimension.

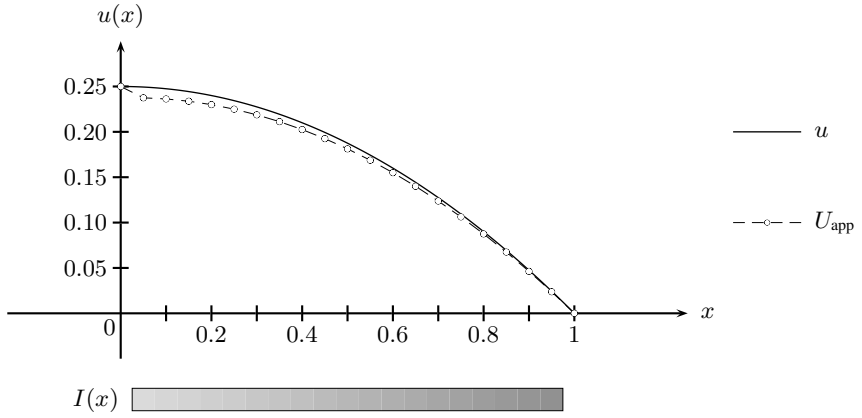

Figure 41 – Graphical representations of I, u and U_{app} when $(l, l_3) = \left(\dfrac{-1}{2}, \dfrac{\sqrt{3}}{2}\right)$

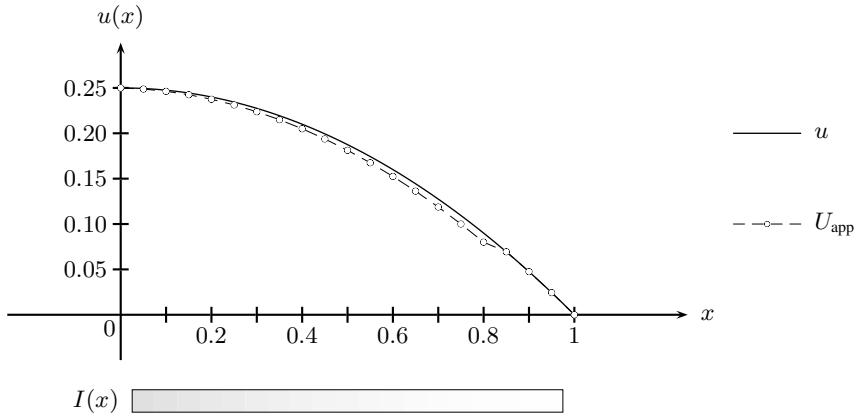

Figure 42 – Graphical representations of I, u and U_{app} when $(l, l_3) = \left(\dfrac{1}{2}, \dfrac{\sqrt{3}}{2}\right)$

6.5.4 — *An example with a shadow area.*

In order to illustrate the behaviour of our algorithm when there exists a shadow area, let us consider that the shape is defined by the following piecewise continuous function $u : [0, 4] \to \mathbb{R}$ defined, for all $x \in [0, 4]$, as follows:

$$u(x) = \begin{cases} 2 & \text{if} \quad x \in [0, 1[, \\ -2x + 4 & \text{if} \quad x \in [1, 2[, \\ 0 & \text{if} \quad x \in [2, 4]. \end{cases}$$

We thus have $\Omega =]0, 4[$. Now let us suppose that

$$(l, l_3) = \left(\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right).$$

Then, from a theoretical point of view, for all $x \in]0,4[$, the brightness equation (II.1) implies

$$I(x) = \begin{cases} \sqrt{2}/2 & \text{if} \quad x \notin [1,2[, \\ 0 & \text{if} \quad x \in [1,2[. \end{cases}$$

But as shown by the figure 43, from a physical point of view, the orientation of the light vector makes that a shadow area can be observed on the interval $[1,3[$, and not only $[1,2[$ (to better visualize it, we can for instance suppose that u describes the profile of a mountain).

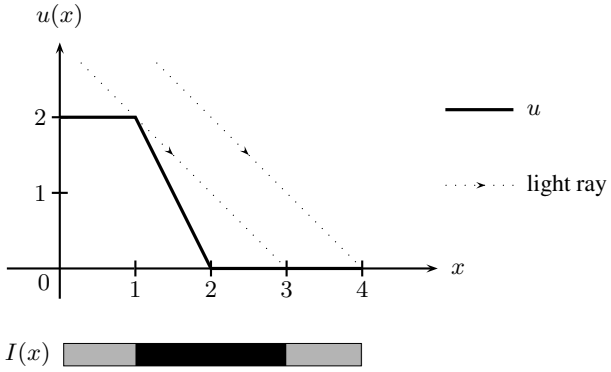

Figure 43 – Graphical representation of u and I from a physical point of view

That is the reason why we will finally suppose, for all $x \in [0,4]$, that

$$I(x) = \begin{cases} \sqrt{2}/2 & \text{if} \quad x \notin [1,3[, \\ 0 & \text{if} \quad x \in [1,3[. \end{cases}$$

As described in paragraph 6.3.2, over the shadow area, we expect to recover a shape that has the same direction than the light source in output of our algorithm. By fixing $N = 20$, this is precisely what we get in practice, as shown by the figure 44.

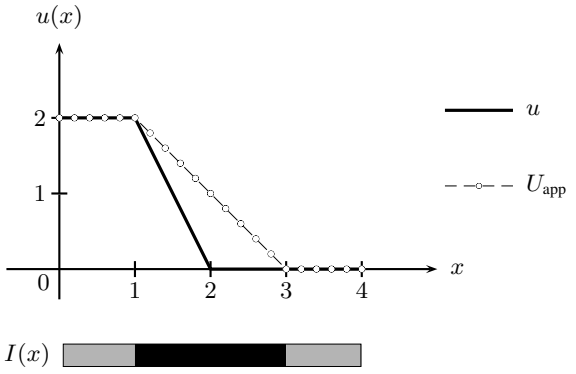

Figure 44 – Example of a reconstructed shape in the presence of a shadow area

Conclusion

In this chapter, we have mainly proved how to obtain an explicit formulation of an approximation scheme associated to the s.f.s. problem. This was manifestly considered as implicit until now, in the one-dimensional case. By using this explicit expression of the approximation scheme, we have been able to implement an effective and efficient algorithm that enables to recover continuous as well as discontinuous solutions of various s.f.s. problems, practically instantaneously.
If the one-dimensional s.f.s. problem has no interest in practice, its study could be useful to understand how to obtain an explicit formulation of the approximation scheme we have considered. As we will see in the following chapter, it will also be possible to introduce an analogous scheme in two-dimension, and we will be able to obtain an explicit expression of it by using mainly the same tricks.

Chapter 7

Numerical resolution of the bi-dimensionnal shape from shading problem

Introduction

In this new chapter, we will essentially extend to the bi-dimensional case what we have done in chapter 6 with the shape from shading problem in the one-dimensional case. Thus from now on, we will assume that $n = 2$ and $\Omega =]0, \alpha_1[\times]0, \alpha_2$ for such $\alpha_1, \alpha_2 > 0$, and we will consider the s.f.s. problem:

$$\left\{ \begin{array}{lll} H(x, \nabla(x)) \ = \ 0 & \text{if} & x \in \Omega, \\ u(x) \ = \ \varphi(x) & \text{if} & x \in \partial\Omega, \end{array} \right. \tag{7.1}$$

where H is given, for all $(x, p) \in \Omega \times \mathbb{R}$, by

$$H(x, p) \ = \ I(x)\sqrt{1 + |p|^2} + l \cdot p - l_3.$$

As a reminder, I is a function from $\overline{\Omega}$ to $[0, 1]$ corresponding to the brightness intensity, $l = (l_1, l_2) \in \mathbb{R}^2$ and $l_3 > 0$ are such that (l, l_3) corresponds to the unit vector oriented to the light source, and $\varphi : \partial\Omega \to \mathbb{R}$ a function that is supposed to be continuous. In addition, by introducing

$$\Upsilon \ : \ [-1, 1] \times \big(\overline{\Omega} \times \mathbb{R}\big) \to \mathbb{R}, \ (b\,; x, p) \mapsto I(x)\sqrt{1 - |b|^2} + \big(I(x)b + l\big) \cdot p - l_3,$$

thanks to 5.5.7, for all $(x, p) \in \Omega \times \mathbb{R}$, we now have

$$H(x, p) \ = \ \sup_{b \in \mathrm{B}'(0,1)} \Upsilon(b\,; x, p).$$

As in chapter 6, everything we will present in this chapter will be available when the set of points from Ω where I is equal to 1 corresponds to a finite union of disjoint connected compact sets by replacing Ω by $\Omega' = \{x \in \Omega \mid I(x) < 1\}$. But in order to simplify our talk, we will not do it here.
So here, on the same principle as in chapter 6, we will first see how to obtain an approximation scheme associated to our s.f.s. problem (section 7.1). Then we will give an explicit expression of this approximation scheme (section 7.2) and of its solutions (section 7.3). We will present a numerical algorithm that will enable to solve (7.1) numerically (section 7.4) and we will illustrate its performances (section 7.5). We will terminate this chapter by presenting possible extensions of our work (section 7.6).

7.1 An approximation scheme associated to the s.f.s. problem

7.1.1 — The approximation scheme associated to the bi-dimensional s.f.s. problem is constructed in a totally similar way than in the one-dimensional case. We start by fixing $h = (h_1, h_2) \in \left(\mathbb{R}_+^*\right)^2$ and by introducing

$$\Omega_h \;=\; \left\{x \in \Omega \;\middle|\; \forall\,(j,s) \in \{1,2\} \times \{-1,1\},\; x + s h_j e_j \in \overline{\Omega}\right\} \qquad \text{and} \qquad \partial\Omega_h \;=\; \overline{\Omega} \setminus \Omega_h.$$

Let us remind that (e_1, e_2) corresponds to the standard basis from $\mathbb{R}^2$. By doing this, we still have $\overline{\Omega} = \Omega_h \cup \partial\Omega_h$, and $\partial\Omega_h$ does not correspond rigorously to the boundary of Ω_h, but can be morally considered as a boundary of Ω_h that has h to diameter, as shown by the following figure:

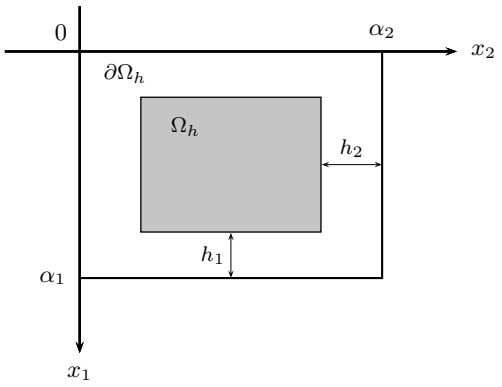

Figure 45 – Representation of the domains Ω_h and $\partial\Omega_h$

Now by fixing $x \in \overline{\Omega}$, $t \in \mathbb{R}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, the approximation scheme $S : \left(\mathbb{R}_+^*\right)^2 \times \overline{\Omega} \times \mathbb{R} \times \mathcal{B}\big(\overline{\Omega}\big) \to \mathbb{R}$ is constructed as follows:

- If $x \in \Omega_h$, for all $b \in \mathrm{B}'(0,1)$, by introducing the vector

$$q(b\,;h,x,t,u) \;=\; \big(q_1(b\,;h,x,t,u),\; q_2(b\,;h,x,t,u)\big)$$

that is defined, for all $j \in \{1,2\}$, by

$$q_j(b\,;h,x,t,u) \;=\; s_j\,\frac{t - u(x - s_j h_j e_j)}{h_j} \qquad \text{where} \qquad \begin{cases} s_j = 1 & \text{if} \quad I(x)b_j + l_j \geqslant 0, \\ s_j = -1 & \text{if} \quad I(x)b_j + l_j < 0, \end{cases}$$

we can set

$$S(h,x,t,u) \;=\; \sup_{b \in [-1,1]} \Upsilon\big(b\,;x,\; q(b\,;h,x,t,u)\big).$$

- If $x \in \partial\Omega_h$, by extending continuously φ on $\partial\Omega_h$ in order to insure the continuity of $S(h,x,\,\cdot\,,\,\cdot\,)$, we set

$$S(h,x,t,u) \;=\; t - \varphi(x).$$

As in the one-dimensional case, we thus obtain an approximation scheme

$$S \;:\; \mathbb{R}_+^* \times [-1,1] \times \mathbb{R} \times \mathcal{B}([-1,1]) \to \mathbb{R},\; (h,x,t,u) \mapsto S(h,x,t,u)$$

that is non-increasing with respect to its last variable ($u \in \mathcal{B}([0,1])$) and non-decreasing with respect to its third variable ($t \in \mathbb{R}$).

REMARK. In figure 45, we have represented the axes in an non-usual way. Even if our choice can be surprising, it is motivated by an ease of presentation that will be revealed in section 7.4.

7.1.2 — The following result extend to the bi-dimensional case lemma 6.1.2:

Proposition. *For all $h \in \mathbb{R}_+^*$, the function*

$$u_0 \,:\, \overline{\Omega} \to \mathbb{R},\ x \mapsto \frac{-l\cdot x}{l_3} + C,$$

where $C \in \mathbb{R}$ is chosen so that $u_0 \leqslant \min_{\partial\Omega_h} \varphi$, is a subsolution of $S(h,\,\cdot\,,\,\cdot\,,\,\cdot\,)$.

7.1.3 — As in paragraph 6.1.3, we can determine conditions that insure that the approximation scheme defined in 7.1.1 is stable:

- S is non-increasing with respect to its last variable ($u \in \mathcal{B}\big(\overline{\Omega}\big)$), thus it always satisfies (A_{12}).
- S is non-decreasing with respect to its third variable ($t \in \mathbb{R}$), and for fixed $h \in \big(\mathbb{R}_+^*\big)^2$, $x \in \overline{\Omega}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, when $t \to +\infty$, we have $S(h,x,t,u) \to +\infty$ as soon as we do not have $I(x) = l = 0$. Thus (A_{13}) is satisfied as soon as we do not have $I(x) = l = 0$.
- Let us fix $h = (h_1, h_2) \in \big(\mathbb{R}_+^*\big)^2$ and $x \in \overline{\Omega}$.
  - If $x \in \partial\Omega_h$, for all $t \in \mathbb{R}$ and $U \in \mathbb{R}^4$, let us set
  $$\sigma_{h,x}(t,U) \;=\; t - \varphi(x).$$
  - If $x \in \Omega_h$, for all $t \in \mathbb{R}$ and $U = (U_1,U_2,U_3,U_4) \in \mathbb{R}^4$, by setting
  $$\sigma_{h,x}(t,U) \;=\; \sup_{b\in \mathrm{B}'(0,1)} \left\{ I(x)\sqrt{1-|b|^2} + \big(I(x)b_1 + l_1\big)s_1 \cdot \frac{t-\alpha_1}{h_1} + \big(I(x)b_2 + l_2\big)s_2 \cdot \frac{t-\alpha_2}{h_2} - l_3 \right\},$$
  where
  $$\left\{ \begin{array}{llll} s_1 = 1, & \alpha_1 = U_1, & \text{if} & I(x)b_1 + l_1 \geqslant 0, \\ s_1 = -1, & \alpha_1 = U_3, & \text{if} & I(x)b_1 + l_1 < 0, \end{array} \right. \text{and} \left\{ \begin{array}{llll} s_2 = 1, & \alpha_2 = U_2, & \text{if} & I(x)b_2 + l_2 \geqslant 0, \\ s_2 = -1, & \alpha_2 = U_4, & \text{if} & I(x)b_2 + l_2 < 0, \end{array} \right.$$
  we can easily observe that
  $$S(h,x,t,u) \;=\; \sigma_{h,x}\Big(t, \big(u(x-h_1e_1),\ u(x-h_2e_2),\ u(x+h_1e_1),\ u(x+h_2e_2)\big)\Big).$$
  So by extending continuously φ to $\partial\Omega_h$, we can prove by refering to lemma 4 from [18] that
  $$\sigma_{h,x} \,:\, \mathbb{R}\times\mathbb{R}^2 \to \mathbb{R},\ (t,U) \mapsto \sigma_{h,x}(t,U)$$
  is continuous, and that (A_{14}) happens with $d = 4$ and $\Xi_{h,x} = \{x-h_1e_1,\, x-h_2e_2,\, x+h_1e_1,\, x+h_2e_2\}$.
- According to proposition 7.1.2, assumption (A_{15}) is satisfied.
- Thanks to proposition 6 from [18], assumption (A_{16}) is also satisfied.

Finally, conditions that insure that the solutions of S converge to the viscosity solution of (7.1) can be found identically than in paragraph 6.1.4 with the one-dimensional case.

7.2 An explicit expression of the approximation scheme

7.2.1 — Now on the same principle as in section 6.2 with the one-dimensional case, given $h \in \left(\mathbb{R}_+^*\right)^2$, $x \in \Omega_h$ such that $I(x) > 0$, $t \in \mathbb{R}$ and $u \in \mathcal{B}(\overline{\Omega})$, we will determine an explicit expression of the approximation scheme S constructed in the previous section in (h, x, t, u). Since h, x, t and u are fixed, we will more simply note $\Psi(b)$ the quantity $\Upsilon\big(b; x, q(b; h, ; x, t, u)\big)$, hence a function $\Psi : \mathrm{B}'(0,1) \to \mathbb{R}$. In the one-dimensional, the exact expression of $S(h, x, t, u)$ was essentially found by considering the regions R_{-1} and R_1 from $[-1, 1]$ where the expression of the quantity $q(b\,; h, x, t, u)$ was changing. In the bi-dimensional, we need to do this for each component of $q(b\,; h, x, t, u)$, thus we need to consider:

$$\begin{aligned}
R_{(1,1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 > 0,\ I(x)b_2 + l_2 > 0\}, \\
R_{(1,-1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 > 0,\ I(x)b_2 + l_2 < 0\}, \\
R_{(-1,1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 < 0,\ I(x)b_2 + l_2 > 0\}, \\
R_{(-1,-1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 < 0,\ I(x)b_2 + l_2 < 0\},
\end{aligned}$$

and the regions that corresponds, possibly also with the point $-l/I(x)$, to their boundary in $\mathrm{B}'(0,1)$:

$$\begin{aligned}
R_{(0,1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 = 0,\ I(x)b_2 + l_2 > 0\}, \\
R_{(0,-1)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 = 0,\ I(x)b_2 + l_2 < 0\}, \\
R_{(1,0)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 > 0,\ I(x)b_2 + l_2 = 0\}, \\
R_{(-1,0)} &= \{(b_1, b_2) \in \mathrm{B}'(0,1) \mid I(x)b_1 + l_1 < 0,\ I(x)b_2 + l_2 = 0\}.
\end{aligned}$$

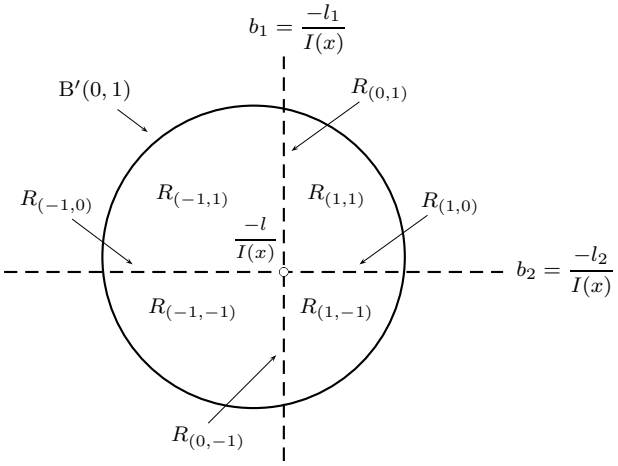

Figure 46 – Representation of the regions $R_{(\pm1,\pm1)}$, $R_{(0,\pm1)}$ and $R_{(\pm1,0)}$

By the definition of S, we thus have

$$S(h, x, t, u) = \sup_{\mathrm{B}'(0,1)} \Psi = \max\left\{ \sup_{R_{(1,1)}} \Psi,\ \sup_{R_{(1,-1)}} \Psi,\ \sup_{R_{(-1,1)}} \Psi,\ \sup_{R_{(-1,-1)}} \Psi\} \right\}. \tag{7.2}$$

7.2.2 — To find an explicit expression of $S(h, x, t, u)$, we will use the following result, that can be seen as an extension in the bi-dimensional case of lemma 6.2.2:

Lemma. *For all $\delta > 0$, $w \in \mathbb{R}^2$ and $c \in \mathbb{R}$, the function*

$$\psi \,:\, \mathrm{B}'(0,1) \to \mathbb{R},\ b \mapsto \delta\sqrt{1-|b|^2} + w\cdot b + c$$

is concave. It has an unique critical point b^* on $\mathrm{B}'(0,1)$, which corresponds to its global maximiser and which satisfies

$$b^* \;=\; \frac{w}{\sqrt{\delta^2+|w|^2}} \qquad \textit{and} \qquad \psi(b^*) \;=\; c + \sqrt{\delta^2+|w|^2}.$$

7.2.3 — Let us start by supposing $|l| \leqslant I(x)$, which is equivalent to say that $-l/I(x) \in \mathrm{B}'(0,1)$.

- Let us fix $s = (s_1, s_2) \in \{(1,1),(1,-1),(-1,1),(-1,-1)\}$. For all $b = (b_1, b_2) \in R_s$, we have

$$\begin{aligned}\Psi(b) \;&=\; I(x)\sqrt{1-|b|^2} + \big(I(x)b_1 + l_1\big)s_1\cdot\frac{t-u(x-s_1h_1e_1)}{h_1} + \big(I(x)b_2+l_2\big)s_2\cdot\frac{t-u(x-s_2h_2e_2)}{h_2} - l_3\\
&=\; \underbrace{I(x)}_{\delta_s}\sqrt{1-|b|^2} + \underbrace{I(x)\left(s_1\,\frac{t-u(x-s_1h_1e_1)}{h_1},\ s_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}\right)}_{w_s}\cdot(b_1,b_2)\\
&\qquad + \underbrace{l_1\,\frac{t-u(x-s_1h_1e_1)}{h_1} + l_2\,\frac{t-u(x-s_2h_2e_2)}{h_2} - l_3}_{c_s}\,,\end{aligned}$$

  which proves that Ψ coincides with $\psi_s : \mathrm{B}'(0,1) \to \mathbb{R}, b \mapsto \delta_s\sqrt{1-|b|^2} + w_s b + c_s$ on R_s. Therefore if Ψ has a critical point on R_s, then according to lemma 7.2.2, it necessarily corresponds to the maximiser b_s^* of ψ_s, hence

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s + \sqrt{\delta_s + |w_s|^2}.$$

  But if Ψ has no critical point on R_s, then Ψ reaches its supremum on the boundary of R_s in $\mathrm{B}'(0,1)$, which is included in $R_{(0,1)} \cup R_{(0,-1)} \cup R_{(1,0)} \cup R_{(-1,0)} \cup \{-l/I(x)\}$.

- Now let us fix $s_2 \in \{-1,1\}$ and let us set $s = (0, s_2)$. For all $b = (b_1, b_2) \in R_s$, we have

$$\begin{aligned}\Psi(b) \;&=\; I(x)\sqrt{1-\frac{l_1^2}{I^2(x)} - b_2^2} + \big(I(x)b_2+l_2\big)s_2\,\frac{t-u(x-s_2h_2e_2)}{h_2} - l_3\\
&=\; \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1-\frac{l_1^2}{I^2(x)}}_{\alpha_s} - b_2^2} + \underbrace{I(x)s_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}}_{w_s}\cdot b_2 + \underbrace{s_2l_2\,\frac{t-u(x-s_2h_2e_2)}{h_2} - l_3}_{c_s}\,,\end{aligned}$$

  which proves that Ψ coincides with $\psi_s : [-\alpha_s, \alpha_s] \to \mathbb{R},\ b_2 \mapsto \delta_s\sqrt{1-b_2^2} + w_s b_2 + c_s$ on R_s. Therefore if Ψ has a critical point on R_s, then according to lemma 6.2.2, it necessarily corresponds to the maximiser b_s^* of ψ_s, hence

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s + \sqrt{\delta_s + |w_s|^2}.$$

  But if Ψ has no critical point on R_s, then Ψ reaches its supremum on the boundary of R_s, and so necessarily in $-l/I(x)$ since $\Psi_{|R_s} = \psi_{s\,|R_s}$ is concave.

- Similarly, if $s_1 \in \{-1,1\}$ and $s=(s_1,0)$, for all $b=(b_1,b_2)\in R_s$, we have

$$\begin{aligned}\Psi(b) &= I(x)\sqrt{1-\frac{l_2^2}{I^2(x)}-b_1^2}+\big(I(x)b_1+l_1\big)s_1\,\frac{t-u(x-s_1h_1e_1)}{h_2}-l_3\\ &= \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1-\frac{l_2^2}{I^2(x)}}_{\alpha_s}-b_1^2}+\underbrace{I(x)s_2\,\frac{t-u(x-s_1h_1e_1)}{h_1}}_{w_s}\cdot b_1+\underbrace{s_1l_1\,\frac{t-u(x-s_1h_1e_1)}{h_1}-l_3}_{c_s},\end{aligned}$$

  which proves that Ψ coincides with $\psi_s : [-\alpha_s,\alpha_s]\to\mathbb{R},\ b_1\mapsto \delta_s\sqrt{1-b_1^2}+w_sb_1+c_s$ on R_s. Therefore if Ψ has a critical point on R_s, then according to lemma 6.2.2, it necessarily corresponds to the maximiser b_s^* of ψ_s, hence

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s+\sqrt{\delta_s+|w_s|^2}.$$

  But if Ψ has no critical point on R_s, then Ψ reaches its supremum on the boundary of R_s, and so necessarily in $-l/I(x)$ since $\Psi_{|R_s}=\psi_{s\,|R_s}$ is concave.

Consequently, by reusing the previous notations, by introducing $\Sigma=\{(s_1,s_2)\mid s_1,s_2\in\{-1,0,1\}\}$, and by setting

$$a_{(0,0)}(h,x,t,u) \;=\; \Psi\left(\frac{-l}{I(x)}\right)$$

and, if $s\neq(0,0)$,

$$a_s(h,x,t,u) \;=\; \begin{cases}\psi_s(b_s^*) & \text{if}\quad b_s^*\in R_s,\\ \Psi\big(-l/I(x)\big) & \text{if}\quad b_s^*\notin R_s,\end{cases}$$

thanks to (7.2), we have proved that

$$S(h,x,t,u) \;=\; \max\left\{a_s(h,x,t,u)\mid s\in\Sigma\right\}.$$

7.2.4 — Now let us suppose that $|l|>I(x)$. In this case, we have $-l/I(x)\notin \mathrm{B}'(0,1)$, which leads to some differences in comparison to the previous case, studied in 7.2.3.

- If $s=(s_1,s_2)\in\{(1,1),(1,-1),(-1,1),(-1,-1)\}$, for all $b=(b_1,b_2)\in R_s$, we still have

$$\begin{aligned}\Psi(b) \;=\; \underbrace{I(x)}_{\delta_s}\sqrt{1-|b|^2}+\underbrace{I(x)\left(s_1\,\frac{t-u(x-s_1h_1e_1)}{h_1},\ s_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}\right)}_{w_s}\cdot(b_1,b_2)\\ \underbrace{+\,l_1\,\frac{t-u(x-s_1h_1e_1)}{h_1}+l_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}-l_3}_{c_s},\end{aligned}$$

  and Ψ coincides with $\psi_s:\mathrm{B}'(0,1)\to\mathbb{R},\ b\mapsto\delta_s\sqrt{1-|b|^2}+w_s\cdot b+c_s$ on R_s. Thus if Ψ has a critical point on R_s, then it still corresponds to the critical point b_s^* of ψ_s, hence

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s+\sqrt{\delta_s+|w_s|^2}.$$

  But if Ψ has no critical point on R_s, then Ψ reaches its supremum on the boundary of R_s in $\mathrm{B}'(0,1)$, which is now only included in $R_{(0,1)}\cup R_{(0,-1)}\cup R_{(1,0)}\cup R_{(-1,0)}$.

- If $s_2 \in \{-1,1\}$ and $s=(0,s_2)$, for all $b=(b_1,b_2)\in R_s$, we still have

$$\Psi(b) \;=\; \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1-\frac{l_1^2}{I^2(x)}}_{\alpha_s}-b_2^2}+\underbrace{I(x)s_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}}_{w_s}\cdot b_2+\underbrace{s_2l_2\,\frac{t-u(x-s_2h_2e_2)}{h_2}-l_3}_{c_s},$$

  which proves that Ψ coincides with $\psi_s:[-\alpha_s,\alpha_s]\to\mathbb{R},\ b_2\mapsto\delta_s\sqrt{1-b_2^2}+w_sb_2+c_s$ on R_s. But since $-l/I(x)\notin \mathrm{B}'(0,1)$, we now have $R_s=[-\alpha_s,\alpha_s]$. Thus denoting by b_s^* the maximiser of ψ_s, now we always have

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s+\sqrt{\delta_s+|w_s|^2}.$$

- Similarly, if $s_1\in\{-1,1\}$ and $s=(s_1,0)$, for all $b=(b_1,b_2)\in R_s$, we have

$$\Psi(b) \;=\; \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1-\frac{l_2^2}{I^2(x)}}_{\alpha_s}-b_1^2}+\underbrace{I(x)s_2\,\frac{t-u(x-s_1h_1e_1)}{h_1}}_{w_s}\cdot b_1+\underbrace{s_1l_1\,\frac{t-u(x-s_1h_1e_1)}{h_1}-l_3}_{c_s},$$

  which proves that Ψ coincides with $\psi_s:[-\alpha_s,\alpha_s]\to\mathbb{R},\ b_1\mapsto\delta_s\sqrt{1-b_1^2}+w_sb_1+c_s$ on R_s. And since $-l/I(x)\notin \mathrm{B}'(0,1)$, then $R_s=[-\alpha_s,\alpha_s]$, thus denoting by b_s^* the maximiser of ψ_s, we obtain

$$\sup_{R_s}\Psi \;=\; \psi_s(b_s^*) \;=\; c_s+\sqrt{\delta_s+|w_s|^2}.$$

Consequently, by reusing the previous notations, by introducing $\Sigma=\{(s_1,s_2)\mid s_1,s_2\in\{-1,0,1\}\}$, and by setting, for all $s\in\Sigma\smallsetminus\{(0,0)\}$,

$$a_s(h,x,t,u) \;=\; \left\{\begin{array}{cc}\psi_s(b_s^*) & \text{if}\quad b_s^*\in R_s,\\ -\infty & \text{if}\quad b_s^*\notin R_s,\end{array}\right.$$

thanks to (7.2), we have proved that

$$S(h,x,t,u) \;=\; \max\left\{a_s(h,x,t,u)\mid s\in\Sigma\smallsetminus\{(0,0)\}\right\}.$$

REMARK. It is maybe important to highlight that when $|l|>I(x)$, there still exists $s\in\Sigma\smallsetminus\{(0,0)\}$ such that $b_s^*\in R_s$, that is to say $S(h,x,t,u)\neq-\infty$. Indeed, we the previous notations:

– If $|l_1|>I(x)$ and $|l_2|>I(x)$, then there exists $s\in\{(1,1),(1,-1),(-1,1),(-1,-1)\}$ such that $R_s=\mathrm{B}'(0,1)$, that obviously implies $b_s^*\in R_s$.

– If $|l_1|\leqslant I(x)$ or $|l_2|\leqslant I(x)$, then there exists $s\in\{(1,0),(-1,0),(0,1),(0,-1)\}$ such that $R_s\neq\varnothing$, and for such a s, we have $R_s=[-\alpha_s,\alpha_s]$, hence $b_s^*\in R_s$.

7.2.5 — By taking the notations from 7.2.3 and 7.2.4 back, results from these two paragraphs can be summarized as follows:

Theorem. *Let us fix $h=(h_1,h_2)\in\left(\mathbb{R}_+^*\right)^2$, $x\in\Omega_h$ such that $I(x)>0$, $t\in\mathbb{R}$ and $u\in\mathcal{B}\big(\overline{\Omega}\big)$. If*

$$a_{(0,0)}(h,x,t,u) \;=\; \left\{\begin{array}{cc}\Psi\big(-l/I(x)\big) & \textit{if}\quad |l|\leqslant I(x),\\ -\infty & \textit{if}\quad |l|>I(x),\end{array}\right.,$$

and, for all $s\in\Sigma\smallsetminus\{(0,0)\}$,

$$a_s(h,x,t,u) \;=\; \left\{\begin{array}{cc}\psi_s(b_s^*) & \textit{if}\quad b_s^*\in R_s,\\ a_{(0,0)}(h,x,t,u) & \textit{if}\quad b_s^*\notin R_s,\end{array}\right.$$

then we have

$$S(h,x,t,u) \;=\; \max\left\{a_s(h,x,t,u)\mid s\in\Sigma\right\}.$$

7.3 Explicit solutions of the approximation scheme

7.3.1 — Given $h = (h_1, h_2) \in \left(\mathbb{R}_+^*\right)^2$, $x \in \Omega_h$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, we are now interested in the equation of unknown $t \in \mathbb{R}$:

$$S(h, x, t, u) \;=\; 0. \tag{7.3}$$

As in section 6.3, equation (7.3) does not necessarily has a solution. Assuming that this equation has a solution, our goal will consist in finding an explicit expression of such a solution. Here again, if $x \in \partial\Omega_h$, it is obvious that $t = \varphi(x)$ is its only solution.
In addition, when $x \in \Omega_h$ and $I(x) = 0$, the definition of S easily implies

$$S(h, x, t, u) \;=\; \left(\frac{s_1 l_1}{h_1} + \frac{s_2 l_2}{h_2}\right) t - \frac{s_1 l_1 u(x - s_1 h_1 e_1)}{h_1} - \frac{s_2 l_2 u(x - s_2 h_2 e_2)}{h_2} - l_3,$$

where, for all $j \in \{1, 2\}$, the value of s_j only depends on the sign of l_j:

$$s_j \;=\; \begin{cases} 1 & \text{if} \quad l_j \geqslant 0, \\ -1 & \text{if} \quad l_j < 0. \end{cases}$$

In this case, if the quantity

$$\lambda \;=\; \left(\frac{s_1 l_1}{h_1} + \frac{s_2 l_2}{h_2}\right)$$

is null, then we can not conclude anything. But if λ is non-null, then

$$t \;=\; \frac{1}{\lambda}\left(\frac{s_1 l_1 u(x - s_1 h_1 e_1)}{h_1} + \frac{s_2 l_2 u(x - s_2 h_2 e_2)}{h_2} + l_3\right)$$

is the only solution of equation (7.3).
In the following, we will thus only consider that $x \in \Omega_h$ and $I(x) > 0$. We will take the notations introduced in section 7.2 back, but in a similar way as in section 6.3, we will more precisely denote by:

– $\Psi(\cdot\,; t)$ the function from $[-1, 1]$ to $\mathbb{R}$ defined, for all $b \in [-1, 1]$, by $\Psi(b; t) = \Upsilon\big(b; x, q(b; h, x, t, u)\big)$,

– for all $s \in \Sigma \smallsetminus \{(0, 0)\}$, by $\psi_s(\,\cdot\,; t)$ the function ψ_s, and by $b_s^*(t)$ its global maximiser.

7.3.2 — For all $s \in \Sigma \smallsetminus \{(0, 0)\}$, let us introduce:

$$\Theta_s \;=\; \left\{t \in \mathbb{R} \;\middle|\; \psi_s\big(b_s^*(t)\,; t\big) = 0,\; b_s^*(t) \in R_s\right\}.$$

As for the one-dimensional problem, it seems natural to determine the elements that belongs to such a set Θ_s. To begin with, let us fix $s = (s_1, s_2) \in \{(1, 1), (1, -1), (-1, 1), (-1, -1)\}$ and $t \in \Theta_s$. In order to simplify our notations, let us set $q(t) = \big(q_1(t),\; q_2(t)\big)$ where, for all $j \in \{1, 2\}$,

$$q_j(t) \;=\; s_j\,\frac{t - u_j}{h_j} \qquad \text{with} \qquad u_j \;=\; u(x - s_j h_j e_j).$$

Therefore, for all $b \in \mathrm{B}'(0, 1)$,

$$\psi_s(b\,; t) \;=\; \underbrace{I(x)}_{\delta_s}\sqrt{1 - |b|^2} + \underbrace{I(x)q(t)}_{w_s} \cdot b + \underbrace{I(x)q(t)\cdot l - l_3}_{c_s}.$$

Thus thanks to lemma 7.2.2, we have

$$\psi_s\big(b_s^*(t)\,;t\big) \;=\; 0 \quad\Longrightarrow\quad I^2(x)[1+q_1^2(t)+q_2^2(t)] \;=\; [l_3 - l_1 q_1(t) - l_2 q_2(t)]. \tag{7.4}$$

By the definition of $q(t)$, we have

$$I^2(x)[1+q_1^2(t)+q_2^2(t)] \;=\; \underbrace{I^2(x)\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)}_{a}t^2 \underbrace{-\,2I^2(x)\left(\frac{u_1}{h_1^2}+\frac{u_2}{h_2^2}\right)}_{b}t + \underbrace{I^2(x)\left(1+\frac{u_1^2}{h_1^2}+\frac{u_2^2}{h_2^2}\right)}_{c}$$

$$l_3 - l_1 q_1(t) - l_2 q_2(t) \;=\; \underbrace{-\left(\frac{s_1 l_1}{h_1}+\frac{s_2 l_2}{h_2}\right)}_{\alpha}t + \underbrace{\left(l_3+\frac{s_1 l_1 u_1}{h_1}+\frac{s_2 l_2 u_2}{h_2}\right)}_{\beta}.$$

Therefore, implication (7.4) can be rewritten

$$\psi_s\big(b_s^*(t)\,;t\big) \;=\; 0 \quad\Longrightarrow\quad \underbrace{(\alpha^2-a)}_{A}t^2 + \underbrace{(2\alpha\beta-b)}_{B}t + \underbrace{(\beta^2-c)}_{C} \;=\; 0.$$

With the previous notations (and especially those of A, B and C), we now immediately obtain the following result:

Proposition. *Let us assume that $I(x) > 0$ and let us fix $s \in \{(1,1),(1,-1),(-1,1),(-1,-1)\}$, $t \in \Theta_s$.*

(*i*) *If $A = 0$ and $B \neq 0$, then*

$$t \;=\; \frac{-C}{B}.$$

(*ii*) *If $A \neq 0$, then we have $\Delta = B^2 - 4AC \geqslant 0$ and existence of $\sigma \in \{-1,1\}$ such that*

$$t \;=\; \frac{-B+\sigma\sqrt{\Delta}}{2A}.$$

REMARK. This result can be linked to proposition 6.3.4. In the case (*ii*) of this last proposition, we have proved algebraically that Δ is always non-negative, whereas here, we can just say that Δ is non-negative since we have supposed that (there exists) $t \in \Theta_s$. In a general way, there is no reason for such a discriminant Δ be always non-negative, but it is maybe possible to prove that it is always the case.

7.3.3 — Now let us suppose that $s_2 \in \{-1,1\}$, and let us set $s = (0,s_2)$. By observing that $R_s = \varnothing$ as soon as $|l_1| > I(x)$, we will in addition suppose that $|l_1| \leqslant I(x)$. Therefore by setting

$$q_2(t) \;=\; s_2\,\frac{t - u(x - s_2 h_2 e_2)}{h_2},$$

we obtain, for all $b \in \mathrm{B}'(0,1)$,

$$\psi_s(b\,;t) \;=\; \underbrace{I(x)}_{\delta_s}\sqrt{\underbrace{1-\frac{l_1^2}{I^2(x)}}_{\alpha_s} - b_2^2} + \underbrace{I(x)\,q_2(t)}_{w_s}\,b_2 + \underbrace{l_2\,q_2(t) - l_3}_{c_s}.$$

Thus thanks to lemma 7.2.2,

$$\begin{aligned}\psi_s\big(b_s^*(t)\,;t\big) \;=\; 0 \quad&\Longleftrightarrow\quad l_3 - l_2\,q_2(t) \;=\; I(x)\sqrt{\left(1-\frac{l_1^2}{I^2(x)}\right)\big(1+q_2^2(t)\big)}\\ &\Longrightarrow\quad \underbrace{[I^2(x)-l_1^2-l_3^2]}_{A}q_2^2(t) + \underbrace{2l_2l_3}_{B}\,q_2(t) + \underbrace{I^2(x)-l_1^2-l_2^2}_{C} \;=\; 0.\end{aligned} \tag{7.5}$$

Consequently:

- If $A = 0$, then $I^2(x) = l_1^2 + l_3^2$, and since $(l, l_3) = (l_1, l_2, l_3)$ is an unit vector, we can thus observe that $l_2 = 0$ if and only if $I(x) = 1$. So if we also have $I(x) < 1$, then we get $B \neq 0$, and

$$q_2(t) \;=\; \frac{-C}{B}.$$

- If $A \neq 0$, by setting $\Delta = B^2 - 4AC$ and knowing that (l, l_3) is an unit vector, we can observe that

$$\begin{aligned}\Delta \;&=\; 4l_2^2l_3^2 - 4\left[\left(I^2(x) - l_1^2\right) - l_2^2\right]\left[\left(I^2(x) - l_1^2\right) - l_3^2\right]\\ &=\; 4l_2^2l_3^2 - 4\left(I^2(x) - l_1^2\right)^2 + (l_2^2 + l_3^2)\left(I^2(x) - l_1^2\right) - 4l_2^2l_3^2\\ &=\; 4\left(I^2(x) - l_1^2\right)\big(\underbrace{l_1^2 + l_2^2 + l_3^2}_{=\,1} - I^2(x)\big) \;\geqslant\; 0.\end{aligned}$$

  Thus, according to (7.5), we have existence of $\sigma \in \{-1, 1\}$ such that

$$q_2(t) \;=\; \frac{-B + \sigma\sqrt{\Delta}}{2A}.$$

Finally, by using the expression of $q_2(t)$, we obtain the following result:

Proposition. *Let us suppose that $I(x) > 0$, and let fix $s_2 \in \{-1, 1\}$, $s = (0, s_2)$ and $t \in \Theta_s$.*

(*i*) *If $A = 0$ and $I(x) < 1$, then*

$$t \;=\; u(x - sh) - \frac{shC}{B}.$$

(*ii*) *If $A \neq 0$, then we have $\Delta = B^2 - 4AC \geqslant 0$ and existence of $\sigma \in \{-1, 1\}$ such that*

$$t \;=\; u(x - sh) + sh\,\frac{-B + \sigma\sqrt{\Delta}}{2A}.$$

REMARK. Here again, we can note that the particular case $I(x) = 1$ is problematic...

7.3.4 — As in 7.3.3, for the regions Θ_s where $s = (0, s_2)$ and $s_2 \in \{-1, 1\}$, by setting

$$A \;=\; I^2(x) - l_2^2 - l_3^2, \quad B \;=\; 2l_1l_3 \quad \text{and} \quad C \;=\; I^2(x) - l_1^2 - l_2^2,$$

we can establish the following result:

Proposition. *Let us suppose that $I(x) > 0$, and let fix $s_1 \in \{-1, 1\}$, $s = (s_1, 0)$ and $t \in \Theta_s$.*

(*i*) *If $A = 0$ and $I(x) < 1$, then*

$$t \;=\; u(x - sh) - \frac{shC}{B}.$$

(*ii*) *If $A \neq 0$, then we have $\Delta = B^2 - 4AC \geqslant 0$ and existence of $\sigma \in \{-1, 1\}$ such that*

$$t \;=\; u(x - sh) + sh\,\frac{-B + \sigma\sqrt{\Delta}}{2A}.$$

7.3.5 — On the same model as proposition 6.3.5, the following result gives a characterization about the existence of a solution to the equation (7.1):

Proposition. *Let us fix $x \in \Omega_h$, and let us suppose that $0 < I(x) < 1$ and that the assumptions from points (i) and (ii) from proposition 7.3.2 are satisfied for all $s \in \{(1,1),(1,-1),(-1,1),(-1,-1)\}$. Then following conditions are equivalent:*

(*i*) *There exists $t \in \mathbb{R}$ such that $S(h,x,t,u) = 0$.*

(*ii*) *There exists $s \in \Sigma \setminus \{(0,0)\}$ such that $\Theta_s \neq \varnothing$.*

If these conditions are satisfied, then by setting

$$\Theta \;=\; \bigcup_{s \in \Sigma \setminus \{(0,0)\}} \Theta_s$$

we have

$$S(h,x,\tau,u) \;=\; 0 \qquad \textit{with} \qquad \tau \;=\; \min \Theta.$$

Proof.— Implication (*i*) $\Rightarrow$ (*ii*) can be proved as those from proposition 6.3.5 in the one-dimensional case, by referring to theorem 7.2.5. For (*ii*) $\Rightarrow$ (*i*), thanks to propositions 7.3.2 to 7.3.4, we can verify that the various assumptions done in our proposition insure that the above quantity τ is well-defined. Then the equality $S(h,x,\tau,u) = 0$ can be established as in 6.3.5. □

7.3.6 — Finally, when the equation (7.3) has a solution, we can explicitly determine one of it as in 6.3.6 in the one-dimensional case: under the assumptions from proposition 7.3.5, we will have $S(h,x,\tau,u) = 0$, and an explicit expression of τ can be obtained by referring to propositions 7.3.2 to 7.3.3.
We can maybe highlight that the hypotheses done in proposition 7.3.5 are pretty heavier and (physically) complicated than those done in proposition 6.3.5 in the one-dimensional case. Until now, we have not found how to simplify them (but is it possible to do it?), or considered what it could happened when they are not satisfied (but can they not be satisfied?).

7.4 Associated algorithm

7.4.1 — In the following, we will denote by $\mathcal{N}$ the Cartesian product $(\mathbb{N} \setminus \{0,1\})^2$.

Definition. *A* mesh *of $\overline{\Omega}$ (resp. Ω, $\partial\Omega$) is a subset of the form:*

$$\begin{aligned} \overline{\Omega}_N \;&=\; \big\{(ih_1, jh_2) \;\big|\; i \in \{0,\dots,N_1\},\; j \in \{0,\dots,N_2\}\big\} \\ \textit{(resp. } \Omega_N \;&=\; \big\{(ih_1, jh_2) \;\big|\; i \in \{1,\dots,N_1-1\},\; j \in \{1,\dots,N_2-1\}\big\}, \\ \partial\Omega_N \;&=\; \overline{\Omega}_N \setminus \Omega_N), \end{aligned}$$

where $N = (N_1, N_2)$ is a couple from $\mathcal{N}$, and h_1 and h_2 are defined by

$$h_1 \;=\; \frac{A_1}{N_1} \qquad \textit{and} \qquad h_2 \;=\; \frac{A_2}{N_2}. \tag{7.6}$$

With these notations, each element of the mesh is called a node*, and the couple $h = (h_1, h_2)$ is called the* mesh-size*.*

7.4.2 — Given $N \in \mathcal{N}$, and in a similar way as in section 6.4:

- We still have $\overline{\Omega}_N = \Omega_N \cup \partial\Omega_N$.

- Notations $\overline{\Omega}_N$ and $\partial\Omega_N$ could be still abusive or confusing, but they are motived by the same reasons as those exposed in 6.4.2 in the one-dimensional case.
- The mesh-size h is also fully determined by N, even if we do not mention this dependence, and for this choice of h, Ω_N still corresponds to the nodes of $\overline{\Omega}_N$ that belongs to the set Ω_h defined in 7.1.1.
- In the following, given $N = (N_1, N_2) \in \mathcal{N}$, $h = (h_1, h_2)$ will automatically corresponds to the mesh size of $\overline{\Omega}_N$, *i.e.* defined by (7.6). In addition, for all $i \in \{0, \dots, N_1\}$ and $j \in \{0, \dots, N_2\}$, we will denote by $\omega_{i,j}$ the node (ih_1, jh_2).

The following figure illustrates all of this:

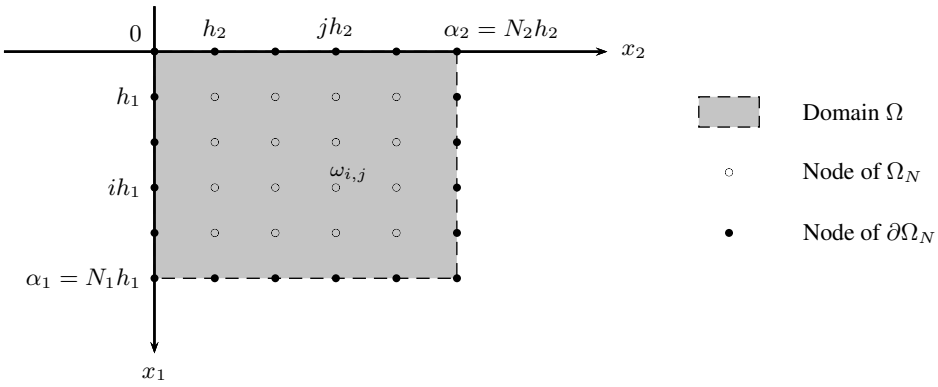

Figure 47 – Representation of the meshes Ω_N and $\partial\Omega_N$ of Ω and $\partial\Omega$

7.4.3 — The algorithm we will use to solve numerically the s.f.s. problem (7.1) will be constructed similarly than those exposed in 6.4.4 for the one-dimensional case. Thus here, thanks to 7.1.3, given $N \in \mathcal{N}$, instead of S, we will consider the function $S_N : \overline{\Omega}_N \times \mathbb{R} \times \mathrm{M}_{N+1}(\mathbb{R})$ defined, for all $x \in \overline{\Omega}_N$, $t \in \mathbb{R}$ and $V \in \mathrm{M}_{N+1}(\mathbb{R})$, by:

– if $x = \omega_{i,j}$ with $(i, j) \in \{1, \dots, N_1 - 1\} \times \{1, \dots, N_2 - 1\}$,

$$S_N(x, t, v) \;=\; \sigma_{h,x}\Big(t,\; \big(V(i-1, j),\, V(i, j-1),\, V(i+1, j),\, V(i, j+1)\big)\Big),$$

– if $x \in \partial\Omega_N$,

$$S_N(x, t, u) \;=\; t - \varphi(x).$$

Such a matrix V will be called a *subsolution* if, for all $(i, j) \in \{0, \dots, N_1\} \times \{0, \dots, N_2\}$, it satisfies $S_N(\omega_{i,j}, V(i, j), V) \leqslant 0$.

REMARK. Given a function $v \in \mathcal{B}\big(\overline{\Omega}\big)$, let us suppose that

$$V \;=\; \big(v(\omega_{i,j})\big)_{\substack{0 \leqslant i \leqslant N_1 \\ 0 \leqslant j \leqslant N_2}} \;=\; \begin{pmatrix} v(\omega_{0,0}) & v(\omega_{0,1}) & \cdots & v(\omega_{0,N_2}) \\ v(\omega_{1,0}) & v(\omega_{1,1}) & \cdots & v(\omega_{1,N_2}) \\ \vdots & \vdots & & \vdots \\ v(\omega_{N_1,0}) & v(\omega_{N_1,1}) & \cdots & v(\omega_{N_1,N_2}) \end{pmatrix} \;\in\; \mathrm{M}_{N+1}(\mathbb{R})\,. \qquad (7.7)$$

The orientation of our axes in figure 47 (and 45) was chosen such that the nodes of the mesh $\overline{\Omega}_N$ can be visually ordered as the values of v (at each node of $\overline{\Omega}_N$) in the matrix V from (7.7). If this choice is maybe not natural, it will be largely comfortable from algorithmic and graphical points of view.

7.4.4 — *Suggested algorithm.*
Given $N = (N_1, N_2) \in \mathcal{N}$, h defined as in (7.6), $\varepsilon > 0$ an accuracy threshold and $n_{\max} \in \mathbb{N}^*$, an approximation $U_{\text{app}} \in \mathbb{R}^{N+1}$ at each node of $\overline{\Omega}_N$ will be computed as follows. Just by knowing the values of I at each nodes of Ω_N, those of the components of the unit vector (l_1, l_2, l_3) oriented to the light source and those of the function φ at each node of $\partial\Omega_N$, U_{app} will be computed as follows:

1) We start with $U_0 \in \mathrm{M}_{N+1}(\mathbb{R})$ a subsolution of S_N satisfying $U_0(i,j) = \varphi(\omega_{i,j})$ for all indexes i and j such that $\omega_{i,j} \in \partial\Omega_N$.

2) Given $n \in \mathbb{N}$ and $U_n \in \mathrm{M}_{N+1}(\mathbb{R})$ a subsoltion of S_N, $U_{n+1} \in \mathrm{M}_{N+1}(\mathbb{R})$ is chosen such that, for all $(i,j) \in \{0, \dots, N_1\} \times \{0, \dots, N_2\}$,

$$S_N(\omega_{i,j}, U_{n+1}(i,j), U_n) \;=\; 0.$$

   In practice, $U_{n+1}(i)$ can be explicitly computed as explained in paragraph 7.3.6.

3) Computations from 2) are done as long as $\|U_{n+1} - U_n\|_\infty \geqslant \varepsilon$ and $n \leqslant n_{\max}$. If there exists $n \leqslant n_{\max}$ such that $\|U_{n+1} - U_n\|_\infty < \varepsilon$, we set $U_{\text{app}} = U_{n+1}$ for the smallest of these integer n. If this is not the case, we set $U_{\text{app}} = U_{n_{\max}}$.

REMARK. As for the one-dimensional case:

- At step 2), let us highlight that for the same reasons as in 6.4.4, U_{n+1} is a subsolution if S_N as soon as U_n is a subsolution of S_N, and it is also possible to lead the computations in parallel.

- If there exists point from Ω such that $I(x) = 1$, then if $\Omega' = \{x \in \Omega \mid I(x) < 1\}$, we can adapt our work by replacing Ω_N by $\Omega'_N \left\{x \in \overline{\Omega}_N \;\middle|\; x \in \Omega'\right\}$ and $\partial\Omega_N$ by $\partial\Omega'_N = \overline{\Omega}_N \setminus \Omega'_N$.

7.5 Numerical simulations

7.5.1 — All the simulations done in this section were done in a total similar way as those performed in section 6.5 in one-dimension. Consequently, by taking the notations from 7.4.4, the hypotheses and notations adopted for our bi-dimensional simulations will be *mutatis mutandis* the same as those exposed in 6.5.1. Let us just explain how to initialize the subsolution U_0 from step 1). By considering

$$\mu \;=\; \min_{x \in \partial\Omega_N} \varphi(x) \quad \text{and} \quad C \;=\; \begin{cases} \mu & \text{if} \quad l_1 \geqslant 0 \;\text{ and }\; l_2 \geqslant 0, \\ \mu + \dfrac{N_2 l_2 h_2}{l_3} & \text{if} \quad l_1 \geqslant 0 \;\text{ and }\; l_2 < 0, \\ \mu + \dfrac{N_1 l_1 h_1}{l_3} & \text{if} \quad l_1 < 0 \;\text{ and }\; l_2 \geqslant 0, \\ \mu + \dfrac{N_1 l_1 h_1 + N_2 l_2 h_2}{l_3} + & \text{if} \quad l_1 < 0 \;\text{ and }\; l_2 < 0, \end{cases}$$

then thanks to proposition 7.1.2, we observe that we can set, for all indexes i and j satisfying $\omega_{i,j} \in \Omega_N$,

$$U_0(i,j) \;=\; C - \frac{i l_1 h_1 + j l_2 h_2}{l_3}.$$

7.5.2 — As in paragraph 6.5.3 in the one-dimensional case, we can consider the surface given by:

$$u \,:\, [0,1]^2 \to \mathbb{R},\ (x,y) \mapsto \frac{x_1^2 + x_2^2}{4}.$$

In this case, we thus have $\Omega =]0,1[^2$, and

$$I \,:\,]0,1[^2 \to \mathbb{R},\ (x_1,x_2) \mapsto \frac{l_3 - l_1 x_1 - l_2 x_2}{\sqrt{4 + x_1^2 + x_2^2}}.$$

For our simulations, we have supposed $N = (20,20)$. The following table indicates, for various choices of l_1, l_2 and l_3, the global performances of each one of them.

l_1	l_2	l_3	I_{min}	I_{max}	η_∞	η_{ave}	ι	n_{fix}	τ
0	0	1	0.82	1	$2.24\cdot10^{-2}$	$8.05\cdot10^{-3}$	38	20	$40.8\,\text{ms}$
$\frac{1}{2}$	0	$\frac{\sqrt{3}}{2}$	0.52	0.85	$1.61\cdot10^{-2}$	$6.21\cdot10^{-3}$	34	0	$32.3\,\text{ms}$
$\frac{-1}{2}$	0	$\frac{\sqrt{3}}{2}$	0.78	1	$2.15\cdot10^{-2}$	$7.91\cdot10^{-3}$	105	12	$112\,\text{ms}$
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	0.18	0.69	$2.10\cdot10^{-2}$	$6.82\cdot10^{-3}$	38	0	$35.5\,\text{ms}$
$\frac{1}{2}$	$\frac{-1}{2}$	$\frac{\sqrt{2}}{2}$	0.42	0.84	$1.89\cdot10^{-2}$	$6.59\cdot10^{-3}$	45	0	$44.1\,\text{ms}$
$\frac{-1}{2}$	$\frac{-1}{2}$	$\frac{\sqrt{2}}{2}$	0.72	0.98	$2.18\cdot10^{-2}$	$7.73\cdot10^{-3}$	86	0	$87.8\,\text{ms}$

Table 20 – Values of I_{min}, I_{max}, η_∞, η_{ave}, ι, n_{fix} and τ for various choices of (l_1,l_2,l_3)

As we can easily observe, the performances of our bi-dimensional algorithm are globally similar to those obtained with the one-dimensional algorithm. We can especially appreciate the very short computation time due to the parallelization of our computations, which is encouraging if we have in mind to use this algorithm with higher values of N, that means with larger and/or refined meshes.

7.5.3 — *When the mesh-size goes to* 0.

Linked to the simulations done in 7.5.2, we can focus on the behaviour of our algorithm for various choices of N, and so mesh-size h. Thus we have deal here with the same u and I as in the previous paragraph, and by supposing

$$(l_1,l_2,l_3) \;=\; \left(\frac{1}{2},\frac{1}{2},\frac{\sqrt{2}}{2}\right).$$

As shown by the table 21, the errors η_∞ and η_{ave} are of order $1/h$, which illustrates the convergence of our approximation to the viscosity solution of our s.f.s. problem.

Moreover, the number of iterations ι seems linear. More precisely, we always have $\iota = N_1 + N_2 - 2$ in these particular cases, which seems to be a strange coincidence. In fact, for other choices of l_1, l_2 and l_3, ι still increases linearly with respect to N_1 and N_2, but there generally does not such an elegant formula. With this same notation of k, the computation time seems to increase quadratically with respect to k. This is totally logical since the number of nodes equals to k^2, and since we have an explicit method to compute U_{app}.

(N_1, N_2)	η_∞	η_{ave}	ι	τ
$(10, 10)$	$3.72{\cdot}10^{-2}$	$1.07{\cdot}10^{-2}$	18	$1.59{\cdot}10^{-2}$ s
$(20, 20)$	$2.10{\cdot}10^{-2}$	$6.82{\cdot}10^{-3}$	38	$3.55{\cdot}10^{-2}$ s
$(30, 30)$	$1.46{\cdot}10^{-2}$	$4.93{\cdot}10^{-3}$	58	$6.80{\cdot}10^{-2}$ s
$(50, 50)$	$9.12{\cdot}10^{-3}$	$3.16{\cdot}10^{-3}$	98	$1.90{\cdot}10^{-1}$ s
$(100, 100)$	$4.71{\cdot}10^{-3}$	$1.66{\cdot}10^{-3}$	198	$9.87{\cdot}10^{-1}$ s
$(200, 200)$	$2.40{\cdot}10^{-3}$	$4.93{\cdot}10^{-4}$	398	$8.14{\cdot}10^{0}$ s

Table 21 – Values of η_∞, η_{ave}, ι, n_{fix} and τ for various choices of (N_1, N_2)

7.5.4 — *About the computation time*

Let us consider the same situation than those considered in the previous paragraph. We will now illustrate the efficiency of the parallelization of the computations done in step 2). In order to do this, we will consider a *punctual version* of our algorithm. Concretically, we have done the computations from step 2) as suggested by E. PRADOS and O. FAUGERAS in [18]:

- We start with $U_{n+1} = U_n$.
- Then the values of U_{n+1} are updated one by one. For i going from 1 to N_1, if i is an odd (resp. even) number, for j going from 1 to N_2 (resp. N_2 to 1), if $\omega_{i,j} \in \overline{\Omega}_N$, $U_{n+1}(i,j)$ is now taken as the solution with respect to t of the equation
$$S(\omega_{i,j}, t, U_{n+1}) \;=\; 0.$$
Consequently, these computations are no more independent, and so, not parallelizable!

This can be illustrated by the following figure:

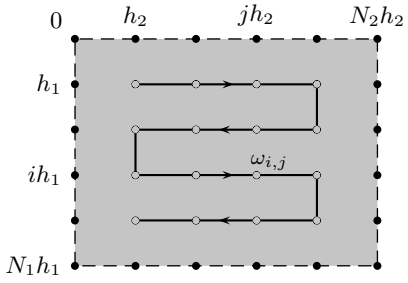

Figure 48 – Computation order from step 2) with a punctual version of our algorithm

The table 22, that must be compared with the table 21, presents the general performance of this modified algorithm. If the errors η_∞ and η_{ave} are mainly identical and the number of iterations is globally divided by two, the computation time is dramatically increasing. It is for instance multiplied by more than 40 when $N_1 = N_2 = 200\ldots$

This illustrates the real efficiency of the parallelization of the computations done in the step 2) of our algorithm. More generally, this shows how we have been able to take advantage of the explicit formulation of our approximation scheme. So even if in the end, from a numerical point of view, we have mainly done adaptations of existent methods, this would probably have been impossible without the theoretical analyses done in sections 7.2 and 7.3.

(N_1, N_2)	η_∞	η_{ave}	ι	τ
$(10, 10)$	$3.72 \cdot 10^{-2}$	$1.07 \cdot 10^{-2}$	10	$4.23 \cdot 10^{-2}$ s
$(20, 20)$	$2.10 \cdot 10^{-2}$	$6.82 \cdot 10^{-3}$	20	$3.44 \cdot 10^{-1}$ s
$(30, 30)$	$1.46 \cdot 10^{-2}$	$4.93 \cdot 10^{-3}$	29	$1.19 \cdot 10^{0}$ s
$(50, 50)$	$1.46 \cdot 10^{-2}$	$4.93 \cdot 10^{-3}$	47	$5.53 \cdot 10^{0}$ s
$(100, 100)$	$4.61 \cdot 10^{-3}$	$1.66 \cdot 10^{-3}$	90	$4.52 \cdot 10^{1}$ s
$(200, 200)$	$2.33 \cdot 10^{-3}$	$8.52 \cdot 10^{-4}$	398	$3.63 \cdot 10^{2}$ s

Table 22 – General performances with a punctual implementation of our algorithm

7.5.5 — It is possible to give more examples like the previous, but we will not do it longer. We will rather focus on an application of the shape from shading we have considered: how to determine the relief of a given countryside just by knowing I and some altitudes? This will constitute the topic of the next section. But before that, let us illustrate graphically the situations considered both in paragraphs 7.5.2 and 7.5.3. In figure 49, we have represented the black and white picture associated to the surface described by the viscosity solution u. The figure 50 shows the surface we have recovered with our approximation U_{app}.

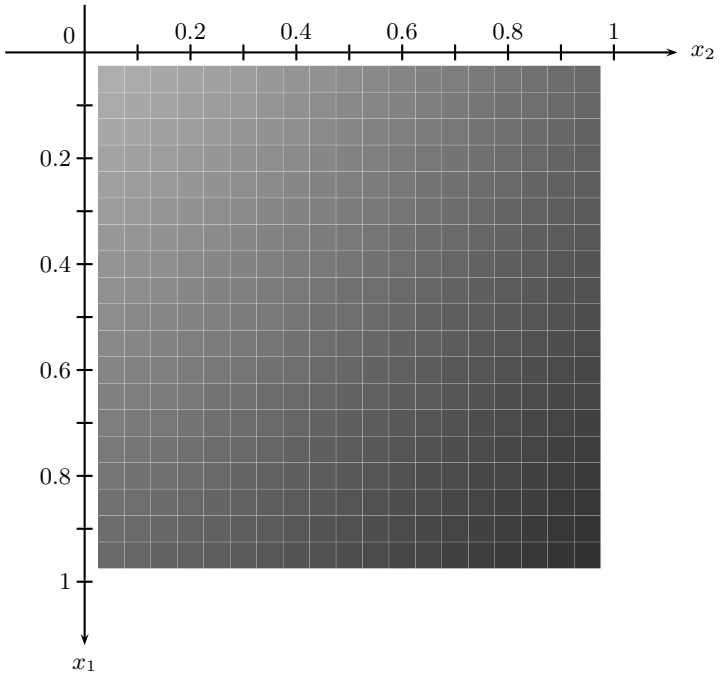

Figure 49 – Initial image corresponding to I when $(l_1, l_2, l_3) = \left(\frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{2}\right)$ and $N_1 = N_2 = 20$

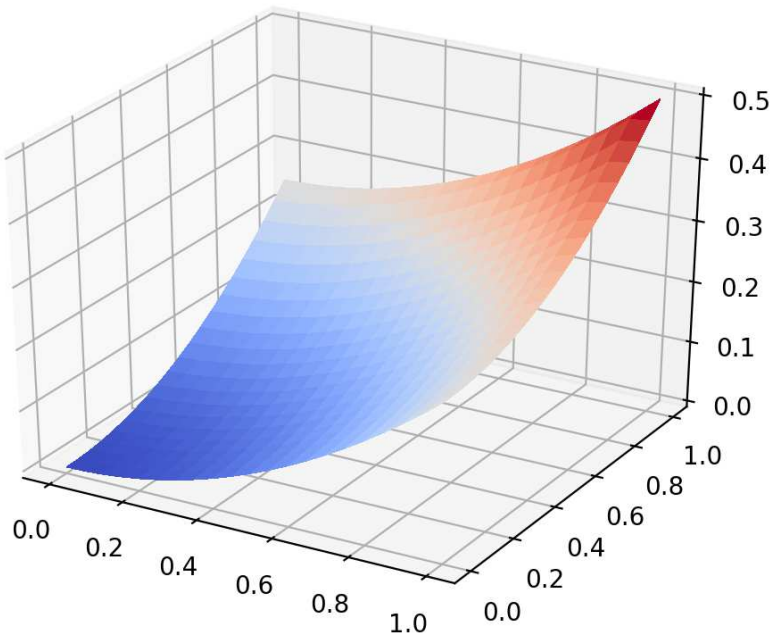

Figure 50 – Graphical representation of U_{app} when $(l_1, l_2, l_3) = \left(\frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{2}\right)$ and $N_1 = N_2 = 20$

7.6 A possible application

7.6.1 — An application we have considered consists in recovering the surface of a given countryside just by using informations associated to *one* black and white virtual images. We can find (*e.g.* on the internet) such images with data sets that indicates an approximation of the brightness intensity at each pixel of the images, and the altitudes at each pixel of the images. In this section, we will present various simulations done by using this kind of data sets. Here are mentioned the main assumptions done for our simulations:

- The unit light vector is given by $(l_1, l_2, l_3) = \left(\frac{-1}{2}, \frac{-1}{2}, \frac{\sqrt{2}}{2}\right)$.
- The altitudes will be expressed in meters. Whatever the value of N, the mesh-size will $h = (h_1, h_2)$ will satisfy $h_1 = h_2 = 1$.
- Given indexes i and j, the brightness intensity $I(\omega_{i,j})$ was computed (and so, the virtual image obtained) by replacing $\nabla u(\omega_{i,j})$ by an approximation $\nabla_{\text{app}} u(\omega_{i,j})$ in the brightness equation (II.1). The approximation $\nabla_{\text{app}} u(\omega_{i,j}) = \big(\partial_{1,\text{app}} u(x),\ \partial_{2,\text{app}} u(x)\big)$ was defined as follows:

$$\partial_{1,\text{app}} u(\omega_{i,j}) = \frac{[u(\omega_{i+1,j-1}) - u(\omega_{i-1,j-1})] + 2[u(\omega_{i+1,j}) - u(\omega_{i-1,j})] + [u(\omega_{i+1,j+1}) - u(\omega_{i-1,j+1})]}{4h_1},$$
$$\partial_{2,\text{app}} u(\omega_{i,j}) = \frac{[u(\omega_{i-1,j+1}) - u(\omega_{i-1,j-1})] + 2[u(\omega_{i,j+1}) - u(\omega_{i,j-1})] + [u(\omega_{i+1,j+1}) - u(\omega_{i+1,j-1})]}{4h_2}.$$

Then, we have performed our algorithm in the following situations:

n° 1: $N = (200, 200)$, and the black and white picture is given by the figure 51. In this case, $I_{\text{min}} = 0.22$, $I_{\text{max}} = 0.82$, and the minimal and maximal altitudes are $u_{\text{min}} = 1087.78\,\text{m}$ and $u_{\text{max}} = 1143.70\,\text{m}$.

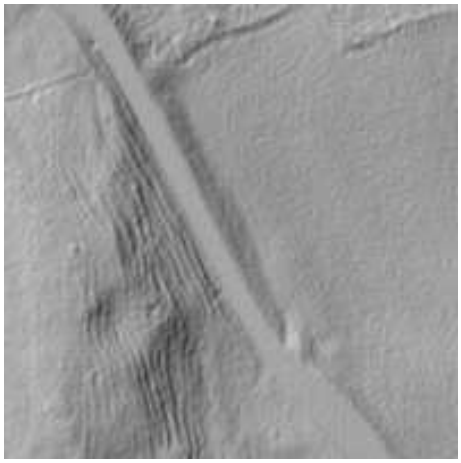

Figure 51 – Black and white image associated to the situation n° 1

n° 2: $N = (200, 200)$, and the black and white picture is given by the figure 52. In this case, $I_{\text{min}} = 0$, $I_{\text{max}} = 0.88$, and the minimal and maximal altitudes are $u_{\text{min}} = 1032.90\,\text{m}$ and $u_{\text{max}} = 1119.98\,\text{m}$.

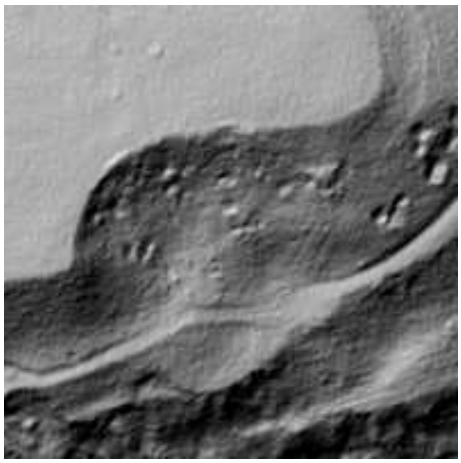

Figure 52 – Black and white image associated to the situation n° 2

n° 3: $N = (200, 200)$, and the black and white picture is given by the figure 53. In this case, $I_{\min} = 0.13$, $I_{\max} = 1$, there are $n_{\text{fix}} = 16$ nodes of Ω_N where the brightness intensity is greater than 0.99, and the minimal and maximal altitudes are $u_{\min} = 978.42\,\text{m}$ and $u_{\max} = 1067.16\,\text{m}$.

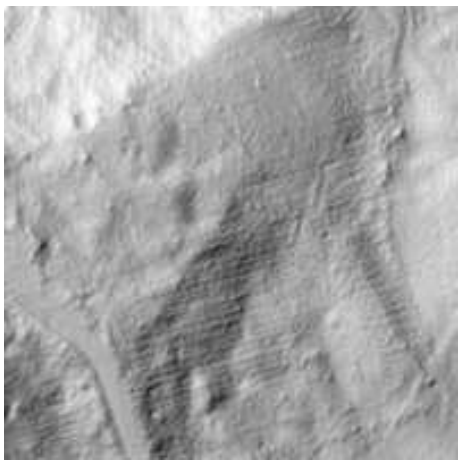

Figure 53 – Black and white image associated to the situation n° 3

REMARK. The previous black and white images are views of moutain ranges from the Canton of Zürich (Switzerland). The associated data sets come from the file `26740_12450.tif` from [1].

7.6.2 — For each situation, we have applied our algorithm as for the simulations done in section 7.5. The following table indicates their associated performances:

situation	η_∞	η_{ave}	ι	τ
n° 1	1.32	0.36	398	11.59 s
n° 2	2.55	0.31	396	11.24 s
n° 3	2.05	0.47	516	15.18 s

Table 23 – Values of η_∞, η_{ave}, ι and τ depending on the considered situation

As we can see, the errors η_∞ and η_{ave} are close to 1, *i.e.* of the order of $1/h_1 = 1/h_2$, which is coherent with the results observed in section 7.5. It also proves that our algorithm tolerates the imprecision on the values of the brightness intensity I (given in input). These errors are also very satisfying in comparison to the associated differences $u_{\max} - u_{\min}$.
We can also note that the number of iterations ι and the computation times τ are equivalent to those obtained in our previous situations.
In order to illustrate graphically our simulations, we have presented in figures 54 to 56 the surfaces recovered with the computed altitudes (stored in U_{app}, given in output of our algorithm) associated to simulations n° 1 to n° 3. In particular with the surface from figure 55 (associated to the situation n° 2), we can observe the plateau that seems to be represented in the left top of figure 52.

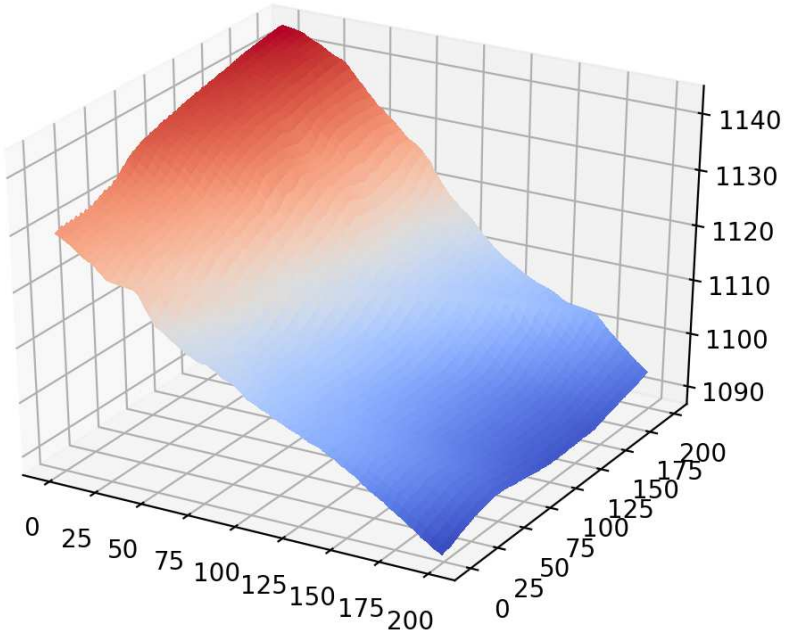

Figure 54 – Graphical representation of the approximation U_{app} associated to the situation n$^{\text{o}}$ 1

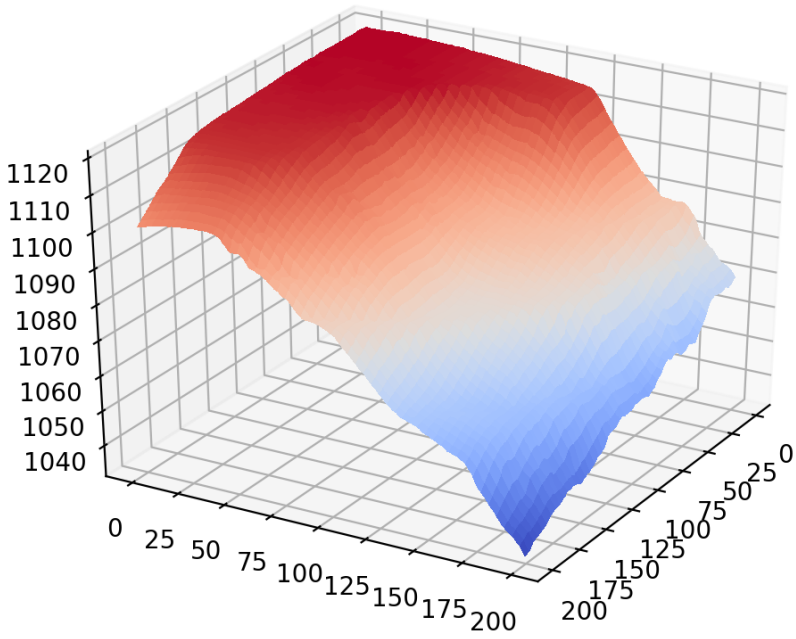

Figure 55 – Graphical representation of the approximation U_{app} associated to the situation n$^{\text{o}}$ 2

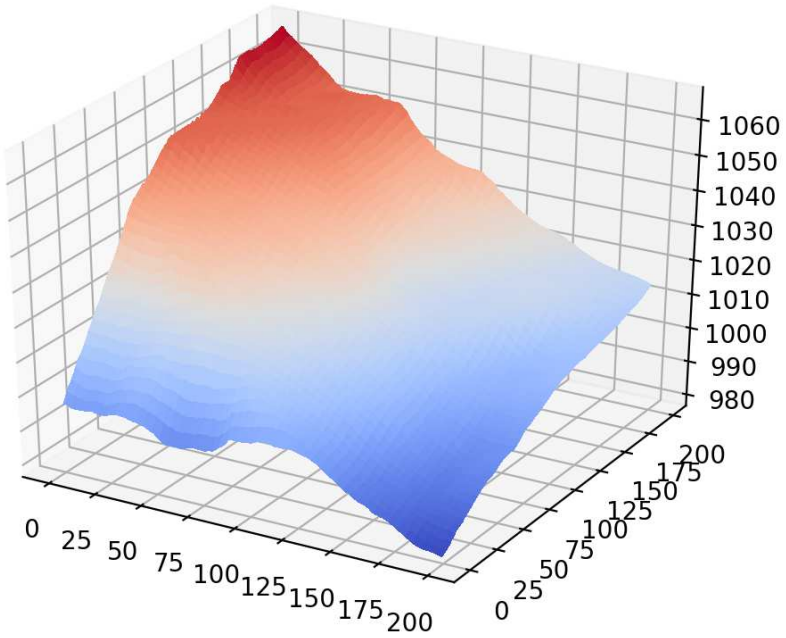

Figure 56 – Graphical representation of the approximation U_{app} associated to the situation n$^{\text{o}}$ 3

7.6.3 — In order to improve the output of the simulations done in paragraph 7.6.2, we have tried to use a Newton's method. Let us explain how we have done this. We still want to determine a matrix

$$U \;=\; (U_{i,j})_{\substack{0\leqslant i\leqslant N_1\\ 0\leqslant j\leqslant N_2}}$$

such that $U_{i,j}$ corresponds to an approximation of $u(\omega_{i,j})$, for all $(i,j)\in\{0,\dots,N_1\}\times\{0,\dots,N_2\}$. As explained in 7.6.1, in our case, for such a couple (i,j), the brightness intensity $I(\omega_{i,j})$ satisfies

$$\frac{-\nabla_{\text{app}}u(\omega_{i,j})\cdot l + l_3}{\sqrt{1+|\nabla_{\text{app}}u(\omega_{i,j})|^2}} - I(\omega_{i,j}) \;=\; 0.$$

By setting

$$\begin{aligned} d_1(U) \;&=\; \frac{[U_{i+1,j-1}-U_{i-1,j-1}]+2[U_{i+1,j}-U_{i-1,j}]+[U_{i+1,j+1}-U_{i-1,j+1}]}{4h_1},\\ d_2(U) \;&=\; \frac{[U_{i-1,j+1}-U_{i-1,j-1}]+2[U_{i,j+1}-U_{i,j-1}]+[U_{i+1,j+1}-U_{i+1,j-1}]}{4h_2}, \end{aligned}$$

then $\big(d_1(U),d_2(U)\big)$ corresponds to an approximation of $\nabla_{\text{app}}u(\omega_{i,j})$. Thus by considering the function

$$F \;:\; \mathrm{M}_{N+1}(\mathbb{R})\to\mathrm{M}_{N+1}(\mathbb{R})\,,\; U=(U_{i,j})_{\substack{0\leqslant i\leqslant N_1\\ 0\leqslant j\leqslant N_2}} \mapsto F(U)=\big(F_{i,j}(U)\big)_{\substack{0\leqslant i\leqslant N_1\\ 0\leqslant j\leqslant N_2}}$$

where, for all $(i,j)\in\{0,\dots,N_1\}\times\{0,\dots,N_2\}$,

$$F_{i,j}(U) \;=\; \frac{-l_1d_1(U)-l_2d_2(U)+l_3}{\sqrt{1+d_1^2(U)+d_2^2(U)}} - I(\omega_{i,j}),$$

we have thus applied the Newton's method in order to solve the equation $F(U) = 0$, by starting with the matrix U_{app} obtained in output of our algorithm.
Unfortunately, it is well-known that the Newton's method is convergent for initial data sufficiently close to the solution. In our case, U_{app} seems still too far from the solution of our s.f.s. problem, so the Newton's method does not converge. At the moment, we have not find a way that enables to overcome this difficulty.

7.6.4 — In comparison to the data sets used as explained in 7.6.1, we can also find sets where we still know (an approximation of) I at each pixel of the corresponding image, but the altitudes only at some pixels *not necessarily located on the boundary of the image*. This raises the question of the adaptability of our models when the brightness equation is not associated to Dirichlet boundary conditions.
For now, we are not able to recover any information on u without Dirichlet boundary conditions. In practice, this constitutes an important limitation of our work. In his P.h.D. thesis [17], E. PRADOS has proposed a new formulation of the s.f.s. problem in order to overcome this difficulty, but we have not considered it for now.

Conclusion

In this chapter, we have shown how to extend our theory and algorithm from chapter 6 to the bi-dimensional case. Therefore, we have been able to implement an efficient algorithm that enables to solve numerically various s.f.s. problems. Unfortunately, our algorithm requires Dirichlet boundary conditions. If this hypothesis was motivated theoretically in chapter 5, it is not realistic for many applications of the s.f.s. problem. For now, we have not tried to overcome this difficulty.

Conclusion et perspectives

Pour rappel, dans cette partie, nous nous intéressions à un problème de *shape from shading*, ou comment reconstituer une surface représentée sur une image en noir et blanc, à partir des seules nuances de gris et d'altitudes connues en certains points. Nous avons vu qu'un tel problème pouvait mathématiquement être modélisé par une équation de Hamilton-Jacobi du premier ordre, que nous avons résolu dans un cadre classique : une équation couplée à des conditions de type Dirichlet.
Ce faisant, au chapitre 5, nous avons montré comment l'utilisation du formalisme des solutions de viscosité permettait d'assurer l'existence et l'unicité d'une solution à un tel problème, de base mal posé. Nous avons ensuite proposé une formulation explicite d'un schéma d'approximation adapté au problème de s.f.s, considéré comme implicite jusqu'alors, dans un cadre unidimensionnel au chapitre 6 puis dans un cadre bidimensionnel au chapitre 7. En nous basant sur ce schéma, et plus particulièrement sur sa formulation explicite, nous avons ainsi pu proposer un algorithme de résolution du problème de s.f.s. qui, dans les cas que nous avons considéré, s'avère significativement plus rapide que ceux proposés jusqu'alors.
À présent, une continuation logique de notre travail consisterait à sortir de ce cadre d'étude classique, où notre équation est couplée à des conditions de type Dirichlet. Si cette approche usuelle peut s'avérer raisonnable dans certains cadres, elle se révèle bien trop contraignante dans le nôtre. En effet, dans bon nombre de cas concrets, l'altitude de la surface que l'on cherche à reconstruire est inconnue au bord, mais connue en un certain nombre de points situés à l'intérieur du domaine. Un premier travail pourrait donc consister à voir comment adapter notre travail, si tenté qu'il puisse l'être, à ce type de situations.

Terminons en signalant une application possible du shape from shading. Comme nous l'avons évoqué à la section 7.6, on trouve aujourd'hui en accès libre de nombreuses images, par exemple satellites ou virtuelles, en noir et blanc avec les jeux de données d'intensité associés, mais pas nécessairement les altitudes correspondantes. Une fois que les difficultés liées à la résolution du problème de s.f.s., précédemment évoquées, seront surmontées, l'information que nous récupérerions sur l'altitude pourrait permettre à des mobiles volants, tels que les drônes, de connaître les éventuels obstacles à éviter. Utilisant en outre les modèles présentés dans la partie I de ce manuscrit, nous pourrions ainsi permettre leur guidage en temps réel.
Même si ce cadre applicatif resterait à préciser et considérer, nous pourrions ainsi unifier de manière concrète les théories présentées dans les deux parties de ce manuscrit. Et nous donnerions de cette façon, à celles et ceux qui en douteraient encore, une nouvelle illustration du caractère applicatif des Mathématiques à de très larges situations.

Main assumptions done in part II

(A_1) There exists a modulus m such that, for all $x, y \in \Omega$ and $p \in \mathbb{R}^n$:

$$H(x,p) - H(y,p) \leqslant m\big(|x-y|(1+|p|)\big).$$

(A_2) For all $x \in \Omega$, the function $H(x,\cdot\,) : \mathbb{R}^n \to \mathbb{R},\ p \mapsto H(x,p)$ is convex.

(A_3) There exists $\alpha < 0$ and $\psi \in \mathcal{C}\big(\overline{\Omega}\big)$ with $\psi_{|\Omega} \in \mathcal{C}^1(\Omega)$ such that, for all $x \in \Omega$:

$$H(x, \nabla\psi(x)) \leqslant \alpha.$$

(A_4) Ω is smooth and connected.

(A_5) $\inf\{H(x,p) \mid p \in \mathbb{R}^n\} \leqslant 0$ for all $x \in \Omega$.

(A_6) For all $x \in \Omega$, function $H(x,\,\cdot\,)$ is coercive, *i.e.* $H(x,p) \to +\infty$ when $|p| \to +\infty$.

(A_7) H can be extended as a continuous function on $\overline{\Omega} \times \mathbb{R}^n$, still noted H.

(A_8) For all $x, y \in \partial\Omega$, function φ satisfies the following compatibility condition:

$$\varphi(x) - \varphi(y) \leqslant L(x,y).$$

(A_9) I is continuous.

(A_{10}) $I(x) > |l|$ for all $x \in \Omega$.

(A_{11}) There exists $\alpha < 1$ such that $I(x) \leqslant \alpha$ for all $x \in \Omega$.

(A_{12}) S is monotonous.

(A_{13}) For all $h \in \big(\mathbb{R}_+^*\big)^n$, $x \in \overline{\Omega}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$, the function $S(h,x,\,\cdot\,,u)$ is non-decreasing and has a positive limit in the neighbourhood of $+\infty$.

(A_{14}) There exists $d \in \mathbb{N}^*$ such that for all $h \in \big(\mathbb{R}_+^*\big)^n$ and $x \in \overline{\Omega}$, there exists $\Xi_{h,x} \subset \overline{\Omega}^d$ and $\sigma_{h,x} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ such that $S(h,x,t,u) = \sigma_{h,x}\big(t, (u(\xi))_{\xi\in\Xi_{h,x}}\big)$ for all $t \in \mathbb{R}$ and $u \in \mathcal{B}\big(\overline{\Omega}\big)$.

(A_{15}) For all $h \in \big(\mathbb{R}_+^*\big)^n$, the scheme S admits a subsolution.

(A_{16}) All the subsolutions of S are upper bounded independently with respect to the first variable of S.

(A_{17}) S is stable.

(A_{18}) S is consistent.

(A_{19}) Equation (5.19) satisfies a *strong uniqueness* property, *i.e.* if u is an u.s.c. solution of (5.19) and v a l.s.c. solution of (5.19), then $u(x) \leqslant v(x)$ for all $x \in \overline{\Omega}$.

General notations and conventions

We specify here the general notations and conventions used in all this document.

Abbreviations

e.g. corresponds to the abbreviation of the Latin locution *exampli gratia*, which means *for instance*.

etc. corresponds to the abbreviation of the Latin locution *et cetera*, which means *and the rest*.

i.e. corresponds to the abbreviation of the Latin locution *id est*, which means *that is*.

l.s.c. corresponds to the abbreviation of *lower semi-continuous*.

resp. corresponds to the abbreviation of *respectively*.

s.f.s. correspond to the abbreviation of *shape from shading*.

s.t. corresponds to the abbreviation of *such that*.

u.s.c. corresponds to the abbreviation of *upper semi-continuous*.

Quantifiers

We denote by $\forall$ the universal quantifier and by $\exists$ the existential quantifier.

Sets theory

- Symbols $\in$, $\subset$, $\cup$ and $\cap$ have their usual signification.
- We denote by $\varnothing$ the empty set.
- If X is a set and A, B two subsets of X, $B \setminus A$ is the set of $x \in X$ such that $x \in B$ and $x \notin B$.
- If n is an integer larger than 2, $X_1, \ldots, X_n$ are n sets, and $x \in X_1 \times \cdots \times X_n$, for all integer j from 1 to n, we denote by x_j or $x(j)$ the j^{th} component of x.

Numbers

- $\mathbb{N}$ and $\mathbb{R}$ respectively correspond to the sets of natural numbers, integers and real numbers.
- We denote by $\leqslant$ the usual order on $\mathbb{N}$ or $\mathbb{R}$. In addition, we define

$$\mathbb{N}^* \,=\, \mathbb{N}\smallsetminus\{0\}, \quad \mathbb{R}^* \,=\, \mathbb{R}\smallsetminus\{0\}, \quad \mathbb{R}_+ \,=\, \{x\in\mathbb{R}\mid x\geqslant 0\} \quad \text{and} \quad \mathbb{R}_+^* \,=\, \mathbb{R}^*\cap\mathbb{R}_+.$$

- We add to $\mathbb{R}$ two symbols, $-\infty$ and $+\infty$, and we extend the usual ordering relation on $\mathbb{R}$ admitting that $-\infty\leqslant x\leqslant+\infty$ for all $x\in\mathbb{R}\cup\{-\infty,+\infty\}$.
- We will adopt the french notation to designate the intervals. Concretely, if $a,b\in\mathbb{R}$, we define

$$\begin{array}{ll} [a,b] \,=\, \{x\in\mathbb{R}\mid a\leqslant x\leqslant b\}, & [a,+\infty[\,=\, \{x\in\mathbb{R}\mid x\geqslant a\},\\ ]a,b[\,=\, \{x\in\mathbb{R}\mid a<x<b\}, &]-\infty,a] \,=\, \{x\in\mathbb{R}\mid x\leqslant a\},\\ [a,b[\,=\, \{x\in\mathbb{R}\mid a\leqslant x<b\}, &]a,+\infty[\,=\, \{x\in\mathbb{R}\mid x>a\},\\ ]a,b] \,=\, \{x\in\mathbb{R}\mid a<x\leqslant b\}, &]-\infty,a[\,=\, \{x\in\mathbb{R}\mid x<a\}. \end{array}$$

- If A is a part of $\mathbb{R}$, we define $-A=\{-a\mid a\in A\}$.
- If A is a non-empty subset (or family of elements) of $\mathbb{R}$ lower (resp. upper) bounded, we denote by $\inf A$ (resp. $\sup A$) the infimum (resp. supremum) of A, and we also denote it by $\min A$ (resp. $\max A$) when there exists a an element of A such that $a=\inf A$ (resp. $a=\sup A$). When A is not lower (resp. upper) bounded, we set $\inf A=-\infty$ (resp. $\sup A=+\infty$).

Applications

- The identity application of a set X is $\mathrm{id}_X : X\to X,\ x\mapsto x$.
- We denote by $f_{|A}$ the restriction of an application f to a set A.
- We denote by $g\circ f$ the application that composes an application g with an application f.
- If $n\in\mathbb{N}\not\in\{0,1\}$, $k\in\{1,\dots,n\}$, $X_1,\dots,X_n,Y$ are $(n+1)$ sets and $f : X_1\times\cdots\times X_n\to Y$ is an application, given $x_j\in X_j$ for all $j\in\{1,\dots,n\}\smallsetminus\{k\}$, we define the k^{th} partial application of f by

$$f(x_1,\dots,x_{k-1},\,\cdot\,,x_{k+1},\dots,x_n) \,:\, X_k\to Y,\ x\mapsto f(x_1,\dots,x_{k-1},x,x_{k+1},\dots,x_n).$$

- If $n\in\mathbb{N}^*$, X, Y_1, …, Y_n are $(n+1)$ sets, $f : X\to Y_1\times\cdots\times Y_n$, $x\in X$ and $j\in\{1,\dots,n\}$, we denote by $f_j(x)$ the j^{th} component of $f(x)$, hence an application $f_j : X\to Y_j,\ x\mapsto f_j(x)$.
- When it makes sense, we write

$$\lim_{x\to a} f(x) \,=\, \lambda, \quad \text{or} \quad f(x)\to\lambda \text{ when } x\to a, \quad \text{or} \quad f(x)\xrightarrow[x\to a]{}\lambda$$

to mean that the limit of $f(x)$ when x approaches a is equal to λ.

Functions

A *function* corresponds to an application that is real valued.

- If X is a set, f and g are two functions defined on X, and $\lambda \in \mathbb{R}$:
  - function $f+g$ is given by $(f+g) : X \to \mathbb{R},\ x \mapsto f(x)+g(x)$,
  - function λf is given by $\lambda f : X \to \mathbb{R},\ x \mapsto \lambda f(x)$, with in particular $-f = (-1)f$,
  - function $f+\lambda$ is given by $f+\lambda : X \to \mathbb{R},\ x \mapsto f(x)+\lambda$,
  - functions $f-g$ and $f-\lambda$ respectively corresponds to functions $f+(-g)$ and $f+(-\lambda)$,
  - we write $f \leqslant g$ to mean that $f(x) \leqslant g(x)$ for all $x \in X$.
- If A is a set and $f(x)$ a real quantity depending on x, we define

$$\inf_{x\in A} f(x) \ = \ \inf\{f(x) \mid x \in A\} \qquad (\text{resp. } \sup_{x\in A} f(x) \ = \ \sup\{f(x) \mid x \in A\}).$$

Derivatives and differentials

- When it makes sense, if f is a function defined on a non-empty interval I, $a \in I$ and $k \in \mathbb{N}^*$, we denote by $f^{(k)}(a)$ the k^{th} derivative of f in a, hence a function $f^{(k)} : I \to \mathbb{R},\ x \mapsto f^{(k)}(x)$. By convention we set $f^{(0)} = f$. Finally we can also write f' instead of $f^{(1)}$, f'' instead of $f^{(2)}$, etc.
  More generally, if $f : I \to \mathbb{R}^n$ ($n \in \mathbb{N}^*$), we extend these notations by setting $f^{(k)} = \left(f_1^{(k)}, \ldots, f_n^{(k)}\right)$.
- If $n \in \mathbb{N}^*$, $k \in \{1, \ldots, n\}$, f is a sufficiently smooth function defined on U an open subset of $\mathbb{R}^n$ and $a \in U$, we denote by $\partial_k f(a)$ the k^{th} partial derivative of f in a and by $\nabla f(a) = (\partial_j f(a))_{1\leqslant j\leqslant n}$ its gradient, hence applications $\partial_k f : U \to \mathbb{R},\ x \mapsto \partial_j f(x)$ and $\nabla f : U \to \mathbb{R}^n,\ x \mapsto \nabla f(x)$. We will also write $\partial_k^2 f$ instead of $\partial_k(\partial_k f)$.

General bibliography

[1] Kanton zürich webpage. `http://maps.zh.ch/download/hoehen/2014/dtm/`.

[2] I. Ayed, E. de Bézenac, A. Pajot, J. Brajard, and P. Gallinari. Learning dynamical systems from partial observations. *arXiv preprint arXiv:1902.11136*, 2019.

[3] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer Science & Business Media, 2008.

[4] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*. Springer-Verlag, 1994.

[5] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic analysis*, 4(3):271–283, 1991.

[6] A. Bressan. Viscosity solutions of hamilton-jacobi equations and optimal control problems. *Lecture notes*, 2011.

[7] M. G. Crandall, L. C. Evans, and P.-L. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American Mathematical Society*, 282(2):487–502, 1984.

[8] R. Farager. Understanding the basis of the kalman filter. *IEEE SIGNAL PROCESSING MAGAZINE*, 2012.

[9] B. K. Horn. Obtaining shape from shading information. *The psychology of computer vision*, pages 115–155, 1975.

[10] H. Ishii. A simple, direct proof of uniqueness for solutions of the Hamilton-Jacobi equations of eikonal type. *Proceedings of the American Mathematical Society*, pages 247–251, 1987.

[11] L. Jaulin. *Le calcul ensembliste par analyse par intervalles et ses applications*. PhD thesis, Université d'Angers, 2000.

[12] R. Kimmel and J. A. Sethian. Optimal algorithm for shape from shading and path planning. *Journal of Mathematical Imaging and Vision*, 14(3):237–244, 2001.

[13] S. K. Lele. Compact finite difference schemes with spectral-like resolution. *Journal of computational physics*, 103(1):16–42, 1992.

[14] P.-L. Lions. *Generalized solutions of Hamilton-Jacobi equations*, volume 69. London Pitman, 1982.

[15] P.-L. Lions. Solutions de viscosité des équations de hamilton-jacobi du premier ordre et applications. *Séminaire Équations aux dérivées partielles (Polytechnique)*, pages 1–12, 1984.

[16] P.-L. Lions, E. Rouy, and A. Tourin. Shape-from-shading, viscosity solutions and edges. *Numerische Mathematik*, 64(1):323–353, 1993.

[17] E. Prados. *Application of the theory of the viscosity solutions to the Shape From Shading problem*. PhD thesis, Université Nice Sophia Antipolis, 2004.

[18] E. Prados and O. Faugeras. A mathematical and algorithmic study of the Lambertian SFS problem for orthographic and pinhole cameras. Technical Report RR-5005, INRIA, 2003.

[19] E. Rouy and A. Tourin. A viscosity solutions approach to shape-from-shading. *SIAM Journal on Numerical Analysis*, 29(3):867–884, 1992.

[20] J. A. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

[21] B. Wu, W. C. Liu, A. Grumpe, and C. Wöhler. Construction of pixel-level resolution dems from monocular images by shape and albedo from shading constrained with low-resolution dem. *ISPRS journal of photogrammetry and remote sensing*, 140:3–19, 2018.

General index

Abstract

The work presented in the first part of this thesis is the result of a collaboration between Alstom and the RATP. We present various models and algorithms that can be used to bound a real-valued function f defined on an interval I and its $(d-1)$ first derivatives by knowing reliable bounds on f in some discrete points and global bounds on its d^{th} derivative. These results are applied to a situation inspired by the railway world. Finally, we present various extensions of our work, and we explain how the previous models can be easily generalized to vector-valued applications defined on an interval.
The second part of this thesis is dedicated to the theoretical and numerical study of a *shape from shading* problem, which consists in a surface reconstitution from a black and white picture, by knowing only the shades of gray and the altitude of the surface at some points. We remind how the viscosity solutions framework allows us to obtain a well-posed formulation of this problem. Then we expose an explicit expression of an approximation scheme associated to this problem, and we propose a significant optimization of some algorithms used to solve numerically such a problem.
In the future, the works presented in the two parts of the thesis could be coupled to allow a real-time guidance of flying objects like drones over a given region.

Keywords. Differentiable functions, bounded functions, certified bounds, Taylor-Lagrange formula, optimisation, real time computations, Hamilton-Jacobi equations, finite differences schemes, shape from shading.

Résumé

Les travaux présentés dans la première partie de ce manuscrit de thèse sont le fruit d'une collaboration entre Alstom et la RATP. Nous y présentons différents modèles et algorithmes permettant de borner une fonction réelle f définie sur un intervalle I et ses $(d-1)$ premières dérivées à partir de bornes sur f en certains points et de bornes globales sur la dérivée d-ième de f. Nous appliquons cela à une situation inspirée du monde ferroviaire. Enfin, nous présentons diverses extensions de nos travaux, et nous montrons comment les résultats précédents peuvent se généraliser à des applications définies sur un intervalle I et à valeurs vectorielles.
La seconde partie de ce manuscrit est consacrée à l'étude théorique et numérique d'un problème de « *shape from shading* », qui consiste à reconstituer une surface représentée sur une image en noir et blanc, par la seule connaissance des nuances de gris et d'altitudes en certains points. Nous y rappelons comment le cadre des solutions de viscosité permet d'obtenir une formulation mathématique bien posée de ce problème. Nous donnons ensuite une formulation explicite d'un schéma d'approximation associé à ce problème, et nous proposons une optimisation notable d'algorithmes permettant de résoudre numériquement un tel problème.
À terme, l'ensemble des travaux présentés dans ce manuscrit pourraient être couplés pour permettre le guidage en temps réel de mobiles volants, tels que des drones, au dessus d'une région donnée.

Mots-clés. Fonctions différentiables, fonctions bornées, bornes certifiées, formule de Taylor-Lagrange, optimisation, calculs en temps réel, équations de Hamilton-Jacobi, schémas aux différences finies, problème de « shape from shading ».