Philosophie, méthodologie et applications de l'analyse non standard

Augustin Fruchard, Veronique Gautheron, Tewfik Sari

Actes du colloque à la mémoire d'Emmanuel Isambert

21-22 décembre 2007, Université de Paris 7

Philosophie, méthodologie et applications de l'analyse non standard

Éditeurs :

Augustin Fruchard
Véronique Gautheron
Tewfik Sari

2010 Mathematics Subject Classification:
01, 03, 28, 34, 35, 39, 41, 49, 60, 92, 93.

Publications de l'Université de Paris 13

Emmanuel Isambert était Professeur à l'université de Paris 13, lorsqu'il est décédé brutalement en septembre 2007.

Logicien et mathématicien, spécialiste d'équations différentielles, il était un membre très actif du réseau Georges Reeb.

La rencontre du réseau de décembre 2007 était dédiée à sa mémoire.

La diversité des thèmes abordés dans ces Actes, philosophie, logique et théorie des ensembles, mathématiques financières, probabilités et équations différentielles, histoire des mathématiques, témoigne de l'ouverture d'esprit d'Emmanuel.

Préface

Emmanuel Isambert allait sur ses soixante ans au moment de sa disparition brutale en septembre 2007. Après une formation et une thèse de logique, puis une assez brève incursion dans les espaces de Banach, un problème de mécanique des solides l'a orienté vers l'utilisation de l'analyse non standard dans l'étude des équations différentielles ordinaires. Il a en particulier étudié de près les fleuves, qui sont des trajectoires de champs de vecteurs où se concentrent d'autres trajectoires, en collaboration avec Michèle Artigue et Véronique Gautheron. Il a aussi étudié les canards d'équations non différentiables et exploré les fondements de l'analyse non standard.

Emmanuel était un pilier du réseau Georges Reeb, participant activement aux rencontres, ayant toujours des questions pertinentes à poser à la fin des exposés, avec une belle énergie et sa bonne humeur inoxydable.

La rencontre du réseau qui a eu lieu à Paris Chevaleret en décembre 2007 était dédiée à sa mémoire. Les Actes contiennent d'une part la version écrite – parfois largement complétée – des exposés et d'autre part des contributions nouvelles. La diversité des thèmes abordés dans ces Actes, allant de questions philosophiques, logiques ou de théorie des ensembles, aux mathématiques financières, probabilités et équations différentielles, en passant par des thèmes relevant de l'histoire des mathématiques, témoigne de l'ouverture d'esprit d'Emmanuel.

Emmanuel est arrivé dans la toute jeune Université Paris 13 en 1971. Il y a toujours pris à cœur son rôle d'enseignant. Elément stable et fiable de l'équipe, il était volontiers prêt à faire différemment pour faire mieux. En particulier il a saisi toutes les opportunités d'introduire et d'expérimenter l'outil informatique dans les formations qui se sont ouvertes, comme en témoignent (page vii) deux de ses collègues du département de mathématiques.

Emmanuel venait de se remarier au printemps 2007, quelques mois avant son décès, et respirait le bonheur, empli de projets d'avenir, comme le montrent les photos de première page prises le jour de ce mariage. Deux autres passions l'habitaient : la musique comme hautboïste d'orchestre, puis choriste d'opérette, et la montagne, qu'il a escaladée sur tous les continents.

Sa fille, physicienne, née d'une vie antérieure, venait de lui annoncer qu'il allait être grand-père, mais il n'a malheureusement pas pu voir la petite Clémence.

Les éditeurs

Liste des participants et contributeurs

Michèle Artigue (Paris) ..... michele.artigue@univ-paris-diderot.fr
Rachid Bebbouchi (Alger) ..... rbebbouchi@hotmail.com
Éric Benoît (La Rochelle) ..... ebenoit@univ-lr.fr
Imme van den Berg (Évora) ..... ivdb@uevora.pt
Gilles Bernot (Sophia-Antipolis) ..... bernot@unice.fr
Jacques Bosgiraud (Paris) ..... jacques.bosgiraud@orange.fr
Bernard Brighi (Mulhouse) ..... bernard.brighi@uha.fr
Agathe Chollet (La Rochelle) ..... achollet@univ-lr.fr
Jean-Paul Comet (Sophia-Antipolis) ..... comet@unice.fr
René Cori (Paris) ..... cori@logique.jussieu.fr
Aparna Das (Nice) ..... Aparna.DAS@unice.fr
Antoine Delcroix (Pointe-à-Pitre) ..... adelcroi@univ-ag.fr
Francine Diener (Nice) ..... Francine.DIENER@unice.fr
Marc Diener (Nice) ..... diener@unice.fr
Frédéric Eyssette (Sophia-Antipolis) ..... Frederic.Eyssette@unice.fr
Thomas Forget (La Rochelle) ..... tforget@univ-lr.fr
Augustin Fruchard (Mulhouse) ..... augustin.fruchard@uha.fr
Véronique Gautheron (Paris) ..... vero.gauth@free.fr
Rémi Goblot (Lille) ..... remi.goblot@nordnet.fr
Luis Gonzaga Albuquerque (Lisbonne) ..... lgalbu@univ-ab.pt
Claude Lobry (Nice) ..... lobrinria@wanadoo.fr
Robert Lutz (Mulhouse) ..... robert.lutz@uha.fr
Gwenola Madec (Paris) ..... madec@math.univ-paris13.fr
Jean-André Marti (Pointe-à Pitre) ..... Jean-Andre.Marti@univ-ag.fr
François Parreau (Paris) ..... parreau@math.univ-paris13.fr
Pheakdei Mauk (Nice) ..... pheak@unice.fr
Yves Péraire (Clermont-Ferrand) ..... peraire2000@yahoo.fr
Yvette Perrin (Clermont-Ferrand) ..... yvette.perrin1@orange.fr
Jean-Michel Salanskis (Paris) ..... jmsalanskis@free.fr
Nadir Sari (La Rochelle) ..... nsari@univ-lr.fr
Tewfik Sari (Mulhouse) ..... tewfik.sari@uha.fr
Dimitris Scarpalezos (Paris) ..... scarpa@math.jussieu.fr
Reinhard Schäfke (Strasbourg) ..... schaefke@math.u-strasbg.fr
Jean-Marie Strelcyn (Paris-Villetaneuse) ..... strelcyn@math.univ-paris13.fr
Guy Wallet (La Rochelle) ..... guy.wallet@gmail.com

Table des matières – Contents

Témoignages

Gwenola Madec et François Parreau

François : Dès le début des années 80, Emmanuel avait compris que l'outil informatique ouvrait un champ considérable de possibilités pour l'enseignement des mathématiques. Il existait alors très peu, sinon pas du tout, de ressources disponibles et tout était à inventer dans ce domaine. Avec un petit groupe de collègues motivés, nous avons appris ensemble à programmer et, sous l'impulsion de Jean-François Méla, nous avons mis au point un diplôme d'université "Mathématiques pour non-spécialistes". Le but était d'aborder un certain nombre de notions à partir de l'expérimentation sur machine et sans grands prérequis. Emmanuel était l'un des piliers de cette petite équipe ; il était toujours disponible et il y a apporté beaucoup d'idées, bien souvent les plus pertinentes. L'expérience n'a duré que quelques années, mais nous avons eu la satisfaction de voir des étudiants peu formés aux mathématiques universitaires partager le plaisir de la découverte et prendre goût à notre discipline.

Ensuite, nous avons participé à la création de la filière MASS à Paris 13, avec un DEUG "Économie, Informatique et Méthodes Mathématiques" pour lequel nous avons encore privilégié une approche pratique des mathématiques autant qu'il était possible. De plus, à l'époque l'Université n'avait pas de département d'Informatique et, en attendant le recrutement de spécialistes, nous avons aussi assuré les cours d'informatique en première année, en nous limitant à des éléments de programmation et d'algorithmique. Ces cours se pratiquaient l'essentiel du temps devant les ordinateurs et il y avait là une approche de l'enseignement très nouvelle pour nous. À vrai dire, l'Université manquait aussi d'ingénieurs et de techniciens et nous allions jusqu'à installer les salles machines.

Je garde un excellent souvenir de cette période, du travail en commun avec Emmanuel, de notre enthousiasme et du plaisir partagé de construire un rapport différent avec les étudiants.

Gwenola : Lorsqu'en 2002 les directives officielles conduisent à la mise en place d'enseignement de méthodologie en première année de licence, j'ai peu d'expérience et de recul sur l'enseignement dans le supérieur. Emmanuel, dont je viens de faire la connaissance puisque nous travaillons ensemble pour la première fois sur un enseignement de premier semestre, me propose d'utiliser la plateforme d'apprentissage en ligne Wims comme support à notre nouvelle formation. Mettre nos étudiants en activités, leurs donner les moyens de contrôler leur travail, sont les objectifs qui nous guident pour l'élaboration de scénarii utilisant l'outil Wims. Emmanuel apportait dans notre collaboration son expérience d'enseignant mais aussi ses questionnements et "dada" de chercheur, ou encore les réflexions sur la construction des connaissances et des compétences, qu'il avait développées en construisant un cours sur internet dont l'objet était les équations différentielles. Emmanuel a toujours écouté avec attention mes réflexions parfois naïves et ne m'a jamais fait sentir que, parfois, je "redécouvrais la lune". Il avait cette générosité dans le rapport humain qui rend faciles les collaborations. Il savait rebondir sur des propositions plus ou moins prometteuses. Il acceptait de prendre du temps, ou de le perdre, pour éclairer la réalité de nos étudiants (conception de tests d'évaluation de compétence à l'entrée de l'université, participation à un groupe de travail mêlant enseignants du secondaire et du supérieur ainsi que des IPR de l'Académie de Créteil, participation au groupe de travail Secondaire – Supérieur de l'IREM Paris Nord).

Tests, rigueur et fantaisie

Véronique Gautheron

Les amis de trente ans, on en ricane lorsqu'il s'agit de politique. Il semble que, dans d'autres cercles que ceux du pouvoir, certaines amitiés perdurent sans être trahies, et même sans tiédir. Je vais essayer d'évoquer celle qui m'a liée à Emmanuel depuis la fin de nos études.

Pendant deux ou trois dizaines d'années, presque sans interruption, nous avons passé chaque semaine une journée délicieuse à faire des maths ensemble, et tout le monde autour de moi savait que « le mardi, c'est sacré ».

Il m'est difficile de parler des intérêts mathématiques d'Emmanuel, parce qu'il s'intéressait à presque tout. Son parcours éclectique dans l'univers mathématique a pourtant une grande cohérence, et je vais essayer de parler de sa vision des maths, où il avait un pied dans la rigueur logique et l'autre dans l'expérimental.

Sa formation initiale, c'est la logique mathématique. Il en a gardé une rigueur sans concession dans son travail, il pinaillait sur les moindres détails de formulation. Ceux qui pensent que c'est un calvaire d'avoir un compagnon de travail qui ne supporte aucune imprécision ignorent à quel point on peut être strict et rigolo, comme il l'a toujours été.

J'ai commencé à travailler avec lui (et Michèle Artigue) sur un sujet de mécanique lié aux mouvements d'une toupie, plus précisément à la détermination des zones topologiques parcourues dans $\mathbb{S}^2$ par l'extrémité d'un vecteur lié à cette toupie. Aucun d'entre nous ne connaissait quoi que ce soit à la mécanique. Nous avons attaqué le problème en utilisant abondamment les possibilités graphiques des ordinateurs balbutiants de l'époque, navigant entre preuve théorique et images expérimentales obtenues sur une table traçante Hewlett-Packard. Ces allers et retours seront une constante dans tous les travaux d'Emmanuel.

Nous étions donc plongés dans des problèmes de mouvement, quand nous avons rencontré une équipe qui, à l'initiative de Georges Reeb et sous l'impulsion de Marc et Francine Diener en développait une approche très neuve : l'analyse non standard, dans sa version IST formalisée par E. Nelson. On y parlait d'ordre de grandeur, de nombres infiniment petits ; les premiers canards étaient nés, E. Benoît et J.-L. Callot en faisaient l'élevage dans les premiers fleuves, l'A.N.S. démarrait très fort.

Emmanuel en a été un des piliers pendant plus de dix ans.

Les ordinateurs avaient alors fait des progrès, entre autres leurs capacités graphiques. Nous en usions et abusions tous énormément. On pourrait presque dire que l'on faisait des mathématiques expérimentales, avant tout visuelles.

Certains matheux contestaient cette approche et son côté « mathématiques pour physiciens », ce qui sous-entendait encore chez certains « peu rigoureux ». Mais on ne pouvait pas attaquer Emmanuel, lui le logicien, sur ce terrain. Il a même contribué à renforcer les fondements théoriques de l'analyse non standard. On se souvient aussi de ses travaux sur les fleuves abstraits, les ombres de trajectoires, les canards anguleux.

Chemin faisant, nous avons visité l'Algérie, les Pays-Bas, la Corse, les soirées qui clôturaient les divers colloques n'étaient pas tristes, je me souviens d'une improvisation théâtrale savoureuse d'Emmanuel et Robert Lutz reconstituant une discussion de comptoir entre « Monsieur Arsène » et « Monsieur Paul ». Il a toujours eu le sens de la scène, qu'il a développé dans plusieurs domaines (musique, chorale, théâtre et j'en passe).

Mais revenons aux maths.

Et à sa conception imagée et dynamique des math en général et des équa-difs en particulier. Il attaquait un phénomène mathématique, aussi bien résolvant des équations (et il était expert en la matière) qu'en le considérant comme un objet vivant dont il voulait cerner la personnalité, toujours avec sa rigueur légendaire.

Il y a quelques années, alors que les activités non-standard proprement dites s'étaient un peu tassées, osons dire qu'elles étaient un peu passées de mode, on nous a proposé de participer à un cours scientifique sur internet. Il s'est jeté dans l'aventure avec autant de passion qu'il mettait à grimper sur l'Himalaya, et nous en avons réalisé la partie équations différentielles. Ça été l'occasion de les présenter de façon résolument géométrique, dynamique et interactive, parfois même nous l'espérons rigolote. Après tout, les systèmes dynamiques, ça gigote par nature. Nous, en tout cas, nous nous sommes bien amusés. Utiliser Internet pour inventer une nouvelle approche a été un défi passionnant.

Mais Emmanuel était capable d'avoir plusieurs fers au feu et, toujours éclectique, il écrivait parallèlement avec J.-M. Strelcyn un livre d'algèbre sur la théorie de Galois, auquel je l'ai vu travailler quelques jours avant sa mort, et dont on peut espérer la parution prochaine.

Je ne peux pas finir sans dire quelques mots du copain de toujours. Toutes les semaines, pendant la pause déjeuner on a testé systématiquement tous les restos chinois du Chinatown parisien, et vu leur turn-over on peut éternellement en goûter de nouveaux. Je ne vois aucun sujet dont nous n'ayons pas discuté, la politique, les amours, la famille, l'esthétique, le boulot, les joies et les peines.

J'aurais aimé parler aussi du musicien (le hautboïste quasi professionnel, le choriste de plusieurs opérettes) et du montagnard (des Alpes à l'Himalaya et à la Cordillère des Andes, de Chamonix au Népal et au Pérou), mais j'ai trop peur de dire des bêtises sur ces sujets si essentiels qui me sont totalement inconnus.

Terminons par quelques anecdotes.

Emmanuel au restaurant chinois : l'implacable logicien.

Acte I, rue du Château des Rentiers :

Emmanuel commande un dessert inconnu, on lui sert un verre contenant deux étages de gelée bizarre, l'un rouge et l'autre vert.

Il goute le vert : « Je n'ai jamais rien mangé de plus mauvais. »

Il arrive au rouge : « Si, maintenant, j'ai mangé quelque chose de plus mauvais. »

Acte II, avenue d'Ivry :

Emmanuel hésite sur le choix d'un dessert, et s'inquiète : « Ce n'est pas celui avec un étage rouge et l'autre vert ? »

Réponse du serveur : « Oh, je vois que Monsieur est un connaisseur, d'habitude nous n'en faisons pas, mais je vais le confectionner spécialement pour vous. »

Emmanuel aux champignons : l'expérimentateur stoïque.

Nous étions tous deux amateurs de champignons, et la virée automnale en forêt de Fontainebleau faisait partie intégrante de notre collaboration (on recherche ce qu'on peut...). Il était, en la matière comme en bien d'autres, beaucoup plus savant que moi, et m'a appris à déguster sans crainte les pleurotes, les pied-bleu, les tricholomes. Mais il a longtemps hésité avant de me faire confiance et manger des golmottes, ou amanites rougeâtres, que certains craignent de confondre avec la très vénéneuse amanite panthère.

Adresse de l'auteur : 34 rue Nationale, 75013 Paris

Courriel : vero.gauth@free.fr

Souvenirs...

Michèle Artigue

J'ai fait la connaissance d'Emmanuel lors d'un congrès de logique à Orléans au début des années 70. Il venait de soutenir ou allait soutenir sa thèse, je ne me souviens plus très bien, la mienne était récente, elle aussi. Avec Marie Jeanne Perrin et Anne Strauss, nous avons décidé de constituer un petit groupe et de travailler sur les modèles non standard de l'arithmétique du second ordre. Emmanuel travaillait déjà sur les extensions élémentaires non-standard de modèles de théorie des ensembles, le sujet était en revanche nouveau pour moi car ma thèse avait porté sur des questions liées à la récursivité. Nous travaillions ensemble tous les quatre au moins une fois par semaine. Les premières années, ce travail mené dans une ambiance détendue et amicale a été productif. L'étude d'extensions de modèles a débouché sur la notion de bicommutabilité, une relation d'équivalence entre théories généralisant la relation connue entre A_2 et $ZFC^- + V = HC$, et sur la recherche de sous-systèmes de la théorie des ensembles et de l'arithmétique du second ordre bicommutables [1]. Nous en avons exhibé différents exemples et montré aussi que la théorie KP des ensembles admissibles n'a pas d'équivalent par bicommutation dans l'arithmétique du second ordre. Considérant ensuite l'arithmétique du troisième ordre, nous avons montré finalement que, dans ce cas, la réponse dépend de l'interprétation choisie du langage de l'arithmétique vers le langage de la théorie des ensembles, et nous avons relié ce fait à la satisfaction de différentes formes faibles de l'axiome du choix. Emmanuel a joué un rôle clef dans toutes ces constructions.

Au bout de quelques années cependant, cette problématique s'est épuisée. Nous avions l'impression de tourner en rond, notre travail ne rebondissait pas sur de nouvelles questions. Le groupe s'est finalement dissous et chacun a suivi son chemin propre. Pour moi, c'est le travail que je menais parallèlement dans le cadre de l'IREM Paris 7 à l'école élémentaire expérimentale de l'Almont près de Melun, puis les sections expérimentales maths-physique montées avec les physiciens à l'université Paris 7 qui sont devenus mes centres principaux d'intérêt, suscitant à leur tour de multiples questions de recherche, didactique cette fois.

La collaboration avec Emmanuel s'est renouée quelques années plus tard. Avec Véronique Gautheron, après avoir exploité une des ressources technologiques de l'IREM (une superbe table traçante HP), pour nous initier à l'étude qualitative des systèmes différentiels autonomes, nous travaillions sur une conjecture de Jean-Louis Verdier que le frère de Véronique, Adrien Douady, nous avait suggéré d'explorer. Emmanuel nous a rejointes et son appui a été décisif. Il s'agissait d'étudier la nature topologique des variétés intégrales liées au mouvement d'un solide pesant autour d'un point fixe, en caractérisant l'ensemble de bifurcation correspondant ou diagramme de Cerf. Nous sommes arrivés à caractériser ces diagrammes dans le cas où le centre de gravité du solide est situé dans un des plans d'inertie passant par le point fixe ou à son voisinage, ce qui fut suffisant pour montrer que la conjecture initiale était erronée. La situation était plus complexe qu'initialement prévu. Les nombreux tracés effectués nous

ont aussi conduits à conjecturer que le cas général ne pouvait donner lieu à des diagrammes plus complexes que ceux déjà obtenus mais nous n'avons pas réussi à le démontrer [2].

C'est aussi à cette époque où nous avons appris que des chercheurs autour de Georges Reeb avaient développé une approche non-standard des concentrations de trajectoires que nous observions dans les portraits de phase de systèmes différentiels. Ce fut la découverte des fleuves et des canards, les contacts noués avec le petit monde de l'analyse non standard. Ces contacts perdureront pour Véronique et Emmanuel. De mon côté, après avoir soutenu ma thèse d'état, il me sera de plus en plus difficile de concilier recherche mathématique et recherche didactique, et nos chemins une fois de plus divergeront. Ils se recroiseront à diverses reprises et je suivrai avec beaucoup d'intérêt le travail que Véronique et lui réaliseront sur les équations différentielles pour Université en ligne, mais nous n'aurons plus directement l'occasion de collaborer.

Mais ce que je voudrais aussi dire c'est qu'Emmanuel pour moi c'est beaucoup plus que le mathématicien avec qui j'ai collaboré pendant de nombreuses années et dont j'ai apprécié les qualités scientifiques. C'est le souvenir d'une longue amitié, les balades en moto, c'est aussi les randonnées en montagne dans les Pyrénées françaises et espagnoles, la découverte des canyons de la Sierra de Guarra dans le Haut Aragon à un moment où ils étaient encore très peu connus et explorés, l'excitation des premières descentes dans l'eau souvent glacée sans savoir si cela allait passer ou non, l'exaltation quand on réussissait tout aussi forte que celle que nous procurait la preuve d'une conjecture. Je me souviens aussi de notre dernière course en montagne, une arête sur un sommet du cirque de Gavarnie, une course sans grande difficulté dont je pensais expédier l'escalade en trois heures en sautant la première partie moins intéressante et en grimpant à corde tendue les parties faciles. Nous n'étions donc pas partis aux aurores. Mais Emmanuel ne l'entendait pas de cette oreille. Pas question de sauter le début, pas question non plus de sacrifier à la sécurité. Exit l'escalade à corde tendue, place aux relais qu'il réorganisait régulièrement chaque fois qu'il me rejoignait jugeant mes choix peu orthodoxes. Quand nous sommes finalement arrivés au sommet, l'ombre du soir baignait les sommets du cirque et il n'y avait pas âme qui vive aussi loin que portât notre regard. La montagne était à nous seuls. Un moment d'intense bonheur que nous avons partagé en silence.

Références

[1] M. Artigue, E. Isambert, M. J. Perrin, A. Zalc, Some remarks on bicommutability, *Fundamenta Mathematicae* (1978) 345–364.

[2] M. Artigue, V. Gautheron, E. Isambert, Ensemble de bifurcation et topologie des variétés intégrales dans le problème du solide pesant, *Journal de Mécanique théorique et appliquée*, Vol. 5, No 3 (1986) 429–469.

Adresse de l'auteur :

U.F.R. de Mathématiques
Université Paris Diderot - Paris 7
175 rue du Chevaleret, 75013 Paris

Courriel : michele.artigue@univ-paris-diderot.fr

Les vicissitudes d'un raisonnement par analogie

Yvette Perrin

Le problème de la mesure des surfaces a parcouru presque toute l'histoire des Mathématiques, d'Archimède à nos jours. Un des moments capitaux de cette histoire se situe à la fin du 19ème siècle lorsque dans son *Cours de calcul différentiel et intégral* de 1879 [15], le mathématicien français J.A. Serret propose une définition de l'aire d'une surface gauche inacceptable. Ceci a incité un certain nombre de grands mathématiciens à se pencher sur ce problème et à proposer à leur tour leur propre définition.

C'est un raisonnement par analogie qui est la cause de l'erreur de Serret, analogie entre rectification d'une courbe et quadrature d'une surface. En 1880, simultanément mais indépendamment, H. Schwarz [14] et G. Peano [13] fournissent un contre-exemple montrant qu'avec la définition de Serret on peut attribuer à l'aire d'un cylindre droit, de révolution, toute valeur finie ou infinie supérieure à sa valeur exacte. G. Bachelard choisit ce contre-exemple dans *Essai sur la connaissance approchée* [1] pour illustrer sa thèse d'une limitation essentielle de la connaissance intuitive et le danger qu'il y a parfois à procéder par analogie. Or en revisitant l'histoire du concept d'aire [5], nous avons constaté que tous les mathématiciens rencontrés se sont appliqués à appuyer leur définition sur celle de la longueur d'une courbe mais en la réinterprétant, chacun d'une manière différente. En y regardant de près, le diagnostique de Bachelard paraît historiquement erroné.

C'est à Archimède (-287, -212) que nous devons les premières notions de longueur d'une courbe et d'aire d'une surface. On peut les résumer de la façon suivante :

1. La longueur d'un arc curviligne plan convexe est la valeur commune de la limite supérieure des longueurs des lignes polygonales inscrites, et de la limite inférieure des circonscrites.
2. L'aire d'une surface convexe est la valeur commune de la limite supérieure des aires des surfaces polyèdrales inscrites et de la limite inférieures des circonscrites.

Ces deux définitions, dont l'une découle de l'autre par analogie, ne s'appliquent pas aux courbes et aux surfaces gauches quelconques car, aussi bien pour les courbes que pour les surfaces gauches, la notion de courbes ou de surfaces circonscrites n'a pas de sens. Dans les exemples étudiés par Archimède, les limites des éléments inscrits et circonscrits sont égales. D'où la tentation de ne garder dans les définitions de longueur et d'aire que la première partie des définitions d'Archimède, à savoir celle concernant les éléments inscrits. Or si cette démarche est valable pour les courbes, elle ne l'est plus pour les surfaces. Nous allons voir comment ceci a induit en erreur ceux qui ont cherché à donner une définition géométrique générale de l'aire d'une surface quelconque.

Bien avant Serret, des mathématiciens se sont intéressés au problème de la mesure des surfaces, mais ils ont surtout fourni des procédés de calcul relevant de l'analyse, sans donner une définition purement géométrique de l'aire, c'est-à-dire qui ne fait pas intervenir d'éléments extérieurs à la surface : repères, systèmes de coordonnées, équations. Deux des plus célèbres sont J.L. Lagrange et A. Cauchy qui ont démontré les formules intégrales de la rectification des courbes et de la quadrature des surfaces, unanimement acceptées aujourd'hui. Tous les deux

ont utilisé l'analogie entre courbes et surfaces et il est intéressant, pour notre propos, de voir comment ils l'ont fait.

Prenons d'abord le cas de Lagrange. En 1813, dans *Théorie des fonctions analytiques* [8], il montre que l'aire d'une surface gauche régulière définie, dans un repère orthonormé par une équation $z = f(x, y)$ est donnée par la formule :

$$A = \int\int_D \sqrt{1 + p^2 + q^2}dxdy \quad \text{où} \quad p = \frac{\partial f}{\partial x} \quad \text{et} \quad q = \frac{\partial f}{\partial y} \tag{1}$$

Il commence par calculer la longueur d'une courbe plane convexe définie par une équation $y = f(x)$ dans un repère orthonormé. Il considère la longueur $F(x)$ d'un arc de cette courbe compris entre un point fixe donné et un point variable d'abscisse x. Il montre que la longueur de l'arc de courbe d'extrémités A, B d'abscisses respectives x et $x + h$ est comprise entre la plus petite et la plus grande longueur des deux portions de tangentes en A et B situées entre les deux parallèles à l'axe des ordonnées d'abscisses x et $x + h$. A l'aide de développements limités il en déduit que $F'(x) = \sqrt{1 + f'^2(x)}$.

Pour évaluer l'aire d'une surface, il procède de la même façon. La surface est définie dans un repère orthonormé par une équation $z = f(x, y)$; Il appelle $F(x, y)$ l'aire de la portion de surface comprise entre deux plans fixes d'équations $X = a$, $Y = b$ et les deux plans d'équations $X = x$, $Y = y$. Il considère la portion de surface comprise entre les quatre faces du prisme droit qui a pour base le rectangle $[x, x + h] \cdot [y, y + k]$ ainsi que les quatre plans tangents à la surface aux points situés sur les arrêtes du prisme, et il écrit : "On pourra prouver, par un raisonnement analogue à celui relatif aux tangentes, que la portion de surface qui forme la base supérieure du prisme sera comprise entre la plus grande et la plus petite section du prisme, faites par les quatre plans tangents de la surface courbe". Or il est facile de montrer que ce résultat est faux en général. A l'aide de développements limités Lagrange arrivera quand même à la conclusion :

$$\frac{\partial^2 F}{\partial x \partial y} = \sqrt{1 + p^2 + q^2} \quad \text{où} \quad p = \frac{\partial f}{\partial x} \quad \text{et} \quad q = \frac{\partial f}{\partial y}$$

mais à partir d'un résultat erroné dû à l'application d'une analogie grossière, voir Figure 1.

Dans son cours de 1826 *Application du Calcul intégral à la géométrie* [3], A. Cauchy redémontre la formule intégrale (1) en calculant, comme Lagrange, la dérivée partielle de F $\frac{\partial^2 F}{\partial x \partial y}$ mais par une autre méthode : en utilisant les infinitésimaux et une équivalence d'infinitésimaux qu'il ne démontre pas. Lui aussi s'appuie sur l'analogie entre longueur de courbe et aire de surface. Citons-le : "Nous avons admis qu'un très petit arc de courbe se confond sensiblement avec sa projection sur la tangente menée par l'un de ses points... Nous aurons recours, pour la quadrature des surfaces courbes, à un principe analogue ; et nous ferons servir à la mesure d'une petite portion de surface courbe, passant par un point donné, le plan qui se rapproche le plus de la surface dans le voisinage de ce point, en admettant qu'un élément de surface courbe dont les deux dimensions sont très petites se confond sensiblement avec sa projection sur le plan tangent mené par un de ses points."

Revenons à Serret. Dans son *Cours de calcul différentiel et intégral* de 1879 [15], il donne une définition purement géométrique de l'aire d'une surface et pour la justifier il fait appel à la longueur d'une courbe.

"On ne peut comparer à une ligne droite qu'une autre ligne droite ou une somme de telles lignes ; aussi nous avons dû définir avec précision, dans le Calcul différentiel la longueur rectiligne qu'on nomme longueur d'un arc de courbe. Nous emploierons ici des considérations analogues pour définir ce que nous entendons par aire de surface courbe...

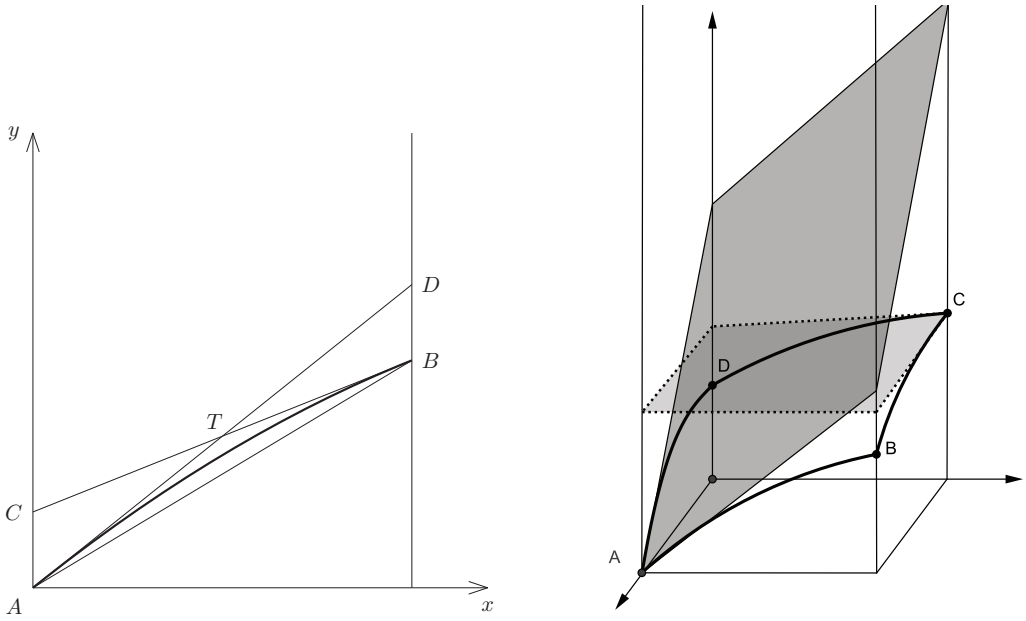


FIG. 1 – À gauche, la portion de courbe AB et les sections des deux tangentes aux points. À droite, la portion de surface $ABCD$ et deux sections planaires aux points A et C.

Soit une portion de surface courbe terminée par un contour C ; nous nommerons aire de cette surface la limite S vers laquelle tend l'aire d'une surface polyédrique inscrite formée de faces triangulaires et terminée par un contour polygonal G ayant pour limite le contour C.

Il faut démontrer que la limite S existe et qu'elle est indépendante de la loi suivant laquelle décroissent les faces de la surface polyédrale inscrite."

Serret croit le démontrer, mais sa démonstration est fausse comme le montre le contre-exemple de Schwarz-Peano.

C. Hermite est le premier à diffuser ce contre-exemple dans son cours de 1882 [6] ; il y propose une définition de l'aire d'une surface qui rompt avec la stratégie de Serret et reprend celle de Cauchy. Il approche la surface cible par des surfaces polyédrales discontinues tangentes, réinterprétant l'analogie entre courbes et surfaces de la façon suivante :

"Nous suivons une autre analogie à laquelle conduit la remarque [...] qu'on peut substituer aux côtés [d'un polygone inscrit dans une courbe C] la série des segments non contigus [...], ces segments étant des portions comprises entre les ordonnés [d'un morceau de la projection de C sur l'axe des x] d'une tangente en un point quelconque [de l'arc de C considéré]."

En 1890, Peano explique dans sa note *Sulla definizione dell'area d'una superficie* [13] pourquoi la définition de Serret ne marche pas et il présente à son tour sa définition de l'aire. Les surfaces polyédrales approchantes considérées par Serret tendent en dimension vers la surface cible mais pas en direction, i.e. les plans des faces des polyèdres ne tendent pas vers les plans tangents à la surface. Si pour les courbes les lignes polygonales inscrites qui tendent en dimension vers la courbe tendent également en direction, il n'en est pas de même pour les surfaces polyédrales inscrites. Cela explique que l'analogie entre mesure des courbes et mesure des surfaces proposée par Serret ne marche pas.

Cependant Peano reste attaché à cette analogie. Tout imprégné du calcul géométrique de Grassmann qu'il a traduit et clarifié en 1888 dans son *Calcolo geometrico...* [12], il trouve dans les formes grassmanniennes de degrés un et deux les outils lui permettant de respecter cette analogie. "La rigueur et l'analogie entre les définitions reliant arc et surface peuvent être toutes les deux préservées en faisant usage non seulement du concept de segment linéaire considéré comme ayant une longueur et une direction (vecteur), mais également du concept dual de région planaire considérée comme ayant une taille et une orientation".

Peano réinterprète les lignes polygonales inscrites dans une courbe comme une succession de vecteurs caractérisés non seulement par leur longueur, mais aussi par leur direction et leur sens. Inscrire une ligne polygonale dans une courbe, c'est diviser la courbes en morceaux contigus, c'est-à-dire en arcs, et associer à chaque arc de la subdivision un vecteur ayant même origine et même extrémité. Quand la maille de la ligne polygonale inscrite tend vers 0 les vecteurs correspondants tendent en direction vers les tangentes à la courbe. Pour les surfaces Peano fait jouer aux formes grassmanniennes de degré deux, interprétées en terme de bivecteurs (ou produits de deux vecteurs) le même rôle que les vecteurs pour les courbes.

Le bivecteur de deux vecteurs OA, OB non colinéaires est caractérisé par sa direction – celle du plan défini par les deux vecteurs –, par sa grandeur – l'aire du triangle OAB –, et par son sens – le bivecteur de OB, OA est de sens opposé au bivecteur de OA, OB . Peano définit le bivecteur d'une ligne triangulaire OAB : c'est le bivecteur des vecteurs OA, OB. Puis par additivité, il définit le bivecteur d'une ligne polygonale fermée plane ou gauche et enfin il étend cette notion à une ligne courbe fermée régulière : c'est la limite des bivecteurs des lignes polygonales inscrites, lorsque la maille de ces polygones tend vers zéro. (Si la courbe est plane et convexe son bivecteur a pour direction celle du plan de la courbe et sa grandeur l'aire plane intérieure à la courbe. Si la courbe est une courbe gauche quelconque il est plus difficile de donner une interprétation géométrique simple de son bivecteur si ce n'est que les projections de la courbe et de son bivecteur sur un plan quelconque et suivant une direction quelconque ont même aire).

Peano insiste encore sur l'analogie entre vecteur et bivecteur. Il écrit : "Entre le vecteur d'un arc de courbe et le bivecteur d'une portion de surface, l'analogie est complète. Ainsi, sous certaines conditions, à la proposition :

– La direction du vecteur d'un arc infinitésimal de courbe est celle de la tangente ; et le rapport entre sa grandeur et la longueur de l'arc est l'unité.

correspond, sous des conditions analogues, la proposition :

– La direction du bivecteur d'une portion infinitésimale de surface est celle du plan tangent ; et le rapport entre sa grandeur et l'aire de cette portion est l'unité." (Peano [13] : 56-57).

Pour calculer l'aire d'une surface, Peano la divise en morceaux limités par des courbes fermées. À chaque morceau il associe le bivecteur de sa frontière qu'il interprète comme un triangle orienté défini par sa direction, sa grandeur et son sens. La réunion de ces triangles constitue une surface polyédrale non continue dont l'aire est égale à la somme des aires de ces triangles. Appelons approximation polyédrale peanienne une telle surface. Peano définit alors l'aire de la surface cible comme la limite supérieure de toutes ses approximations peaniennes.

Il est facile de faire le parallèle avec le calcul de la longueur d'une courbe. Ainsi Peano donne une définition purement géométrique de l'aire d'une surface qui respecte l'analogie avec celle de la longueur d'une courbe et qui corrige l'erreur de Serret. Mais pour cela, il a dû abandonner l'antique concept de figure pour le remplacer par la notion de formes grassmanniennes plus apte à supporter les analogies naturelles que les anciens géomètres, Archimède notamment, avaient forgées. Aux notions familières de segments ou de triangles inscrits ou circonscrits, il substitue celles de vecteurs et de bivecteurs.

Ce point de vue avant-gardiste, unanimement adopté aujourd'hui ne l'a pas été immédiatement. Bien après Peano, certains mathématiciens resteront attachés à l'idée que pour mesurer une surface il faut l'approcher par des polyèdres. C'est le cas de H. Lebesgue, E. Cartan, M. Fréchet pour n'en citer que quelques uns. Chez tous, la préoccupation première est le respect de l'analogie entre rectification et quadrature. Citons d'abord E. Cartan et M. Fréchet pour revenir ensuite, plus en détail, à H. Lebesgue.

En 1907, dans une note aux C.R.A.S., E. Cartan écrit [2] :

"Dans les définitions habituellement données de l'aire d'une portion de surface courbe continue (et admettant un plan tangent variant d'une manière continue), on fait intervenir des sommes de parallélogrammes situés dans des plans tangents à la surface ; Il semblerait plus naturel de considérer, par analogie avec ce que l'on fait dans la définition de la longueur d'un arc de courbe, des sommes d'aires de triangles inscrits dans la surface. Il est, en effet, possible de procéder ainsi, mais à condition de prendre une précaution qui me paraît devoir être signalée."

Et M. Fréchet, en 1939, commence sa note *Sur une définition intrinsèque de l'aire d'une surface courbe comme limite d'aires polyédrales inscrites* [4] par cette phrase :

"Nous développerons des considérations sur la longueur d'une courbe qui n'ont rien d'essentiellement nouveau, mais qui pourtant sont négligées même dans des ouvrages rédigés par certains des plus éminents spécialistes des questions de longueurs et d'aires et sans lesquelles certains raisonnements qui y figurent prêtent à objection. Outre leur intérêt propre, ces considérations auront l'avantage de préparer l'examen du cas des surfaces."

Cartan comme Fréchet approchent la surface à mesurer par des polyèdres inscrits à faces triangulaires et imposent à ces faces des conditions du type : les angles restent supérieurs à un nombre fixé strictement positif.

Terminons ce parcours par l'un des plus célèbres acteurs de l'élaboration des notions de mesure et d'intégration, H. Lebesgue. Dans sa thèse de 1902 : *Intégrales, longueurs, aires* [9] Lebesgue commence par poser le problème de la mesure des courbes.

"Attacher à chaque courbe un nombre fini ou infini que l'on appellera sa longueur et satisfaisant aux conditions suivantes :

1. Il existe des courbes planes ayant des longueurs finies.
2. Deux courbes égales ont même longueur.
3. Une courbe somme de plusieurs autres a pour longueur la somme des longueurs des courbes composantes.
4. La longueur d'une courbe C est la plus petite limite des longueurs des courbes polygonales dont C est la limite."

L'idée originale de Lebesgue ici est de considérer toutes les lignes polygonales qui tendent vers la courbe à mesurer et pas seulement les lignes inscrites.

Il pose ensuite le problème de la mesure des surfaces et le formule exactement de la même façon, remplaçant simplement les mots "courbe", "longueur", "polygonale" par "surface", "aire", "polyédrale". Il définit ainsi l'aire d'une surface S comme la plus petite limite des aires des surfaces polyédrales dont S est la limite.

Après avoir donné cette définition, Lebesgue propose un procédé de calcul effectif de l'aire d'une surface S. Il considère une suite de subdivisions $D(i)$ de la surface par des courbes quarrables, les diamètres de ces courbes tendant vers zéro avec $1/i$. A chaque subdivision $D(i)$, il associe la somme $S(i)$ des aires minimales des courbes de cette subdivision et montre que l'aire de la surface est la limite des $S(i)$ quand i tend vers l'infini ; et il conclut :

"Remarquons que cette définition de l'aire d'une surface est analogue à la définition de la longueur d'une courbe comme limite des périmètres des polygones inscrits. Un polygone inscrit

définit en effet une division de la courbe à laquelle nous faisons correspondre une division de la surface à l'aide de courbes quarrables. À la longueur d'un côté ab d'un polygone, c'est-à-dire à la limite inférieure des longueurs des courbes qui joignent les deux points de division consécutifs a, b, nous faisons correspondre la limite inférieure des aires des surfaces limitées par C l'un des contours quarrables qui intervient dans la division de la surface. L'analogie se poursuit encore plus loin car il est possible de démontrer qu'étant donnée une courbe fermée C, il existe une surface limitée à C ayant pour aire l'aire minima de C. Ces surfaces correspondent aux côtés des polygones inscrits."

Toutefois, Camille Jordan, dans une lettre dont Lebesgue cite des passages [10] (Lebesgue, 1925 :163–164), remarque que cette analogie entre rectification et quadrature n'est pas parfaite, et qu'elle aurait dû être poussée plus loin :

"Je ne suis pas satisfait par ce que vous avez dit de l'aire des surfaces", m'avait déclaré Jordan lorsque je lui portais ma Thèse. Et, sur ma demande, il me fit des objections, que d'ailleurs je m'étais faites moi-même (voir par exemple le §70 de ma Thèse), et que je peux résumer ainsi : "Vous dites que vous édifiez, pour la mesure des aires des surfaces, une théorie entièrement analogue à celle de la mesure des longueurs des courbes mais, pourtant, vous laissez sans réponses des problèmes essentiels alors que ces problèmes sont résolus dans le cas des courbes. Une courbe $x = f(t)$, $y = g(t)$, $z = h(t)$ étant donnée, nous savons quelle suite d'opérations il nous faut effectuer sur f, g, h pour reconnaître si la courbe est rectifiable et pour calculer sa longueur finie ou infinie ; vous ne dites rien du problème analogue pour les surfaces données par trois équations $x = f(u, v)$, $y = g(u, v)$, $z = h(u, v)$. De là résulte aussi que tandis que l'on sait construire les formes les plus générales des fonctions $f(t)$, $g(t)$, $h(t)$ relatives aux courbes rectifiables, vous ne nous apprenez pas à former les fonctions $f(u, v)$, $g(, v)$, $h(u, v)$ donnant des surfaces quarrables."

Il faudra attendre les travaux du mathématicien italien Leonida Tonelli en 1926 pour avoir les réponses à ces questions. Tonelli [16] donne une nouvelle définition des fonctions de deux variables à variation bornée, et des fonctions absolument continues. Il montre qu'une surface S définie par $z = f(x, y)$ sur le carré $D = [0, 1] \times [0, 1]$ est quarrable si et seulement si f est à variation bornée. S'il en est ainsi f admet presque partout des dérivées partielles $p = \frac{\partial f}{\partial x}$, $q = \frac{\partial f}{\partial y}$, et l'aire de S est égale à l'intégrale (1) si et seulement si f est absolument continue.

Dans son article *Comment mesurer les surfaces ?* de 2006, Yves Meyer [11] fait remarquer que l'analogie entre rectification et quadrature explique pourquoi il a fallu autant de temps pour trouver une solution aux problèmes posés par Jordan. En effet, les fonctions d'une variable, à variation bornée, sont caractérisées par le fait qu'elles sont la différence de deux fonctions croissantes. Or, cette propriété, que l'on cherchait à généraliser aux fonctions de deux variables, ne permet pas de caractériser les surfaces quarrables. C'est précisément parce que Tonelli définit les fonctions à variation bornée sans faire référence à cette caractérisation qu'il parvient à résoudre la question laissée en suspens par Lebesgue.

Ce brève parcours dans l'histoire de la définition de l'aire d'une surface nous conduit à nuancer les propos de G. Bachelard évoqués plus haut :

" La clarté de l'intuition ne s'étend pas au-delà de son domaine d'origine. C'est là seul, à son propre centre, qu'elle est un guide certain. Plus loin, elle s'estompe dans la pénombre des analogies ; elle peut même devenir un obstacle à la connaissance précise. Une connaissance intuitive est tenace, mais elle est fixe. Elle entrave finalement la liberté de l'esprit."

Selon Bachelard [1], l'histoire des définitions de l'aire manifesterait de façon exemplaire les dangers de l'analogie en science, et la mésaventure de Serret constituerait une mise en garde spectaculaire à l'encontre de ce genre de raisonnement. Ce diagnostic n'est certes pas faux, mais il se fonde sur une partie seulement de l'histoire que nous venons de raconter - la séquence certainement la mieux connue, celle qui court de la définition de Serret au contre-exemple de Schwarz.

De la suite des événements, une leçon un peu différente nous paraît devoir être dégagée. Ce qui est frappant, en effet, c'est que, loin d'abandonner le recours à l'analogie, tous les mathématiciens venant après Schwarz [14] ont cherché à réparer le raisonnement de Serret en se basant sur l'analogie entre rectification et quadrature. Le problème auquel ils ont été confrontés est que cette analogie peut être construite de plus d'une manière, et qu'en conséquence le concept d'aire pouvait être caractérisé de plus d'une façon. Si nous avons quelques réserves sur les conclusions épistémologiques que Bachelard tire de cette histoire, nous partageons le sentiment que cette histoire possède une certaine forme d'exemplarité. En réalité, chaque mathématicien revisite la définition de la longueur d'une courbe de façon à l'ajuster à sa définition de l'aire. L'analogie dimensionnelle entre rectification et quadrature est extrêmement enracinée dans notre esprit : nous ne pouvons que difficilement nous déprendre de l'idée que ce qui doit valoir pour la longueur d'une courbe doit valoir pour l'aire d'une surface. Cette analogie n'est cependant pas suffisamment forte pour nous permettre de déterminer à elle seule un concept univoque d'aire.

Références

[1] Bachelard, G. (1927). Essai sur la connaissance approchée, Paris, Vrin.

[2] Cartan, E. (1907). Sur la définition de l'aire d'une portion de surface courbe, Comptes Rendus de l'Académie des Sciences.

[3] Cauchy, A. (1826). Applications du calcul infinitésimal à la géométrie, 0euvres Complètes, seconde série, tome V, Gauthier-Villars, 1903.

[4] Fréchet, M. (1939). Sur une définition intrinsèque de l'aire d'une surface courbe comme limite d'aires polyédrales inscrites, Annali delle Scuola Normale Superiore di Pisa, 8 ; 3 : 285–300.

[5] Gandon, S. et Perrin, Y. (2009). Le problème de la définition de l'aire d'une surface gauche : Peano et Lebesgue, Archives for History of Exact Sciences 63 : 665–704.

[6] Hermite, C. (1882). Cours de M. Hermite rédigé en 1882 par M. Andoyer, Hermann, Paris.

[7] Jordan, C. (1882). Cours d'analyse, Gauthier-Villars, Paris.

[8] Lagrange, J.-L. (1813). Théorie des fonctions analytiques, in Oeuvres de Lagrange, publié par Serret, tome 9, Gauthier-Villars, 1881.

[9] Lebesgue, H. (1902). Intégrale, Longueur, Aire, In Oeuvres, vol. I : 102–231, L'enseignement mathématique.

[10] Lebesgue, H. (1925). Quelques remarques sur la définition de l'aire d'une surface (extrait d'une lettre à M. W. Sierpinski). In Oeuvre, IV, 163–164.

[11] Meyer, Y. (2006). Comment mesurer les surfaces ?, Gazette de la Société Mathématique de France, 109 : 23–36.

[12] Peano, G. (1888). Calcolo geometrico secondo l'Ausdehnungslehre di Hermann Grassmann, p receduto dalle operazioni della logica deduttiva. Bocca, Turin.

[13] Peano, G. (1890). Sulla definizione dell'area d'una superficie, Atti della Reale Accademia dei Lincei : Rendiconti, 4 : 54–57.

[14] Schwarz, H. (1890). Sur une définition erronée de l'aire d'une surface gauche, Gesammelte mathematische Abhandlungen : 309–311.

[15] Serret, J.-A. (1879). Cours de calcul différentiel et intégral, Gauthier-Villars, Paris.

[16] Tonelli, L. (1926). Sur la quadrature des surfaces, Comptes Rendus de l'Académie des Sciences, 10 mai 1926.

Adresse de l'auteur :

31 avenue Phelut
63130 Royat

Courriel : `yvette.perrin1@orange.fr`

Le concept métaphysique de relation dynamique

Robert Lutz

Depuis une dizaine d'années j'ai été amené à sortir du cadre des mathématiques pour m'intéresser avec mon ami Jean-François Froger à un point de vue nouveau concernant la métaphysique en tant que support d'un discours rationnel à propos du monde physique, biologique, économique, social ou autre. La métaphysique classique est la branche de la philosophie qui traite de la substance, c'est-à-dire de l'existence dans son sens le plus large et le plus général. Elle s'avère insuffisante comme cadre conceptuel de la pensée scientifique, bien que les philosophes aient cherché pendant plus de deux millénaires à lui faire jouer ce rôle. La difficulté tient à la trop grande pauvreté du concept d'existence lorsqu'il est limité à l'alternative exister/ne pas exister, faute de négation plus élaborée. Pour aller plus loin, il s'est avéré fécond d'introduire un concept métaphysique plus subtil, celui de relation dynamique. Je me propose ici de présenter ce concept et d'en illustrer quelques utilisations tirées de nos récents ouvrages.

1 Définition axiomatique du concept

Le concept de relation dynamique à n termes est régi par les propriétés axiomatiques suivantes :

i. Une *relation dynamique à n termes* est un acte qui produit *n termes* à partir de n substrats distincts ou non que cet acte modifie de manière permanente.

ii. Chaque terme actualise la relation, c'est-à-dire contribue à chaque instant à son état interne.

iii. Tout ce qui existe est terme d'une relation.

iv. Les termes d'une relation peuvent servir de substrats pour d'autres relations.

v. Aucun terme n'est égal au substrat correspondant.

Les concepts métaphysiques d'acte, de substrat, de produire, de terme sont considérés comme premiers. C'est pourquoi il n'y a pas lieu d'en donner une définition. Il faudra se contenter d'en comprendre le sens d'après les exemples que nous allons donner au paragraphe 2.

Le nombre de termes d'une relation vaut au moins 1. Pour $n = 1$ on parlera de relation unaire, pour $n = 2$ de relation binaire. Le nombre de termes n'est pas limité.

Nous appellerons *esse ad produit par la relation en un terme* ce qui, dans ce terme, résulte de la relation. Et nous appellerons *esse in induit sur la relation par l'ensemble des termes* la manière dont la relation est actualisée par les termes. Il y a donc n *esse ad* et un seul *esse in*. Ces vocables latins d'origine scolastique signifient littéralement "être vers" et "être dans", ce qui correspond bien à la façon dont la relation se comporte vers les termes et à la façon dont les termes se comportent dans la relation.

Ainsi une relation dynamique est caractérisée par ses *esse ad* et son *esse in* qui constituent un processus de type action-réaction. En voici quelques exemples.

2 Exemples de relations dynamiques

- Considérons une personne désignée par une élection pour présider une université pendant cinq ans. Cette personne est le substrat d'une relation unaire dont le terme est le président de cette université. En tant que tel, elle est munie, par un acte constamment renouvelé, de pouvoirs qui constituent l'*esse ad* de la relation vers le terme. La manière dont elle les exerce au jour le jour constitue l'*esse in* du terme dans la relation.
- Une personne désignée à la suite d'un concours pour enseigner pendant quarante ans dans une école est le substrat d'une relation unaire dont le terme est un enseignant de cette école. Celui-ci est chargé par un acte administratif constamment renouvelé de la mission d'enseigner, dont les modalités constituent l'*esse ad* de la relation vers le terme. La manière dont cet enseignant assume cette mission constitue l'*esse in* du terme dans la relation.

Dans ces deux exemples la relation est bien un acte qui détermine une dynamique. Il est facile de donner d'autres exemples sur le même modèle : une personne servant de substrat à un acte qui lui confie une responsabilité à exercer pendant une certaine durée. Mais on peut aussi penser à un objet matériel qui sert d'instrument pour réaliser durablement un objectif : le substrat est cet instrument et l'acte relationnel est le fait de s'en servir pour réaliser cet objectif.

Les relations dynamiques unaires paraissent quelque peu insolites mais le concept devient plus « parlant » dans le cas binaire. Considérons par exemple l'amitié née entre deux personnes à la suite d'une rencontre fortuite. Une fois qu'elle est installée durablement, il s'agit d'une relation dynamique binaire dont les deux personnes sont les substrats. En effet, cette amitié est un acte constamment renouvelé qui enrichit les deux personnes en les maintenant amis, ce qui se traduit par deux *esse ad*. L'*esse in* est ici l'ensemble des contributions combinées des deux partenaires à la vie de leur amitié.

Un autre exemple typique est la relation pédagogique dont les termes sont un maître et son élève. Elle est instituée lorsque le maître est en train de transmettre un savoir à un élève. En l'absence de cet acte, il n'y a ni maître ni élève, mais seulement deux personnes qui sont des substrats potentiels pour une éventuelle relation pédagogique.

Cet exemple se généralise aisément en une relation n-aire dont les termes sont un maître et $n-1$ élèves. L'*esse in* est alors la manière dont l'ensemble des élèves accueille globalement le savoir transmis par le maître.

On constate sur ces exemples que le concept de relation dynamique n'a pas grand chose à voir avec la notion mathématique de relation donnée par une partie d'un produit de n ensembles. Ici c'est le caractère dynamique qui joue un rôle central, et non la liste de n-uples reliés par un trait commun.

3 Structures relationnelles dynamiques

Les relations dynamiques sont les briques de base sur lesquelles repose un discours scientifique cohérent. Elles servent à construire des *structures relationnelles*. Une structure relationnelle d'ordre p est la donnée de p relations dynamiques dont les nombres de termes respectifs peuvent être distincts. Nous dirons qu'une telle structure est *close* si tous les substrats sont confondus en un seul.

Considérons par exemple une structure relationnelle close d'ordre deux constituée de deux relations dynamiques unaires A et B de même substrat. Il est facile de construire des exemples où les deux termes n'ont rien à voir l'un avec l'autre ; par exemple un même homme peut être enseignant et officier de réserve. Mais si l'on exige que le terme de A soit B et que celui de B soit

A on tombe sur une impossibilité, comme on le voit aisément. Une telle structure auto-suffisante d'ordre deux n'existe pas ! Par contre nous allons fixer notre attention sur des structures spéciales d'ordre quatre, les quaternités, qui apparaissent comme des structures fondamentales du monde réel. Leurs termes sont cette fois les quatre relations elles-mêmes.

4 Structures de quaternité

Nous appellerons quaternité une structure relationnelle close d'ordre quatre constituée de deux relations dynamiques unaires A et D et de deux relations binaires B et C telles que :
- B soit le terme de A,
- A et C soient les termes de B,
- B et D soient les termes de C,
- C soit le terme de D.

Les décomptes des nombres de termes impliquent les trois « interdits » suivants :
- A n'est pas un terme de D et D n'est pas un terme de A,
- A n'est pas un terme de C et C n'est pas un terme de A,
- B n'est pas un terme de D et D n'est pas un terme de B.

Ils impliquent également les trois « permis » suivants :
- A est terme de B et B est terme de A,
- B est terme de C et C est terme de B,
- C est terme de D et D est terme de C.

On peut représenter géométriquement une quaternité par un tétraèdre de sommets $ABCD$ dont les six arêtes sont dites permises ou interdites selon la règle suivante :

AB, BC et CD sont permises alors que AC, BD et AD sont interdites.

Afin de montrer qu'il existe des quaternités et qu'elles sont présentes dans divers domaines de la connaissance, je propose un exemple qui concerne la science économique et qui ne se trouve pas dans les ouvrages cités en références.

5 La quaternité de l'économie

La pensée économique repose traditionnellement sur l'observation du comportement de deux catégories d'êtres humains : les producteurs et les consommateurs. Ces catégories sont mal définies, car chaque homme est à la fois producteur de quelque chose et consom-mateur de quelque chose, ne serais-ce qu'en raison de son fonctionnement biologique. Par contre on peut distinguer conceptuellement les actes de l'homme soit en tant que producteur, soit en tant que consommateur. Il s'agit d'une distinction typiquement binaire. Mais une nouvelle difficulté surgit : pour définir le concept de producteur, on utilise celui de consommateur et vice-versa. En effet l'un se définit par rapport à l'autre, car un produit n'est tel que si quelqu'un peut le consommer et un consommateur n'est tel que si quelqu'un peut lui fournir un produit. Il manque donc des termes pour obtenir une véritable différenciation !

Or produire et consommer sont des actes humains inséparables de deux autres actes : recevoir des données naturelles et désirer des choses matérielles ou immatérielles. Ces quatre actes différencient la capacité d'action de l'*homo economicus*. Nous allons voir qu'ils sont organisés en quaternité.

Notons A l'acte de recevoir des données naturelles, B celui de produire, C celui de consommer et D celui de désirer. Ces actes transforment tous les quatre les capacités physiques et intellectuelles des hommes, à la manière de quatre relations de même substrat S. Les résultats de ces transformations sont encore A, B, C, D qui apparaissent ainsi comme les termes des quatre relations. Plus précisément, l'acte A transforme S en l'acte B de produire des choses par le travail (par exemple un outil à partir d'un silex) ; ainsi A est une relation unaire de terme B. L'acte B de produire transforme S en l'acte A (par exemple lors de l'extraction de matières premières par le travail), mais aussi en l'acte C de consommer. Ainsi B est une relation binaire de termes A et C. L'acte C de consommer transforme S en l'acte B de produire et aussi en l'acte D de désirer. Ainsi C est une relation binaire de termes B et D. Enfin l'acte D de désirer transforme S en l'acte C de consommer. C'est donc une relation unaire de terme D.

Les six *esse ad* qui déterminent la dynamique se la structure s'expriment de la manière suivante :

- de A vers B, il s'agit de l'appropriation des données,
- de B vers A, il s'agit de l'extraction de matières premières,
- de B vers C, il s'agit de l'offre,
- de C vers B, il s'agit de la demande,
- de C vers D, il s'agit de la stimulation,
- de D vers C, il s'agit de la satisfaction.

Cette description met en évidence, en dehors de toute valorisation par le prix, donc de marché, les concepts d'offre et de demande qui fondent le discours économique classique. Les autres aspects restent implicites alors qu'ici ils font partie de la structure. Le travail est un aspect de l'acte de production ; il consiste à transformer les données naturelles en choses désirables par l'homme.

D'autres quaternités interviennent en économie, par exemple celle du travail et celle de l'entreprise. Elles offrent des outils qui peuvent servir à structurer ce domaine qui reste à développer.

Références

[1] J.-F. Froger et R. Lutz, *Structure de la connaissance*, DésIris, 2003.

[2] J.-F. Froger et R. Lutz, *Fondements logiques de la physique*, DésIris, 2007.

[3] J.-F. Froger et R. Lutz, *La structure cachée du réel-The hidden structure within reality*, Désiris-Ara, 2009.

Adresse de l'auteur :

Laboratoire de Mathématiques, Informatique et Applications, EA3993
Faculté des Sciences et Techniques, Université de Haute Alsace
4, rue des Frères Lumière, F-68093 Mulhouse cedex, France

Courriel : `Robert.Lutz@uha.fr`

Les nombres entiers naturels dans la théorie constructive des types

Guy Wallet

Résumé : L'objet premier de ce texte est de montrer comment sont présentés les nombres entiers dans la théorie intuitionniste des types de Per. Martin-Löf. En conséquence, ce travail est aussi un exposé d'introduction à la théorie constructive des types. Le but plus lointain de l'auteur est de préparer l'étude des travaux de Martin-Löf sur une extension non standard du type entier naturel [3].

Mots-clés : nombres entiers naturels, théorie constructive des types, mathématiques constructives.

La *Théorie constructive des types* (notée TCT dans la suite) ou *Théorie intuitionniste des types* est une construction théorique introduite au tournant des années 70-80 par Per Martin-Löf [1, 2], logicien-mathématicien suédois. Cette théorie se situe au carrefour de trois disciplines : les mathématiques constructives, la logique et la science de la programmation. De fait, la TCT intègre des aspects novateurs forts issus de chacun de ces secteurs. De plus, la TCT elle-même peut être interprétée au choix comme une formalisation des mathématiques constructives [1], comme une branche de la logique ou comme un langage de programmation.

La TCT présente quelques variantes subtiles dont il est difficile pour le néophyte de saisir tout le sens lors d'une première approche. Dans le but d'éviter des comparaisons délicates entre ces variantes, le texte qui suit est basé uniquement sur la présentation qui en est donnée par Martin-Löf dans son traité [2] *Intuitionnistic Type Theory* (Bibliopolis, 1984). Cette approche de la TCT présente pour le mathématicien l'avantage de ne posséder essentiellement que deux espèces d'objets, les ensembles et les éléments de ces ensembles, ce qui rappelle le cadre familier des mathématiques classiques. Le lecteur intéressé pourra se reporter aux autres références [4, 5, 6] pour découvrir d'autres versions de cette théorie.

1 Propositions, jugements et ensembles dans la TCT

Dans les présentations usuelles et exhaustives de la TCT, les nombres entiers ne sont présentés que vers la fin de l'exposé, après l'introduction de toutes les structures générales (les $\prod$-ensembles, les $\sum$-ensembles et les ensembles dérivés, les ensembles finis) et juste avant les types inductifs généraux et les univers. Cependant, puisqu'il ne s'appuie sur aucun autre ensemble, l'ensemble des nombres entiers peut être introduit en premier. Néanmoins, il faut commencer par préciser quelques aspects généraux de la TCT.

1.1 Différence entre jugements et propositions

Une distinction essentielle à la base de la TCT est celle entre proposition et jugement. Une jugement est un morceau de connaissance attesté par une preuve, alors qu'une proposition est la

[1] Signalons sans en dire plus au niveau de ce texte, que la TCT est une théorie prédicative, c'est-à-dire qu'il y est impossible de définir un objet de manière non prédicative.

formulation d'une propriété hypothétique ou d'un problème. Typiquement, si $\mathcal{P}$ est une proposition, alors l'affirmation "$\mathcal{P}$ est vraie" soutenue par une preuve est un jugement. Non clairement explicitée dans les mathématiques classiques, cette distinction va être soigneusement représentée et utilisée dans la TCT.

1.2 Les jugements de la TCT

Les jugements de la TCT concernent les ensembles et leurs éléments. Les formes de jugement découlent de la conception des ensembles portée par cette théorie, conception orientée par le fait que la sémantique des objets est purement calculatoire.

Relativement à la théorie des ensembles classiques, le premier point nouveau est que, par convention, un ensemble possède des éléments qualifiés de canoniques. Ce sont des éléments d'une forme particulière qui manifeste immédiatement l'appartenance à l'ensemble considéré. Cette forme canonique est assez générale pour caractériser l'ensemble lui-même. A partir de là, on peut décrire la conception des ensembles de la TCT par les explications suivantes :

1. *Un ensemble A est défini en indiquant comment sont formés ses éléments canoniques et comment est définie l'égalité des éléments canoniques.*
2. *Deux ensembles A et B sont égaux s'ils ont les mêmes éléments canoniques et la même égalité des éléments canonique.*
3. *Un élément d'un ensemble A est une méthode, un algorithme qui, une fois exécuté, fournit un élément canonique de A.*
4. *Deux éléments a et b d'un ensemble A sont égaux si, lorsque l'on exécute a et b, on obtient deux éléments canoniques de A qui sont égaux.*

Chacune de ces explications donne le sens des quatre seules formes de jugements définies dans la TCT :

1. *A est un ensemble* (noté formellement A Ens.).
2. *Les ensembles A et B sont égaux* (noté formellement $A = B$).
3. *a est un élément de l'ensemble A* (noté formellement $a \in A$).
4. *a et b sont deux éléments égaux de l'ensemble A* (noté formellement $a = b \in A$).

L'un des buts de la TCT est de formuler des règles d'inférence formelle entre jugements. Ces règles sont présentées dans le style de la déduction naturelle. En voici un exemple :

$$\frac{A \text{ Ens.} \quad B \text{ Ens.} \quad A = B \quad a = b \in A}{a = b \in B}$$

Au-dessus du trait horizontal la prémisse est constituée d'un ou plusieurs jugements ; sous le trait horizontal se trouve la conclusion sous la forme d'un jugement validé par la règle formelle.

Remarquons par ailleurs que la notion d'ensemble est une notion ouverte : contrairement à ce qui se passe dans les mathématiques classiques, on ne prétend pas dans la TCT avoir circonscrit à l'avance le mode de fabrication des ensembles dont nous aurons besoin un jour. Mais une fois qu'elle est correctement introduite, on sait reconnaître qu'une entité donnée est un ensemble.

1.3 Jugements dépendants et famille d'ensembles

Un autre point fort de la TCT est que la notion de fonction est primitive. Cela correspond aux jugements dépendants (jugements sous condition ou dans un contexte) dont l'exemple le plus simple est celui d'un famille d'ensembles

$$B(x) \text{ Ens. } (x \in A)$$

dont le sens est que l'on dispose d'un ensemble A et que, de tout jugement $a \in A$ nous pouvons inférer que $B(a)$ est lui-même un ensemble.

Dans le cadre de cette brève introduction à la TCT, nous ne développerons pas plus cet aspect par ailleurs important de cette théorie.

1.4 Les propositions

Il découle de la conception des jugements que cela n'a aucun sens d'en nier un. Ce à quoi on peut appliquer l'opérateur de négation $\neg$ (et plus généralement les constantes logiques $\&$, $\vee$, $\supset$ et les deux quantificateurs $\forall$ et $\exists$) sont les propositions. Rompant avec la conception classique des propositions comme valeur de vérité qui pose des problèmes en cas de quantification sur des ensembles infinis, les intuitionnistes mettent en avant la notion de preuve (et de preuve canonique) en affirmant que :

1. *Une proposition est définie en fixant ce qui peut être vu comme une preuve de cette proposition.*
2. *Une proposition est vraie si elle a une preuve, c'est-à-dire si une preuve peut en être donnée.*
3. *On peut considérer la proposition $\perp$ qui, par définition, n'admet pas de preuve.*
4. *Une preuve canonique de $A\&B$ est la donnée de (a, b) où a est une preuve de A et b est une preuve de B.*
5. *Une preuve canonique de $A \vee B$ est de la forme $i(a)$ ou de la forme $j(b)$ où a est une preuve de A et où b est une preuve de B.*
6. *Une preuve canonique de $A \supset B$ est de la forme $(\lambda x)b(x)$ où $b(a)$ est une preuve de B sachant que a est une preuve de A.*
7. *Une preuve canonique de $(\forall x)B(x)$ est de la forme $(\lambda x)b(x)$ où $b(a)$ est une preuve de $B(a)$ pourvu que a soit un individu.*
8. *Une preuve canonique de $(\exists x)B(x)$ est de la forme (a, b) où a est un individu et b est une preuve de $B(a)$.*
9. *Une preuve arbitraire d'une proposition A est une méthode dont la mise en œuvre fournit une preuve canonique de A.*

Il apparaît ainsi une forte analogie entre les propositions et les ensembles. *L'isomorphisme de Curry-Howard* précise cette analogie en faisant correspondre à toute proposition l'ensemble de ses preuves. Le miracle de Curry-Howard est qu'aux modes de constructions des propositions correspondent les modes de constructions des ensembles. Il ne reste plus qu'à poser que les propositions et les ensembles sont des notions que l'on peut confondre : c'est l'interprétation *des propositions comme ensembles* sur laquelle la TCT est construite. Il en découle que les objets de la TCT sont indissociablement les ensembles et leurs éléments et/ou les propositions et leurs preuves.

1.5 L'introduction des ensembles dans la TCT

La TCT est une théorie ouverte qui permet d'introduire pas à pas les ensembles dont on a besoin en en contrôlant soigneusement le sens et le mode opératoire. L'introduction d'un nouvel ensemble se fait en respectant systématiquement quatres règles formelles :

1. *La règle de formation.* Elle énonce que l'on peut constituer un ensemble (ou une proposition) à partir d'autres ensembles (ou propositions) déjà définies ou à partir de familles d'ensembles (ou de fonctions propositionnelles) déjà définies.
2. *La règle d'introduction.* Elle décrit comment sont obtenues les éléments canoniques de cet ensemble, donnant par là même la signification de l'ensemble introduit.
3. *La règle d'élimination.* Elle introduit généralement une fonction d'importance structurelle pour cet ensemble (fonction qui est souvent de nature inductive). Cette règle est soutenue par une explication sémantique basée sur le contenu calculatoire des éléments de l'ensemble considéré. Cette explication justifie la règle suivante.
4. *La règle d'égalité.* Elle établit un lien entre les règles d'introduction et d'élimination en décrivant comment la fonction donnée par la règle d'élimination opère sur les éléments canoniques de l'ensemble.

Ce sont des règles d'inférence immédiate : on ne peut pas les analyser ou les décomposer plus mais on peut (et on doit) les expliquer, c'est-à-dire en donner le sens. Cependant, les explications ont une fin, et finalement, aucune explication ne peut se substituer à la compréhension pratique de chacune de ces règles. Signalons enfin que ces règles introduisent des constantes dont le statut est parfaitement défini comme nous allons le remarquer dans la suite.

2 La $\mathbb{N}$-formation

L'ensemble des entiers naturels est introduit sans utiliser d'autre ensemble. La règle de formation se réduit à annoncer la constitution d'un ensemble en donnant le symbole $\mathbb{N}$ qui le représentera. Elle s'écrit formellement de la manière suivante

$$\boxed{\mathbb{N} \qquad \text{Ens.}}$$

ce qui signifie que $\mathbb{N}$ est un ensemble que l'on est en train de définir.

Cela se fait donc au moyen de l'introduction de la constante $\mathbb{N}$. Cette dernière est une *constante primitive* (on dit aussi aussi un *constructeur*), ce qui signifie que sa valeur est elle-même ; elle n'a pas de définition mais seulement un type (elle est du type ou de la catégorie des ensembles) et son sens est donné par la sémantique de la théorie.

3 La $\mathbb{N}$-introduction

Cette règle s'énonce au moyen de deux nouvelles constantes primitives, d'une part 0 qui désigne un élément distingué de $\mathbb{N}$ et d'autre part S qui désigne la fonction successeur (de type $\mathbb{N} \to \mathbb{N}$). La règle d'introduction s'écrit formellement de la manière suivante.

$$\boxed{0 \in \mathbb{N} \qquad \frac{k \in \mathbb{N}}{S(k) \in \mathbb{N}}}$$

Cette règle postule que les éléments canoniques de $\mathbb{N}$ sont soit l'élément 0, soit un élément de la forme $S(k)$ où k désigne un élément (non nécessairement canonique) de $\mathbb{N}$ déjà obtenus. Par exemple, $S(2+3)$ est un élément canonique de $\mathbb{N}$ alors que $2+3$ est un élément non canonique ; la valeur de ce dernier est l'élément canonique $S(2+2)$. On remarque que c'est la forme externe d'un nombre entier qui permet de savoir s'il est canonique ou non. Cela constitue un concept apparenté à l'évaluation paresseuse en informatique.

La sémantique d'un jugement du type $k \in \mathbb{N}$ est que k a une valeur qui est soit 0 soit de la forme $S(k_1)$, et dans ce dernier cas, k_1 a une valeur qui est soit 0 soit de la forme $S(k_2)$, et ainsi de suite jusqu'à ce que l'on atteigne finalement la valeur 0.

4 La $\mathbb{N}$-élimination

Elle s'énonce de la manière suivante, dans laquelle on se donne une famille d'ensembles $C(x)$ Ens. $(x \in \mathbb{N})$.

$$\frac{\begin{array}{ccc} & & (x \in \mathbb{N}, y \in C(x)) \\ n \in \mathbb{N} & a \in C(0) & e(x,y) \in C(S(x)) \end{array}}{\mathrm{Rec}(n,a,e) \in C(n)}$$

La prémisse consiste à se donner un entier $n \in \mathbb{N}$, un élément $a \in C(0)$ et une famille d'éléments e (c'est-à-dire $e(x,y) \in C(S(x))$ sous l'hypothèse $x \in \mathbb{N}, y \in C(x)$). La conclusion est que l'on dispose d'un élément $\mathrm{Rec}(n,a,e) \in C(n)$. On introduit ainsi une nouvelle constante, l'opérateur Rec, qui formalise une forme très générale de la récursivité. Ce n'est pas une constante primitive mais une *constante implicitement définie* (on dit aussi un *sélecteur*). Cela signifie que l'on va en expliquer le sens en montrant ce qu'elle donne lorsque l'on applique cette constante à des arguments correctement typés. En quelque sorte, la régle d'élimination donne seulement le type de Rec.

L'explication de ce qu'est l'élément $\mathrm{Rec}(n,a,e)$ de $C(n)$ est la suivante :

- On commence par exécuter n, ce qui donne un élément canonique de $\mathbb{N}$. Deux cas se présentent pour ce dernier élément, (1) il est égal à 0 auquel cas $n = 0 \in \mathbb{N}$, (2) il est de la forme $S(k)$ avec $k \in \mathbb{N}$ auquel cas $n = S(k) \in \mathbb{N}$.
- Dans le cas (1), on exécute a ce qui fournit un élément canonique $f \in C(0)$; puisque $n = 0 \in \mathbb{N}$, il vient que f est aussi un élément canonique de $C(n)$. Alors, $\mathrm{Rec}(n,a,e)$ admet pour valeur l'élément $f \in C(n)$.
- Dans le cas (2) et sous l'hypothèse que $\mathrm{Rec}(k,a,e) \in C(k)$, on pose que la valeur de $\mathrm{Rec}(n,a,e)$ est celle de $e(k, \mathrm{Rec}(k,a,e)) \in C(S(k))$. Pour déterminer la valeur de $e(k, \mathrm{Rec}(k,a,e))$, il suffit de savoir calculer celle de $\mathrm{Rec}(k,a,e)$, ce qui est expliqué par les points suivants.
- Lorsque $k = 0 \in \mathbb{N}$, on sait calculer $\mathrm{Rec}(k,a,e) \in C(k)$ d'après le cas (1).
- Lorsque k a une valeur $S(k_1)$, on recommence le processus précédent, et ceci jusqu'à atteindre la valeur 0.

5 La $\mathbb{N}$-égalité

Elle se décline en deux formes d'inférence justifiées par l'explication sémantique de ce qu'est l'opérateur Rec.

$$\frac{a \in C(0) \qquad \begin{array}{c}(x \in \mathbb{N}, y \in C(x)) \\ e(x,y) \in C(S(x))\end{array}}{\mathrm{Rec}(0,a,e) = a \in C(0)}$$

$$\frac{n \in \mathbb{N} \qquad a \in C(0) \qquad \begin{array}{c}(x \in \mathbb{N}, y \in C(x)) \\ e(x,y) \in C(S(x))\end{array}}{\mathrm{Rec}(S(n),a,e) = e(n,\mathrm{Rec}(n,a,e)) \in C(S(n))}$$

6 Substitution d'entités égales dans les règles précédentes

En toute généralité dans la TCT, à chacune des trois règles de formation, d'introduction et d'élimination, est associée une règle d'égalité qui dit que la substitution d'entités égales dans ces règles conduit à des entités égales. Ces règles ne sont pas systématiquement explicitées.

La première est associée à la $\mathbb{N}$-formation et prend la forme triviale

$$\mathbb{N} = \mathbb{N}$$

qui ne fait que répéter la propriété de réflexivité de l'égalité des ensembles.

La seconde est liée à la $\mathbb{N}$-introduction et s'énonce

$$0 = 0 \in \mathbb{N} \qquad \frac{k = l \in \mathbb{N}}{S(k) = S(l) \in \mathbb{N}}$$

Il faut non seulement la comprendre comme donnant une propriété d'extensionnalité du constructeur S [2] mais aussi et surtout comment on forme des éléments canoniques égaux. C'est un point essentiel dans la compréhension de ce qu'est un ensemble, à mettre en parallèle avec la constitution des éléments canoniques.

La dernière est associée à la $\mathbb{N}$-élimination et elle montre que le sélecteur Rec est extensionnel en ses arguments.

$$\frac{n = m \in \mathbb{N} \quad a = b \in C(0) \quad \begin{array}{c}(x \in \mathbb{N}, y \in C(x)) \\ e(x,y) = f(x,y) \in C(S(x))\end{array}}{\mathrm{Rec}(n,a,e) = \mathrm{Rec}(m,b,f) \in C(n)}$$

[2] Des éléments égaux ont des images par S qui sont égales.

7 La récurrence mathématique

On va maintenant faire une lecture de la règle d'élimination de $\mathbb{N}$ dans le cas où la famille $C(x)$ ($x \in \mathbb{N}$) est interprétée comme une famille de propositions (plus correctement nommée une fonction propositionnelle).

D'après le principe d'identification des propositions aux ensembles, tout énoncé du type $u \in C(v)$ signifie que u est une preuve de $C(v)$, auquel cas si l'on supprime la mention explicite de la preuve u, il reste l'information selon laquelle $C(v)$ est vraie. En supprimant de cette manière les preuves dans la régle d'élimination

$$\frac{\begin{array}{ccc} & & (x \in \mathbb{N}, y \in C(x)) \\ n \in \mathbb{N} & a \in C(0) & e(x,y) \in C(S(x)) \end{array}}{\mathrm{Rec}(n,a,e) \in C(n)}$$

on obtient ainsi la forme simplifiée

$$\frac{\begin{array}{ccc} & & (x \in \mathbb{N}, C(x) \text{ vraie}) \\ n \in \mathbb{N} & C(0) \text{ vraie} & C(S(x)) \text{ vraie} \end{array}}{C(n) \text{ vraie}}$$

qui est la représentation formelle dans notre cadre du raisonnement par récurrence. Si l'on veut expliciter une preuve de $C(n)$, il faut revenir à la forme générale de la règle d'élimination et donc utiliser l'opérateur de récursivité Rec. En conséquence, récurrence et récursivité sont deux faces d'un même concept lorsque l'on interprète les propositions comme des ensembles.

8 La fonction prédécesseur

On veut disposer d'une fonction prédécesseur $\mathrm{Pred} : \mathbb{N} \to \mathbb{N}$ qui soit telle que

$$\left\{ \begin{array}{rcll} \mathrm{Pred}(0) & = & 0 \in \mathbb{N} & \\ \mathrm{Pred}(S(k)) & = & k \in \mathbb{N} & \text{pour tout } k \in \mathbb{N} \end{array} \right. \tag{1}$$

Contrairement à ce qui se passe en mathématiques classiques, la prescription (1) ne constitue pas une définition valable dans la TCT, en particulier pour la raison que le symbole d'égalité qui intervient représente un jugement, c'est-à-dire une connaissance attestée par une preuve, ce qui la rend impropre à fonder une définition. Heureusement, la TCT ne possède pas que l'égalité des jugements ; en particulier, Martin-Löf a introduit l'égalité définitionnelle $\equiv$ qui désigne l'égalité du sens entre deux expressions linguistiques.

On peut introduire la fonction prédécesseur comme une constante explicitement définie à l'aide du sélecteur de récursivité en posant

$$\mathrm{Pred}(n) \equiv \mathrm{Rec}(n, 0, f)$$

dans laquelle la famille d'ensembles $C(x)$ doit être interprétée comme la famille constante de valeur $\mathbb{N}$ et f désigne la famille $f(x,y) \equiv x$ pour $x \in \mathbb{N}$ $y \in \mathbb{N}$. Pour analyser les propriétés de cette fonction maintenant correctement définie, il suffit d'utiliser les règles d'élimination et

d'égalité précédentes. La première règle d'égalité valide le jugement $\mathrm{Rec}(0,0,f)=0\in\mathbb{N}$, c'est-à-dire

$$\mathrm{Pred}(0)=0\in\mathbb{N}$$

et la seconde donne le jugement $\mathrm{Rec}(S(n),0,f)=n\in\mathbb{N}$, ce qui signifie

$$\mathrm{Pred}(S(n))=n\in\mathbb{N}$$

On est maintenant en mesure de valider ce qui constitue habituellement le troisième axiome de Peano, à savoir l'inférence suivante

$$\frac{S(k)=S(l)\in\mathbb{N}}{k=l\in\mathbb{N}} \tag{2}$$

En effet, de $S(k)=S(l)\in\mathbb{N}$ on déduit que $\mathrm{Pred}(S(k))=\mathrm{Pred}(S(l))\in\mathbb{N}$, ce qui donne le résultat puisque $\mathrm{Pred}(S(k))=k\in\mathbb{N}$ et $\mathrm{Pred}(S(l))=l\in\mathbb{N}$.

Dans une appendice, on introduira l'égalité propositionnelle et on démontrera une forme propositionnelle de cet axiome.

9 L'addition et la multiplication dans $\mathbb{N}$

On définit l'addition de deux nombres entiers a et b par

$$a+b\equiv\mathrm{Rec}(b,a,g)$$

dans laquelle à nouveau la famille d'ensembles $C(x)$ doit être interprétée comme la famille constante de valeur $\mathbb{N}$ et g désigne la famille $g(x,y)\equiv S(y)$ pour $x\in\mathbb{N}$ et $y\in\mathbb{N}$. Le sens de $a+b$ est que l'on applique b fois l'opérateur successeur S en partant de a. On obtient immédiatement les propriétés usuelles de l'addition

$$\frac{a\in\mathbb{N}\quad b\in\mathbb{N}}{a+b\in\mathbb{N}}$$

$$\frac{a\in\mathbb{N}}{a+0=a\in\mathbb{N}}\qquad\frac{a\in\mathbb{N}\quad b\in\mathbb{N}}{a+S(b)=S(a+b)\in\mathbb{N}}$$

De même, la multiplication de deux nombres entiers a et b est définie par

$$a.b\equiv\mathrm{Rec}(b,0,h)$$

dans laquelle la famille d'ensembles $C(x)$ doit être interprétée comme la famille constante de valeur $\mathbb{N}$ et h désigne la famille $h(x,y)\equiv y+a$ pour $x\in\mathbb{N}$ et $y\in\mathbb{N}$. Le sens de $a.b$ est que l'on ajoute b fois a en partant de 0. Comme pour l'addition, on obtient les propriétés usuelles de la multiplication

$$\frac{a\in\mathbb{N}\quad b\in\mathbb{N}}{a.b\in\mathbb{N}}$$

$$\frac{a\in\mathbb{N}}{a.0=0\in\mathbb{N}}\qquad\frac{a\in\mathbb{N}\quad b\in\mathbb{N}}{a.S(b)=(a.b)+a\in\mathbb{N}}$$

10 Appendices

Pour obtenir d'autres propriétés des entiers naturels, il est nécessaire d'utiliser d'autres ensembles (propositions) de la théorie constructive des types. Il s'agit de l'égalité propositionnelle, des ensembles finis $\mathbb{N}_n$ (particulièrement de l'ensemble vide $\mathbb{N}_0$), et de l'univers U.

10.1 L'égalité propositionnelle

Pour l'instant, nous disposons dans la TCT de trois formes d'égalité : l'égalité des ensembles ($A = B$), l'égalité des éléments d'un ensemble ($a = b \in A$), l'égalité définitionnelle $\equiv$; les deux premières sont des égalités entre objets et apparaissent dans des jugements, la troisième établit une relation d'équivalence sémantique entre des entités linguistiques. Dans les raisonnements, le besoin se fait aussi sentir d'une forme d'égalité adaptée aux propositions et sur laquelle on pourrait opérer à l'aide des opérations logiques. Pour cela, Martin-Löf a introduit pour chaque ensemble A, l'égalité propositionnelle sur A que l'on peut voir comme un nouvel ensemble que l'on définit dans la TCT selon le schéma des quatre règles habituelles.

1. *I-formation*

$$\frac{A \text{ Ens.} \quad a \in A \quad b \in A}{I(A, a, b) \text{ Ens.}}$$

2. *I-introduction*

$$\frac{a = b \in A}{\text{eq} \in I(A, a, b)}$$

3. *I-élimination*

$$\frac{c \in I(A, a, b)}{a = b \in A}$$

4. *I-égalité*

$$\frac{c \in I(A, a, b)}{c = \text{eq} \in I(A, a, b)}$$

Remarquons que l'élément $\text{eq} \in I(A, a, b)$ donné dans la régle d'introduction est un élément (une preuve) canonique de $I(A, a, b)$ qui ne dépend pas de a, b et A. Martin-Löf explique que cela n'a pas d'importance de savoir quel élément canonique $I(A, a, b)$ a pourvu qu'il en ait un lorsque $a = b \in A$. Dans ce cas de l'égalité propositionnelle, les règles d'élimination et d'égalité ont une forme particulière qui ne fait pas intervenir un sélecteur [3].

A titre d'exemple d'application aux nombre entiers, nous allons considérer la proposition suivante qui représente dans notre cadre le troisième axiome de Peano

$$(\forall x \in \mathbb{N})(\forall y \in \mathbb{N})(I(\mathbb{N}, S(x), S(y)) \supset I(\mathbb{N}, x, y))$$

[3] Il en résulte que cette forme de l'égalité propositionnelle (qualifiée d'extentionnelle) peut être interprétée comme présentant un défaut de prédicativité. Il existe une autre forme d'égalité, appelée égalité propositionnelle intentionnelle, qui possède une régle d'élimination de la forme habituelle et qui ne présente plus ce défaut. Néanmoins, cette deuxième égalité propositionnelle ne permet pas de prouver les mêmes résultats que l'égalité intentionnelle. C'est le cas des propriétés de $\mathbb{N}$ que nous allons donner dans cet appendice.

où $I(\mathbb{N}, a, b)$ représente l'égalité propositionnelle (de a et b dans $\mathbb{N}$) et $\supset$ l'implication. Une preuve de cette proposition est une fonction qui, à tout $x \in \mathbb{N}$, tout $y \in \mathbb{N}$ et tout $z \in I(\mathbb{N}, S(x), S(y))$, fait correspondre un élément de $I(\mathbb{N}, x, y)$. Pour cela, on va s'appuyer sur (2). En effet, pour tout $x \in \mathbb{N}$, tout $y \in \mathbb{N}$ et tout $z \in I(\mathbb{N}, S(x), S(y))$, par I-élimination on obtient $S(x) = S(y) \in \mathbb{N}$, d'où nous déduisons $x = y \in \mathbb{N}$ puis eq $\in I(\mathbb{N}, x, y)$ par I-introduction. Alors, la fonction $(\lambda x)(\lambda y)(\lambda z)$ eq est une preuve (une construction) de la proposition considérée.

10.2 Les ensembles finis N_n

Pour chaque $n = 0, 1, \ldots$ (entiers naïfs du langage non formalisé), on introduit un ensemble fini N_n par les règles et explications suivantes. Comme pour $\mathbb{N}$, ces ensembles sont définis directement sans s'appuyer sur d'autres ensembles, ce qui explique que les règles de formation correspondantes n'ont pas de prémisse.

1. *N_n-formation*
$$N_n \text{ Ens.}$$
2. *N_n-introduction*
$$0_n \in N_n \quad 1_n \in N_n \quad \ldots (n-1)_n \in N_n$$
Donc, N_0 n'a aucun élément (canonique ou pas), N_1 a un unique élément canonique 0_1, N_2 a deux éléments canoniques 0_2 et 1_2, etc..
3. *N_n-élimination*
$$\frac{c \in N_n \quad c_0 \in C(0_n) \quad \ldots \quad c_{n-1} \in C((n-1)_n)}{R_n(c, c_0, \ldots, c_{n-1}) \in C(c)}$$
dans laquelle, on dispose d'une famille d'ensembles $C(z)$ Ens. ($z \in N_n$). Cette règle introduit un nouveau sélecteur R_n. L'élément $R_n(c, c_0, \ldots, c_{n-1}) \in C(c)$ est expliqué de la manière suivante. L'exécution de c donne un élément canonique $k_n \in N_n$ pour un certain $k = 0, 1, \ldots, n-1$. Cela permet de sélectionner l'élément correspondant $c_k \in C(k_n)$. On exécute ce dernier, ce qui donne un élément canonique $d \in C(k_n)$ qui est aussi élément canonique de $C(c)$. Alors, d est la valeur de $R_n(c, c_0, \ldots, c_{n-1})$. L'opérateur R_n est donc une forme de définition par cas. Cette explication justifie la règle suivante.
4. *N_n-égalité*
$$\frac{c_0 \in C(0_n) \quad \ldots \quad c_{n-1} \in C((n-1)_n)}{R_n(k_n, c_0, \ldots, c_{n-1}) = c_k \in C(k_n)}$$
Ce qui correspond en fait à n règles, une pour chaque choix de $k = 0, 1, \ldots, n-1$.

L'ensemble N_0 n'ayant pas d'éléments, il est naturel de poser $\emptyset \equiv N_0$. On peut aussi définir la proposition sans preuve $\perp \equiv N_0$. Pour l'ensemble N_0, la règle d'élimination prend la forme limite suivante
$$\frac{c \in N_0}{R_0(c) \in C(c)}$$
dans laquelle il faut comprendre que, puisque nous n'aurons jamais un élément $c \in N_0$, nous n'aurons jamais à exécuter $R_0(c)$. Ce dernier est un programme vide sans instruction. Dans le

cas où $C(z)$ ne dépend pas de z et est interprétée comme une proposition, on peut supprimer la mention de la preuve dans la prémisse et la conclusion de la règle précédente, ce qui donne la régle générale *ex falso quolibet*.

$$\frac{\bot \text{ vraie}}{C \text{ vraie}}$$

10.3 L'univers U

Tant que l'on ne dispose pas d'un univers dans la TCT, les différents constructeurs ne peuvent être itérés qu'un nombre fini de fois. Dans le but d'obtenir un langage formel plus puissant permettant la construction d'ensembles complexes, Martin-Löf a introduit un ensemble univers reflétant au niveau de ses éléments les ensembles dont la construction a été déjà faite. Dans le cadre de cet exposé introductif, cela ne concerne que l'égalité propositionnelle, les ensembles finis N_0, N_1, etc. et l'ensemble des entiers $\mathbb{N}$. L'univers consiste en un ensemble U et une famille $T(x)$ $(x \in U)$ définis par les deux seules règles de formation et d'introduction suivantes.

1. *U-formation*

$$U \text{ Ens.} \qquad \frac{a \in U}{T(a) \text{ Ens.}} \qquad \frac{a = b \in U}{T(a) = T(b)}$$

2. *U-introduction*

$$\frac{a \in U \quad b \in T(a) \quad c \in T(a)}{i(a,b,c) \in U} \qquad \frac{a = b \in U \quad b = d \in T(a) \quad c = e \in T(a)}{i(a,b,c) = i(b,d,e) \in U}$$

$$\frac{a \in U \quad b \in T(a) \quad c \in T(a)}{T(i(a,b,c)) = I(T(a),b,c)}$$

$$n_0 \in U \quad n_1 \in U \quad \ldots \quad T(n_0) = N_0 \quad T(n_1) = N_1 \quad \ldots$$

$$n \in U \qquad T(n) = \mathbb{N}$$

A titre d'exemple d'application, on peut considérer le quatrième axiome de Peano

$$(\forall x \in \mathbb{N})\ \neg I(\mathbb{N}, 0, S(x))$$

dont la preuve dans la TCT s'appuie sur l'univers U et les règles correspondantes. Pour cela, on part de $y \in I(\mathbb{N}, 0, S(x))$ avec $x \in \mathbb{N}$. Par I-élimination, on déduit $0 = S(x) \in \mathbb{N}$. A partir de la famille d'ensembles constante de valeur U, la règle de $\mathbb{N}$-élimination permet de définir sur $\mathbb{N}$ la fonction

$$f(k) \equiv \mathrm{Rec}(k, n_0, u)$$

dans laquelle u désigne la famille $u(x,y) \equiv n_1$ pour $x \in \mathbb{N}$ $y \in U$ (famille constante de valeur n_1). De la définition du sélecteur Rec, on déduit que $f(0) = n_0$ et $f(S(k)) = n_1$ pourvu que $k \in \mathbb{N}$. Puis, de $0 = S(x) \in \mathbb{N}$ et de la partie égalité de la règle de $\mathbb{N}$-élimination on déduit que

$$\mathrm{Rec}(0, n_0, u) = \mathrm{Rec}(S(x), n_0, u) \in U$$

Comme $f(0) = n_0$ et $f(S(x)) = n_1$, il vient $n_0 = n_1 \in U$ et donc aussi $N_0 = N_1$. On obtient que l'élément canonique 0_1 de N_1 appartient aussi à N_0. Cela permet de considérer la fonction

$$(\lambda y)0_1 \in I(\mathbb{N}, 0, S(x)) \to N_0$$

Enfin, par abstraction sur la variable x, on obtient une fonction

$$(\lambda x)((\lambda y)0_1) \in \mathbb{N} \to (I(\mathbb{N}, 0, S(x)) \to N_0)$$

qui, puisque $\neg I(\mathbb{N}, 0, S(x)) \equiv I(\mathbb{N}, 0, S(x)) \to N_0$, s'interprète comme une preuve de la proposition

$$(\forall x \in \mathbb{N})\neg I(\mathbb{N}, 0, S(x))$$

10.4 Présentation informelle des axiomes de Peano

Voici une présentation informelle de la définition axiomatique par Peano de $\mathbb{N}$.

1. 0 est une entier naturel.
2. Tout entier naturel n a un unique successeur $S(n)$.
3. Deux entiers naturels ayant même successeur sont égaux.
4. Aucun entier naturel n'admet 0 pour successeur.
5. Si une proposition $\mathcal{P}(x)$ est telle que $\mathcal{P}(0)$ est vraie et $\mathcal{P}(S(n))$ est vraie chaque fois que $\mathcal{P}(n)$ est vraie, alors $\mathcal{P}(n)$ est vraie pour tout entier n.

Références

[1] P. Martin-Löf, Constructive mathematics and computer programming, *Logic, Methodology and Philosophy of Science VI*, (1980) 153–175.

[2] P. Martin-Löf, *Intuitionnistic Type Theory*, Bibliopolis, Napoli, 1984.

[3] P. Martin-Löf, Mathematics of infinity, Lecture Notes in Computer Science, COLOG-88 Computer Logic, Springer-Verlag Berlin, (1990) 146–197.

[4] B. Nordström, K. Petersson, J. M. Smith, *Programming in Martin-Löf's Type Theory. An Introduction*, Clarendon Press, Oxford, 1990.

[5] B. Nordström, K. Petersson, J. M. Smith, *Martin-Löf's Type Theory*, Handbook of Logic in Computer Logic, Oxford Sciences Publications (2000) 1–38.

[6] G. Sambin, J. Smith Ed., *Twenty-Five Years of Constructive Type Theory*, Oxford Sciences Publications, Oxford, 1998.

Adresse de l'auteur :

11, avenue de la Petite Borde

17340 Châtelaillon-Plage

France

Courriel : guy.wallet@gmail.com

CONTEXTUAL APPROACH OF AUTOMATIC DEDUCTION THEORY. APPLICATION TO ANALYSIS

YVES PÉRAIRE

Introduction.

This conference is devoted to Emmanuel Isambert. I remember that, among the very first, he had been interested in relative mathematics; thus I can imagine that, perhaps, Emmanuel would have appreciated my conference.
Researchs about proof theory and analysis are currently in progress, using more classical methods and concepts (see for example [2]). Now my point of view is quite different, in many aspects.

1 - Concerning the matter itself, analysis, the principal difference is that I am not interested in the theorems of analysis such as they are formulated traditionally, through the metaphor of sets (functional spaces, classically). I am interested rather in the formal demonstration of the facts which are hidden behind these formulations.

2 - For the language also I will diverge from what is usually done. The language is a little richer than that of ZF and consequently, concerning semantics, I am not obliged to resort systematically to this kind of metaphor. The language that I use makes it possible to choose my style, more or less metaphorical, more or less direct. Of course the direct style will be more efficient in automatization of the deductions. Moreover, I think that as well the formulations as the proofs should be more fluent if I allowed that the significance of a symbol depends on the context, as in natural languages.

CONTEXTUAL ANALYSIS.

After *Relative Non Standard Analysis* and *Relative analysis*, I now tend to name *contextual analysis* my old mathematical practice, although nothing is changed in the fundations developped in [5],[6] and [7]. This recent preference for the denomination “contextual analysis” is partly encouraged by the reading of the pedagogical work of Oliver Lessmann and Richard O’Donovan, *Analysis with relative infinitesimals*, [1]. In this book the authors make sometimes use of the same symbol of infinitesimalness for distinct relations, in various situations, and they introduce the concept of context explicitly. This practice certainly would be considered scandalous for many, not for me. Another argument for the introduction of the term ”contextual analysis” lies in my recurrent experiments in logical programing, which made me understand the interest to seek formulations as natural as possible.

Representative cases of contextuality.

In natural language. If I speak about a small cat in a small house, then the definition of small is different for a cat and for a house. So, a mathematician, or a logician would propose to make more rigorous this text by putting an index at the predicate "small": small_{cat} having a distinct definition that small_{house}. In the extreme it could seem extremely rigourous, indexing ANY object, predicate constant or function, in a mathematical text, and even more in a computing program, by types. However, the texts, in vernacular languages, are not so ambiguous. We actually succeed to communicate, with a lot of nuances sometimes, when we perfectly dominate our language, why? We shall only remark two thinks.

1- "small" is not so distinct in its definition when we speak about a cat, or about a house : The definition in either case is of the form "less than ..."
2- Actually the predicate "small" has a hidden index, which consist in *the whole text around.*

In mathematical language. It seems that the concept of context does not appear in the formulations of traditional mathematics ... but it turn out that when we practice relative methods, the introduction of the context is natural, and gives a formal speech which resembles the natural language. I will illustrate this through one example. I enonce first a theorem of relative analysis, with a stratified formulation.

Theorem

$$\forall g \in \mathbb{R}^{\mathbb{R}}\ \forall^{\alpha} f \in \mathbb{R}^{\mathbb{R}}\ \big[\ (cont_{\alpha} g \wedge\ g\ {}^{\alpha}\!\sim\ f\ ) \Rightarrow cont(f))\big]$$

[*for the significance of the index α, le reader is referred to* [5]. *The formula $cont_{\alpha} g$ means that, observed from the level α, g looks continuous Formally:* $\forall^{\alpha} x \in \mathbb{R}\ \forall y \in \mathbb{R}\ \ (y {}^{\alpha}\!\sim\ x\ \Rightarrow g(y) {}^{\alpha}\!\sim g(x))$. *$cont(f)$ means that f is continuous, $g\ {}^{\alpha}\!\sim\ f$ means $\forall^{\alpha} x\ g(x)\ {}^{\alpha}\!\sim\ f(x)$.*]

Proof
Let be y and x such that $y {}^{\alpha}\!\sim x$ and x dominated by α.
Let be $\beta = [(x, y)]$ then there exists $h \in \mathbb{R}^{\mathbb{R}}$ such that:

$$cont_{\alpha} h \wedge\ h {}^{\beta}\!\sim\ f.$$

This is a consequence of a theorem of Relative Set Theory, the theorem of the transfer partial. Now I can write the line :

$$f(y) {}^{\beta}\!\sim h(y) {}^{\alpha}\!\sim h(x) {}^{\alpha}\!\sim f(x)$$

which imply $f(y) {}^{\alpha}\!\sim f(x)$. So f is continuous.

This theorem says more about the world of facts than the classical Ascoli theorem. However the proof is direct and we can begin to think to the possibility of an automatical deduction. Now we have the feeling that the introduction of the levels α and β is somewhat artificial, may be we could avoid their introduction *a priori* ant get a more natural demonstration. More precisely, the proof that I gave previously raises several questions.

1- In this proof we made use of two relations of infinitesimalness. Can we fix once for all a quantity of relation of infinitesimallness available for ANY proof? I will show that this is not necessary, because the differents definitions, for the more or less tight ${}^{\alpha}\!\sim$ are identical up to the context. So we can say that there is only ONE relation of infinitesimalness, and various contexts. We could modify this proof using only one symbol $\sim$ the precise signification of which, depends on the context. The introduction of the contexts into a proof is called naturally during

its development. We will require that, at each step of the deduction, the context introduced are that of objects *already introduced* into the former steps: *no new context, falling from the moon and introduced by means of an existential quantifier, must be accepted.*

2 - So, how to introduce the function h? h is a new function, introduced through an existential quantifier. I affirm that it is not necessary to introduce a new symbol, we can keep the same one, g. In another context, g will take a more precise significance, and I claim that will make emerge no contradiction. Now I could write, in the place of the stratified demonstration of the preceding stratified statement, a contextual proof of a contextual statement, having the same significance. The exercice is let to the reader. I prefer from now on to investigate directly the question of automatisation of contextual deduction for contextual theorems.

In logics. Before any attempt at programming, I will use tables of natural deduction, in the style of Gentzen. However, my tables will be more general because, among the clauses appearing in these tables, there will be, in addition to the classical statements, some symbolical indications of the context, of the form $> f$, which we can read "under the context of f".An important part of this investigation is

Point out the correct syntactical rules in the construction of tables of deductions which use indicators of context.

Some rules of syntax.

A table of deduction is built from symbols of relations, indicator of context and bars of deduction. The syntax of the relations is similar to the syntax of prolog relations. We use capital letter, for variables, and lower case for constants. The bar is similar to the symbol **:-** of the language prolog. We use also indicators of context, > F, > c ... that we can read "in the (variable) context of F, "in the (fixed) context of c ... "

Il would be nice to give inductive rules of construction for the tables, but I will give only some examples, and their transcriptions both in extended prolog language and in predicate classical language.

Examples.

This first series of example use no indicators of context. *var* in a table is a meta-predicate. In a table $var(X)$ indicate that X is variable. We shall transcript it by the prolog built-in predicate **var**. To questions with the predicate **var**, the prolog interpreter answer as follows.

?- **var**(X). % A capital letter is a variable
true.
?- **var**(x). % A small letter is a constant
false.

We chose this way to deal with universal quantifiers in a formula of the form $(\forall X p(X)) \Rightarrow q_{>X<}$, preferably to the approah of Dominique Pastre [4] who prefer define quantifiers like operators.

Tables	**language of predicates**	**prolog**
$\dfrac{p(x)}{q(b)}$	$p(x) \Rightarrow q(b)$	$q(b)$ **:-** $p(x)$.
$\dfrac{p(X)}{q_{>X<}}$	$\begin{cases} \forall X(p(X) \Rightarrow q) \\ (\exists X\, p(X)) \Rightarrow q \end{cases}$	q **:-** $p(X)$.
$\begin{cases} \dfrac{p(x), q(a)}{r(c)} \\ otherform, \\ p(x) \\ \dfrac{q(a)}{r(c)} \end{cases}$	$(p(x) \wedge q(a)) \Rightarrow r(c)$	$r(c)$ **:-** $p(x), q(a)$.
$\dfrac{\mathrm{var}(X), p(X)}{q_{>X<}}$	$\begin{cases} (\forall X\, p(X)) \Rightarrow q \\ \exists X\, (p(X)) \Rightarrow q) \end{cases}$	q **:-** **var**(X), $p(X)$.
.....	.....	.....

In these examples the formulae $p(X)$, $q(a)$, $r(c)$ ... could have hidden variables. Some of these hidden variables indicate variables levels of quantification. For example $\sim(X, Y)$ could be a simplification for the formula $\sim(Z; X, Y) = \forall^Z a \in \mathbb{R}^\star\ (|X - Y| \leq |a|)$, which define the ordinary stratified relation of infinitesimal closeness. In the pratice we shall prefer replace the prefix identifier $\sim$ by an infix one. Shortly : $X \sim Y$ is equivalent to $\sim(Z; X, Y)$, the variable Z is hidden, the identifier is infix.

In the following examples we use operator of context and also the predicate of visibility "v". These two predicates are related, because the visibility of an "object" is anytime defined in a given context. Consider a table such that

$$\frac{\begin{array}{c} > c \\ v(X) \end{array}}{p(X)}$$

in wich $p(X)$ denotes some $p(Z; X..)$ with possible hidden variables and, among them, one level variable Z, everytime followed by a semicolon. It can be understood: for any X, if X is visible in the context of c, then $p(c; X, ...)$. The transcription in the language of Relative Set Theory (RST) is

$$\forall^c X\ p(c; X, ...)$$

The transcription in the language of prolog is

$$p(X) \textbf{ :- } >(c), \textbf{visible}(X).$$

We remark that, in this transcription, the hidden variables of the tabular form remain hidden.

Now above a bar of deduction there can be more than one indicator of context. As in the examples below.

Tables	**language of predicates**	**prolog**
$\begin{array}{l} > c \\ p(X) \\ > c' \\ q(Y) \\ \hline r(X,Y) \end{array}$	$\forall X \forall Y\, [p(c; X...) \wedge q((c,c'); Y...) \Rightarrow r(c; X, Y, ...)]$	$r(X,Y)$**:-** $> (c),\ p(X),$ $> (c'),\ q(Y).$
$\begin{array}{l} > c \\ v(X) \\ p(X) \\ > X \\ q(Y) \\ \hline r(X,Y) \end{array}$	$\forall^c X \forall Y\, [p(c; X...) \wedge q((c,X); Y...) \Rightarrow r(c; X, Y, ...)]$	$r(X,Y)$**:-** $> (c),\ \mathbf{visible}(X),$ $p(X),$ $> (X),\ q(Y).$
$\begin{array}{l} > c \\ v(X) \\ p(X) \\ > X \\ v(Y) \\ q(Y) \\ \hline r(X,Y) \end{array}$	$\forall^c X \forall^{(c,X)} Y((p(c; X) \wedge q((c,X); Y)) \Rightarrow r(c,X,Y))$	$r(X,Y)$**:-** $> (c), \mathbf{visible}(X),$ $p(X), > (X), \mathbf{visible}(Y),$ $q(Y).$

We call *global context* the first context at the top a table or subtable. The other contexts in the table will be named *local contexts*.

Remarks.

1 - The addition of a local context in the clauses, above a bar, modifies the meaning of the formulas which follow, above the bar, by changing the levels of quantification. On the contrary, the level of quantification in the conclusion, under the bar, is anytime the level of the global context.
2- The changes of meaning are not visible, as well in the tables as in the prolog program, in the writing of the formulas themselves but only considering their position relatively to the indications of context.

Some tabular definitions for basic concepts.

A proof of analysis need some elementary tabular definitions.

Contextual continuity.

I introduced the relation $cont(F, G)$ in order to express that the function G looks continuous in the context in wich F is visible. Obviously $cont(F, F)$ indicates that F is continuous.

Direct Rule

$$\frac{\begin{array}{l} cont(G,F) \\ > F \\ v(X) \\ Y \sim X \end{array}}{G(Y) \sim G(X)}$$

Inverse Rule

$$\frac{\text{var}(X), \quad \text{var}(Y) \quad \dfrac{\begin{array}{c} > F \\ v(X) \\ Y \sim X \end{array}}{G(Y) \sim G(X)}}{cont(G,F)}$$

In a complete computer program we should certainly need to declare than F and G are functions from a set in an other set. In order not to hide the essential, I neglect this declaration of types, in my tables.

The facts wich, in non relative mathematics, are expressed with help of punctual and uniform convergences and equi-convergences of families of functions .... are described in relative mathematics through contextual, pointwise or uniform, proximities. They are defined by the following rules:

Contextual pointwise proximity: $\sim$.

Direct Rule

$$\frac{\begin{array}{c} > F \\ G \sim F \\ v(X) \end{array}}{G(X) \sim F(X)}$$

Inverse Rule

$$\frac{var(X), \dfrac{\begin{array}{c} > F \\ v(X) \end{array}}{G(X) \sim F(X)}}{G \sim F}$$

Contextual uniform proximity: $\approx$.

Direct Rule Inverse Rule

$$\frac{\begin{array}{c} > F \\ G \approx F \end{array}}{G(X) \approx F(X)} \qquad \frac{>_F,\ var(X),\ G(X) \approx F(X)}{G \approx F}$$

This list of elementary tables is not exhaustive. In particular we need rules to deal with the transitivity and the commutativity of the relations of infinitesimal proximity, and which do not cause loopsĚ it is not so easy!

Construction of the tables of deduction.

>From now on, we can begin to compose complex tables starting from elementary tabular rules, as in examples 1 and 2 at the end of this presentation. Although I have not for the moment systematically defined the rules of composition for the tables, we can already discover some of them.

For example It could happen that in a deduction from several conditions a new condition is introduced above the bar of deduction. Is this introduction valid? This dont cause problem if the new condition is a sentence of RST because, whatever the context, if we can deduce a formula r from a formula p, we can a fortiori deduce it from two propositions p and q. Now what happens if the new condition is an indication of context?

It does not cause problem too. Precisely, if $p(X)$, $q(X)$, $r(X)$ are stratified formulae with all parameters visible in the context c, then the next rules are avalable.

Rule 1.

$$\text{If the deduction } \frac{\begin{array}{c} > c \\ p(X) \\ q(X) \end{array}}{r(X)} \text{ is allowed, so the deductions}$$

$$\frac{\begin{array}{c} > c \\ > c' \\ p(X) \\ q(X) \end{array}}{r(X)} \text{ and } \frac{\begin{array}{c} > c \\ p(X) \\ > c' \\ q(X) \end{array}}{r(X)} \text{ are authorized too.}$$

Rule 2.

$$\text{If the deduction } \frac{\begin{array}{c} > c \\ p(x) \\ .... \\ > c' \\ q(x) \end{array}}{r} \text{ is allowed,}$$

$$\text{then} \quad \text{the deduction} \quad \frac{\begin{array}{c} > c \\ p(x) \\ \dots \\ q(x) \end{array}}{r} \quad \text{is authorized too.}$$

I can add a local context c' to the general context c, and I can rewrite, in the enlarged context (c, c'), without changing any notation, the statement $q(x)$ already formulated in the context c. (We can suppose that $q(x)$ is still true in the more constraining context (c, c').)

Justification.

I justify the rules by showing that their transcription in the language of RST produces a theorem of RST. The key of justification is the theorem of partial transfer,proved in [6].

Rule 1.

$$\begin{array}{c} \forall X(p((c,c');X) \wedge q((c,c');X) \Rightarrow r(c;X) \\ \Downarrow \text{ (partial transfer)} \\ \forall X(p(c;X) \wedge q(c;X) \Rightarrow r(c;X)) \end{array}$$

$$\begin{array}{c} \forall X(p(c;X) \wedge q((c,c');X) \Rightarrow r(c;X)) \\ \Downarrow \text{ (partial transfer)} \\ \forall X(p(c;X) \wedge q(c;X) \Rightarrow r(c;X)) \end{array}$$

Rule 2.

The justification is again in local transfer because

$$\exists x \ [\ p(c;x) \wedge q(c;x)\ ] \Rightarrow \exists x \ [\ p(c;x) \wedge q((c,c');x)\ ]$$

Two examples.

Example 1: *The theorem of the continuous shadow.*

The next table proves that if g looks continuous and is infinitly close from f in the context of f, then f is continuous.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{cont(g,f)}{>f}}{g\sim f}}{v(X)}}{Y\sim X}}{var(X),var(Y),\dfrac{g(X)\sim f(X)}{f(X)\sim g(X)}}}{f(X)\sim g(Y)}\quad\dfrac{g(Y)\sim g(X)}{g(X)\sim g(Y)}}{f(X)\sim g(Y)}\qquad\dfrac{\dfrac{>Y}{g\sim f}}{v(Y)}}{g(Y)\sim f(Y)}}{f(X)\sim f(Y)}}{f(Y)\sim f(X)}}{continuous(f)}$$

Example 2: The theorem of the continuous uniform shadow.

The table below proves that if g is continuous and is uniformly close from f in the context of f, then $f$is continuous.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{cont(g)}{>f}}{g\approx f}}{v(X)}}{Y\sim X}}{\dfrac{g(Y)\sim f(Y)}{f(Y)\sim g(Y)}}}{\mathrm{var}(X),\mathrm{var}(Y),\dfrac{f(Y)\sim g(X)}{f(Y)\sim f(X)}}}{cont(f)}\quad\dfrac{\dfrac{>g}{Y\sim X}}{g(Y)\sim g(X)}\qquad\dfrac{....}{....}}{g(X)\sim f(X)}$$

Conclusion.

What does it remain to do? Much! Of course, check again the coherence of this methods of tables, and then, to translate all that by effective programs of automatic deduction ..... It is a long-term work! But the way seems open.

References

[1] **Richard O'Donovan, Olivier Lessmann**, *Analyse avec infinitésimaux relatifs* Manuel de l'enseignant, preprint, Ressource et Développement, Département de l'instruction publique Genève, 2 juillet 1988.

[2] **U.Kholenbach**, Applied Proof Theory: Proof Interpretations and their Use in Mathematics Series: Springer Monographs in Mathematics, 2008, XX, 536 p.

[3] **E. Nelson,** Internal set theory : a new approach to nonstandard.

[4] **D. Pastre**, Automated Theorem Proving in Mathematics, Annals on Artificial Intelligence and Mathematics 8(3-4) (1993), 425-447.

[5] **Y.Péraire**, Théorie relative des ensembles internes, Osaka J.Math. 29 (1992), 267-297.

[6] **Y.Péraire**, Formules absolues dans la théorie relative des ensembles internes. Rivista di matematica pura ed applicata n°19, 1996.

[7] **Y.Péraire**, Some extensions of the principles of idealization transfer and choice in the relative internal set theory. Archive for Mathematical Logic nř 34 p. 269-277 (1995).

[8] **P.Vopenka**, Mathematics in the Alternative Set Theory, Teubner, Lipzig,(1979).

Université Blaise Pascal Laboratoire de Mathématiques Pures
E-mail address: yves.peraire@math.univ-bpclermont.fr

Functions of limited accumulation

Imme van den Berg

Abstract

Measures on the real line may be decomposed into a regular, singular and atomic part. The objective of the present article is to provide analogous decompositions for a class of nonstandard, discrete functions defined on an infinitesimal discretization of the real line.

Keywords: limited accumulation, atomic, singular and regular discrete functions, decomposition, nonstandard analysis.

AMS classification: 03H05, 26A30, 46F30.

1 Introduction

Functions of limited accumulation are nonstandard, discrete functions defined on an infinitesimal discretization of the real line, such that their partial sums over discrete intervals of infinitesimal length are limited. In fact this local property of limited accumulation is semi-global, for it will be proved that also their partial sums over discrete intervals of limited length are limited. Functions of limited accumulation may be decomposed into a regular, singular and atomic part, imitating on a lower level of complexity the well-known analogous decomposition of measures on the real line.

We consider five types of decompositions. Firstly, we identify atomic, singular and regular contributions to the discrete integral of a function of limited accumulation. Secondly, respectively thirdly, we identify internal and external subsets where the contributions are realized. The fourth decomposition concerns a decomposition of the function of limited accumulation itself, in an atomic, singular and regular function. The last decomposition links a given decomposition of the function of limited accumulation to a decomposition of its discrete primitive into a jump-function, a sort of discrete Cantor function - locally constant almost everywhere, but not globally constant - and a sort of discrete absolutely continuous function.

For classical results on the decomposition of measures we refer for example to [1]. Various nonstandard authors recognized the possibility to obtain decompositions of measures or alternatively distributions by means of internal functions, for instance [10][6][16][15]. Still a comprehensive treatment of the decomposition using uniquely discrete internal functions seems novel. Being devoted to discrete integration, in a sense the present article is a complement to the article on discrete differentiation [4] in this same volume. Both articles attempt to share the spirit of simplicity due to the use of external concepts of the "radically elementary probability theory" of Nelson [14] and the "approximate analysis" of Callot [5].

Section 2 contains some preliminary remarks and notations. In Section 3 we define the functions of limited accumulation. We consider some examples, the most characteristic being a discrete Dirac function, a sort of discrete Δ-function. We investigate some basic properties, like accumulation number, accumulation point and domain of accumulation. Within the class of functions of limited accumulation we identify subclasses: atomic functions, functions of infinitesimal accumulation, singular and regular functions.

In Section 4 we investigate the behavior of the functions of limited accumulation under elementary operations. We prove that the discrete integral of a function of limited accumulation over the whole domain of definition is limited. We treat rather thoroughly the class of atomic functions, which have no counterpart in the classical theory of real functions.

The existence of the various types of decomposition of functions of limited accumulation is proved in Section 5.

The setting of this article is the axiomatic form of nonstandard analysis IST of Nelson. Introductions and notations are contained in [7], [12], [8] and [9].

The article uses material from the Masters Thesis "Funções de acumulação limitada" of Cristina Canelas and a previous unpublished manuscript of the author; the notion of limited accumulation was also used in [3]. I thank Cristina Canelas for agreeing to include the material of her thesis and Jacques Bosgiraud (University Paris VIII) for careful reading of the manuscript and suggestions for improvement.

2 Preliminaries

In this article we consider functions defined on a discrete subset of $\mathbb{R}$, consisting of successive points at an infinitesimal distance.

Definition 2.1 *We let δx always be a positive non-zero infinitesimal and $\mathbb{X} = \{k\delta x \,|\, k \in \mathbb{Z}\}$. Let $a, b \in \mathbb{X}$ be limited with $a < b$. We define the* near-interval $[a...b]$ *by*

$$[a...b] = \{x \in \mathbb{X} \,|\, a \leq x \leq b\}\,.$$

We define also

$$[a...b[= \{x \in \mathbb{X} \,|\, a \leq x < b\}\,.$$

The set $\mathbb{X}$ is called an *equidistant near-continuum* in [4]. We use here three dots to indicate near-intervals, instead of two. We will reserve the notation $[a..b]$ for a near-interval with equally spaced points at distance $\sqrt{\delta x}$.

The choice to develop discrete integration on an equidistant near-continuum is by convenience. Mutatis mutandis a general *near-continuum* can be used; we recall here its definition. Let $\mathbb{Y} \subset \mathbb{R}$ be internal; it is called a *near-continuum* if it is the image of an equidistant near-continuum by a strictly monotone function $\varphi : \mathbb{X} \to \mathbb{R}$ of class S^0 (a definition of the property of being of class S^0 can also be found in [4]).

A simple measure on the internal subsets of a discrete interval is given by the counting measure, as follows.

Definition 2.2 *Let $[a...b]$ be a near-interval. Let $A \subseteq [a...b]$ be internal. We define the* measure *λA of A by*

$$\lambda A = \sharp A \cdot \delta x.$$

Observe that $\lambda[a...b[= b - a$, its Lebesgue-measure.

We will use some basic properties of nonstandard analysis like the Cauchy Principle (no external set is internal) and the Fehrele Principle (no halo is a galaxy) and the Extension Principle (an external function with internal values defined only on the standard elements of a standard set has an extension which is an internal function defined on this set); also Robinson's Lemma (which may be seen as a particular instance of the Fehrele Principle) and the Lemma of Dominated Approximation [2]. The latter is a consequence of Robinson's Lemma and says that if

f, g are Riemann-integrable real functions such that $\int_a^b f(x)dx \simeq \int_a^b g(x)dx$ for all limited $a, b \in \mathbb{R}$, and $|f|, |g|$ are bounded by a standard integrable function h, then $\int_{-\omega}^{\omega} f(x)dx \simeq \int_{-\omega}^{\omega} g(x)dx$ for all $\omega \simeq +\infty$. The lemma also holds for Riemann-sums instead of integrals.

3 Functions of limited accumulation: definitions and examples

We define the class of functions of limited accumulation and three definite subclasses: atomic functions, singular functions and regular functions. We introduce some related notions and present some examples.

Definition 3.1 *Let $[a...b]$ be a near interval. A function $\varphi : [a...b] \to \mathbb{R}^+$ is said to be of* limited accumulation *if $\sum_{y \le x \le z} \varphi(x)\,\delta x$ is limited for all $y, z \in [a...b]$ with $y \simeq z$, $y \le z$.*

Clearly functions of class S^0 are of limited accumulation. We give here an example of a function of limited accumulation which takes unlimited values and is not S-continuous.

Example 3.2 *Let $[a...b]$ be a near interval and $c \in \mathbb{X}$ with $a < c < b$. Let $\Delta_c(x) : [a\ ...b] \to \mathbb{R}$ be defined by*

$$\triangle_c(x) = \begin{cases} \frac{1}{\delta x} & x = c \\ 0 & x \neq c. \end{cases}$$

Then Δ_c is a function of limited accumulation, since

$$\sum_{a \le x \le b} \triangle_c(x) = \sum_{a \le x < c} 0\delta x + \frac{1}{\delta x}\delta x + \sum_{c < x \le b} 0\delta x = 1.$$

We call $\triangle_c$ the discrete Dirac function *associated to c.*

Definition 3.3 *Let $\varphi : [a...b] \to \mathbb{R}^+$ be a function of limited accumulation and $h \in {}^\circ[a, b]$ be standard. The* accumulation number *α_h of φ at h is defined by*

$$\alpha_h = sup^{st}\left\{0\left(\sum_{y \le x \le z} \varphi(x)\,\delta x\right) | y, z \simeq h\right\}.$$

If $\alpha_h > 0$, the point h is called an accumulation point *of φ. The set $H = {}^{st}\{h\,|\alpha_h > 0\}$ is called the* accumulation domain *of φ.*

As for examples, let $[a...b]$ be a near interval and $c \in \mathbb{X}$ with $a < c < b$. Then ${}^\circ c$ is an accumulation point of the discrete Dirac function $\triangle_c$, with accumulation number $\alpha_{{}^\circ c} = 1$. Now let $c_1 \simeq c_2 \in [a...b]$, $a < c_1 < c_2 < b$, and $f : [a...b] \to \mathbb{R}^+$ be defined by

$$f(x) = \begin{cases} \frac{1}{2\delta x} & x = c_1 \text{ or } x = c_2 \\ 0 & x \neq c_1, c_2. \end{cases}$$

Then f is of limited accumulation, with accumulation point ${}^\circ c_1 = {}^\circ c_2$ and accumulation number

$$\alpha_{{}^\circ c_1} = sup^{st}\left\{0\left(\sum_{a_1 \le x \le b_1} h(x)\,\delta x\right) | a_1 \simeq b_1 \simeq c_1\right\} = \sup\{0, \frac{1}{2}, 1\} = 1.$$

A third example is given by a very concentrated Gaussian function, as follows. Let $a \in \mathbb{X}^+$ be appreciable and $g : [-a...a] \to \mathbb{R}$ be defined by

$$g(x) = \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}. \tag{1}$$

To see that g is of limited accumulation, put $\xi = x/\delta x^{\frac{1}{4}}$. Then $\delta x = \delta\xi \cdot \delta x^{\frac{1}{4}}$ and

$$\sum_{-a\le x\le a} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}\delta x = \sum_{\frac{-a}{\delta x^{\frac{1}{4}}}\le\xi\le\frac{a}{\delta x^{\frac{1}{4}}}} \frac{e^{\frac{-\xi^2}{2}}}{\sqrt{2\pi}}\delta\xi.$$

It is easily seen by the Lemma of Dominated Approximation that

$$\sum_{\frac{-a}{\delta x^{\frac{1}{4}}}\le\xi\le\frac{a}{\delta x^{\frac{1}{4}}}} \frac{e^{\frac{-\xi^2}{2}}}{\sqrt{2\pi}}\delta\xi \simeq \int_{-\infty}^{+\infty} \frac{e^{\frac{-\eta^2}{2}}}{\sqrt{2\pi}}d\eta = 1.$$

This implies that g is of limited accumulation. Further, let ω be unlimited such that $\omega\delta x^{\frac{1}{4}} \simeq 0$. Then

$$\sum_{-\omega\delta x^{\frac{1}{4}}\le x\le\omega\delta x^{\frac{1}{4}}} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}\delta x = \sum_{-\omega\le\xi\le\omega} \frac{e^{\frac{-\xi^2}{2}}}{\sqrt{2\pi}}\delta\xi \simeq \int_{-\omega}^{\omega} \frac{e^{\frac{-\eta^2}{2}}}{\sqrt{2\pi}}d\eta \simeq 1, \tag{2}$$

so 0 is an accumulation point of g. Its accumulation number is given by

$$\begin{aligned}
\alpha_0 &= \sup{}^{st}\left\{0\left(\sum_{y\le x\le z} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}\delta x\right) | y, z \simeq 0\right\} \\
&= \sup{}^{st}\left\{0\left(\sum_{\frac{y}{\delta x^{\frac{1}{4}}}\le\xi\le\frac{z}{\delta x^{\frac{1}{4}}}} \frac{e^{\frac{-\xi^2}{2}}}{\sqrt{2\pi}}\delta\xi\right) | y, z \simeq 0\right\} \\
&= \sup{}^{st}\left\{0\left(\int_{y/\delta x^{\frac{1}{4}}}^{z/\delta x^{\frac{1}{4}}} \frac{e^{\frac{-\eta^2}{2}}}{\sqrt{2\pi}}d\eta\right) | y, z \simeq 0\right\} \\
&= \sup{}^{st}\{stw \,|w \in [0,1]\} = \sup[0,1] = 1.
\end{aligned}$$

We define now two stronger notions. The second is rather common within nonstandard analysis.

Definition 3.4 *Let $[a...b]$ be a near interval. A function $\varphi : [a...b] \to \mathbb{R}^+$ is said to be of* infinitesimal accumulation *if $\sum_{y\le x\le z} \varphi(x)\,\delta x \simeq 0$ for all $y, z \in [a...b]$ with $y \simeq z$, $y \le z$.*

Definition 3.5 (See also [10][6]). *Let $[a...b]$ be a near interval. Let $D \subseteq [a...b]$. A function $\varphi : [a...b] \to \mathbb{R}^+$ is said to be S-*integrable *or* regular *on D if for all internal subsets $N \subseteq D$*

$$\lambda N \simeq 0 \Rightarrow \sum_{x\in N} \varphi(x)\,\delta x \simeq 0.$$

Clearly an S-integrable function is of infinitesimal accumulation. Obvious examples of S-integrable functions are limited functions and in particular functions of class S^0. Still, an S-integrable function may take unlimited values, as is shown by the function $\varphi(x) : [-1...1] \to \mathbb{R}^+$ defined by

$$\varphi(x) = \begin{cases} \sqrt{\frac{1}{\delta x}} & x = 0 \\ 0 & x \neq 0. \end{cases}$$

Next example exhibits a function of infinitesimal accumulation which is not S-integrable.

Example 3.6 *Assume for convenience that* $\frac{1}{\sqrt{\delta x}} \in \mathbb{N}$. *Define* $f : [0...1[\to \mathbb{R}^+$ *by*

$$f(x) = \begin{cases} \frac{1}{\sqrt{\delta x}} & \frac{x}{\sqrt{\delta x}} \in \mathbb{N} \\ 0 & \textit{else.} \end{cases}$$

Let $\eta = \left\{0, \sqrt{\delta x}, 2\sqrt{\delta x}, 3\sqrt{\delta x}, ..., \left(\frac{1}{\sqrt{\delta x}} - 1\right)\sqrt{\delta x}\right\}$. *Then*

$$\lambda\eta = \frac{1}{\sqrt{\delta x}}\delta x = \sqrt{\delta x} \simeq 0$$

and

$$\sum_{x\in\eta} f(x)\,\delta x = \sum_{x\in\eta} \sqrt{\delta x} = \frac{1}{\sqrt{\delta x}}\sqrt{\delta x} = 1.$$

A discrete Dirac function is an example of a function of limited accumulation which is not of infinitesimal accumulation.

The notion of function of limited accumulation has been defined for positive functions. The notion may not be extended as such to alternate functions without undesirable consequences. This is illustrated by the next example.

Let $f : [0...1] \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} \frac{1}{\delta x} & \frac{x}{\delta x} \text{ even} \\ -\frac{1}{\delta x} & \frac{x}{\delta x} \text{ odd.} \end{cases}$$

Let y, z be such that $y < z$, $y \simeq z$. Then $-1 \leq \sum_{y\leq x\leq z} f(x)\,\delta x \leq 1$, so f is of limited accumulation. But $f^{\,+}$ and $f^{\,-}$ are not of limited accumulation. Indeed, if $z - y = k\delta x$ with $k \simeq +\infty$ even, one has

$$\sum_{y\leq x\leq z} f^{\,+}(x)\,\delta x = \left[\frac{1}{\delta x}\left(\frac{z-y}{2\delta x}\right)\delta x\right] = \frac{k}{2},$$

which is unlimited. In the same way we find $\sum_{y\leq x\leq z} f^{\,-}(x)\,\delta x = -\frac{k}{2} \simeq -\infty$.

We might consider adaptations of the notion of φ being of limited accumulation, by applying it to $|\varphi|$, or by asking that $\sum_{x\in\eta} \varphi(x)\,\delta x$ is limited for any set $\eta \subset \mathbb{X}$ of infinitesimal measure, but we do not pursue this here.

Next to regular functions we distinguish two more types of functions of limited accumulation. Our goal will be to show that every function of limited accumulation can be decomposed in functions of the types in question.

Definition 3.7 *Let $\varphi : [a...b] \to \mathbb{R}^+$ be a function of limited accumulation. The function φ will be called* atomic *if there exists an internal set $C \subseteq [a...b]$ of infinitesimal measure such that $\varphi(x) = 0$ for $x \in [a...b] \setminus C$ and such that $\sum_{x\in D} \varphi(x)\,\delta x \simeq 0$ for every internal set $D \subseteq C$ such that $D \cap hal(h) = \emptyset$ for every accumulation point h of φ.*

Examples of atomic functions are the discrete Dirac-functions, standard finite sums of them, and the function g defined in (1), when restricted to an interval of infinitesimal length. Observe that a function of limited accumulation without accumulation points is necessarily of infinitesimal accumulation.

Definition 3.8 *Let $\varphi : [a...b] \to \mathbb{R}^+$ be a function of infinitesimal accumulation. The function φ will be called* singular *if there exists a set η of infinitesimal measure such that $\varphi(x) = 0$ for $x \in [a...b] \setminus \eta$.*

The function φ defined in Example 3.6 is an example of a singular function.

4 Properties of functions of limited accumulation

We will always consider non-negative functions defined on a given near-interval $[a...b]$. We show that functions of limited accumulation have a particular bound. As a consequence the class of functions of limited accumulation is closed under addition, but not under multiplication and discrete differentiation. Then we turn to discrete integrals of functions of limited accumulation. We prove a fundamental property of functions of limited accumulation: the total sum over the near-interval $[a...b]$ is limited. This implies that the discrete primitive of a function of limited accumulation is of limited accumulation. We investigate in particular the properties of the accumulation points with respect to the discrete integral.

Proposition 4.1 *If f is a function of limited accumulation, it is bounded by $\frac{c}{\delta x}$ for some limited c.*

Proof. Suppose not. Then by the Cauchy Principle there exists $x_0 \in [a...b]$ such that $f(x_0)\,\delta x \simeq \infty$. Then $\sum_{x_0 \le x < x_0+\delta x} f(x)\,\delta x = f(x_0)\,\delta x$ is unlimited. So we have a contradiction. ■

Proposition 4.2 *The sum of two functions of limited accumulation is a function of limited accumulation.*

Proof. Let f and g be two functions of limited accumulation and $y, z \in [a...b]$ be such that $y \le z, y \simeq z$. Then

$$\sum_{y\le x\le z} (f(x) + g(x))\delta x = \sum_{y\le x\le z} f(x)\,\delta x + \sum_{y\le x\le z} g(x)\,\delta x,$$

which is limited, being the sum of two limited numbers. ■

Proposition 4.3 *The product of two functions of limited accumulation is not necessarily a function of limited accumulation.*

Proof. A counterexample is given by the square of the discrete Dirac function Δ_0. Indeed, let $y, z \in \mathbb{X}$ be such that $y < 0$, $z > 0$ and $y \simeq z$. Then

$$\sum_{y\le x\le z} \Delta_0^2(x)\delta x = \Delta_0^2(0)\,\delta x = \frac{1}{\delta x^2}\delta x = \frac{1}{\delta x},$$

which is unlimited. ■

Proposition 4.4 *The discrete derivative of a discrete Dirac function is not a function of limited accumulation.*

Proof. Let $y \in [a...b[$. Then

$$\frac{\delta\Delta_y(y-\delta x)}{\delta x} = \frac{\Delta_y(y)-\Delta_y(y-\delta x)}{\delta x} = \frac{\frac{1}{\delta x}-0}{\delta x} = \frac{1}{\delta x^2},$$

which is not of the form $c/\delta x$ with limited c. ■

We will now study discrete primitives and integrals of functions of limited accumulation.

Theorem 4.5 *Let φ be a function of limited accumulation. Let $I = \sum_{a\leq x<b}\varphi(x)\,dx$. Then I is limited.*

Proof. Let G be the external set of all $\epsilon \geq 0$ such that ϵ is the length of a discrete subinterval η of $[a...b[$ and $\sum_{x\in\eta}\varphi(x)\,\delta x$ is limited. Then G is a galaxy or is internal. Clearly G contains the halo, say H, formed by all multiples of δx which are infinitesimal. One has $H \subseteq G$ by Definition 3.1, hence $H \subsetneqq G$ by the Fehrele Principle. Let $\epsilon \in G\backslash H$, then ϵ is appreciable. Hence there exists a standard $n \in \mathbb{N}$ and $x_0, x_1, ..., x_n \in [a...b]$ such that $x_0 = a$, $x_0 < x_1 < ... < x_n$, $x_n = b$ and $x_{i+1} - x_i \leq \epsilon$, for all i with $0 \leq i \leq n-1$. Let

$$M = \max\left\{\sum_{y\leq x\leq y+\epsilon}\varphi(x)\,\delta x | a \leq y, y+\epsilon \leq b\right\}.$$

Then M is limited. Hence

$$I = \sum_{a\leq x<b}\varphi(x)\,\delta x = \sum_{i=0}^{n-1}\left(\sum_{x_i\leq x<x_{i+1}}\varphi(x)\,\delta x\right) \leq \sum_{i=0}^{n-1} M = nM$$

is limited. ■

Definition 4.6 *Let $\varphi : [a...b] \to \mathbb{R}^+$ be a function. The function $\Phi : [a...b[\to \mathbb{R}^+$ given by $\Phi(x) = \sum_{a\leq y<x}\varphi(y)\delta x$ will be called its* discrete primitive.

We derive the following consequence of Theorem 4.5.

Proposition 4.7 *The discrete primitive of a function of limited accumulation is a limited non-decreasing function.*

Proof. Let $\varphi : [a...b] \longrightarrow \mathbb{R}^+$ be of limited accumulation and Φ its discrete primitive. Then Φ is non-decreasing. Moreover $\Phi(a) = 0$, and $\Phi(b)$ is limited by Theorem 4.5. Hence Φ is limited. ■

Discrete primitives of functions of limited accumulation are *functions of limited steps* in the sense [8]: these are discrete functions such that the difference of two successive values is at most limited. The steps of the discrete primitives which are truly limited and not infinitesimal correspond to the accumulation points of the functions of limited accumulation.

We study now the behavior of a function of limited accumulation with respect to its accumulation points.

Proposition 4.8 *Let φ be a function of limited accumulation. Let h be an accumulation point of φ and α_h be the accumulation value of φ at h. Then there exist $y, z \simeq h$ with $y \leq z$ such that for all $\eta, \zeta \simeq h$ with $\eta \leq y$ and $\zeta \geq z$*

$$\sum_{\eta \leq x \leq \zeta} \varphi(x)\, \delta x \simeq \alpha_h. \tag{3}$$

Proof. Let $\eta, \zeta \simeq h$ be such that $\eta \leq \zeta$. We show firstly that

$$\sum_{\eta \leq x \leq \zeta} \varphi(x)\, \delta x \lesssim \alpha_h. \tag{4}$$

Then

$$\begin{aligned} \sum_{\eta \leq x \leq \zeta} \varphi(x)\, \delta x &\simeq {}^0\left(\sum_{\eta \leq x \leq \zeta} \varphi(x)\, \delta x \right) \\ &\leq \sup{}^{st} \left\{ {}^0\left(\sum_{\beta \leq x \leq \gamma} \varphi(x)\, \delta x \right) \mid \beta, \gamma \simeq h,\ \beta \leq \gamma \right\} \\ &= \alpha_h. \end{aligned}$$

This implies (4).

Secondly we show that there exist $y, z \simeq h$, $y \leq z$ such that

$$\sum_{y \leq x \leq z} \varphi(x)\, \delta x \gtrsim \alpha_h. \tag{5}$$

Let $n \in \mathbb{N}$ be standard. There exist $\beta, \gamma \in [a...b[$ such that $\beta < \gamma$, $\beta \simeq h \simeq \gamma$ and

$$\sum_{\beta \leq x \leq \gamma} \varphi(x)\, \delta x > \alpha_h - \frac{1}{n}.$$

By the Extension Principle, there exist internal sequences $(\beta_n)_{n \in \mathbb{N}}$ and $(\gamma_n)_{n \in \mathbb{N}}$ such that

$$\sum_{\beta_n \leq x \leq \gamma_n} \varphi(x)\, \delta x > \alpha_h - \frac{1}{n} \tag{6}$$

for all standard $n \in \mathbb{N}$. By Robinson's Lemma and the Cauchy Principle there exists unlimited $\nu \in \mathbb{N}$ such that still $\beta_\nu \simeq h \simeq \gamma_\nu$, and (6) holds for $n = \nu$. Then

$$\sum_{\beta_\nu \leq x \leq \gamma_\nu} \varphi(x)\, \delta x \gtrsim \alpha_h. \tag{7}$$

Put $y = \beta_\nu$ and $z = \gamma_\nu$. Then (3) follows from (7) and (4). ■

Example 4.9 Let $\varphi : [a...b] \to \mathbb{R}^+$ be defined by

$$\begin{cases} \varphi(0) &= \frac{1}{\delta x} \\ \varphi(\delta x) &= \delta x \\ \varphi(-\delta x) &= \delta x \\ \varphi(x) &= 0 \qquad \text{if } x \notin \{-\delta x, 0, \delta x\}. \end{cases}$$

Then φ is atomic, with accumulation number $\alpha_0 = 1$. Putting $y = -\delta x$, $z = \delta x$ we have for all $\eta, \zeta \simeq 0$ with $\eta \leq y$ and $\zeta \geq z$

$$\sum_{\eta \leq x \leq \zeta} \varphi(x)\, \delta x = \delta x \cdot \delta x + \frac{1}{\delta x} \cdot \delta x + \delta x \cdot \delta x = 1 + 2\delta x^2 \simeq 1.$$

Example 4.10 Consider the function g defined in (1). Applying the substitution $x = y \cdot (\delta x)^{\frac{1}{4}}$ we derive from (2) with $\omega = \delta x^{-\frac{1}{8}}$ that

$$\sum_{-\delta x^{\frac{1}{8}} \leq x \leq \delta x^{\frac{1}{8}}} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}} \delta x \simeq 1.$$

Hence with $\eta, \zeta \simeq 0$, $\eta \leq -\delta x^{\frac{1}{8}}$ and $\zeta \geq \delta x^{\frac{1}{8}}$, also

$$\sum_{\eta \leq x \leq \zeta} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}} \delta x \simeq 1.$$

As a consequence of Theorem 4.5 and Proposition 4.8 we obtain that the external set of accumulation points of a function of limited accumulation is externally countable, i.e. in one-to-one correspondence with (a subset of) the external set of standard elements of $\mathbb{N}$.

Theorem 4.11 *The external set of accumulation points of a function of limited accumulation is at most externally countable.*

Proof. Let $\varphi : [a...b] \longrightarrow \mathbb{R}^+$ be a function of limited accumulation and Φ its discrete primitive. Let h be an accumulation point of φ. By Proposition 4.8 there exist $y, z \in [a...b[$ with $y, z \simeq h$ and $y < z$ such that $\Phi(y) \lessapprox \Phi(z)$. Now $\Phi(y)$ and $\Phi(z)$ are limited by Proposition 4.7. Then there exists a standard rational number q such that ${}^\circ\Phi(y) < q < {}^\circ\Phi(z)$. Applying the Standardization Principle we may associate a standard rational number q_h to every accumulation point h of φ; because Φ is non-decreasing, different accumulation points correspond to different standard rational numbers. Because the external set of standard rational numbers is externally countable, the external set of accumulation numbers is at most externally countable. ■

5 Decompositions

We consider five types of decompositions. Firstly, we identify atomic, singular and regular contributions to the discrete integral of a function of limited accumulation. Secondly, respectively thirdly, we identify internal and external subsets of the domain of definition where the contributions are realized. The fourth decomposition concerns a decomposition of the function of limited accumulation itself, into an atomic, singular and regular function. The last decomposition relates a given decomposition of the function of limited accumulation to a decomposition of its discrete primitive into a kind of jump-function, a sort of discrete Cantor-function and a sort of discrete absolutely continuous function.

5.1 Contributions to the discrete integral

Again we consider functions φ of limited accumulation defined on a near-interval $[a...b]$. By Theorem 4.11 the accumulation domain H of φ may be arranged into a sequence. With the help of H we identify three definite types of contribution to the value of the discrete integral $I = \sum_{a \leq x < b} \varphi(x)\, \delta x$ of φ.

Notation 5.1 *Let φ be a function of limited accumulation. We will write the accumulation domain of φ in the form $H = \{h_n \,|n \leq m\}$ in case H has m elements, for some standard*

$m \in \mathbb{N}$, and in case H is infinite we write $H = \{h_n \,|n \in \mathbb{N}\}$, where h_n is an accumulation point of φ for every standard $n \in \mathbb{N}$. Whenever h_n is an accumulation point of φ, we let α_{h_n} be the accumulation number associated to h_n. If H has m elements for some standard $m \in \mathbb{N}$, we put $\alpha_{h_n} = 0$ for $n > m$.

Definition 5.2 *Let φ be a function of limited accumulation. Let $(\alpha_{h_n})_{n\in\mathbb{N}}$ be the standardized of the (possibly) external sequence $(\alpha_{h_n})_{st\ n\in\mathbb{N}}$ of its accumulation numbers. Then the series A defined by*

$$A = \sum_{n=0}^{\infty} \alpha_{h_n}$$

is called the accumulated contribution *of φ to I.*

Definition 5.3 *Let φ be a function of limited accumulation. We define the* singular contribution *S of φ to I by*

$$S = \sup{}^{st} \left\{ {}^{0}\left(\sum_{x\in N} \varphi(x)\,\delta x \right) |N \subset [a...b] - \bigcup_{sth\in H} hal(h),\ N \text{ internal, } \lambda N \simeq 0 \right\}.$$

is called the singular contribution *of φ to I.*

Definition 5.4 *Let φ be a function of limited accumulation. We define the* regular contribution *R of φ to I by*

$$R = \sup{}^{st} \left\{ {}^{0}\left(\sum_{x\in D} \varphi(x)\,\delta x \right) |D \subset [a...b] \text{ internal, } \varphi \ S-\text{integrable on } D \right\}.$$

In order to justify the definitions we start by showing that the series $\sum_{n=0}^{\infty} \alpha_{h_n}$ converges indeed.

Lemma 5.5 *Let φ be a function of limited accumulation. Let $m \in \mathbb{N}$ be standard and $h_0, ..., h_m$ be accumulation points of φ. Then there exist disjoint internal intervals $J_0, ..., J_m \subseteq [a...b]$ such that $J_0 \subset hal(h_0), ..., J_m \subset hal(h_m)$ and $\sum_{n=0}^{m} \alpha_{h_n} \simeq \sum_{n=0}^{m} \sum_{x\in J_n} \varphi(x)\,\delta x$.*

The lemma is an immediate consequence of Proposition 4.8.

Theorem 5.6 *Let φ be a function of limited accumulation. Then its accumulated contribution $\sum_{n=0}^{\infty} \alpha_{h_n}$ converges.*

Proof. It is only needed to consider the case where H is infinite. Let $m \in \mathbb{N}$ be standard. By Lemma 5.5 there exist internal intervals $J_0, ..., J_m \subseteq [a...b]$ such that $J_0 \subset hal(h_0)$, ..., $J_m \subset hal(h_m)$ and $\sum_{n=0}^{m} \alpha_{h_n} \simeq \sum_{n=0}^{m} \sum_{x\in J_n} \varphi(x)\,\delta x \leq I$. Then $\sum_{n=0}^{m} \alpha_{h_n} \leq {}^{o}I$ for all standard $m \in \mathbb{N}$. By Transfer $\sum_{n=0}^{m} \alpha_{h_n} \leq {}^{o}I$ for all $m \in \mathbb{N}$. Hence $\sum_{n=0}^{\infty} \alpha_{h_n}$ converges, the sequence of its partial sums being non-decreasing. ■

Theorem 5.7 *Let φ be a function of limited accumulation.*

1. *There exist $\nu \in \mathbb{N}$ and an internal sequence $(J_n)_{n\leq\nu}$ of disjoint intervals such that, with $C = \bigcup_{n\leq\nu} J_n$,*

(a) *For every standard* $n \in \mathbb{N}$ *such that* $\alpha_{h_n} \neq 0$ *it holds that* $J_n \subset hal(h_n)$ *and* $\sum_{x\in J_n} \varphi(x)\,\delta x \simeq \alpha_{h_n}$.

(b) $\lambda C \simeq 0$.

(c) $\sum_{x\in C} \varphi(x)\,\delta x \simeq A$.

2. *If* $\nu' \in \mathbb{N}$ *and* $(J'_n)_{n\leq\nu'}$ *is an internal sequence with union* $C' = \bigcup_{n\leq\nu'} J'_n$, *having the same properties, one has* $\sum_{x\in C\Delta C'} \varphi(x)\,\delta x \simeq 0$.

3. *If* $I \subseteq C$ *is internal and* $I \cap \cup\{hal(h_n)\,|stn \in \mathbb{N}, \alpha_{h_n} \neq 0\} = \emptyset$, *one has* $\sum_{x\in I} \varphi(x)\delta x \simeq 0$.

Proof.

1. If H is standard finite, say of the form $\{h_0, ..., h_m\}$ with standard m, by Proposition 4.8 and Lemma 5.5 there exist disjoint internal intervals $J_0, ..., J_m \subseteq [a...b]$ such that $J_0 \subset hal(h_0), ..., J_m \subset hal(h_m)$ and $\sum_{x\in J_n} \varphi(x)\,\delta x \simeq \alpha_{h_n}$ for all n with $1 \leq n \leq m$. Let $C = \bigcup_{1\leq n\leq m} J_n$. Then $\lambda C = \lambda\left(\bigcup_{n\leq m} J_n\right) = \sum_{n=0}^{m} \lambda J_n \simeq 0$ and $\sum_{x\in C} \varphi(x)\,\delta x = \sum_{n=0}^{m} \sum_{x\in J_n} \varphi(x)\,\delta x \simeq \sum_{n=0}^{m} \alpha_{h_n} = A$.
Now suppose that $H = \{h_n\,|n \in \mathbb{N}\}$ is infinite. By the above method we obtain an external sequence $(J_n)_{stn\in\mathbb{N}}$ of internal disjoint intervals of infinitesimal length such that $J_n \subset hal(h_n)$, $\sum_{x\in J_n} \varphi(x)\,\delta x \simeq \alpha_{h_n}$ for all standard n, $\sum_{n=0}^{m} \sum_{x\in J_n} \varphi(x)\,\delta x \simeq \sum_{n=0}^{m} \alpha_{h_n}$ for all standard m, and $\lambda\left(\bigcup_{1\leq n\leq m} J_n\right) \simeq 0$. By the Extension Principle there exists an internal sequence $(J_n)_{n\in\mathbb{N}}$ extending the external sequence $(J_n)_{stn\in\mathbb{N}}$. By the Cauchy Principle there exists $\nu \in \mathbb{N}$ such that all intervals J_n, $n \leq \nu$, are disjoint. Also, applying Robinson's Lemma we may assume, with $C = \bigcup_{n\leq\nu} J_n$, that still $\lambda C \simeq 0$ and $\sum_{x\in C} \varphi(x)\,\delta x = \sum_{n=0}^{\nu} \sum_{x\in J_n} \varphi(x)\,\delta x \simeq \sum_{n=0}^{\nu} \alpha_{h_n}$. By Theorem 5.6 the series $\sum_{n=0}^{\infty} \alpha_{h_n}$ converges to A, and because it is standard $\sum_{n=0}^{\nu} \alpha_{h_n} \simeq A$. Hence $\sum_{x\in C} \varphi(x)\,\delta x \simeq A$.

2. Assume $C' = \bigcup_{n\leq\nu'} J'_n$ has the same properties. It follows from Proposition 4.8 that $\sum_{x\in J_n\Delta J'_n} \varphi(x)\,\delta x \simeq 0$ for all standard n. It follows from the Cauchy Principle and Robinson's Lemma that there exists $\mu \simeq +\infty$ with $\mu \leq \min(\nu, \nu')$ such that $J_m \bigcap J'_n \neq \emptyset$, once $m \neq n$, $m, n \leq \mu$, and

$$\sum_{n\leq\mu} \left(\sum_{x\in J_n\Delta J'_n} \varphi(x)\,\delta x \right) \simeq 0. \tag{8}$$

Now

$$C\Delta C' \subseteq \left(\bigcup_{n\leq\mu} J_n\Delta J'_n\right) \cup \left(\bigcup_{\mu\leq n\leq\nu} J_n\right) \cup \left(\bigcup_{\mu\leq n\leq\nu'} J'_n\right). \tag{9}$$

Because

$$\sum_{0\leq n\leq\mu} \left(\sum_{x\in J_n} \varphi(x)\,\delta x \right) \simeq \sum_{0\leq n\leq\mu} \left(\sum_{x\in J'_n} \varphi(x)\,\delta x \right) \simeq A,$$

one has

$$\sum_{\mu\leq n\leq\nu}\left(\sum_{x\in J_n}\varphi\left(x\right)\delta x\right)\simeq 0 \tag{10}$$

and

$$\sum_{\mu\leq n\leq\nu'}\left(\sum_{x\in J'_n}\varphi\left(x\right)\delta x\right)\simeq 0. \tag{11}$$

Combining (9), (8), (10) and (11), we prove that $\sum_{x\in C\Delta C'}\varphi\left(x\right)\delta x\simeq 0$.

3. Let $I\subseteq C$ be internal and $I\cap\cup\{hal(h_n)\,|stn\in\mathbb{N},\alpha_{h_n}\neq 0\}=\emptyset$. Put $C'=C\backslash I$ and $J'_n=J_n\backslash I$ for all $n\leq\nu$. Then $J'_n=J_n$ for all standard $n\in\mathbb{N}$. Then Part 2 implies that $\sum_{x\in I}\varphi(x)\delta x\simeq 0$.

■

Observe that the restriction of φ to the set C of the above theorem is an atomic function.

Next theorem relates the singular contribution and the regular contribution to the accumulated contribution to the value of the discrete integral.

Theorem 5.8 *Let φ be a function of limited accumulation. Let C as been given by Theorem 5.7.*

1. *There exists an internal set $M\subseteq[a...b]\backslash C$, with $\lambda M\simeq 0$, such that φ is a function of infinitesimal accumulation on M and $\sum_{x\in M}\varphi\left(x\right)\delta x\simeq S$. If M' has the same properties, with respect to a set C' given by Theorem 5.7, one has $\sum_{x\in M\triangle M'}\varphi\left(x\right)\delta x\simeq 0$.*

2. *There exists an internal set $Q\subseteq[a...b]\backslash(C\cup M)$, with $\lambda Q\simeq b-a$, such that $\varphi_{|Q}$ is S-integrable and $\sum_{x\in Q}\varphi\left(x\right)\delta x\simeq R$. If Q' has the same properties with respect to a set C' given by Theorem 5.7 and a set M' given by Part 1, one has $\sum_{x\in Q\triangle Q'}\varphi\left(x\right)\delta x\simeq 0$.*

Proof. Let $C\subseteq[a...b]$ be an internal set such that $\lambda C\simeq 0$ and $\sum_{x\in C}\varphi\left(x\right)\delta x\simeq A$, such as given by 5.7.

1. For every standard $n\in\mathbb{N}$ there exists an internal set $N_n\subseteq[a...b]\backslash\bigcup_{sth\in H}hal(h)$ such that $\lambda N_n\simeq 0$ and $S\geq\sum_{x\in N_n}\varphi\left(x\right)\delta x\geq S-\frac{2}{n}$. For such n, put $M_n=N_n\backslash C$. Because $(N_n\cap C)\cap\bigcup_{sth\in H}hal(h)=\emptyset$, Theorem 5.7.3 implies that $\sum_{x\in N_n\cap C}\varphi\left(x\right)\delta x\simeq 0$. Hence $S\geq\sum_{x\in M_n}\varphi\left(x\right)\delta x=\sum_{x\in N_n}\varphi\left(x\right)\delta x-\sum_{x\in N_n\cap C}\varphi\left(x\right)\delta x\geq S-\frac{1}{n}$. By the Extension Principle there exists an internal sequence $(M_n)_{n\in\mathbb{N}}$ extending the external sequence $(M_n)_{st\ n\in\mathbb{N}}$. Applying Robinson's Lemma and the Cauchy Principle we see that there exists $\nu\simeq\infty$ such that still $\lambda M_\nu\simeq 0$, $M_\nu\subseteq[a...b]\backslash C$ and $S\geq\sum_{x\in M_\nu}\varphi\left(x\right)\delta x\geq S-\frac{1}{\nu}$. Put $M=M_\nu$. Because $M\subseteq[a...b]\backslash C$, the function φ is of infinitesimal accumulation on M. Moreover, $\lambda M\simeq 0$ and $\sum_{x\in M}\varphi\left(x\right)\delta x\simeq S$.
Assume that M' has the same properties, as prescribed. Because $\lambda(M\cup M')\simeq 0$, the definition of S implies that $\sum_{x\in M\cup M'}\varphi\left(x\right)\delta x\lessapprox S$, hence $\sum_{x\in M\cup M'}\varphi\left(x\right)\delta x\simeq S$. So

$$S\lessapprox\sum_{x\in M}\varphi\left(x\right)\delta x+\sum_{x\in M'\backslash M}\varphi\left(x\right)\delta x\simeq\sum_{x\in M\cup M'}\varphi\left(x\right)\delta x\simeq S,$$

which implies that $\sum_{x\in M'\backslash M}\varphi\left(x\right)\delta x\simeq 0$. Similarly we prove that $\sum_{x\in M\backslash M'}\varphi\left(x\right)\delta x\simeq 0$. Hence $\sum_{x\in M\triangle M'}\varphi\left(x\right)\delta x\simeq 0$.

2. Let $Q = [a...b] \setminus (C \bigcup M)$. Then Q is an internal set and $\lambda Q \simeq b - a$. Let $\eta \subseteq Q$ such that $\lambda\eta \simeq 0$. Put $\sigma = \sum_{x\in\eta} \varphi(x)\,\delta x$. Suppose $\sigma \gtrapprox 0$. If there exists a standard $h_n \in H$ and an internal interval $J \subset hal(h_n)$ such that $\sum_{x\in\eta\cap J} \varphi(x)\,\delta x \not\approx 0$, the accumulated contribution of φ to its discrete integral would be larger than A, a contradiction. Hence $\sum_{x\in\eta\cap J} \varphi(x)\,\delta x \simeq 0$ for every internal interval $J \subset hal(h_n)$ such that $h_n \in H$ is standard. Applying the Principle of Cauchy we may find for each standard $n \in \mathbb{N}$ an internal interval $K \supset hal(h_n)$ such that $\sum_{x\in K\cap\eta} \varphi(x)\,\delta x \leq \sigma/2^{n+1}$. By the Extension Principle and the Principle of Cauchy we may find an unlimited integer ν and an internal sequence $(K_n)_{n\leq\nu}$ such that $K_n \supset hal(h_n)$ for all standard $n \in \mathbb{N}$ and $\sum_{n\leq\nu}\sum_{x\in K_n\cap\eta} \varphi(x)\,\delta x \leq \sum_{n\leq\nu} \sigma/2^{n+1} \leq \sigma/2$. Hence $\sum_{x\in(\cup_{n\leq\nu}K_n)\cap\eta} \varphi(x)\,\delta x \leq \sigma/2$, which implies that $\sum_{x\in\cap\eta\setminus(\cup_{n\leq\nu}K_n)} \varphi(x)\,\delta x \geq \sigma/2$. Hence the singular contribution to the discrete integral I of φ would be larger than S, a contradiction. We conclude that φ is S-integrable on Q.
   By definition of R, one has $\sum_{x\in Q} \varphi(x)\,\delta x \lessapprox R$. Suppose $\sum_{x\in Q} \varphi(x)\,\delta x \lessapprox R$. Then there exists some internal set $D \subseteq [a...b]$ such that φ is S-integrable on D and $\sum_{x\in Q} \varphi(x)\,\delta x \lessapprox \sum_{x\in D} \varphi(x)\,\delta x$. Because $\lambda Q \simeq b - a$, it holds that $\lambda(D\setminus Q) \simeq 0$. Because φ is S-integrable on $D\setminus Q$,

$$\sum_{x\in D} \varphi(x)\,\delta x = \sum_{x\in D\cap Q} \varphi(x)\,\delta x + \sum_{x\in D\setminus Q} \varphi(x)\,\delta x \lessapprox \sum_{x\in Q} \varphi(x)\,\delta x.$$

   So we derived a contradiction. Hence $\sum_{x\in Q} \varphi(x)\,\delta x \simeq R$.
   Assume $Q' \subseteq [a...b]$ has the same properties, as prescribed. Since $Q \triangle Q' = Q \cap (M' \cup C') \cup Q' \cap (M \cup C)$ it is the union of a set of infinitesimal measure within Q and a set of infinitesimal measure within Q'. By the above $\sum_{x\in Q\triangle Q'} \varphi(x)\,\delta x$ is the sum of two infinitesimals, hence is infinitesimal.

■

Theorem 5.9 *Let φ be a function of limited accumulation. Then its discrete integral I may be written in the form $I \simeq A + S + R$, where A is the accumulated contribution of φ, S is the singular contribution of φ and R is the accumulated contribution of φ.*

Proof. Let C, M and Q be as defined in Theorem 5.7 and 5.8. Because $[a...b] = C \cup M \cup Q$ and C, M and Q are two-by-two disjoint,

$$I = \sum_{x\in C} \varphi(x)\delta x + \sum_{x\in M} \varphi(x)\delta x + \sum_{x\in Q} \varphi(x)\delta x.$$

The near-equalities $\sum_{x\in C} \varphi(x)\delta x \simeq A$, $\sum_{x\in M} \varphi(x)\delta x \simeq S$ and $\sum_{x\in Q} \varphi(x)\delta x \simeq R$ follow also from Theorem 5.7 and 5.8. ■

5.2 Decompositions of the domain

We decompose the interval of definition $[a...b]$ of the function of limited accumulation in a way which corresponds to the decomposition in values of Theorem 5.9. The decomposition of $[a...b]$ will de done in two ways: one into internal sets and one into external sets. As for the first, we adapt the decomposition $[a...b] = C \cup M \cup Q$, in order to obtain a slightly more natural decomposition.

Theorem 5.10 *Let φ be a function of limited accumulation. Then $[a...b] = C \cup M \cup Q$, where C, M and Q are internal and two-by-two disjoint, $\lambda C \simeq \lambda M \simeq 0$ and φ takes only unlimited values on $C \cup M$, such that $\sum_{x\in C} \varphi(x)\delta x \simeq A$, $\sum_{x\in M} \varphi(x)\delta x \simeq S$ and $\sum_{x\in Q} \varphi(x)\delta x \simeq R$.*

Proof. By Theorem 5.9 there exist internal sets C, M and Q, which are two-by-two disjoint, satisfy $\lambda C \simeq \lambda M \simeq 0$ and are such that $\sum_{x\in C} \varphi(x)\delta x \simeq A$, $\sum_{x\in M} \varphi(x)\delta x \simeq S$ and $\sum_{x\in Q} \varphi(x)\delta x \simeq R$. Define for $n \in \mathbb{N}$

$$M_n = \{x \in M | \varphi(x) \geq n\}.$$

Then $\sum_{x\in M_n} \varphi(x)\,\delta x \simeq S$ for all standard $n \in \mathbb{N}$. By Robinson's Lemma there exists unlimited $\nu \in \mathbb{N}$ such that still $\sum_{x\in M_\nu} \varphi(x)\,\delta x \simeq S$**.** Put $M' = M_\nu$. Then φ takes only unlimited values on M' and $\sum_{x\in M'} \varphi(x)\,\delta x \simeq \sum_{x\in M} \varphi(x)\,\delta x \simeq S$. In the same way we obtain $C' \subseteq C$ such that φ takes only unlimited values on C' and $\sum_{x\in C'} \varphi(x)\,\delta x \simeq \sum_{x\in C} \varphi(x)\,\delta x \simeq A$. Put $Q' = [a...b]\backslash M' \cup C'$. Then $\sum_{x\in Q'} \varphi(x)\,\delta x \simeq \sum_{x\in Q} \varphi(x)\,\delta x \simeq R$. By construction C', M' and Q' are two-by-two disjoint with $\lambda C' \simeq \lambda M' \simeq 0$. ■

The internal decomposition of Theorem 5.10 is unique up to sets of infinitesimal measure. The decomposition may be transformed into a "canonical" external decomposition, which is unique. Generically spoken, the contributions to the discrete integral I on these sets cannot be represented by real numbers. However, they may be given in terms of the external numbers of [11]. An example of an external number is the external set of all infinitesimals $\oslash$. The external set of external numbers is an ordered structure of particular external intervals of $\mathbb{R}$, with strong algebraic properties. Using the notation of external numbers the following theorem holds.

Theorem 5.11 *Let φ be a function of limited accumulation. Define*

$$\begin{aligned}
\gamma &= \left\{ x \in [a...b] \,\middle|\, \sum_{y\simeq x} \varphi(y)\,\delta x > \oslash \right\} \\
\mu &= \left\{ x \in [a...b] \,\middle|\, \sum_{y\simeq x} \varphi(y)\,\delta x = \oslash,\ \varphi(x) \simeq +\infty \right\} \\
\theta &= \left\{ x \in [a...b] \,\middle|\, \sum_{y\simeq x} \varphi(y)\,\delta x = \oslash,\ \varphi(x) \ \textit{limited.} \right\}
\end{aligned}$$

Then γ, μ and θ are two-by-two disjoint and $[a...b] = \gamma \cup \mu \cup \theta$. Moreover there exists an internal set $C \subseteq \gamma$ such that $\sum_{x\in C} \varphi(x)\delta x \simeq A$, an internal set $M \subseteq \mu$ such that $\sum_{x\in M} \varphi(x)\delta x \simeq S$ and for all standard $\epsilon > 0$ there exists an internal set $P \subseteq \theta$ such that $\sum_{x\in P} \varphi(x)\delta x \geq R - \epsilon$.

Proof. By definition the external sets are two-by-two disjoint and fill up the interval $[a...b]$. The set C of Theorem 5.10 is contained in γ and it holds that $\sum_{x\in C} \varphi(x)\delta x \simeq A$. Similarly, the set M of Theorem 5.10 is contained in μ and it holds that $\sum_{x\in M} \varphi(x)\delta x \simeq S$. For the set $Q = [a...b]\backslash(C \cup M)$ of Theorem 5.10 it holds that $\sum_{x\in Q} \varphi(x)\delta x \simeq R$. We define $\psi = \varphi_{|Q}$ and for each $n \in \mathbb{N}$ we let $\psi^{(n)}$ be the function ψ truncated at n and $P_n \subseteq Q$ be the support of $\psi^{(n)}$. Observe that ψ is regular on Q. We claim that $\sum_{x\in Q} \psi^{(\nu)}(x)\delta x \simeq R$ for all $\nu \simeq +\infty$. If not, let $\mu \simeq +\infty$ be such that $\sum_{x\in Q} \psi^{(\mu)}(x)\delta x \lessapprox R$. Then $\sum_{x\in Q}(\psi(x) - \psi^{(\mu)}(x))\delta x \geq \mu\lambda(Q - P_\mu) \gtrapprox 0$. If $\lambda(Q - P_\mu) \gtrapprox 0$, $\sum_{x\in Q}(\psi(x) - \psi^{(\mu)}(x))\delta x \simeq +\infty$, a contradiction with the fact that $\psi - \psi^{(\mu)} \leq \psi$ is of limited accumulation. If $\lambda(Q - P_\mu) \simeq 0$, the function $\psi - \psi^{(\mu)}$ is singular on $Q - P_\mu$, hence also $\psi \geq \psi - \psi^{(\mu)}$, in contradiction with the fact that ψ is assumed to be regular. This proves

the claim. Let $\epsilon > 0$ be standard. One obtains with the aid of the Principle of Cauchy that $\sum_{x\in Q}\psi^{(n)}(x)\delta x = \sum_{x\in P_n}\psi(x)\delta x \geq S-\epsilon$ for some standard $n\in\mathbb{N}$. Noting that $P_n\subseteq\theta$, this proves the theorem. ■

Observe that in general there may not exist an internal set $P\subseteq\theta$ such that $\sum_{x\in P}\varphi(x)\delta x\simeq R$. A counterexample is given by the function $f:]0...1]\to\mathbb{R}^+$ defined by $f(x)=1/\sqrt{x}$.

5.3 Decomposition of the function

We write the function φ of limited accumulation itself as the sum of three functions which correspond to the accumulation part, the singular part and the regular part of its discrete integral.

Theorem 5.12 *Let $\varphi:[a...b]\to\mathbb{R}^+$ be a function of limited accumulation. Then there exist functions φ_A, φ_S, $\varphi_R:[a...b]\to\mathbb{R}^+$ with disjoint supports such that φ_A is atomic, φ_S is singular, φ_R is regular and $\varphi=\varphi_A+\varphi_S+\varphi_R$ and its discrete integrals satisfy*

$$\begin{aligned}\sum_{x\in[a...b]}\varphi_A(x)\,\delta x &\simeq A\\ \sum_{x\in[a...b]}\varphi_S(x)\,\delta x &\simeq S\\ \sum_{x\in[a...b]}\varphi_R(x)\,\delta x &\simeq R.\end{aligned}$$

Moreover, let φ'_A, φ'_S, φ'_R be a second decomposition into an atomic function φ'_A, a singular function φ'_S and a regular function φ'_R with disjoints supports. Then

$$\begin{aligned}\sum_{x\in[a...b]}\left|\varphi'_A(x)-\varphi_A(x)\right|\delta x &\simeq \sum_{x\in[a...b]}\left|\varphi'_S(x)-\varphi_S(x)\right|\delta x\\ &\simeq \sum_{x\in[a...b]}\left|\varphi'_R(x)-\varphi_R(x)\right|\delta x\simeq 0.\end{aligned}$$

Proof. Consider the decomposition of $[a...b]$ into internal sets C, M and Q as given by Theorem 5.7 and 5.8. Define $\varphi_A=\varphi|_C$, $\varphi_S=\varphi|_M$ and $\varphi_R=\varphi|_Q$. Then φ_A is atomic, φ_S is singular and φ_R is regular. Hence φ_A, φ_S, φ_R is the required decomposition. Let φ'_A, φ'_S, φ'_R be a second decomposition satisfying the properties of the theorem. Let C' be the support of φ'_A, M' be the support of φ'_S and Q' be the support of φ'_R, supposed disjoint and covering $[a...b]$. Observe that C' and M' have infinitesimal measure. Clearly $\sum_{x\in C'}\varphi'_A(x)\,\delta x\lesssim A$ and $\sum_{x\in Q'}\varphi'_R(x)\,\delta x\lesssim R$.

Suppose firstly that $\sum_{x\in C'}\varphi'_A(x)\,\delta x\lessapprox A$. Then there exists an accumulation point h and an internal interval $J\subset hal(h)\cap C$ such that $\sum_{x\in C'\cap J}\varphi'_A(x)\,\delta x\lessapprox\alpha_h$, while $\sum_{x\in J}\varphi(x)\,\delta x\simeq\alpha_h$. Then $\sum_{x\in J}(\varphi'_S(x)+\varphi'_R(x))\delta x\gtrapprox 0$, meaning that $\varphi'_S(x)$ or $\varphi'_R(x)$ cannot be of infinitesimal accumulation, a contradiction. Hence

$$\sum_{x\in C'}\varphi'_A(x)\,\delta x\simeq A; \tag{12}$$

observe that if h is an accumulation point, we also may conclude that $\sum_{x\in C'\cap J}\varphi'_A(x)\,\delta x\simeq\alpha_h$ for sufficiently large internal intervals $J\subseteq hal(h)\cap C$. As a consequence of the latter

$$\sum_{x\in C'\cap C}\varphi(x)\,\delta x\simeq\sum_{h\in H}\alpha_h=A. \tag{13}$$

Secondly,

$$\sum_{x\in Q'}\varphi'_R(x)\,\delta x=\sum_{x\in Q'\cap C}\varphi'_R(x)\,\delta x+\sum_{x\in Q'\cap M}\varphi'_R(x)\,\delta x+\sum_{x\in Q'\cap Q}\varphi'_R(x)\,\delta x,$$

where $\sum_{x\in Q'\cap C}\varphi'_R(x)\,\delta x\simeq\sum_{x\in Q'\cap M}\varphi'_R(x)\,\delta x\simeq 0$, being discrete integrals of a regular function over sets with infinitesimal measure. Hence

$$\sum_{x\in Q'}\varphi'_R(x)\,\delta x\simeq\sum_{x\in Q'\cap Q}\varphi'_R(x)\,\delta x\simeq\sum_{x\in Q'\cap Q}\varphi_R(x)\,\delta x\lessapprox\sum_{x\in Q}\varphi_R(x)\,\delta x.$$

In the same way we show that $\sum_{x\in Q}\varphi_R(x)\,\delta x\lessapprox\sum_{x\in Q'}\varphi'_R(x)\,\delta x$. Hence

$$\sum_{x\in Q'}\varphi'_R(x)\,\delta x\simeq\sum_{x\in Q}\varphi_R(x)\,\delta x\simeq R. \tag{14}$$

Then it follows from (12), (14) and Theorem 5.9 that $\sum_{x\in M'}\varphi'_S(x)\,\delta x\simeq S$.

To prove the remaining part of the theorem, we prove first the near-equalities

$$\begin{aligned}\sum_{x\in C\triangle C'}\varphi_A(x)\,\delta x &\simeq \sum_{x\in C\triangle C'}\varphi'_A(x)\,\delta x\simeq 0\\ \sum_{x\in M\triangle M'}\varphi_S(x)\,\delta x &\simeq \sum_{x\in M\triangle M'}\varphi'_S(x)\,\delta x\simeq 0\\ \sum_{x\in Q\triangle Q'}\varphi_R(x)\,\delta x &\simeq \sum_{x\in Q\triangle Q'}\varphi'_R(x)\,\delta x\simeq 0.\end{aligned}$$

As for the first near-equalities, applying (13) we find

$$\begin{aligned}\sum_{x\in C\triangle C'}\varphi_A(x)\,\delta x &= \sum_{x\in C\setminus C'}\varphi_A(x)\,\delta x\\ &= \sum_{x\in C}\varphi_A(x)\,\delta x-\sum_{x\in C\cap C'}\varphi_A(x)\,\delta x\\ &\simeq A-\sum_{x\in C\cap C'}\varphi(x)\,\delta x\simeq A-A=0.\end{aligned}$$

In an analogous way we derive that $\sum_{x\in C\triangle C'}\varphi'_A(x)\,\delta x\simeq 0$. As for the second inequalities, we use the fact that $\lambda M\simeq 0$ and φ'_R is regular to obtain

$$\begin{aligned}\sum_{x\in M\triangle M'}\varphi_S(x)\,\delta x &= \sum_{x\in M\setminus M'}\varphi_S(x)\,\delta x\\ &= \sum_{x\in M\setminus M'}\varphi'_A(x)\,\delta x+\sum_{x\in M\setminus M'}\varphi'_R(x)\,\delta x\\ &\simeq \sum_{x\in C'\setminus(C\cup R)}\varphi'_A(x)\,\delta x\\ &\le \sum_{x\in C'\setminus C}\varphi'_A(x)\,\delta x\\ &= \sum_{x\in C'\Delta C}\varphi'_A(x)\,\delta x\simeq 0.\end{aligned}$$

The near-equality $\sum_{x\in M\triangle M'}\varphi'_S(x)\,\delta x$ is derived analogously. Finally, we use the previous near-equalities to prove that

$$\begin{aligned}
\sum_{x\in Q\triangle Q'}\varphi_R(x)\,\delta x &= \sum_{x\in Q\setminus Q'}\varphi_R(x)\,\delta x\\
&= \sum_{x\in Q\cap C'}\varphi_R(x)\,\delta x+\sum_{x\in Q\cap M'}\varphi_R(x)\,\delta x\\
&= \sum_{x\in Q\cap C'}\varphi'_A(x)\,\delta x+\sum_{x\in Q\cap M'}\varphi'_S(x)\,\delta x\\
&= \sum_{x\in C'\setminus(M\cup C)}\varphi'_A(x)\,\delta x+\sum_{x\in M'\setminus(M\cup C)}\varphi'_S(x)\,\delta x\\
&\lessapprox \sum_{x\in C'\setminus C}\varphi'_A(x)\,\delta x+\sum_{x\in M'\setminus M}\varphi'_S(x)\,\delta x\simeq 0,
\end{aligned}$$

with an anologous derivation of $\sum_{x\in Q\triangle Q'}\varphi'_R(x)\,\delta x\simeq 0$.

To finish the proof, using the first of the above near-qualities we show that

$$\begin{aligned}
&\sum_{x\in[a...b]}\left|\varphi'_A(x)-\varphi_A(x)\right|\delta x\\
=\ &\sum_{x\in C\cap C'}\left|\varphi'_A(x)-\varphi_A(x)\right|\delta x+\sum_{x\in C\triangle C'}\left|\varphi'_A(x)-\varphi_A(x)\right|\delta x\\
=\ &\sum_{x\in C\triangle C'}\left|\varphi'_A(x)-\varphi_A(x)\right|\delta x\\
\leq\ &\sum_{x\in C\triangle C'}\left|\varphi'_A(x)\right|\delta x+\sum_{x\in C\triangle C'}\left|\varphi_A(x)\right|\delta x\\
=\ &\sum_{x\in C\triangle C'}\varphi'_A(x)\,\delta x+\sum_{x\in C\triangle C'}\varphi_A(x)\,\delta x\\
\simeq\ &0.
\end{aligned}$$

On shows in analogous way that

$$\sum_{x\in[a...b]}\left|\varphi'_S(x)-\varphi_S(x)\right|\delta x\simeq 0$$

and

$$\sum_{x\in[a...b]}\left|\varphi'_R(x)-\varphi_R(x)\right|\delta x\simeq 0.$$

■

Example 5.13 Assume $\frac{1}{\sqrt{\delta x}}\in\mathbb{N}$. Consider the function $f:[0...1[\to\mathbb{R}^+$ defined by

$$f(x)=\begin{cases}\frac{2x}{\sqrt{\delta x}} & \frac{x}{\sqrt{\delta x}}\in\mathbb{N}\\ 0 & \textit{else.}\end{cases}$$

Here we may take $C=\emptyset$, $M=\mathbb{N}\sqrt{\delta x}\cap[0...1[$ and $Q=[0...1[\setminus M$. Then only f_S is non-zero. Let us consider M as a near-interval with equally spaced points at distance $\sqrt{\delta x}$ and denote it by $M=[0..1[$. Then

$$S\simeq\sum_{x\in M}f_S(x)\,\delta x=\sum_{x\in M}\frac{2x}{\sqrt{\delta x}}\delta x=\sum_{x\in[0..1[}2x\sqrt{\delta x}\simeq\int_0^1 2x dx=1,$$

hence $S=1$, being standard.

Example 5.14 Let $\varphi : [0...2] \to \mathbb{R}$ be defined by

$$\varphi(x) = \begin{cases} \frac{1}{\delta x} & x = 1 \\ x & x \neq 1. \end{cases}$$

We may take $C = \{1\}$, $M = \emptyset$ and $R = [0...1[\,\backslash\{1\}$, and we have

$$A = \varphi_A(1)\delta x = \triangle_1(1)\,\delta x = \frac{1}{\delta x}\delta x = 1.$$

Also

$$R \simeq \sum_{x\in Q} \varphi_R(x)\,\delta x = \sum_{x\in[0...2[\backslash\{1\}} \varphi_R(x)\,\delta x \simeq \sum_{x\in[0...2[} x\delta x \simeq \int_0^2 x dx = 2,$$

hence $R = 2$, being standard.

Example 5.15 Assume $\frac{1}{\delta x^{1/8}} \in \mathbb{N}$. Let $\varphi :]0...1[\to \mathbb{R}^+$ be defined by $\varphi(x) = g(x) + f(x) + w(x)$, with g defined as in (1), f defined as in Example 5.13 and w defined by $w(x) = 1/\sqrt{x}$. We may take $C =]0...\delta x^{1/8}[$, $M = \mathbb{N}\sqrt{\delta x} \cap \left[\delta x^{1/8}...1\right[$, and $Q = \left[\delta x^{1/8}...1\right[\,\backslash M$. Then it follows from Example 4.10 that

$$\begin{aligned} A &\simeq \sum_{]0...\delta x^{1/8}[} \varphi_A(x)\delta x = \sum_{0<x<\delta x^{\frac{1}{8}}} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}\delta x \\ &\simeq \frac{1}{2} \sum_{-\delta x^{\frac{1}{8}}<x<\delta x^{\frac{1}{8}}} \frac{e^{\frac{-x^2}{2\sqrt{\delta x}}}}{\sqrt{2\pi\sqrt{\delta x}}}\delta x \simeq \frac{1}{2}. \end{aligned}$$

Let us use the notation $[c..d]$ for near intervals from limited numbers c to d with steps $\sqrt{\delta x}$, as suggested by Example 5.13. Then

$$S \simeq \sum_{x\in M} \varphi_S(x)\delta x = \sum_{x\in[\delta x^{1/8}..1[} 2x\sqrt{\delta x} \simeq \sum_{x\in]0..1[} 2x\sqrt{\delta x} \simeq \int_0^1 2x dx = 1.$$

Finally

$$R \simeq \sum_{x\in Q} \varphi_R(x)\delta x \simeq \sum_{x\in[\delta x^{1/8}...1[\backslash M} \frac{1}{\sqrt{x}}\delta x \simeq \sum_{x\in]0...1[} \frac{1}{\sqrt{x}}\delta x \simeq \int_0^1 \frac{1}{\sqrt{x}} dx = 2.$$

We have $A = \frac{1}{2}$, $S = 1$ and $R = 2$, because A, S and R are standard. The discrete integral I of φ over $]0...1[$ amounts to $I \simeq A + S + R = 3\frac{1}{2}$.

The above examples illustrate that a singular or an atomic function may very well be regular after an appropriate rescaling, with discrete integral nearly equal to a Riemann-integral.

5.4 Decomposition of the discrete primitive

In this final section we add to the previous decompositions a decomposition of the discrete primitive of a function of limited accumulation. We start by recalling some definitions. As before we suppose that the discrete functions φ are non-negative and defined on a near-interval $[a...b]$. For $x \in [a...b[$ we define the *difference function* $\delta\varphi$ by $\delta\varphi(x) = \varphi(x + \delta x) - \varphi(x)$.

Definition 5.16 *A function φ is said to be* absolutely S-continuous *if for all $N \subseteq [a...b[$ with $\lambda N \simeq 0$*

$$\sum_{x \in N} \delta\varphi(x) \simeq 0.$$

The shadow of an absolutely S-continuous function with limited values is well-defined and absolutely continuous [10].

Definition 5.17 *A function φ is said to be* nearly everywhere constant*, if there exists a set $N \subseteq [a...b]$ with $\lambda N \simeq b - a$ such that $\varphi(x) = \varphi(x + \delta x)$ for all x such that $x, x + \delta x \in N$.*

A discrete Dirac function is an example of a nearly everywhere constant function.

Definition 5.18 *Let $r \in {}^0[a, b]$ and $s_r \neq 0$ be standard. A function φ* has a jump in r of width s_r *if there exist $y, z \in [a...b]$ with $y, z \simeq r$ and $y < x$, such that*

$$\varphi(\zeta) - \varphi(\eta) \simeq s_r$$

for all $\eta, \zeta \in [a...b]$ such that $\eta, \zeta \simeq r$, $\eta \leq y$ and $\zeta \geq z$.

Definition 5.19 *A function φ with exactly one jump s_r in some point $r \in {}^0[a, b]$, of type*

$$\varphi(x) = \begin{cases} 0 & a \leq x < \rho \\ \sigma & \rho \leq x \leq b, \end{cases}$$

where $\rho \in [a...b], \rho \simeq r$ and $\sigma \in \mathbb{R}, \sigma \simeq s_r$, will be called a discrete function of Heaviside of width s_r.

Definition 5.20 *A limited function $\varphi : [a...b] \to \mathbb{R}$ such that there exists an internal set $C \subseteq [a...b]$ of infinitesimal measure such that $\delta\Phi(x) = 0$ for $x \in [a...b] \setminus C$ and such that it is absolutely S-continuous on every internal set $D \subseteq C$ of $[a...b]$ which does not touch any halo of the jumping points, will be called a* jump-function.

Observe that the discrete primitive of a discrete Dirac function is a discrete Heaviside function with width of jump equal to 1.

Generally spoken an accumulation point h of a function of limited accumulation, with accumulation value α_h, corresponds to a jump of its discrete primitive at h with width $s_h = \alpha_h$ and vice-versa. So next proposition is a consequence of Theorem 4.11.

Proposition 5.21 *Let Φ be a discrete primitive of a function of limited accumulation. Then the external set of its jumps is at most externally countable.*

Proposition 5.22 *Let Φ be the discrete primitive of an atomic function φ on $[a...b[$. Then Φ is a jump-function with $\Phi(b) \simeq A$. Conversely, if Φ is a jump-function on $[a...b]$, with $\Phi(a) = 0$, its discrete derivative φ is atomic, with accumulation value $A \simeq \Phi(b)$.*

Proof. Let Φ be the discrete primitive of an atomic function φ. Let Γ be the support of φ. By Theorem 5.7 there exist $\nu \in \mathbb{N}$ and an internal sequence $(J_n)_{n \leq \nu}$ of disjoint intervals such that, with $J_n = [j_n...k_n[$ and $C = \bigcup_{n \leq \nu} J_n$, firstly for all standard $n \in \mathbb{N}$ such $\alpha_{h_n} \neq 0$ it holds that

$\Phi(k_n) - \Phi(j_n) \simeq \alpha_{h_n}$, secondly $\lambda C \simeq 0$, and thirdly $\sum_{x\in C} \varphi(x)\delta x \simeq A$. Notice that also $\varphi_{|C}$ is atomic, so

$$\sum_{x\in\Gamma\setminus C} \varphi(x)\delta x \simeq 0 \tag{15}$$

by Theorem 5.12. Hence $\Phi(b) = \sum_{x\in\Gamma} \varphi(x)\delta x \simeq A$. Moreover, let $D \subseteq \Gamma\setminus \cup_{stn\in\mathbb{N}} hal(h_n)$ be internal. Then (15) and Theorem 5.7 imply that

$$\sum_{x\in D} \delta\Phi(x) = \sum_{x\in D\setminus C} \varphi(x)\delta x + \sum_{x\in D\cap C} \varphi(x)\delta x \simeq 0.$$

We conclude that Φ is a jump-function.

Conversely, let Φ be a jump-function, with $\Phi(a) = 0$, and φ be its discrete derivative. Let Γ be the support of φ. Then $\lambda\Gamma \simeq 0$. Because $\Phi(b)$ is limited, the function φ is of limited accumulation. The jumping points, say h, with width s_h of Φ correspond to accumulation points of φ with accumulation value $\alpha_h = s_h$, and the standardized H of the set of jumping points may be arranged into a sequence, so we may use the notation α_{h_n} of Notation 5.1. By Theorem 5.7 we may find $\nu \in \mathbb{N}$ and an internal sequence $(J_n)_{n\leq\nu}$ of disjoint intervals such that, with $J_n = [j_n...k_n[$ and $C = \bigcup_{n\leq\nu} J_n$, firstly for all standard $n \in \mathbb{N}$ such $\alpha_{h_n} \neq 0$ it holds that $\sum_{x\in[j_n...k_n[} \varphi(x)\delta x = \Phi(k_n) - \Phi(j_n) \simeq \alpha_{h_n}$, secondly $\lambda C \simeq 0$, and thirdly $\sum_{x\in C} \varphi(x)\delta x \simeq A$. Now $\sum_{x\in\Gamma} \varphi(x)\delta x = \Phi(b) \gtrapprox \sum_{n\in\mathbb{N}} s_{h_n} = \sum_{n\in\mathbb{N}} \alpha_{h_n} = A$. Let $D \subseteq \Gamma\setminus \cup_{stn\in\mathbb{N}} hal(h_n)$. Then

$$\sum_{x\in D} \varphi(x)\delta x = \sum_{x\in D} \delta\Phi(x) \simeq 0. \tag{16}$$

As a consequence φ is atomic. Let $\epsilon > 0$ be standard. If $n \in \mathbb{N}$ is standard, because $\sum_{x\in J} \varphi(x)\delta x \simeq \alpha_{h_n}$ for every internal interval J with $J_n \subseteq J \subseteq hal(h_n)$, there exists, an internal interval $K \supset hal(h_n)$ such that $\sum_{x\in K} \varphi(x)\delta x < \alpha_{h_n} + \frac{\epsilon}{2^n}$. By the Principle of Extension and the Principle of Cauchy there exists some unlimited $\mu \in \mathbb{N}$ and an internal sequence $(K_n)_{n\leq\mu}$ such that $K_n \supset hal(h_n)$ for all standard n and $\sum_{x\in K_n} \varphi(x)\delta x < \alpha_{h_n} + \frac{\epsilon}{2^n}$ for all $n \leq \mu$. Applying (16) we find

$$\begin{aligned}
\sum_{x\in\Gamma} \varphi(x)\delta x &= \sum_{x\in\Gamma\cap\cup_{n\leq\mu}K_n} \varphi(x)\delta x + \sum_{x\in\Gamma\setminus\cup_{n\leq\mu}K_n} \varphi(x)\delta x \\
&\simeq \sum_{x\in\Gamma\cap\cup_{n\leq\mu}K_n} \varphi(x)\delta x \\
&\lessapprox A + \epsilon.
\end{aligned}$$

Because ϵ is arbitrary, we derive that $\sum_{x\in\Gamma} \varphi(x)\delta x \lessapprox A$. Combining, we conclude that $\sum_{x\in\Gamma} \varphi(x)\delta x \simeq A \simeq \Phi(b)$. ■

Proposition 5.23 *Let φ be a function of limited accumulation and Φ be its discrete primitive. Then*

1. *φ is atomic with accumulated contribution $A > 0$ if and only if Φ is a jump-function with $\Phi(a) = 0$ and $\Phi(b) \simeq A$.*
2. *φ is singular with singular contribution $S > 0$ if and only if Φ is S-continuous and nearly always constant with $\Phi(a) = 0$ and $\Phi(b) \simeq S$.*

3. *φ is regular with regular contribution $R > 0$ if and only if Φ is absolutely S-continuous with $\Phi(a) = 0$ and $\Phi(b) \simeq R$.*

Conversely, the above equivalences continue to hold if $\Phi : [a...b] \to \mathbb{R}$ is non-decreasing such that $\Phi(a) = 0$ and $\Phi(b)$ is limited and $\varphi : [a...b[\to \mathbb{R}$ defined by $\varphi = \frac{\delta\Phi}{\delta x}$ is its discrete derivative.
Proof.

1. This part follows from Proposition 5.22.

2. Let φ be singular and let $\eta \subseteq [a...b[$ be such that $\lambda\eta \simeq 0$, $\sum_{x\in\eta} \varphi(x)\delta x \simeq S$ and φ is zero on $[a...b]\backslash\eta$. This implies that Φ is nearly always constant, with $\Phi(b) = \Phi(b) - \Phi(a) \simeq S$. Let $y, z \in [a...b]$ with $y < z$ and $y \simeq z$. Because φ is of infinitesimal accumulation
$$\Phi(z) - \Phi(y) = \sum_{y\leq x<z} \varphi(x)\delta x \simeq 0. \tag{17}$$
hence Φ is S-continuous.
Conversely, if $\Phi(x) = \Phi(x + \delta x)$ for all x such that $x, x + \delta x \in [a...b]\backslash\eta$, with $\lambda\eta \simeq 0$, its discrete derivative φ is nearly always zero. Because Φ is S-continuous, by (17) the function φ is of infinitesimal accumulation. The function φ is singular, for
$$\sum_{x\in\eta} \varphi(x)\delta x = \Phi(b) - \Phi(a) = \Phi(b) \simeq S.$$
.

3. Let φ be regular and $\eta \subseteq [a...b[$ be such that $\lambda\eta \simeq 0$. Then $\sum_{x\in\eta} \varphi(x)\delta x \simeq 0$. So $\sum_{x\in\eta} \delta\Phi(x) \simeq 0$. Hence Φ is absolutely S-continuous on $[a...b]$. Moreover, $\Phi(a) = 0$ and
$$\Phi(b) = \Phi(b) - \Phi(a) = \sum_{x\in[a...b[} \varphi(x)\delta x \simeq R.$$
Conversely, if Φ is absolutely S-continuous, one has $\sum_{x\in\eta} \varphi(x)\delta x \simeq \sum_{x\in\eta} \delta\Phi(x) \simeq 0$ for all $\eta \subseteq [a...b[$ such that $\lambda\eta \simeq 0$. Hence φ is regular, with $\sum_{x\in[a...b[} \varphi(x)\delta x = \Phi(b) - \Phi(a) = \Phi(b) \simeq R$.

The second part of the proposition amounts to a reformulation of the first part of the proposition. ■

We end with a decomposition theorem for non-decreasing functions.

Theorem 5.24 *Let $\Phi : [\ a...b] \to \mathbb{R}$ be a non-decreasing function such that $\Phi(a)$ and $\Phi(b)$ are limited. Then there exist non-negative non-decreasing functions Φ_A, Φ_S and Φ_R such that Φ_A is a jump-function, Φ_S is an S-continuous function which is nearly everywhere constant, and Φ_R is absolutely S-continuous, such that*
$$\Phi - \Phi(a) = \Phi_A + \Phi_S + \Phi_R.$$
Moreover, let $I = \Phi(b) - \Phi(a)$, A be the accumulated contribution of $\frac{\delta\Phi}{\delta x}$ to I, S its singular contribution to I and R its regular contribution to I. Then $\Phi_A(b) \simeq A$, $\Phi_S(b) \simeq S$ and $\Phi_R(b) \simeq R$. If Φ'_A, Φ'_S, Φ'_R is a second decomposition of $\Phi - \Phi(a)$ into a jump-function Φ'_A, an S-continuous function which is nearly everywhere constant Φ'_S, and an absolutely S-continuous function Φ'_R, all non-negative and non-decreasing, then $\Phi'_A(x) \simeq \Phi_A(x)$, $\Phi'_S(x) \simeq \Phi_S(x)$ and $\Phi'_R(x) \simeq \Phi_R(x)$ for all $x \in [a...b]$.

Proof. Let φ_A, φ_S, φ_R be the decomposition of $\varphi = \frac{\delta\Phi}{\delta x}$ given by Theorem 5.12. Let Φ_A, Φ_S, respectively Φ_R be the discrete primitive of φ_A, φ_S, respectively φ_R. By Proposition 5.23 the function Φ_A is a jump-function, the function Φ_R is S-continuous and nearly everywhere constant and the function Φ_R is absolutely S-continuous. Because the supports of φ_A, φ_S and φ_R are disjoint, one has $\Phi - \Phi(a) = \Phi_A + \Phi_S + \Phi_R$. Again by Proposition 5.23 we have $\Phi_A(b) \simeq A$, $\Phi_S(b) \simeq S$ and $\Phi_R(b) \simeq R$. Let Φ'_A, Φ'_S, Φ'_R be a second decomposition of Φ into a jump-function Φ'_A, an S-continuous function which is nearly everywhere constant Φ'_S, and an absolutely S-continuous function Φ'_R, all non-negative and non-decreasing. Observe that $\Phi_A(a) = \Phi'_A(a) = \Phi_S(a) = \Phi'_A(a) = \Phi_R(a) = \Phi'_R(a) = 0$. Let C, C', M, M' and Q, Q' be the internal sets given by Proposition 5.8. Let $x \in [a...b[$. Put $\varphi'_A = \frac{\delta\Phi'_A}{\delta x}$, $\varphi'_S = \frac{\delta\Phi'_S}{\delta x}$ and $\varphi'_R = \frac{\delta\Phi'_R}{\delta x}$. Then by Theorem 5.12

$$\begin{aligned}
\left|\Phi'_A(x) - \Phi_A(x)\right| &= \left|\sum_{a\le y<x} (\varphi'_A(y) - \varphi_A(y))\delta x\right| \\
&= \left|\sum_{a\le y<x, y\in C\Delta C'} (\varphi'_A(y) - \varphi_A(y))\delta x\right| \\
&\le \sum_{y\in C\Delta C'} \left|\varphi'_A(y) - \varphi_A(y)\right| \delta x \\
&\simeq 0.
\end{aligned}$$

Hence $\Phi'_A(x) \simeq \Phi_A(x)$. One shows in the same way that $\Phi'_S(x) \simeq \Phi_S(x)$ and $\Phi'_R(x) \simeq \Phi_R(x)$. ■

References

[1] R.G. Bartle, *A Modern Theory of Integration,* Graduate Studies in Mathematics 32, American Mathematical Society (2001).

[2] I.P. van den Berg, *Nonstandard Asymptotic Analysis*, Springer Lecture Notes in Mathematics 1249 (1987).

[3] I.P. van den Berg, *Equations paraboliques et intégrales de chemins finies avec applications financières*, Publication pédagogique 32, Université de Nice Sophia-Antipolis, http://math.unice.fr/~rgr (1998).

[4] I.P. van den Berg, *Discretizations of higher order and the theorems of Faà di Bruno and DeMoivre-Laplace*, this volume.

[5] J.-L. Callot, *Analyse grossière et analyse infinitésimale*, in: Actes du Colloque Trajectorien, Obernai, publ. IRMA Strasbourg (1995) 229-237.

[6] P. Cartier and Y. Perrin, *Integration over finite sets*, in: F. and M. Diener eds., Nonstandard Analysis in Practice, Springer Universitext (1995) 185-204.

[7] F. Diener and M. Diener, eds., *Nonstandard Analysis in Practice*, Springer Universitext (1995).

[8] F. Diener and G. Reeb, *Analyse Non Standard*, Hermann, Paris (1989).

[9] A.J. Franco de Oliveira and I.P. van den Berg, *Matemática Não Standard, Uma introdução com aplicações*, Edição Gulbenkian, Lisbon (2007).

[10] A. Hurd and P. Loeb, *An introduction to nonstandard real analysis,* Pure and Applied Mathematics 118, Academic Press (1985).

[11] F. Koudjeti and I.P. van den Berg, *Neutrices, external numbers and external calculus*, in: F. and M. Diener eds., Nonstandard Analysis in Practice, Springer Universitext (1995 145-170).

[12] A. Robert, *Nonstandard Analysis*, John Wiley & Sons (1988).

[13] E. Nelson, *Internal Set Theory*, Bull. Amer. Math. Soc. 83, no. 6 (1977) 1165–1198.

[14] E. Nelson, *Radically Elementary Probability Theory*, Princ. Univ. Press (1987).

[15] J. Sousa Pinto, *Métodos Infinitesimais de Análise Matemática,* Gulbenkian, Lisbon (2000). English translation by R.F. Hoskins: *Infinitesimal methods for Mathematical Analysis*, Horwood, Chichester (2004).

[16] K.D. Stroyan and J.M. Bayod, *Foundations of infinitesimal stochastic analysis.* Studies in Logic and the Foundations of Mathematics, 119, North-Holland Publishing Co., Amsterdam (1986).

Address of the author:

University of Évora, Department of Mathematics
Colégio Luís Verney, Rua Romão Ramalho 59
7000-671 Évora, Portugal

E-mail: `ivdb@uevora.pt.`

Discretisations of higher order and the theorems of Faà di Bruno and DeMoivre-Laplace

I.P. van den Berg

Abstract

We study discrete functions on equidistant and non-equidistant infinitesimal grids. We consider its difference quotients of higher order and give conditions for their near-equality to the corresponding derivatives. Important tools are the formula of Faà di Bruno for higher order derivatives and a discrete version of it. As an application of such transitions from the discrete to the continuous we extend the DeMoivre-Laplace Theorem to higher order: n-th order difference quotients of the binomial probability distribution tend to the corresponding n-th order partial differential quotients of the Gaussian distribution.

Keywords: Difference quotients, chain rule, Faà di Bruno Theorem, DeMoivre-Laplace Theorem, nonstandard analysis.

AMS classification: 03H05, 39A10, 39A12, 60F05.

1 Introduction

Let $\delta\xi > 0$ be infinitesimal in the sense of nonstandard analysis. We let $\mathbb{X}$ be the set of al multiples of $\delta\xi$ by some integer. We study discrete "quasi-continuous" functions f on this set of equally spaced points, but also on more irregular grids, which are images of $\mathbb{X}$ by some near-standard function, say ϕ. Such grids will be called *near-continua*.

On the near-continua we consider difference quotients $\frac{\delta^n f}{\delta\xi^n}$ of standard order n and conditions for their near-equality to the derivatives $\frac{d^n {}^0 f}{dx^n}$ of the shadow ${}^0 f$ of the function f: the unique standard real function (if it exists) infinitely close to f on the limited domain.

On equidistant near-continua $\mathbb{X}$ the condition will be the property to be "of class S^n", a sort of generalization of the notions of S-continuity and S-differentiability to higher order.

For functions defined on non-equidistant near-continua $\phi(\mathbb{X})$ such near-equality does not need to hold - even for functions as elementary as quadratic functions - and appears to depend on the nature of the function ϕ. As is to be expected the Chain Rule will have some importance, in particular its version for higher order derivatives, known as the Formula of Faà di Bruno. We establish an approximate, discrete version of this formula. With the help of this formula we show that the difference quotients of standard order n of a function $\psi : \phi(\mathbb{X}) \to \mathbb{R}$ are infinitely close to the corresponding derivatives of its shadow, provided that ϕ and ψ are both of class S^n.

The present article extends the results on transitions between the discrete and the continuous on equidistant near-continua of [3], [10] and [5] to transitions on general near-continua. These studies ([10] representing a Masters Thesis supervised by the author) apply such results to the transition of the discrete binomial probability distribution to the continuous Gaussian distribution, extending the DeMoivre-Laplace theorem in a sense to difference quotients and (partial) derivatives of higher order. We end the present article by a short proof of this extension.

The article has the following structure. Nonstandard analysis disposes of a terminology common to both a class of discrete functions and a class of continuous functions; this terminology facilitates the transition between discreteness and continuity and will be recalled in Section 2. In Section 3 we define the notion of function of class S^n for general near-continua. We show how these functions behave for discrete differentiation and integration. Also, the class S^n is stable for algebraic operations on equidistant near-continua, but we indicate an elementary counterexample for non-equidistant near-continua. Later on, at the end of Section 5, conditions will be given for algebraic operations to hold on (suitable parts of) such near-continua. In Section 4 we recall the tools for the transition from discreteness to continuity (the class of standard functions of class C^n) at arbitrary order on equidistant near-continua. In Section 5 we develop the tools for the transition from discreteness to continuity at arbitrary order on general near-continua. It is here that we derive a nonstandard, discrete and approximative version of the formula of Faà di Bruno. In Section 6 we extend the process of continuization to functions of two variables.

As an application we consider a higher order version of the DeMoivre-Laplace theorem. The Section 7 contains a shortened version of material presented earlier in [2] and [5]. We introduce first (Section 7.1) a convenient rescaling for the Pascal triangle and the binomial coefficients, leading to the notion of *binomial function* $b(t,x)$. The higher order DeMoivre-Laplace theorem in question states that the difference quotients of the binomial function are infinitely close to the corresponding partial derivatives of the Gaussian function, to all standard orders and for all limited non-infinitesimal t and all limited x.

In Section 7.2 we establish a first-order difference equation for the binomial function and show that the function satisfies also a discrete heat equation. It appears to be convenient to extend this second-order partial difference equation to all orders. In the final Section 7.3 we complete the higher order DeMoivre-Laplace theorem, using the material on the transition from the discrete to the continuous developed in the earlier sections.

The article is written in the nonstandard axiomatic system IST of Nelson. A mayor difference with respect to Robinson's nonstandard analysis [14] is the existence of nonstandard elements within infinite standard sets. Introductions to IST are for example contained in [12], [8], [6], and [7]. A thorough introduction to discrete probability theory in a setting similar to ours is contained in [13]. For a classical proof of the DeMoivre-Laplace theorem, see for instance [9].

2 Functions of class S^0 and near-continua

Functions of class S^0 where introduced by F. Diener in [8], see also [7]. We recall here some basic properties. They are closely related to functions of class C^0.

We start by recalling the notion of S-continuity. This property is shared by the class of standard continuous functions and also a class of nonstandard discrete functions. As such the property is important for the interplay between discreteness and continuity.

Definition 2.1 *Let $A \subset \mathbb{R}$ and $f : A \to \mathbb{R}$. The function f is said to be S-continuous on A if for all $x, y \in A$*

$$x \simeq y \Rightarrow f(x) \simeq f(y).$$

The notion of S-continuity may also be defined for functions on other spaces than $\mathbb{R}$. Relevant for this article are functions defined on (subsets of) $\mathbb{R}^2$.

Definition 2.2 *Let $A \subset \mathbb{R}^2$ and $f : A \to \mathbb{R}$. The function f is said to be S-continuous on A if for all $(x_1, x_2), (y_1, y_2) \in A$*

$$x_1 \simeq y_1, x_2 \simeq y_2 \Rightarrow f(x_1, x_2) \simeq f(y_1, y_2).$$

Definition 2.3 *Let $A \subset \mathbb{R}$ and $f : A \to \mathbb{R}$. The function f is said to be* of class S^0 *on A if $f(x)$ is limited and S-continuous at all limited $x \in A$.*

A class of obvious examples of functions of class S^0 are the standard everywhere continuous functions.

By no means functions of class S^0 need to be limited for unlimited x: standard unbounded functions such as standard polynomials, as well as the exponential function, are all of class S^0. Standard rational functions with poles, like $f(x) = \frac{1}{1+x}$ are of class S^0 on all sets A which do not contain elements infinitely close to the poles.

The next simple proposition enables often to verify that nonstandard functions are of class S^0.

Proposition 2.4 *Let $A \subset \mathbb{R}$ and $f, g : A \to \mathbb{R}$. If f is of class S^0 and $f(x) \simeq g(x)$ for all limited $x \in A$, then g is also of class S^0.*

Proof. Clearly $g(x)$ is limited for all limited $x \in A$. Let $x \in A$ be limited and $y \in A, y \simeq x$. Then

$$g(y) \simeq f(y) \simeq f(x) \simeq g(x).$$

Hence g is S-continuous at x. ■

In this article we consider usually functions defined on a discrete subset of $\mathbb{R}$, consisting of successive points at an infinitesimal distance.

Definition 2.5 *We let $\delta\xi$ always be a positive non-zero infinitesimal and $\mathbb{X} = \{k\delta\xi \,|k \in \mathbb{Z}\}$. The set $\mathbb{X}$ is called an* equidistant near-continuum. *Let $\mathbb{Y} \subset \mathbb{R}$ be internal. The set $\mathbb{Y}$ is called a* near-continuum *if it is the image of an equidistant near-continuum by a strictly monotone function $\phi : \mathbb{X} \to \mathbb{R}$ of class S^0.*

For convenience we suppose, unless otherwise said, that ϕ is increasing, unbounded from the below and unbounded from the above.

Definition 2.6 *Let $\mathbb{Y} = \phi(\mathbb{X})$ be a near-continuum, where $\phi : \mathbb{X} \to \mathbb{R}$ is a strictly increasing function of class S^0. Let $a, b \in \mathbb{Y}$ be limited with $a < b$. Then*

$$[a \cdot\cdot b] \equiv \{\eta \in \mathbb{Y} \,|a \leq \eta \leq b\}$$

is called a near-interval. *We define also*

$$[a \cdot\cdot b[\equiv \{\eta \in \mathbb{Y} \,|a \leq \eta < b\}\,.$$

Let $\eta \in \mathbb{Y}$ and let $\xi \in \mathbb{X}$ be such that $\eta = \phi(\xi)$. Sometimes we write $\eta_\xi = \phi(\xi)$ and $\delta\eta_\xi = \delta\phi(\xi)$, and we may even write with abuse of language $\delta\eta$ instead of $\delta\eta_\xi$; note that generally spoken $\delta\eta$ is not a constant. Given a subset $A \subseteq \mathbb{Y}$ we define

$$A^{(n)} = \left\{\eta_\xi \,\middle|(\forall i)(0 \leq i \leq n \Rightarrow \eta_{\xi+i\delta x} \in A)\right\}.$$

Notice that if f is a real function defined on $A \subseteq \mathbb{X}$, the n^{th}-order difference quotient or n^{th}*-order discrete derivative* of f

$$\frac{\delta^n f}{\delta\xi^n}(\xi) \quad \equiv \quad \frac{1}{\delta\xi^n} \sum_{j=0}^{n} (-1)^j \binom{n}{j} f(\xi + (n-j)\delta\xi) \tag{1}$$

is defined on $A^{(n)}$.

We give two examples of functions of class S^0 defined on an equidistant near-continuum $\mathbb{X} \equiv \{k\delta\xi \,|k \in \mathbb{Z}\}$. The nonstandard function $\mathcal{E} : \mathbb{X} \to \mathbb{R}$ defined by

$$\mathcal{E}(\xi) = (1 + \delta\xi)^{\frac{\xi}{\delta\xi}} .$$

is limited and S-continuous for all limited x. Indeed, one proves easily that the Euler formula $\mathcal{E}(\xi) \simeq e^{\xi}$ holds for all limited $\xi \in \mathbb{X}$, and then one may apply Proposition 2.4. A second nonstandard example is the function $\mathcal{F}$ defined on $\mathbb{X}^+$ by

$$\mathcal{F}(\xi) = \prod_{0 \leq \eta < \xi} (1 + \eta\delta\xi) .$$

Then it is easy to verify that $\mathcal{F}(\xi) \simeq e^{\xi^2/2}$ for all limited $\xi \in \mathbb{X}^+$. Again it follows from Proposition 2.4 that $\mathcal{F}$ is of class S^0 on $\mathbb{X}^+$. The discrete functions $\mathcal{E}$ and $\mathcal{F}$ have even more regularity, as will be shown in Section 3.

The class of functions of class S^0 is closed under the usual algebraic operations, provided one does not divide by infinitesimal values. The proof of the next proposition is straightforward.

Proposition 2.7 *Let $A \subset \mathbb{R}$ and $f, g : A \to \mathbb{R}$ be functions of class S^0. Then $-f$, $f + g$ and $f \cdot g$ are of class S^0, and f/g is of class S^0 on all sets $B \subset A$ such that $g \not\simeq 0$ on the limited part of B.*

To functions of class S^0, defined on a possibly discrete subset of $\mathbb{R}$, may be associated standard continuous functions. This transition from the discrete to the continuous, or *continuization*, uses the notion of shadow.

Any limited number y is nearly-equal to a unique standard number x called the *standard part* or the *shadow* of y, and we write $x = {}^\circ y$. For instance, if $\epsilon \simeq 0$, one has ${}^\circ\epsilon = 0$ and ${}^\circ e^{\epsilon} = 1$. To each function f of class S^0 defined on $\mathbb{R}$ one may associate a unique standard function ${}^\circ f$ such that $f(x) \simeq {}^\circ f(x)$ for all limited x. The function ${}^\circ f$ is called the *standard part* or *shadow* of f. In addition, the function ${}^\circ f$ is continuous. This property was already known to Robinson [14]. We state a version of this theorem for discrete functions on a near-continuum.

Theorem 2.8 (Theorem of the continuous shadow, one variable) *Let $\mathbb{Y}$ be a near-continuum. Let $f : \mathbb{Y} \to \mathbb{R}$ be of class S^0. Then there exists a unique standard function ${}^\circ f : \mathbb{R} \to \mathbb{R}$ such that $f(y) \simeq {}^\circ f(y)$ for all limited $y \in \mathbb{Y}$. In fact f is everywhere continuous.*

It is this theorem which yields a general procedure for the transformation of discrete functions into continuous functions. For a proof of the theorem we refer to the literature [14][12][7].

3 Functions of class S^n of one variable

We extend the notion of functions of class S^0 to functions of class S^n for all standard $n \in \mathbb{N}$. Functions of class S^n are not only limited and S-continuous themselves, but also their difference quotients of order m for $m \leq n$. We show that discrete derivation and integration relate functions of class S^n and S^{n+1} in a similar way as ordinary derivation and integration relate functions of class C^n and C^{n+1}. Stability of the (external) set of functions of class S^n appears to hold for all the usual algebraic operations on equidistant near-continua. On general near-continua stability may depend on the nature of the near-continuum. Hence on equidistant near-continua polynomials of standard degree and limited coefficients are of class S^n, and rational functions which are quotients of such polynomials are of class S^n whenever the arguments are not infinitely close to a singularity of the denominator, while such properties may not hold on all near-continua.

Notation 3.1 *We let* $\mathbb{Y} = \phi(\mathbb{X})$ *be a near-continuum, where* $\phi : \mathbb{X} \to \mathbb{R}$ *is a strictly increasing function of class* S^0. *Let* $\psi : \mathbb{Y} \to \mathbb{R}$. *Let* $\eta \in \mathbb{Y}$ *and let* $\xi \in \mathbb{X}$ *be such that* $\eta = \phi(\xi)$. *Depending on the context we use the following notations for the differences of* ψ:

$$\delta\psi(\eta) \equiv \delta\psi(\eta_\xi) \equiv \psi(\eta_\xi + \delta\eta_\xi) - \psi(\eta_\xi) \equiv \psi(\phi(\xi + \delta\xi)) - \psi(\phi(\xi)).$$

In the same spirit we use the following notations for difference quotients of ψ:

$$\begin{aligned} \psi^{[1]}(\eta) &\equiv \frac{\delta\psi}{\delta\eta}(\eta) \equiv \frac{\delta\psi}{\delta\eta_\xi}(\eta_\xi) \equiv \frac{\psi(\eta_\xi + \delta\eta_\xi) - \psi(\eta_\xi)}{\delta\eta_\xi} \\ &\equiv \frac{\psi(\phi(\xi + \delta\xi)) - \psi(\phi(\xi))}{\phi(\xi + \delta\xi) - \phi(\xi)}. \end{aligned}$$

Let $n \in \mathbb{N}, n \geq 1$. *Assume* $\psi^{[n-1]}(\eta)$ *is defined. Then we define*

$$\begin{aligned} \psi^{[n]}(\eta) &\equiv \frac{\delta\psi^{[n-1]}}{\delta\eta}(\eta) \equiv \frac{\delta\psi^{[n-1]}}{\delta\eta_\xi}(\eta_\xi) \equiv \frac{\psi^{[n-1]}(\eta_\xi + \delta\eta_\xi) - \psi^{[n-1]}(\eta_\xi)}{\delta\eta_\xi} \\ &\equiv \frac{\psi^{[n-1]}(\phi(\xi + \delta\xi)) - \psi(\phi(\xi))}{\phi(\xi + \delta\xi) - \phi(\xi)}. \end{aligned}$$

Observe that, if ψ is defined on $A \subseteq \mathbb{Y}$, its n^{th}-order difference quotient $\psi^{[n]}$ is defined on $A^{(n)}$.

Definition 3.2 *Let* $A \subset \mathbb{Y}$ *and* $\psi : A \to \mathbb{R}$, *and let* $n \in \mathbb{N}, n \geq 1$ *be standard. We say that* ψ *is* of class S^n *if* ψ *is of class* S^{n-1} *and* $\psi^{[n]} : A^{(n)} \to \mathbb{R}$ *is of class* S^0.

The next proposition is a first consequence.

Proposition 3.3 *Let* $A \subset \mathbb{Y}$ *and let* $n \in \mathbb{N}, n \geq 1$ *be standard. Let* $\psi : A \to \mathbb{R}$ *be of class* S^n *and assume* $0 \leq m < n$. *Then* ψ *is of class* S^m.

We now show that the difference quotient $\psi^{[1]}$ of a function ψ of class S^n is of class S^{n-1}, implying that the difference quotient $\psi^{[m]}$ of order $m < n$ is a function of class S^{n-m}. Conversely, if the difference quotient of a function ψ of class S^0 is of class S^{n-1}, the function ψ will be of class S^n.

For convenience we consider functions defined on the whole of $\mathbb{Y}$.

Lemma 3.4 (Lemma of the discrete derivative) *Let* $n \in \mathbb{N}$ *be standard and* $\psi : \mathbb{Y} \to \mathbb{R}$ *be of class* S^n. *Then* $\psi^{[1]}$ *is of class* S^{n-1}.

Proof. By external induction. By definition, if ψ is of class S^1, its difference quotient $\psi^{[1]}$ is of class S^0. Assume the property is valid for some standard integer n. Let ψ be a function of class S^{n+1}. By definition ψ is of class S^n, hence by the induction hypothesis $\psi^{[1]}$ is of class S^{n-1}. Moreover $\left(\psi^{[1]}\right)^{[n]} = \psi^{[n+1]}$ is of class S^0. We conclude that $\psi^{[1]}$ is of class S^n. ■

Proposition 3.5 *Let* $m, n \in \mathbb{N}$ *be standard with* $m \leq n$ *and let* $\psi : \mathbb{Y} \to \mathbb{R}$ *be of class* S^n. *Then* $\psi^{[m]}$ *is a function of class* S^{n-m}.

Next lemma is preparatory to a lemma on discrete integration which is analogous to Lemma 3.4, relating functions of class S^n to functions of class S^{n+1}.

Lemma 3.6 *Let $n \in \mathbb{N}$ be standard and $\psi : \mathbb{Y} \to \mathbb{R}$. If $\psi^{[1]}$ is of class S^n and ψ is of class S^0, then ψ is of class S^n.*

Proof. By external induction. The case $n = 0$ is trivial. Assume the property is valid for some standard integer n. Suppose $\psi^{[1]}$ is of class S^{n+1} and ψ is of class S^0. By definition $\psi^{[1]}$ is of class S^n. By the induction hypothesis ψ is of class S^n. Also $\psi^{[n+1]} = \left(\psi^{[1]}\right)^{[n]}$ is of class S^0. We conclude that ψ is of class S^{n+1}. ■

Lemma 3.7 (Lemma of the discrete integral) *Let $n \in \mathbb{N}$ be standard and $\psi : \mathbb{Y} \to \mathbb{R}$. If $\psi^{[1]}$ is of class S^n and ψ is of class S^0, then ψ is of class S^{n+1}.*

Proof. By external induction. Let $n = 0$. If ψ and $\psi^{[1]}$ are of class S^0, by definition ψ is of class S^1. Assume the property is valid for some standard integer n. Suppose $\psi^{[1]}$ is of class S^{n+1} and ψ is of class S^0. By Lemma 3.6 the function ψ is of class S^{n+1}. Moreover $\psi^{[n+2]} = \left(\psi^{[1]}\right)^{[n+1]}$ is of class S^0. We conclude that ψ is of class S^{n+2}. ■

We now consider the stability of the class S^n under algebraic operations. The proof of the first proposition is immediate.

Proposition 3.8 *Let $n \in \mathbb{N}$ be standard and $f, g : \mathbb{Y} \to \mathbb{R}$ be functions of class S^n. Let $a \in \mathbb{R}$ be limited. Then $f + g$ and af are of class S^n.*

We formulate the product rule and the division rule first for functions defined on an equidistant near-continuum $\mathbb{X}$. Then we discuss its validity on general near-continua. Note that

$$\begin{aligned} \frac{\delta(f \cdot g)}{\delta\xi}(\xi) &= \frac{\delta f}{\delta\xi}(\xi) \cdot g(\xi + \delta\xi) + f(\xi) \cdot \frac{\delta g}{\delta\xi}(\xi) \\ &= \frac{\delta f}{\delta\xi}(\xi) \cdot g(\xi) + \frac{\delta f}{\delta\xi}(\xi) \cdot \frac{\delta g}{\delta\xi}(\xi)\,\delta\xi + f(\xi) \cdot \frac{\delta g}{\delta\xi}(\xi). \end{aligned} \tag{2}$$

Proposition 3.9 *Let $n \in \mathbb{N}$ be standard and $f, g : \mathbb{X} \to \mathbb{R}$ be functions of class S^n. Then $f \cdot g$ is of class S^n.*

Proof. By external induction. The case $n = 0$ follows from Proposition 2.7. Assume the property is valid for some standard integer n. Suppose f and g are of class S^{n+1}. By definition f and g are of class S^n. By the Lemma of the discrete derivative $\frac{\delta f}{\delta\xi}$ and $\frac{\delta g}{\delta\xi}$ are of class S^n. Then $\frac{\delta f}{\delta\xi} \cdot g$, $\frac{\delta f}{\delta\xi} \cdot \frac{\delta g}{\delta\xi}$ and $f \cdot \frac{\delta g}{\delta\xi}$ are of class S^n by the induction hypothesis. By (2) and Proposition 3.8 the difference quotient $\frac{\delta(f \cdot g)}{\delta\xi}$ is of class S^n. Because $f \cdot g$ is of class S^0 we deduce from the Lemma of the discrete integral that $f \cdot g$ is of class S^{n+1}. ■

Lemma 3.10 *Let $n \in \mathbb{N}$ be standard and $f : \mathbb{X} \to \mathbb{R}$ be a function of class S^n with appreciable values for limited arguments. Then $\frac{1}{f}$ is of class S^n on all subsets $A^{(n)} \subseteq \mathbb{X}$ such that $f \not\simeq 0$ on the limited part of A.*

Proof. By external induction. The case $n = 0$ follows from Proposition 2.7. Assume the property is valid for some standard integer n. Let f be a function of class S^{n+1} and $A \subseteq \mathbb{X}$ be such that f takes appreciable values for limited arguments $\xi \in A$. By definition f is of class S^n on $A^{(n)}$, so $\dfrac{1}{f}$ is of class S^n on $A^{(n)}$ by the induction hypothesis. Now

$$\frac{\delta\,(1/f)}{\delta\xi}(\xi) \quad = \quad -f^{[1]}(\xi) \cdot \frac{1}{f^2(\xi) - f(\xi)f^{[1]}(\xi)\delta\xi}.$$

Then the Lemma of the discrete derivative, Proposition 3.9 and the fact that $ff^{[1]}\delta\xi \simeq 0$ on the limited part of A imply that $(1/f)^{[1]}$ is of class S^n on $A^{(n)}$. Because $\dfrac{1}{f}$ is of class S^0 on A, the function $\dfrac{1}{f}$ is of class S^{n+1} by the Lemma of the discrete integral. ■

Proposition 3.11 *Let $n \in \mathbb{N}$ be standard and $f, g : \mathbb{X} \to \mathbb{R}$ be functions of class S^n. Then f/g is of class S^n on all subsets $A^{(n)} \subset \mathbb{X}$ such that $g \not\simeq 0$ on the limited part of A.*

Somewhat surprisingly the quadratic function $\psi : \mathbb{Y} \to \mathbb{R}$ defined by $\psi(\eta) = \eta^2$ does not need to be of class S^2 for arbitrary $\mathbb{Y}$ which is the image of $\mathbb{X}$ by some strictly increasing function ϕ of class S^0. To see this, let such a function ϕ satisfy $\phi(0) = 1$, $\phi(\delta\xi) = 1 + \sqrt{\delta\xi}$ and $\phi(2\delta\xi) = 1 + \sqrt{\delta\xi} + \delta\xi$. The function ψ is of class S^0 on $\left\{1, 1 + \sqrt{\delta\xi}, 1 + \sqrt{\delta\xi} + \delta\xi\right\}$, and also of class S^1, since

$$\begin{aligned}\frac{\psi\left(1 + \sqrt{\delta\xi} + \delta\xi\right) - \psi\left(1 + \sqrt{\delta\xi}\right)}{\delta\xi} &= 2 + 2\sqrt{\delta\xi} + \delta\xi \simeq 2 + \sqrt{\delta\xi} \\ &= \frac{\psi\left(1 + \sqrt{\delta\xi}\right) - \psi\left(1\right)}{\sqrt{\delta\xi}}.\end{aligned}$$

But ψ is not of class S^2, for

$$\begin{aligned}&\frac{\psi\left(1 + \sqrt{\delta\xi} + \delta\xi\right) - \psi\left(1 + \sqrt{\delta\xi}\right)}{\delta\xi} - \frac{\psi\left(1 + \sqrt{\delta\xi}\right) - \psi\left(1\right)}{\sqrt{\delta\xi}} \\ &\qquad = \frac{2 + 2\sqrt{\delta\xi} + \delta\xi - (2 + \sqrt{\delta\xi})}{\delta\xi} = \frac{1}{\sqrt{\delta\xi}} + 1,\end{aligned}$$

which is unlimited. This example only apparently contradicts Proposition 3.7, since $\psi^{[1]}(\phi(\xi)) = 2\phi(\xi) + \delta\phi(\xi)$. This shows that ψ is a discrete primitive of $\eta \longmapsto 2\eta + \delta\eta$, which is here not of class S^1, and not a discrete primitive of $\eta \longmapsto 2\eta$, which is of class S^n for all standard n.

This example shows also that the product rule does not hold in all generality: if f and g are of class S^n on some set $\mathbb{Y}$ for some limited n, the product $f \cdot g$ does not need to be of class S^n. Is is easy to verify that the division rule also does not hold in general. We will show in Section 5.2 that the rules do hold provided $\mathbb{Y}$ is the image of $\mathbb{X}$ by some function ϕ which is of class S^n itself.

On the other hand it follows from the Propositions 3.8 and 3.9 that polynomials of standard degree with limited coefficients defined on the equidistant continuum $\mathbb{X}$ are of class S^n for all standard n. Then by Proposition 3.11 rational functions which are quotients of polynomials of standard degree and limited coefficients are of class S^n on all subsets A of $\mathbb{X}$, whose elements are not infinitely close to a limited singularity of the denominator.

We verify that the two examples of discrete functions $\mathcal{E}$ and $\mathcal{F}$ defined on (part of) $\mathbb{X}$ introduced in Section 2 are of class S^n for all standard n.

Indeed, because

$$\frac{\delta \mathcal{E}(\xi)}{\delta \xi} = \mathcal{E}(\xi)$$

and $\mathcal{E}$ is of class S^0, its discrete derivative $\frac{\delta \mathcal{E}}{\delta \xi}$ is also of class S^0. This implies that $\mathcal{E}$ is of class S^1. Then $\frac{\delta \mathcal{E}}{\delta \xi}$ is also class S^1, so by the Lemma of the discrete integral $\mathcal{E}$ is of class S^2. With external induction, applying this procedure to the induction step, one shows that the function $\mathcal{E}$ is of class S^n for all standard $n \in \mathbb{N}$.

Secondly, for $\xi \in \mathbb{X}^+$

$$\frac{\delta \mathcal{F}(\xi)}{\delta \xi} = \xi \cdot \mathcal{F}(\xi).$$

We already saw that $\mathcal{F}$ is of class S^0 on $\mathbb{X}^+$. Because the monomial ξ is of class S^n for all standard n and these classes are stable under multiplication, we may apply the same method as above to show that $\mathcal{F}$ is of class S^n on $\mathbb{X}^+$, for all standard n.

4 Transition from the discrete to the continuous in one variable on equidistant near-continua.

Let $\mathbb{X}$ be an equidistant near-continuum as above. We consider first discrete functions of class S^n on the whole of $\mathbb{X}$ and extend the Theorem of the continuous shadow to transitions from the discrete to the continuous of higher order of regularity.

The case for differentiability of first order had already been proved earlier [7] [3].

Theorem 4.1 (Theorem of the differentiable shadow) *Let $\phi : \mathbb{X} \to \mathbb{R}$ be of class S^1. Then its shadow f is a real function of class C^1 and $\phi^{[1]}(\xi) \simeq f'(\xi)$ for all limited $\xi \in \mathbb{X}$.*

N.B. The definition of function of class S^1 in [7] is different of ours. The essential difference is that it concerns only functions defined on $\mathbb{R}$. However, this difference is not very important since a discrete function defined on $\mathbb{X}$ can always be appropriately extended to a real function $\overline{f}$ of class S^1 defined on the whole of $\mathbb{R}$.

The proof that the shadow of a function of class S^n defined on $\mathbb{X}$ is a function of class C^n is contained in [10] and [5] and will be repeated here.

Theorem 4.2 *Let $n \in \mathbb{N}$ be standard and let $\phi : \mathbb{X} \to \mathbb{R}$ be of class S^n. Then its shadow is a function $f : \mathbb{R} \to \mathbb{R}$ of class C^n and $\phi^{[n]}(\xi) \simeq f^{(n)}(\xi)$ for all limited $\xi \in \mathbb{X}$.*

Proof. By external induction. If $n = 0$, by the Theorem of the continuous shadow ${}^\circ f$ is a real function of class C^0 and $\phi(\xi) \simeq f(\xi)$ for all limited $\xi \in \mathbb{X}$. Assume the property is valid for some standard integer n. Let ϕ be a function of class S^{n+1}. By the Lemma of the discrete derivative the function $\phi^{[1]}$ is of class S^n. By definition $\phi^{[n+1]}$ is of class S^0. By Proposition 3.5 the function $\phi^{[n]}$ is of class S^1. By the induction hypothesis its shadow equals $f^{(n)}$. By the Theorem of the differentiable shadow $f^{(n)}$ is continuously differentiable and $\left(\phi^{[n]}\right)^{[1]}(\xi) \simeq \left(f^{(n)}\right)'(\xi)$ for all limited $\xi \in \mathbb{X}$. Hence $\phi^{[n+1]}(\xi) \simeq f^{(n+1)}(\xi)$ for all limited $\xi \in \mathbb{X}$. ■

As a consequence we obtain that for pairs of functions of class S^n the property of near-equality is hereditary to discrete derivatives up to order n.

Theorem 4.3 *Let $n \in \mathbb{N}$ be standard and $f, g : \mathbb{X} \to \mathbb{R}$ be two functions of class S^n such that $f(\xi) \simeq g(\xi)$ for all limited $\xi \in \mathbb{X}$. Then $f^{[n]}(\xi) \simeq g^{[n]}(\xi)$ for all limited $\xi \in \mathbb{X}$.*

Proof. By the Theorem of the continuous shadow ${}^\circ f = {}^\circ g$. Let $n > 0$, else nothing has to be proved. Then by Theorem 4.2

$$f^{[n]}(\xi) \simeq {}^\circ f^{(n)}(\xi) = {}^\circ g^{(n)}(\xi) \simeq g^{[n]}(\xi).$$

■

5 Transition from the discrete to the continuous on general near-continua

We consider now functions of class S^n on possibly non-equidistant near-continua. Next theorem gives conditions for near-equality of such functions and their difference quotients to their shadows and the respective derivatives of these shadows.

Theorem 5.1 *Let $n \in \mathbb{N}$ be standard. Let $\overline{\xi} \in \mathbb{X}$ be limited. Let $\phi : \mathbb{X} \to \mathbb{R}$ be a function of class S^n such that $\phi^{[1]}(\overline{\xi}) \not\simeq 0$. Let $\psi : \phi(\mathbb{X}) \to \mathbb{R}$ be a function of class S^n. Let $\chi = \psi \circ \phi$. Then there exists a quasi-interval $[\alpha \cdot\cdot \beta]$ with $\alpha, \beta \in \mathbb{X}$ limited and $\alpha \lessapprox \overline{\xi} \lessapprox \beta$ such that $\phi^{[1]}(\xi) \not\simeq 0$ for all $\xi \in [\alpha \cdot\cdot \beta]$, and the function χ is of class S^n on $[\alpha \cdot\cdot \beta]$, with for all $\xi \in [\alpha \cdot\cdot \beta]$*

$$\chi^{[n]}(\xi) \simeq \sum_{k_1+2k_2+\ldots+nk_n=n} \frac{n!}{k_1!k_2!\ldots k_n!} \psi^{[k_1+k_2+\ldots+k_n]}(\phi(\xi)) \prod_{i=1}^{n} \left(\frac{\phi^{[i]}(\xi)}{i!} \right)^{k_i}. \tag{3}$$

Moreover, let $f = {}^\circ\phi$, $g = {}^\circ\psi$ and $h = {}^\circ\chi$. Let $a = {}^\circ\alpha$ and $b = {}^\circ\beta$. Then g is of class C^n on $[f(a), f(b)]$ with $g^{(n)} = {}^\circ\left(\psi^{[n]}\right)$ and h is of class C^n on $[a, b]$, with for all $x \in [a, b]$

$$h^{(n)}(x) = \sum_{k_1+2k_2+\ldots+nk_n=n} \frac{n!}{k_1!k_2!\ldots k_n!} g^{(k_1+k_2+\ldots+k_n)}(f(x)) \prod_{i=1}^{n} \left(\frac{f^{(i)}(x)}{i!} \right)^{k_i}. \tag{4}$$

Formula (4) is known as the formula of Faà di Bruno. It extends the Chain Rule to derivatives of higher order. Thus formula (3) is a discrete, approximative version of the formula of Faà di Bruno.

One of the consequences of the theorem above is that a sufficient condition for a function ψ to be of class S^n on some near-continuum $\mathbb{Y} \subset \mathbb{R}$, with the n^{th}-order difference quotient infinitely close to the n^{th}-order derivative of its shadow, is that $\mathbb{Y}$ is the image of an equidistant near-continuum $\mathbb{X} \subset \mathbb{R}$ by a (locally) strictly monotone function ϕ which is itself of class S^n. We already saw that absence of such a regularity condition may even lead to difference quotients with infinitely large values, which certainly are not nearly-equal to derivatives of ${}^\circ\psi$. A near-continuum $\mathbb{Y} \subset \mathbb{R}$ which is the image of an equidistant near-continuum $\mathbb{X} \subset \mathbb{R}$ by a monotone function ϕ of class S^n will be called *a near-continuum of class S^n*.

Below we prove formula (3) along the lines of the proof of De la Vallée-Poussin (see also [11]) of the usual formula of Faà di Bruno (4).

The proof of De la Vallée-Poussin is by induction. The first step consists in proving that the coefficients of $g^{(k_1+k_2+\ldots+k_n)}(f(x)) \prod_{i=1}^{n} \left(\frac{f^{(i)}(x)}{i!} \right)^{k_i}$ are integers, and independent of f and g. This means that the coefficients may be determined for convenient special functions, in fact

powers of polynomials of the form $(a_1x + a_2x^2 + ... + a_nx^n)^k$. This step uses the multinomial expansion

$$(x_1 + x_2 + ... + x_n)^k = \sum_{k_1+k_2+...+k_n=k} \frac{k!}{k_1!k_2!...k_n!} x_1^{k_1} x_2^{k_2} ... x_n^{k_n}.$$

Finally the powers of polynomials are repeatedly differentiated.

In the discrete case some complications arise from the fact that difference quotients are not always taken at ξ, but also at points $\xi + \theta\delta\xi$, with $\theta > 0$. Hence the product of powers at the end of the formula of Faà di Bruno transform in products of products; but we will show that the latter products are infinitely close to the corresponding powers taken at ξ.

5.1 Properties of the composition function

Convention, notations. Let $n \in \mathbb{N}$ be standard. Let $\overline{\xi} \in \mathbb{X}$ be limited. Let $\phi : \mathbb{X} \to \mathbb{R}$ be a function of class S^n such that $\phi^{[1]}(\overline{\xi}) \not\approx 0$. As a consequence of the principle of Fehrele [7] there exists a quasi-interval $[\alpha \cdot \cdot \beta]$ with $\alpha, \beta \in \mathbb{X}$ limited and $\alpha \lessapprox \overline{\xi} \lessapprox \beta$ such that $\phi^{[1]}(\xi) \not\approx 0$ for all $\xi \in [\alpha \cdot \cdot \beta]$. Let $1 \leq m \leq n$. Without restriction of generality we assume that $\phi^{[1]}(\xi) \gtrapprox 0$ for all $\xi \in [\alpha \cdot \cdot \beta]$ and that $\phi^{[m]}(\xi)$ is defined on $[\alpha \cdot \cdot \beta]$ instead of $[\alpha \cdot \cdot \beta]^{(m-1)}$. Notice that $\phi(\alpha) \lessapprox \phi(\overline{\xi}) \lessapprox \phi(\beta)$.

We use the following notations. We let $\psi : \phi(\mathbb{X}) \to \mathbb{R}$ be a function of class S^n and let $\chi = \psi \circ \phi$. All functions being of class S^0, their shadows are well-defined and we let $f = {}^\circ\phi$, $g = {}^\circ\psi$ and $h = {}^\circ\chi$. We let $a = {}^\circ\alpha$ and $b = {}^\circ\beta$, and $c \equiv f(a) = {}^\circ(\phi(\alpha))$ and $d \equiv f(b) = {}^\circ(\phi(\beta))$. Then f is of class C^n on $[a, b]$, the function g is at least of class C^0 on $[f(a), f(b)]$ and $h = f \circ g$ is at least of class C^0 on $[a, b]$.

Lemma 5.2 *Let $n \in \mathbb{N}$ be standard and $\xi \in [\alpha \cdot \cdot \beta]$. Then $\chi^{[n]}(\xi)$ is the sum of a standard finite number of terms consisting of products with a standard finite number of factors of the form*

$$\psi^{[k]}(\phi(\xi)) \prod_{j=1}^{k_1} \phi^{[1]}(\xi + \theta_{1j}\delta\xi) \prod_{j=1}^{k_2} \phi^{[2]}(\xi + \theta_{2j}\delta\xi) ... \prod_{j=1}^{k_n} \phi^{[n]}(\xi + \theta_{nj}\delta\xi), \tag{5}$$

where $k_1 + k_2 + ... + k_n \equiv k$ and $k_1 + 2k_2 + ... + nk_n = n$, and $0 \leq \theta_{ij} \leq n$ for all i, j with $1 \leq j \leq k_i$ and $1 \leq i \leq n$.

Proof. By external induction. Firstly, let $n = 1$. Then $k = k_1 = 1 \cdot k_1 = n = 1$ and

$$\chi^{[1]}(\xi) = \psi^{[1]}(\phi(\xi))\phi^{[1]}(\xi) = \psi^{[k]}(\phi(\xi))\phi^{[k_1]}(\xi + \theta_{11}\delta\xi), \tag{6}$$

with $\theta_{11} = 0 \leq n$.

Assume the property has been proved up to some standard n. Applying the product rule (2), starting with the discrete derivative of $\psi^{[k]}(\phi(\xi))$, the first term, say $\tau_{k+1}(\xi)$, becomes

$$\begin{aligned} \tau_{k+1}(\xi) \quad = \quad & \psi^{[k+1]}(\phi(\xi)) \left(\phi^{[1]}(\xi) \prod_{j=1}^{k_1} \phi^{[1]}(\xi + (\theta_{1j} + 1)\delta\xi) \right) \times \\ & \prod_{j=1}^{k_2} \phi^{[2]}(\xi + (\theta_{2j} + 1)\delta\xi) ... \prod_{j=1}^{k_n} \phi^{[n]}(\xi + (\theta_{nj} + 1)\delta\xi). \end{aligned}$$

We define $k'_1 = k_1 + 1, k'_2 = k_2, ..., k'_n = k_n$ and $k'_{n+1} = 0$. We let $k' = k + 1$ and we define $\theta'_{11} = 0$, $\theta'_{1j} = \theta_{1j-1} + 1$ for all j with $2 \leq j \leq k_1$, $\theta'_{ij} = \theta_{ij} + 1$ for all i, j with $1 \leq j \leq k_i$ and

$2 \leq i \leq n$ and $\theta'_{n+1j} = 0$. Then $k'_1 + k'_2 + ... + k'_n \equiv k'$ and $k'_1 + 2k'_2 + ... + (n+1)k'_{n+1} = n+1$, and $0 \leq \theta'_{ij} \leq n+1$ for all i, j with $1 \leq j \leq k'_i$ and $1 \leq i \leq n+1$. Further

$$\begin{aligned}\tau_{k+1}(\xi) &= \psi^{[k']}(\phi(\xi)) \prod_{j=1}^{k'_1} \phi^{[1]}(\xi + \theta'_{1j}\delta\xi) \prod_{j=1}^{k'_2} \phi^{[2]}(\xi + \theta'_{2j}\delta\xi) \times \\ &\quad ... \times \prod_{j=1}^{k'_{n+1}} \phi^{[n+1]}(\xi + \theta'_{n+1j}\delta\xi).\end{aligned}$$

Continuing to apply the product rule, let $1 \leq m \leq k$, and suppose it is the turn of the m^{th} factor of

$$\Pi(\xi) \equiv \prod_{j=1}^{k_1} \phi^{[1]}(\xi + \theta_{1j}\delta\xi) \prod_{j=1}^{k_2} \phi^{[2]}(\xi + \theta_{2j}\delta\xi) ... \prod_{j=1}^{k_n} \phi^{[n]}(\xi + \theta_{nj}\delta\xi)$$

to be differentiated discretely, which is, say, of the form $\phi^{[i]}(\xi + \theta_{ij}\delta\xi)$. Its discrete derivative is $\phi^{[i+1]}(\xi + \theta_{ij}\delta\xi)$. We define $k'_1 = k_1, ..., k'_{i-1} = k_{i-1}, k'_i = k_i - 1, k'_{i+1} = k_{i+1} + 1, k'_{i+2} = k_{i+2}, ..., k'_n = k_n$, and $k'_{n+1} = 0$ whenever $m \leq k - k_n$ and $k'_{n+1} = 1$ whenever $m > k - k_n$. We let $k' = k$ and we define new θ'_{ij} by adding nothing to the θ_{ij} occurring up to the m^{th} factor, and adding 1 to to the θ_{ij} occurring after this factor, with $\theta'_{n+1j} = 0$. The term corresponding to the discrete derivative of the m^{th} factor becomes

$$\psi^{[k']}(\phi(\xi)) \prod_{j=1}^{k'_1} \phi^{[1]}(\xi + \theta'_{1j}\delta\xi) \prod_{j=1}^{k'_2} \phi^{[2]}(\xi + \theta'_{2j}\delta\xi) ... \prod_{j=1}^{k'_{n+1}} \phi^{[n+1]}(\xi + \theta_{n+1j}\delta\xi),$$

with $k'_1 + k'_2 + ... + k'_n \equiv k'$,

$$\begin{aligned}k'_1 + 2k'_2 + ... + ik'_i + (i+1)k'_{i+1} + (i+2)k'_{i+2} + ... + (n+1)k'_{n+1} = \\ k_1 + 2k_2 + ... + ik_i - i + (i+1)k_{i+1} + i + 1 + (i+2)k_{i+1} + ... \\ + (n+1)k_{n+1} = n+1,\end{aligned}$$

and $0 \leq \theta'_{ij} \leq n+1$ for all $1 \leq j \leq k'_i$ and $1 \leq i \leq n+1$. Every product has at most $n+1$ factors, and the number of terms in the induction step increases with at most $n+1$. In conclusion, the number of terms remains standard finite, each term having at most a standard number of factors. ■

Proposition 5.3 *Let $n \in \mathbb{N}$ be standard. Let $\phi : [\alpha \cdot \cdot \beta] \to \mathbb{R}$ and $\psi : [\phi(\alpha) \cdot \cdot \phi(\beta)] \to \mathbb{R}$ be functions of class S^n. Then χ is of class S^n on $[\alpha \cdot \cdot \beta]$ and h is of class C^n on $[a, b]$.*

Proof. With external induction. The composition of two functions of class S^0 is clearly of class S^0. Assume the proposition is proved for $n-1$. Then χ is of class S^{n-1} on $[\alpha \cdot \beta]$. For each k, the first factor $\psi^{[k]} \circ \phi$ of (5) is the composition of two functions of class S^0, hence is of class S^0 in ξ. All the remaining factors are of class S^0. Because there is a standard finite number of such factors, their product is of class S^0. Because $\chi^{[n]}$ is a standard finite sum of such products, it is also of class S^0. Hence χ is of class S^n on $[\alpha \cdot \beta]$. By Theorem 4.2 its shadow h is of class C^n on $[a, b]$. ■

Lemma 5.4 *Let $n \in \mathbb{N}$ be standard. Let $\phi : [\alpha \cdot\cdot \beta] \to \mathbb{R}$ and $\psi : [\phi(\alpha) \cdot\cdot \phi(\beta)] \to \mathbb{R}$ be functions of class S^n. Let $\xi \in [\alpha \cdot\cdot \beta]$. Then there are constants $C_{n,k_1,...,k_n}$, which are standard integers independent from ϕ and ψ, such that $\chi^{[n]}(\xi)$ is of the form*

$$\chi^{[n]}(\xi) \simeq \sum_{k_1+2k_2+...+nk_n=n} C_{n,k_1,...,k_n} \psi^{[k_1+k_2+...+k_n]}(\phi(\xi)) \prod_{i=1}^{n} (\phi^{[i]}(\xi))^{k_i}, \tag{7}$$

whenever $\xi \in [\alpha \cdot\cdot \beta]$.

Proof. Lemma 5.2 states that $\chi^{[n]}(\xi)$ is a standard finite sum of standard finite products of the form

$$\psi^{[k]}(\phi(\xi)) \prod_{j=1}^{k_1} \phi^{[1]}(\xi + \theta_{1j}\delta\xi) \prod_{j=1}^{k_2} \phi^{[2]}(\xi + \theta_{2j}\delta\xi) ... \prod_{j=1}^{k_n} \phi^{[n]}(\xi + \theta_{nj}\delta\xi),$$

where $k_1 + k_2 + ... + k_n = k$ and $k_1 + 2k_2 + ... + nk_n = n$, and $0 \leq \theta_{ij} \leq n$ for all i, j with $1 \leq j \leq k_i$ and $1 \leq i \leq n$; the form of the products does not depend on the individual properties of ϕ and ψ. Because all θ_{ij} satisfy $0 \leq \theta_{ij} \leq n$, all arguments $\xi + \theta_{ij}\delta\xi$ of the factors of the products above are nearly equal to ξ. Now $\psi^{[k]}(\phi(\xi))$ as a composition of two functions of class S^0 is limited, all functions $\phi^{[j]}$ with $1 \leq j \leq n$ are of class S^0 and there is are only a standard finite number of factors in the products above, hence each product satisfies

$$\begin{aligned} & \psi^{[k]}(\phi(\xi)) \prod_{j=1}^{k_1} \phi^{[1]}(\xi + \theta_{1j}\delta\xi) \prod_{j=1}^{k_2} \phi^{[2]}(\xi + \theta_{2j}\delta\xi) ... \prod_{j=1}^{k_n} \phi^{[n]}(\xi + \theta_{nj}\delta\xi) \\ \simeq \quad & \psi^{[k]}(\phi(\xi) \prod_{i=1}^{n} \left(\phi^{[i]}(\xi)\right)^{k_i}. \end{aligned}$$

This operation has the possible effect of regrouping terms with distinct products in some packet of identical powers, but the number of terms resulting from the regrouping does not depend on ϕ and ψ. Combining, the number of these terms in (7) is a function of n and $k_1, ..., k_n$ only, that we may denote by $C_{n,k_1,...,k_n}$. ■

Observing that the coefficients $C_{n,k_1,...,k_n}$ do not depend on ϕ and ψ, we choose convenient functions to determine them. Like in the proof of the ordinary Faà di Bruno Theorem, we define for $k, n \in \mathbb{N}$ and $a_1, \cdots, a_n \in \mathbb{R}$

$$\left\{ \begin{array}{lcl} p(\xi) & = & a_1\xi + ... + a_n\xi^n \\ m(\eta) & = & \eta^k \\ c(\xi) & = & (m \circ p)(\xi) = (a_1\xi + ... + a_n\xi^n)^k. \end{array} \right. \tag{8}$$

It follows from the multinomial expansion that for all $\xi \in \mathbb{X}$

$$c(\xi) = \sum_{k_1+k_2+...+k_n=k} \frac{k!}{k_1!k_2!...k_n!} \prod_{i=1}^{n} a_i^{k_i} \xi^{k_1+2k_2+...+nk_n}. \tag{9}$$

We start with some lemmas.

Lemma 5.5 *Let $k, n \in \mathbb{N}$ be standard with $k, n \geq 1$. Then*

$$\frac{\delta^k \xi^n}{\delta \xi^k} \left\{ \begin{array}{lll} \simeq & 0 & \xi = 0,\ k < n \\ = & n! & k = n \\ = & 0 & k > n. \end{array} \right.$$

Proof. By external induction in n. If $n = 1$ the results are trivial. Assume the lemma has been proved for $n-1$. Observe that

$$\frac{\delta\xi^n}{\delta\xi} = n\xi^{n-1} + \delta\xi q(\xi),$$

where q is a polynomial of degree $n-2$ with limited coefficients, containing powers of $\delta\xi$. Now

$$\frac{\delta^k\xi^n}{\delta\xi^k} = n\frac{\delta^{k-1}\xi^{n-1}}{\delta\xi^{k-1}} + \delta\xi\frac{\delta^{k-1}q(\xi)}{\delta\xi^{k-1}}.$$

As long as $k < n$ the induction hypothesis yields $\frac{\delta^{k-1}\xi^{n-1}}{\delta\xi^{k-1}} \simeq 0$ in $\xi = 0$, and also $\delta\xi\frac{\delta^{k-1}q(\xi)}{\delta\xi^{k-1}} \simeq 0$ in $\xi = 0$, for the coefficients of $\frac{\delta^{k-1}q(\xi)}{\delta\xi^{k-1}}$ will be at most limited. Hence $\frac{\delta^k\xi^n}{\delta\xi^k} \simeq 0$ in $\xi = 0$. If $k = n$, we find by the induction hypothesis that

$$\frac{\delta^n\xi^n}{\delta\xi^n} = n(n-1)! + \delta\xi \cdot 0 = n!.$$

Because $\frac{\delta^n\xi^n}{\delta\xi^n}$ is a constant, $\frac{\delta^k\xi^n}{\delta\xi^k} = 0$ for $k > n$. ■

Lemma 5.6 *Let $n \in \mathbb{N}, n \geq 1$ and $a_1, ..., a_n \in \mathbb{R}^n$ be standard. Then*

$$\frac{\delta^n c^n}{\delta\xi^n}(0) \simeq \sum_{\substack{k_1+2k_2+...+nk_n=n \\ k_1+k_2+...+k_n=k}} \frac{k!}{k_1!k_2!...k_n!} n! \prod_{i=1}^{n} a_i^{k_i}.$$

Proof. By (9) and Lemma 5.5 non-infinitesimal contributions to $\frac{\delta^n c^n}{\delta\xi^n}(0)$ occur only for $k_1, k_2, ..., k_n$ such that $k_1 + 2k_2 + ... + nk_n = n$; in the latter case the coefficient $\frac{k!}{k_1!k_2!...k_n!}\prod_{i=1}^n a_i^{k_i}$ in (9) corresponding to $k \equiv k_1 + k_2 + ... + k_n$ is multiplied by $n!$. ■

Proposition 5.7 *Let $n \in \mathbb{N}$ be standard and let for $k_1, ..., k_n$ with $k_1 + 2k_2 + ... + nk_n = n$ the constants $C_{n,k_1,...,k_n}$ be given by Lemma 5.4. Then*

$$C_{n,k_1,...,k_n} = \frac{n!}{k_1!k_2!...k_n!}\prod_{i=1}^{n}\frac{1}{i!^{k_i}}.$$

.

Proof. Because the $C(n, k_1, ..., k_n)$ of Lemma 5.4 do not depend on the choice of the functions, we choose to determine them for the functions $p : \mathbb{X} \to \mathbb{R}$, $m : \mathbb{Y} \to \mathbb{R}$ and $c = m \circ p : \mathbb{X} \to \mathbb{R}$ as defined in (8). Let $a_1, ..., a_n \in \mathbb{R}^n$ be standard. Notice that for $1 \leq i \leq n$ one has $p^{[i]}(0) \simeq i!a_i$. Let $\eta = p(0) = 0$. Then, with $k \equiv k_1 + k_2 + ... + k_n$, it holds that $m^{[k]}(\eta) = k!$. Then it follows from Lemma 5.6 that

$$\begin{aligned}
c^{[n]}(0) &\simeq \sum_{\substack{k_1+2k_2+...+nk_n=n \\ k_1+k_2+...+k_n=k}} C_{n,k_1,...,k_n} m^{[k_1+k_2+...+k_n]}(0) \prod_{i=1}^{n} (p^{[i]}(0))^{k_i} \\
&= \sum_{\substack{k_1+2k_2+...+nk_n=n \\ k_1+k_2+...+k_n=k}} C_{n,k_1,...,k_n} k! \prod_{i=1}^{n} (i!a_i)^{k_i} \\
&\simeq \sum_{\substack{k_1+2k_2+...+nk_n=n \\ k_1+k_2+...+k_n=k}} \frac{k!}{k_1!k_2!...k_n!} n! \prod_{i=1}^{n} i!^{k_i} a_i^{k_i}.
\end{aligned}$$

We define two polynomials u and v in the variables $a_1, ..., a_n$ by

$$\begin{cases} u(a_1, ..., a_n) &= \sum_{\substack{k_1+2k_2+...+nk_n=n,\\ k_1+k_2+...+k_n=k}} C_{n,k_1,...,k_n} k! \prod_{i=1}^n i!^{k_i} \prod_{i=1}^n a_i^{k_i} \\ v(a_1, ..., a_n) &= \sum_{\substack{k_1+2k_2+...+nk_n=n,\\ k_1+k_2+...+k_n=k}} \frac{k!}{k_1!k_2!..k_n!} n! \prod_{i=1}^n a_i^{k_i}. \end{cases}$$

Then $u(a_1, ..., a_n) \simeq v(a_1, ..., a_n)$ for all standard $a_1, ..., a_n \in \mathbb{R}$. The polynomial v is clearly standard, and also the polynomial u, because the numbers $C_{n,k_1,...,k_n}$ are standard integers. So $u(a_1, ..., a_n) = v(a_1, ..., a_n)$ by the Principle of Carnot. Then $u = v$ by Transfer. This means that their coefficients must be equal. This proves the proposition. ■

Corollary 5.8 (Infinitesimal Faà di Bruno Theorem). *Let $n \in \mathbb{N}$ be standard. Let $\phi : [\alpha \cdot\cdot \beta] \to \mathbb{R}$ and $\psi : [\phi(\alpha) \cdot\cdot \phi(\beta)] \to \mathbb{R}$ be functions of class S^n. Let $\xi \in [\alpha \cdot\cdot \beta]$. Then under the conventions mentioned above, for all $\xi \in [\alpha \cdot\cdot \beta]$*

$$\chi^{[n]}(\xi) \simeq \sum_{k_1+2k_2+...+nk_n=n} \frac{n!}{k_1!k_2!...k_n!} \psi^{[k_1+k_2+...+k_n]}(\phi(\xi)) \prod_{i=1}^n \left(\frac{\phi^{[i]}(\xi)}{i!} \right)^{k_i}.$$

5.2 Continuization on near-continua of class S^n

We prove the transition discrete-continuous of Theorem 5.1 first for the case that $n = 1$. This corresponds to a sort of Chain Rule for discrete functions.

Theorem 5.9 *Let $\phi : [\alpha \cdot\cdot \beta] \to \mathbb{R}$ be a function of class S^1 such that $\phi^{[1]}(\xi) \not\simeq 0$. for all $\xi \in [\alpha \cdot\cdot \beta]$. Let $\psi : [\phi(\alpha) \cdot\cdot \phi(\beta)] \to \mathbb{R}$ be a function of class S^1. Let $\chi = \psi \circ \phi$. Then χ is of of class S^1 and for all $\xi \in [\alpha \cdot\cdot \beta[$*

$$\chi^{[1]}(\xi) = \psi^{[1]}(\phi(\xi)) \phi^{[1]}(\xi). \tag{10}$$

Moreover, let $f = {}^\circ\phi$, $g = {}^\circ\psi$ and $h = {}^\circ\chi$. Let $a = {}^\circ\alpha$ and $b = {}^\circ\beta$. Then g is of class C^1 on $[f(a), f(b)]$ with $g' = {}^\circ\left(\psi^{[1]}\right)$ and h is of class C^1 on $[a, b]$, with for all $x \in [a, b]$

$$h'(x) = g'(f(x)) f'(x). \tag{11}$$

Proof. Formula (10) is evident (see also (6)). By Proposition 5.3 the function $\chi : [\alpha \cdot\cdot \beta] \to \mathbb{R}$ is of class S^1. By the Theorem of the differentiable shadow the functions f and h are of class C^1 on $[a, b]$, with $f' = {}^\circ\left(\phi^{[1]}\right)$ non-zero and $h' = {}^\circ\left(\chi^{[1]}\right)$. Observe that f^{-1} is of class C^1 on $[f(a), f(b)]$ with $(f^{-1})'(y) = 1/f'(f^{-1}(y)) \neq 0$. By the Theorem of the continuous shadow the function g is of class C^0 on $[f(a), f(b)]$ with $h(x) = g(f(x))$ for all $x \in [a, b]$. Then for all $y \in [f(a), f(b)]$

$$g(y) = h(f^{-1}(y)).$$

By the usual Chain Rule g is of class C^1 on $[f(a), f(b)]$. Then the Chain Rule may also be applied to the composition $h = f \circ g$ and yields (11) for all $x \in [a, b]$.

Let $y \in [f(a), f(b)]$ be standard and $\eta \in [\phi(\alpha) \cdot\cdot \phi(\beta)[$ be such that $\eta \simeq y$. Let $\xi \in [\alpha \cdot\cdot \beta[$ be such that $\phi(\xi) = \eta$. Put $x = f^{-1}(y)$. Then, applying the Theorem of the differentiable shadow to h and f, formula (10), the fact that $\psi^{[1]}$ is of class S^0 and the Theorem of the continuous shadow

$$g'(y) = \frac{h'(x)}{f'(x)} \simeq \frac{\chi^{[1]}(\xi)}{\phi^{[1]}(\xi)} = \psi^{[1]}(\eta) \simeq \left({}^\circ\left(\psi^{[1]}\right)\right)(y).$$

Hence $g'(y) = \left(\circ\left(\psi^{[1]}\right)\right)(y)$ for all $y \in [f(a), f(b)]$ by the Principle of Carnot and by Transfer, hence $g' = \circ\psi^{[1]}$. ■

Next theorem is formulated for arbitrary standard $n \in \mathbb{N}$ and presupposes that the conventions at the beginning of Section 5.1 hold.

Theorem 5.10 *Let $n \in \mathbb{N}$ be standard. Let $\phi : [\alpha \cdot \cdot \beta] \to \mathbb{R}$ and $\psi : [\phi(\alpha) \cdot \cdot \phi(\beta)] \to \mathbb{R}$ be functions of class S^n. Let $\xi \in [\alpha \cdot \cdot \beta]$. Then g is of class C^n on $[f(a), f(b)]$, with $g^{(n)} = \circ\left(\psi^{[n]}\right)$, and h is of class C^n on $[a, b]$, with for all $x \in [a, b]$*

$$h^{(n)}(x) = \sum_{k_1+2k_2+\ldots+nk_n=n} \frac{n!}{k_1!k_2!\ldots k_n!} g^{(k_1+k_2+\ldots+k_n)}(f(x)) \prod_{i=1}^{n} \left(\frac{f^{(i)}(x)}{i!}\right)^{k_i}. \tag{12}$$

Proof. By external induction. The case $n = 1$ is contained in Theorem 5.9. Assume the theorem is proved for $n-1$. We follow, mutatis mutandis, the lines of the proof of Theorem 5.9. We prove first that g is of class C^n. By Theorem 4.2 the functions f and h are of class C^n on $[a, b]$, with $f' = \circ\left(\phi^{[1]}\right)$ non-zero, $f^{(m)} = \circ\left(\phi^{[m]}\right)$ for all m with $2 \le m \le n$ and $h^{(n)} = \circ\left(\chi^{[n]}\right)$. Observe that f^{-1} is of class C^n on $[f(a), f(b)]$ with $(f^{-1})'(y) = 1/f'(f^{-1}(y)) \neq 0$. By the Theorem of the continuous shadow the function g is of class C^0 on $[f(a), f(b)]$ with $h(x) = g(f(x))$ for all $x \in [a, b]$. Then for all $y \in [f(a), f(b)]$

$$g(y) = h(f^{-1}(y)).$$

By the ordinary Theorem of Faà di Bruno g is of class C^n on $[f(a), f(b)]$. Then the ordinary Theorem of Faà di Bruno holds for the composition $h = f \circ g$ and yields (12).

To prove the remaining part of the theorem, let $\xi \in [\alpha \cdot \cdot \beta]$. Then

$$\chi^{[n]}(\xi) \simeq \sum_{k_1+2k_2+\ldots+nk_n=n} \frac{n!}{k_1!k_2!\ldots k_n!} \psi^{[k_1+k_2+\ldots+k_n]}(\phi(\xi)) \prod_{i=1}^{n} \left(\frac{\phi^{[i]}(\xi)}{i!}\right)^{k_i}$$

by Proposition 5.8. We put

$$\tau_n(\xi) = \psi^{[n]}(\phi(\xi))(\phi^{[1]}(\xi))^n$$

and

$$\sigma_n(\xi) = \sum_{k_1+2k_2+\ldots+nk_n=n, k_1<n} \frac{n!}{k_1!k_2!\ldots k_n!} \psi^{[k_1+k_2+\ldots+k_n]}(\phi(\xi)) \prod_{i=1}^{n} \left(\frac{\phi^{[i]}(\xi)}{i!}\right)^{k_i}.$$

Then

$$\tau_n(\xi) \simeq \chi^{[n]}(\xi) - \sigma_n(\xi). \tag{13}$$

Now $\tau_n(\xi)$, $\chi^{[n]}(\xi)$ and $\sigma_n(\xi)$ are of class S^0. In particular they are limited, and $\phi^{[1]}(\xi)$ is positive appreciable, hence also $(\phi^{[1]}(\xi))^n$. So for all $\xi \in [\alpha \cdot \cdot \beta]$

$$\psi^{[n]}(\phi(\xi)) = \frac{\tau_n(\xi)}{(\phi^{[1]}(\xi))^n} \simeq \frac{\chi^{[n]}(\xi) - \sigma_n(\xi)}{(\phi^{[1]}(\xi))^n}.$$

We put $t_n = \circ\tau_n$ and $s_n = \circ\sigma_n$. Then $t_n, s_n : [a, b] \to \mathbb{R}$ are continuous. Notice that $k_1 + 2k_2 + \ldots + nk_n = n$, $k_1 < n$ implies that $k_1 + k_2 + \ldots + k_n < n$, so $\circ(\psi^{[k_1+k_2+\ldots+k_n]}) = g^{(k_1+k_2+\ldots+k_n)}$ by the induction hypothesis. Hence s_n satisfies

$$s_n(x) = \sum_{k_1+2k_2+\ldots+nk_n=n, k_1<n} \frac{n!}{k_1!k_2!\ldots k_n!} g^{(k_1+k_2+\ldots+k_n)}(f(x)) \prod_{i=1}^{n} \left(\frac{f^{(i)}(x)}{i!}\right)^{k_i}.$$

for all $x \in [a,b]$. Then it follows from (13) and (12) that for all $x \in [a,b]$

$$t_n(x) = h^{(n)}(x) - s_n(x) = g^{(n)}(f(x))(f'(x))^n.$$

Let $y \in [f(a), f(b)]$ be standard and $\eta \in [\phi(\alpha)\cdot\cdot\phi(\beta)]$ be such that $\eta \simeq y$. Let $\xi \in [\alpha\cdot\cdot\beta]$ be such that $\phi(\xi) = \eta$. Put $x = f^{-1}(y)$. Then it follows from the previous calculations, the Theorem of the differentiable shadow, the fact that $\psi^{[n]}$ is of class S^0 and the Theorem of the continuous shadow that

$$g^{(n)}(y) = \frac{t_n(x)}{(f'(x))^n} \simeq \frac{t_n(\xi)}{(\phi^{[1]}(\xi))^n} = \psi^{[n]}(\eta) \simeq {}^\circ\left(\psi^{[n]}\right)(y).$$

Hence $g^{(n)}(y) = ({}^\circ\left(\psi^{[n]}\right))(y)$ for all $y \in [f(a), f(b)]$ by the Principle of Carnot and by Transfer, hence $g^{(n)} = ({}^\circ\left(\psi^{[n]}\right)$. ■

The general Theorem 5.1 follows from two observations made beforehand. Firstly, the condition that $\phi^{[1]}(\overline{\xi}) \not\simeq 0$ at some limited point $\overline{\xi} \in \mathbb{X}$ implies the existence of a quasi-interval $[\alpha\cdot\cdot\beta]$ with $\alpha, \beta \in \mathbb{X}$ limited and $\alpha \lessapprox \overline{\xi} \lessapprox \beta$ such that $\phi^{[1]}(\xi) \not\simeq 0$ for all $\xi \in [\alpha\cdot\cdot\beta]$. Secondly, the condition of $\phi^{[1]}(\xi) \gtrapprox 0$ to hold on $[\alpha\cdot\cdot\beta]$ was adopted during the previous discourse by mere convenience, and the proof easily carries over to the case where $\phi^{[1]}(\xi) \lessapprox 0$ on $[\alpha\cdot\cdot\beta]$.

We already saw that not on all near-continua $\phi(\mathbb{X})$ the class S^n is stable under the usual algebraic operations. We show that this stability does hold for functions defined on near-continua of class S^n. We end this section by showing that even if $\phi(\mathbb{X})$ is only of class S^0, a subset of $\phi(\mathbb{X})$ may be determined which is a near-continuum of class S^n for all standard $n \in \mathbb{N}$ indeed.

We continue to assume the conventions which where introduced earlier. We prove first a lemma.

Lemma 5.11 *Let $n \in \mathbb{N}, n \geq 1$ be standard, $\mathbb{Y} \equiv \eta(\mathbb{X})$ be a near-continuum of class S^n and $f : \mathbb{Y} \to \mathbb{R}$ be a function of class S^n. Define $g : \mathbb{Y} \to \mathbb{R}$ by $g(\eta) = f(\eta + \delta\eta)$. Then g is of class S^{n-1} in η.*

Proof. By external induction. The case $n = 1$ follows from the fact that $\delta\eta_\xi$ is infinitesimal: the number $\eta_\xi + \delta\eta_\xi$ is limited whenever η_ξ is limited. Assume the property is valid for some standard integer n. Suppose f and η are of class S^{n+1}. By definition f is of class S^n in the variable η and the function η is of class S^n in ξ. By the induction hypothesis g is of class S^{n-1} in η. By the Lemma of the discrete derivative $f^{[1]}$ is of class S^n in η and $\eta^{[1]}$ is of class S^n in ξ. Also

$$g(\eta_\xi) = f\left(\eta_\xi + \delta\eta_\xi\right) = f^{[1]}\left(\eta_\xi\right)\eta^{[1]}(\xi)\delta\xi + f\left(\eta_\xi\right).$$

Then by Proposition 5.3 and Proposition 3.9 the function χ defined by $\chi(\xi) = f^{[1]}\left(\eta_\xi\right)\eta^{[1]}(\xi)$ is of class S^n in ξ. Then $\chi^{[n]}(\xi)\delta\xi$ is infinitesimal and

$$g^{[n]}\left(\eta_\xi\right) = \chi^{[n]}(\xi)\delta\xi + f^{[n]}\left(\eta_\xi\right) \simeq f^{[n]}\left(\eta_\xi\right),$$

which is of class S^0 in η. Then $g^{[n]}$ is of class S^0 in η by Proposition 2.4. Hence g is of class S^n in η. ■

Proposition 5.12 *Let $n \in \mathbb{N}$ be standard, $\mathbb{Y}$ be a near-continuum of class S^n and $f, g : \mathbb{Y} \to \mathbb{R}$ be functions of class S^n. Then $f \cdot g$ is of class S^n.*

Proof. By external induction. The case $n = 0$ is contained in Proposition 2.7. Assume the property is valid for some standard integer n. Suppose f, g and η are of class S^{n+1}. By Lemma 5.11 the function $g(\eta + \delta\eta)$ is of class S^n. By the Lemma of the discrete derivative $f^{[1]}$ and $g^{[1]}$ are of class S^n and $\eta^{[1]}$ is of class S^n. Now

$$(f \cdot g)^{[1]}(\eta) = f^{[1]}(\eta)\, g(\eta + \delta\eta) + f(\eta)\, g^{[1]}(\eta).$$

Hence $(f \cdot g)^{[1]}$ is of class S^n. Because $f \cdot g$ is of class S^0, the function $f \cdot g$ is of class S^{n+1} by the Lemma of discrete integration. ■

Next lemma is proved similarly, now using the formula

$$\frac{\delta(1/f)}{\delta\eta}(\eta) \quad = \quad -\frac{\delta f(\eta)}{\delta\eta} \cdot \frac{1}{f(\eta) f(\eta + \delta\eta)}.$$

Lemma 5.13 *Let $n \in \mathbb{N}$ be standard, $\mathbb{Y}$ be a near-continuum of class S^n and $f : \mathbb{Y} \to \mathbb{R}$ be a function of class S^n. Then $\dfrac{1}{f}$ is of class S^n on all subsets $A^{(n)} \subset \mathbb{Y}$ such that $f \not\simeq 0$ on the limited part of A.*

Proposition 5.14 *Let $n \in \mathbb{N}$ be standard, $\mathbb{Y}$ be a near-continuum of class S^n and $f, g : \mathbb{Y} \to \mathbb{R}$ be functions of class S^n. Then f/g is of class S^n on all subsets $A^{(n)} \subset \mathbb{Y}$ such that $g \not\simeq 0$ on the limited part of A.*

Proof. By Proposition 5.12 and Lemma 5.13. ■

The proposition implies that, for all standard n, all polynomials of standard degree with limited coefficients defined on some near-continuum $\mathbb{Y}$ of class S^n are of class S^n, and all rational functions which are quotients of such polynomials are of class S^n on all subsets $A^{(n)}$ of $\mathbb{Y}$, such that the elements of A are not infinitely close to a limited singularity of the denominator.

As already said such functions do not need to be of class S^n for all standard $n \in \mathbb{N}$ on too irregular near-continua $\phi(\mathbb{X})$. Next theorem says that a sufficiently sparse subset $\mathbb{S}$ of a near-continuum $\phi(\mathbb{X})$ of class S^0 may be determined which happens to be a near-continuum of class S^n for all standard $n \in \mathbb{N}$. Then the restrictions of the functions mentioned beforehand to $\mathbb{S}$ will be of class S^n for all standard $n \in \mathbb{N}$ indeed.

Theorem 5.15 *Let $\phi(\mathbb{X})$ be a near-continuum of class S^0. Then there exists $\delta y > 0, \delta y \simeq 0$, which is a multiple of $\delta\xi$, such that $\psi(\mathbb{Y}) \subseteq \phi(\mathbb{X})$ is a near continuum of class S^n for all standard $n \in \mathbb{N}$, where $\mathbb{Y} \equiv \mathbb{Z}\delta y$ and $\psi : \mathbb{Y} \to \phi(\mathbb{X})$ is defined by*

$$\psi(y) = \min\{\eta \in \phi(\mathbb{X}) \,|\, \eta \geq y\}.$$

Proof. By the Fehrele Principle there is $\epsilon > 0, \epsilon \simeq 0$ such that $\delta\phi(\xi) < \epsilon$ for all limited $\xi \in \mathbb{X}$. Also by the Fehrele Principle there is $\omega \simeq +\infty$ such $\delta y \equiv \epsilon^{1/\omega} \simeq 0$; we may suppose that δy is a multiple of $\delta\xi$. Put $\mathbb{Y} = \mathbb{Z}\delta y$. Observe that if $y \in \mathbb{Y}$ is limited

$$\psi(y) - y \leq \epsilon = \delta y^{\omega}. \tag{14}$$

Hence $\psi(y) - y$ is an element of the external set of all infinitely large powers of δy, that we denote by $\pounds\delta y^{\not\infty}$; this external set is stable by divisions by standard powers of δy.

We prove with external induction that $\psi^{[n]}(y) - I^{[n]}(y) \in \pounds\delta y^{\not\infty}$ for all limited $y \in \mathbb{Y}$ and for all standard $n \in \mathbb{N}$, where I is the identity function on $\mathbb{Y}$. For $n = 0$ the property follows from (14). For $n = 1$ one has, for some ϵ_1 with $0 \leq \epsilon_1 \leq \epsilon$

$$\psi^{[1]}(y) - y^{[1]} = \frac{y + \delta y + \epsilon_1 - y}{\delta y} - 1 \in \pounds\delta y^{\not\infty}.$$

For $n = 2$ one has $y^{[2]} = 0$ and $\psi^{[1]}(y), \psi^{[1]}(y + \delta y) \in 1 + \pounds\delta y^{\not\infty}$. Hence

$$\psi^{[2]}(y) - y^{[2]} = \psi^{[2]}(y) \in \frac{1 + \pounds\delta y^{\not\infty} - (1 + \pounds\delta y^{\not\infty})}{\delta y} \subseteq \pounds\delta y^{\not\infty}.$$

Then for $n = 3$

$$\psi^{[3]}(y) - y^{[3]} = \psi^{[3]}(y) \in \frac{\pounds\delta y^{\not\infty} - \pounds\delta y^{\not\infty}}{\delta y} \subseteq \pounds\delta y^{\not\infty},$$

a property which will also hold for the remaining standard natural numbers. We conclude that ψ and $\psi^{[n]}$ are of class S^0 for all standard $n \in \mathbb{N}$. As a consequence ψ is of class S^n for all standard $n \in \mathbb{N}$ and $\psi(\mathbb{Y})$ is a near continuum of class S^n for all standard $n \in \mathbb{N}$. ■

6 Nearly-continuous properties of discrete functions of two variables

We consider discrete functions defined on Cartesian products of two near-continua $\mathbb{X}, \mathbb{Y} \subset \mathbb{R}$. We denote the difference of two successive points of the set $\mathbb{X}$ by δx and the difference of two successive points of the set $\mathbb{Y}$ by δy. The quantities δx and δy may or may not be interrelated and may or may not be constant. We extend the notion of functions of class S^n to functions of two variables. Then we consider the transition from the discrete to the continuous for such functions.

For convenience we consider functions defined on the whole Cartesian product $\mathbb{X} \times \mathbb{Y}$.

Definition 6.1 *Let $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. The function f is said to be of class S^{00} if f is limited and S-continuous at every limited $(x, y) \in \mathbb{X} \times \mathbb{Y}$.*

Just as standard everywhere continuous real functions of one variable are of class S^0, one may show that standard everywhere continuous real functions of two variables are of class S^{00}.

Definition 6.2 *Let $m, n \in \mathbb{N}$ and $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. We note the m^{th} difference quotient with respect to the first variable by $f_1^{[m]}$ and the n^{th} difference quotient with respect to the second variable by $f_2^{[n]}$.*

We may denote the m^{th} difference quotient with respect to the first variable of the n^{th} difference quotient with respect to the second variable of f by $f_{12}^{[m][n]}$ (equal to the n^{th} difference quotient with respect to the second variable of the m^{th} difference quotient with respect to the first variable $f_{21}^{[n][m]}$).

Definition 6.3 *Let $m, n \in \mathbb{N}$ be standard and $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. We define functions of class S^{m0} and S^{0n} by external induction. The case $m = 0$ is defined in the previous definition. Assume the functions of class S^{m0} are defined. We say that f is of class $S^{(m+1)0}$ if f is of class S^{m0} and $f_1^{[m+1]}$ is of class S^{00}. We say that f is of class $S^{0(n+1)}$ if f is of class S^{0n} and $f_2^{[n+1]}$ is of class S^{00}.*

Definition 6.4 *Let $k \in \mathbb{N}$ be standard and $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. The function f is said to be of class S^k if the difference quotients $f_{12}^{[m][n]}$ are of class S^{00} for all $m, n \in \mathbb{N}$ such that $m, n \leq k$.*

Proposition 6.5 *Let $k \in \mathbb{N}$ be standard and $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be of class S^k. Let $m, n \in \mathbb{N}$ be such that $m, n \leq k$. Then $f_1^{[m]}$ is of class S^{0n} and $f_2^{[n]}$ is of class S^{m0}.*

Proof. The function $f_{12}^{[m][n]}$ is of class S^{00}, hence $f_1^{[m]}$ is of class S^n in y, while it is of class S^0 in x. Hence $f_1^{[m]}$ is of class S^{0n}. The case of the difference quotient in the second variable $f_2^{[n]}$ is similar. ■

A function $f : \mathbb{R}^2 \to \mathbb{R}$ is class C^{m0} if it is m times differentiable in the first variable, and all partial derivatives in the first variable up to order m are continuous in both variables. Functions of class C^{0n} are defined analogously. A function is of class C^k if the mixed partial derivatives $\dfrac{\partial^m \partial^n f}{\partial x^m \partial y^n}$ exist and are continuous in both variables for all m, n with $m, n \leq k$. We will relate functions of class S^{m0} to functions of class C^{m0}, functions of class S^{0n} to functions of class C^{0n} and, more in general, functions of two-variables of class S^k to functions of two-variables of class C^k. We start by recalling two results from the literature.

Theorem 6.6 (Theorem of the continuous shadow, two variables) *Let $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be of class S^0. Then there exists a unique standard function $^\circ f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) \simeq {}^\circ f(x, y)$ for all limited $(x, y) \in \mathbb{X} \times \mathbb{Y}$. In fact f is continuous on $\mathbb{R}^2$.*

The theorem is essentially a consequence of a general theorem on metric spaces, which can be found in several textbooks, like [7].

Proposition 6.7 *Let $\mathbb{X}$ be a near-continuum of class S^1 and $\mathbb{Y}$ be a near-continuum of class S^0. If $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is of class S^{10}, its shadow $^\circ f : \mathbb{R}^2 \to \mathbb{R}$ is of class C^{10} and one has*

$$^\circ\left(f_1^{[1]}\right) = \frac{\partial {}^\circ f}{\partial x}. \tag{15}$$

The proposition has been proved in [3] for functions defined on Cartesian products of equidistant near-continua. Due to Theorem 5.9 there is no difficulty extending the proof to the above case.

We have the following extensions to higher order of this transition from the discrete to the continuous.

Proposition 6.8 *Let $m \in \mathbb{N}$ be standard. Let $\mathbb{X}$ be a near-continuum of class S^m and $\mathbb{Y}$ be a near-continuum of class S^0. If $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is of class S^{m0}, its shadow $^\circ f : \mathbb{R}^2 \to \mathbb{R}$ is of class C^{m0} and one has for all i with $0 \leq i \leq m$*

$$^\circ\left(f_1^{[i]}\right) = \frac{\partial^i \left({}^\circ f\right)}{\partial x^i}. \tag{16}$$

Proof. By external induction in i. For $i = 0$ nothing has to be proved and the case $i = 1$ is contained in formula (15). Assume the equality holds for some standard $i < m$. By definition f is of class S^{i0} and of class $S^{(i+1)0}$. Then $f_1^{[i]}$ and $(f_1^{[i]})_1^{[1]} = f_1^{[i+1]}$ are of class S^{00}. Then $f_1^{[i]}$ is of class S^{10}. Hence by formula (15) and the induction hypothesis

$$^\circ\left(f_1^{[i+1]}\right) = {}^\circ\left((f_1^{[i]})_1^{[1]}\right) = \frac{\partial}{\partial x} {}^\circ\left(f_1^{[i]}\right) = \frac{\partial}{\partial x}\left(\frac{\partial^i \left({}^\circ f\right)}{\partial x^i}\right) = \frac{\partial^{i+1} \left({}^\circ f\right)}{\partial x^{i+1}}.$$

■

Proposition 6.9 *Let $n \in \mathbb{N}$ be standard. Let $\mathbb{X}$ be a near-continuum of class S^0 and $\mathbb{Y}$ be a near-continuum of class S^n. If $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ is of class S^{0n}, its shadow $^\circ f : \mathbb{R}^2 \to \mathbb{R}$ is of class C^{0n} and one has for all j with $0 \leq j \leq n$*

$$^\circ\left(f_2^{[j]}\right) = \frac{\partial^j \left({}^\circ f\right)}{\partial y^j}. \tag{17}$$

The following general theorem on mixed difference quotients and their corresponding mixed differential quotients is a consequence of Propositions 6.5, 6.8 and 6.9.

Theorem 6.10 *Let $k \in \mathbb{N}$ be standard and $\mathbb{X}$ and $\mathbb{Y}$ be near-continua of class S^k. Let $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be of class S^k. Then $\circ f$ is of class C^k on $\mathbb{R}^2$ and for all $m, n \leq k$*

$$\circ \left(f_{12}^{[m][n]} \right) = \frac{\partial^m \partial^n (\circ f)}{\partial x^m \partial y^n}.$$

Proof. Let $m, n \in \mathbb{N}$, with $m, n \leq k$. By Proposition 6.5 the difference quotient $f_2^{[n]}$ is of class S^{m0} and f is of class S^{0n}. Then by Proposition 6.8 and 6.9

$$\circ \left(f_{12}^{[m][n]} \right) = \circ \left((f_2^{[n]})_1^{[m]} \right) = \frac{\partial_1^m}{\partial x^m} \circ \left(f_2^{[n]} \right) = \frac{\partial^m \partial^n (\circ f)}{\partial x^m \partial y^n}.$$

■

7 A higher order DeMoivre-Laplace Theorem

As an application we consider a problem of continuization of higher order in two dimensions. We extend the DeMoivre-Laplace Theorem on the transition of the binomial probability distribution $B(N, j)$ to the Gaussian distribution $G(t, x) = \frac{\exp(-x^2/(2t))}{\sqrt{2\pi t}}$ to their successive difference quotients respectively partial derivatives.

To be able to apply the techniques of the previous sections to continuizations on near-continua we effectuate a change of scale in two dimensions to the Pascal Triangle. With respect to space we normalize and centralize around the mean, reducing the distance with respect to the mean by the standard deviation. Thus for infinitely large fixed N, a discrete function defined for positive integers is transformed into a infinitely fine sequence of points equally spaced at a distance of $\delta x \equiv 2/\sqrt{N}$. With respect to time we reduce the increments to $\delta t \equiv 1/N$. The rescaled binomial coefficients may then be represented by a function $b(t, x)$, which happens to be nearly equal to $G(t, x)$ for positive appreciable t (corresponding to infinitely large N). This is a nonstandard version of the DeMoivre-Laplace Theorem, for which we will show indeed that the near-equality extends to partial difference quotients respectively partial derivatives of all standard order.

7.1 Rescaling of the binomial coefficients

Definition 7.1 *Let $\delta t > 0$. We define the binomial cone $C_{\delta t}$ by*

$$C_{\delta t} = \left\{ (t, x) \in \mathbb{R}^2 \,\middle|\, t \geqslant 0, |x| \leqslant t/\sqrt{\delta t},\ \exists \nu, j \in \mathbb{N}\ ,\ t = \nu \delta t,\ x = (2j - \nu)\sqrt{\delta t} \right\}.$$

We write

$$\begin{cases} \mathbb{T} &= \{\nu \delta t \mid \nu \in \mathbb{N}\} \\ \delta x &= 2\sqrt{\delta t}. \end{cases}$$

If $\delta t \simeq 0$ the upper and lower boundaries of $C_{\delta t}$ are nearly vertical "discrete lines" with infinitely large slope $1/\sqrt{\delta t}$. This implies that $C_{\delta t}$ contains all points (t, x) such that t is appreciable and multiple of δt, and x is limited and of the form $x = (2j - \nu)\sqrt{\delta t}$ for some $\nu \in \mathbb{N}$ and $j \in \mathbb{Z}$.

For convenience we assume that $\dfrac{1}{\delta t}$ is an integer.

Definition 7.2 *Let $(t,x) \in C_{\delta t}$, We write*

$$\begin{aligned} \nu_t &= \frac{t}{\delta t} \\ j_{t,x} &= \frac{t}{2\delta t} + \frac{x}{\delta x}. \end{aligned}$$

Conversely, if $\nu \in \mathbb{N}$, and $j \in \mathbb{N}$ is such that $0 \leq j \leq \nu$, we write

$$\begin{aligned} t_\nu &= \nu \delta t \\ x_{\nu,j} &= (j - \nu/2)\, \delta x. \end{aligned}$$

Definition 7.3 *Let ν, $j \in \mathbb{N}$ be such that $0 \leq j \leq \nu$. Then we write*

$$B(\nu, j) = \binom{\nu}{j} \left(\frac{1}{2}\right)^\nu.$$

Definition 7.4 *The* binomial function $b : C_{\delta t} \to \mathbb{R}$ *is defined by*

$$b(t,x) = \frac{1}{\delta x} B(\nu_t, j_{t,x}).$$

Within this setting, we may formulate the following nonstandard version of the DeMoivre-Laplace Theorem.

Theorem 7.5 (DeMoivre-Laplace theorem, nonstandard) *Let $\delta t \simeq 0$. Then for all $(t,x) \in C_{\delta t}$ such that t is appreciable and x is limited,*

$$b(t,x) \simeq \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right). \tag{18}$$

For a nonstandard proof we refer to [4]. The following proposition is a first consequence of Theorem 7.5.

Proposition 7.6 *1. Let t be fixed and appreciable. Then $b(t,\cdot)$ is of class S^0.*

2. Let x be fixed and limited. Then $b(\cdot,x)$ is of class S^0 on every set $I = \{t \in \mathbb{T} \,|\, t_0 \leq t\}$ with t_0 appreciable.

Up to adapting some technical details the notations and results of Section 6 may be extended to the binomial cone $C_{\delta t}$, which in fact is a subset of the union of two Cartesian products with horizontal step $2\delta t$ and vertical step $2\sqrt{\delta t}$. We adopt the convention that the horizontal differences consider steps of $2\delta t$ and that the vertical differences consider steps $\delta x = 2\sqrt{\delta t}$.

Theorem 7.7 (Higher order DeMoivre-Laplace theorem) *Let $\delta t \simeq 0$. For every standard m, $n \in \mathbb{N}$, one has for all $(t,x) \in C_{\delta t}$ such that t is appreciable and x is limited*

$$b_{12}^{[m][n]}(t,x) \simeq \frac{\partial^m \partial^n G(t,x)}{\partial t^m \partial x^n}.$$

The proof of Theorem 7.7 presented below is a shortened version of the proof contained in [10] and [5]. The proof uses some convenient ordinary and partial difference equations for the binomial functions, which will be derived from some combinatorial properties in the next section.

7.2 Difference equations for the binomial function

The first lemma states a first-order difference equation with respect to the space-variable x.

Lemma 7.8 *For all $(t,x) \in C_{\delta t}$ such that $t > 0$, $x < \frac{t}{\sqrt{\delta t}}$*

$$b_2^{[1]}(t,x) = -b(t,x)\frac{x + \frac{1}{2}\delta x}{t + \frac{x}{2}\delta x + \frac{1}{2}\delta x^2}. \tag{19}$$

The lemma is an easy consequence of the combinatorial formula

$$B(\nu, j+1) = B(\nu, j) \cdot \frac{\nu - j}{j+1}.$$

We derive now a partial difference equation for the binomial function, which is a discrete version of the heat equation.

Proposition 7.9 *For all $(t,x) \in C_{\delta t}$*

$$b_1^{[1]}(t,x) \quad = \quad \tfrac{1}{2} b_2^{[2]}(t, x - \delta x). \tag{20}$$

Proof. From the Pascal Triangle one derives

$$B\left(\nu_t + 1, j_{t,x} + 1\right) \quad = \quad \tfrac{1}{2} B\left(\nu_t, j_{t,x}\right) + \tfrac{1}{2} B\left(\nu_t, j_{t,x} + 1\right).$$

Then by Definition 7.2

$$b(t + \delta t, x) \quad = \quad \tfrac{1}{2} b(t, x + \sqrt{\delta t}) + \tfrac{1}{2} b(t, x - \sqrt{\delta t}).$$

Repeating this step, one obtains

$$b(t + 2\delta t, x) \quad = \quad \tfrac{1}{4} b(t, x + \delta x) + \tfrac{1}{2} b(t,x) + \tfrac{1}{4} b(t, x - \delta x).$$

Hence

$$b(t + 2\delta t, x) - b(t,x) \quad = \quad \tfrac{1}{4} b(t, x + \delta x) - \tfrac{1}{2} b(t,x) + \tfrac{1}{4} b(t, x - \delta x).$$

This implies (20). ■

The discrete heat equation (20) extends to higher order. Indeed, one proves by induction:

Theorem 7.10 *For all $(t,x) \in C_{\delta t}$ and all $m \in \mathbb{N}$ one has, as long as $-\frac{t}{\sqrt{\delta t}} \leq x, x + m\delta x \leq \frac{t}{\sqrt{\delta t}}$*

$$b_1^{[m]}(t,x) = \frac{1}{2^m} b_2^{[2m]}(t, x - m\delta x).$$

7.3 Proof of the higher order DeMoivre-Laplace theorem

Theorem 7.10 reduces difference quotients of the binomial function with respect to time to difference quotients with respect to space (of even order). So regularity properties of difference quotients with respect to time may be derived form corresponding properties of difference quotients with respect to space.

The first-order difference equation with respect to space established in Section 7.2 enables to prove that the binomial function is of class S^n in the space-variable for all standard n. Then the higher order difference quotients of the binomial function with respect to space are nearly equal to the corresponding higher order differential quotients of the Gaussian function, given explicitly in [1]. The higher-order heat difference equation of Theorem 7.10 will enable to conclude the proof.

Lemma 7.11 *Let $(t,x) \in C_{\delta t}$ be such that t is fixed and appreciable and x is limited. Let $n \in \mathbb{N}$ be standard and arbitrary. Then*

1. *The function $b_2^{[n]}(t,\cdot)$ is of class S^0.*
2. $$\begin{aligned} b_2^{[n]}(t,x) &\simeq \frac{\partial^n G(t,x)}{\partial x^n} \\ &\simeq (-1)^n \frac{n!}{\sqrt{2\pi t^{n+1}}} e^{-\frac{x^2}{2t}} \sum_{k=0}^{[n/2]} (-1)^k \frac{1}{k!2^k\,(n-2k)!} \left(\frac{x}{\sqrt{t}}\right)^{n-2k}. \end{aligned} \tag{21}$$
3. *The function b is of class S^{0n} on every set $D = \{(t,x) \in C_{\delta t} \,|t \geq t_0\}$ with $t_0 \gtrapprox 0$.*

Proof. We use the difference equation with respect to x given by (19), rewritten to

$$b_2^{[1]}(t,x) = b(t,x) \left(-\frac{x + \frac{1}{2}\delta x}{t + \frac{x}{2}\delta x + \frac{1}{2}\delta x^2} \right). \tag{22}$$

We put

$$c(t,x) = -\frac{x + \frac{1}{2}\delta x}{t + \frac{x}{2}\delta x + \frac{1}{2}\delta x^2}.$$

Notice that $c(t,\cdot)$, as a rational function of polynomials with limited coefficients without limited poles, is of class S^k for all standard k.

1. By external induction. The function $b(t,\cdot)$ is of class S^0 by Proposition 7.6.1. Let $n \in \mathbb{N}$ be standard and assume $b(t,\cdot)$ is of class S^n. Because $c(t,\cdot)$ is also of class S^n, the product $b(t,\cdot)c(t,\cdot)$ is of class S^n by Proposition 3.9. So $b_2^{[1]}(t,\cdot)$ is of class S^n. Again by Proposition 7.6.1, and by the Lemma of the discrete integral $b(t,\cdot)$ is of class S^{n+1}.
2. It follows from Theorem 6.10 that the near-equality $\frac{\delta_2^n b}{\delta x^n}(t,x) \simeq \frac{\partial^n G(t,x)}{\partial x^n}$ is valid for all limited x such that $(t,x) \in C_{\delta t}$. It is known [1] that

   $$\frac{\partial^n G(t,x)}{\partial x^n} = (-1)^n \frac{n!}{\sqrt{2\pi t^{n+1}}} e^{-\frac{x^2}{2t}} \sum_{k=0}^{[n/2]} (-1)^k \frac{1}{k!2^k\,(n-2k)!} \left(\frac{x}{\sqrt{t}}\right)^{n-2k}.$$

   This implies (21).
3. Let $t_0 \gtrapprox 0$ and $D = \{(t,x) \in C_{\delta t} \,|t \geq t_0\}$. By Proposition 7.6 the function b is of class S^{00} on D. Secondly $b_2^{[n]}(t,x) \simeq \frac{\partial^n G(t,x)}{\partial x^n}$ and $\frac{\partial^n G(t,x)}{\partial x^n}$ is of class S^{00} on D. So $b_2^{[n]}$ is also of class S^{00} on D. Hence b is of class S^{0n} on D.

■

As a corollary we obtain that the binomial function is of class S^k for any standard k.

Theorem 7.12 *Let $k \in \mathbb{N}$ be standard. Let $t_0 \gtrapprox 0$ be limited and $D = \{(t,x) \in C_{\delta t} \,|t \geq t_0\}$. Then b is of class S^k on D.*

Proof. Let $m, n \in \mathbb{N}$ be such that $m, n \leq k$. Note that, whenever defined,

$$b_{12}^{[m][n]}(t,x) = \frac{1}{2^m} b_2^{[n+2m]}(t, x - 2m\delta x).$$

By Lemma 7.11 the function $b_2^{[n+2m]}$ is of class S^{00} on D, hence $b_{12}^{[m][n]}$ is also of class S^{00} on D. We conclude that b is of class S^k on D. ■

Proof of the higher order DeMoivre-Laplace Theorem. Let m, $n \in \mathbb{N}$ be standard and $(t,x) \in C_{\delta t}$ be such that t is appreciable and x is limited. Let t_0 be standard with $0 < t_0 \lessapprox t$. By Lemma 7.11 the binomial function b is of class $S^{0(2m+n)}$ on $D \equiv \{(t,x) \in C_{\delta t} \,|t \geq t_0\}$. Now ${}^\circ(b_{|D}) = G_{|[t_0,\infty)\times\mathbb{R}}$ by Theorem 7.5. Then formula (21) implies that for every limited $(t,x) \in D$

$$\begin{aligned} b_{12}^{[m][n]}(t,x) &= \frac{1}{2^m} b_2^{[n+2m]}(t, x - 2m\delta x) \simeq \frac{1}{2^m} \frac{\partial^{2m+n} G(t, x - 2m\delta x)}{\partial x^{2m+n}} \\ &\simeq \frac{1}{2^m} \frac{\partial^{2m+n} G(t,x)}{\partial x^{2m+n}} = \frac{\partial^m \partial^n G(t,x)}{\partial t^m \partial x^n}. \end{aligned}$$

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of mathematical functions*, Dover, New-York (1965).

[2] I.P. van den Berg, *A higher order time-dependent DeMoivre-Laplace theorem*, SOM Res. Rep. 96A18, Univ. Groningen (1996).

[3] I.P. van den Berg, On the relation between elementary partial difference equations and partial differential equations, *Ann. Pure App. Logic* 92 (1998) 235-265.

[4] I.P. van den Berg, *Principles of infinitesimal stochastic and financial analysis*, World Scientific (2000).

[5] I.P. van den Berg, *A DeMoivre-Laplace theorem of all orders of regularity*, in: Communications of the Laufen Colloquium on Science 2007, A. Ruffing, A. Suhrer, J. Suhrer (Eds.), Shaker Publishing, Maastricht/Aachen, p. 335-360 (2007).

[6] F. Diener and M. Diener (eds.), *Nonstandard analysis in practice,* Universitext, Springer (1995).

[7] F. Diener and G. Reeb, *Analyse Non Standard*, Hermann, Paris (1989).

[8] F. Diener, *Cours d'Analyse Non Standard*, Office des Publ. Univ. Alger (1983).

[9] W. Feller, *An introduction to probability theory and its applications,* Vol. 1, 3^{rd} ed., Wiley (1968).

[10] H. Gião, *Um Teorema de DeMoivre-Laplace de Ordem arbitrária*, Masters Thesis, University of Évora, Portugal (2005).

[11] S.G. Krantz and H.R. Parks, *A primer of real analytic functions*, 2^{nd} ed., Birkhäuser (2002).

[12] E. Nelson, *Internal Set Theory*, Bull. Amer. Math. Soc. 83, no. 6 (1977) 1165–1198.

[13] E. Nelson, *Radically Elementary Probability Theory*, Princ. Univ. Press (1987).

[14] A. Robinson, *Non-standard analysis*, 3^{rd} ed., Princ. Univ. Press (1996).

Address of the author:

University of Évora, Department of Mathematics
Colégio Luís Verney, Rua Romão Ramalho 59
7000-671 Évora, Portugal

E-mail: `ivdb@uevora.pt.`

Moderate deviations in $\mathbb{R}^k$

Jacques Bosgiraud

Abstract

We give an original nonstandard proof of a standard result about moderate deviations in $\mathbb{R}^k$.

KEYWORDS: moderate deviations, nonstandard analysis, external calculus.

1 Introduction

1.1 Normal approximation, large and moderate deviations

Let X be a standard real random variable with null expectation and variance σ^2. Let n be an integer and $X_1, ..., X_n$ be independent duplicates of X. They are independent, identically distributed (i.i.d.) random variables. We denote $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$, and the classical results about the deviations to 0 of the empirical mean $\overline{X}$ are the following ones (y is any positive number).

Normal deviation: the central limit theorem yields that $\lim_{n\to\infty} P(\overline{X} > \frac{y}{\sqrt{n}}) = (1 - \Phi(\frac{y}{\sigma}))$, where Φ is the repartition function of the normal $\mathcal{N}(0,1)$ distibution. A more general form is the following one: for any real sequence $(u_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} u_n = 1$, then $\lim_{n\to\infty} P(\overline{X} > \frac{u_n y}{\sqrt{n}}) = (1 - \Phi(\frac{y}{\sigma}))$

Large deviation: $\lim_{n\to\infty} \frac{1}{n} \ln P(\overline{X} > y) = -I(y)$, where I is the Cramér transform of the distribution of X (see §2.2). A more general form is the following one: for any real sequence $(u_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \frac{u_n}{n} = 1$, then $\lim_{n\to\infty} \frac{1}{n} \ln P(\overline{X} > \frac{u_n y}{n}) = -I(y)$.

Moderate deviation: the result can only be set in terms of sequences. For any real sequence $(u_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} u_n = +\infty$ and $\lim_{n\to\infty} \frac{u_n}{\sqrt{n}} = 0$, then $\lim_{n\to\infty} \frac{1}{u_n^2} \ln P(\overline{X} > \frac{u_n y}{\sqrt{n}}) = -\frac{y^2}{2\sigma^2}$.

Using the notations of external calculus (cf. [6]), these standard results can be set on the following form (n is any unlimited integer).

Normal deviation:

$$\forall x \in \frac{\mathcal{L}}{\sqrt{n}}, P(\overline{X} > x) = (1 - \Phi(\frac{\sqrt{n}x}{\sigma}))(1 + \emptyset). \tag{1}$$

Large deviation:

$$\forall x \in @, \frac{1}{n} \ln P(\overline{X} > x) = -I(x)(1 + \emptyset). \tag{2}$$

Moderate deviation:

$$\forall x \in \left] \frac{\mathcal{L}}{\sqrt{n}}, \emptyset \right], \frac{1}{n} \ln P(\overline{X} > x) = -\frac{x^2}{2\sigma^2}(1 + \emptyset). \tag{3}$$

Note that in the cases (1) and (3) the approximation of $P(\overline{X} > x)$ can be given through σ^2, which is not possible in the case (2), where the Cramér transform is specific to the distribution of X.

More precisely, in the case (3), one can write:

$$\forall x \in \left]\frac{\mathcal{L}}{\sqrt{n}}, \emptyset\right], \frac{1}{n}\ln P(\overline{X} > x) = -I(x)(1+\emptyset) = -\frac{x^2}{2\sigma^2}(1+\emptyset). \tag{4}$$

1.2 Multidimentional case

Let now X be a standard random variable taking values in $\mathbb{R}^k$ (where k is standard) with $\mathbb{E}X = \mathbf{0}$ and A be a borelian subset of $\mathbb{R}^k$. One may try to approximate $P(\overline{X} \in A)$. If the origin $\mathbf{0}$ is an inner point of a standard set A, one uses the normal approximation. If the distance $d(\mathbf{0}, A)$ is appreciable, one may use the large deviations theorem (see [7, section 4]), and, for exemple, if A is a standard set with $\overline{A^0} = \overline{A}$ [if E is a subset of $\mathbb{R}^k$, $\overline{E}$ denotes its closure, E^0 its interior and ∂E its boundary], one obtains

$$\frac{1}{n}\ln P(\overline{X} \in A) = -\inf_{x\in A} I(x)(1+\emptyset)$$

where I is the Cramér transform of the distribution of X (see §2.2).

>From this result, it is possible to deduce results concerning moderate deviations, as the following one:

1.3 A standard theorem

In this paper $(. \mid .)$ denotes the scalar product in $\mathbb{R}^k$, $\|.\|$ the euclidian norm, and for $\theta \in \mathbb{R}^k$, let $M(\theta) := \mathbb{E}e^{(\theta|X)}$.

Theorem 1 *Let X be a $\mathbb{R}^k$-valued random variable, such that $\mathbb{E}e^{(\theta|X)}$ is finite in a neighbourhood of the origin, $\mathbb{E}X = \mathbf{0}$ and V the covariance matrix is non singular. Let $X_1, ..., X_n, ...$ be independent identically distribued (i.i.d.) duplicates of X and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let A_0 be a bounded borelian subset of $\mathbb{R}^k$, such that $d(\mathbf{0}, A_0) \neq 0$, and let*

$$J_0 = \inf_{x\in\partial A_0} J(x) \text{ where } J(x) = \frac{1}{2}xV^{-1}x^t$$

If there exists a point $a_0 \in \partial A_0 \cap \overline{A^0}$ such that $J(a_0) = J_0$, then for any real sequence $(u_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} u_n = +\infty \text{ and } \lim_{n\to\infty} \frac{u_n}{\sqrt{n}} = 0$$

we have

$$\lim_{n\to\infty} \frac{1}{u_n^2}\ln P^n(\overline{X}_n \in \frac{u_n}{\sqrt{n}}A_0) = -J_0$$

This result can be easily deduced from theorem 3.7.1 in [4] (for example). Our purpose is to deduce it by transfert from a nonstandard theorem, the proof of which uses original methods.

2 A nonstandard proof

2.1 Notations

To define the probability space $(\Omega, \mathcal{B}, P)$, the simpliest way is to suppose that Ω is a borelian subset of $\mathbb{R}^k$ and P a probability defined on $(\Omega, \mathcal{B})$ where $\mathcal{B}$ is the family of borelian subsets of Ω. So, $X : \Omega \to \mathbb{R}^k$, $x \to x$ is a random variable with distribution P.

Let n be an integer; $P^n := P^{\otimes n}$ is a probability defined on $(\Omega^n, \mathcal{B}^{\otimes n})$. For $i = 1, ..., n$ we define $X_i : \Omega^n \to \mathbb{R}^k$ by $X_i(x_1, ..., x_n) = x_i$; then $X_1, ..., X_n$ are independent, identically distributed (i.i.d.) random variables with distribution P.

2.2 Cramér transform

In the classical literature about large deviations in $\mathbb{R}^k$ (see, for example, [7, section 4]), the Cramér transform I of P is defined in the following way:

For $\theta \in \mathbb{R}^k$, let $M(\theta) := \mathbb{E}e^{(\theta|X)}$; $M(\theta)$ is supposed to be finite on a neighbourhood $\mathcal{N}$ of the origin. For $y \in \mathbb{R}^k$, let $I(y) := \sup_\theta \{(\theta \mid y) - \ln M(\theta)\}$; then there exists a point $\theta_0 := \theta_0(y)$ such that $I(y) := (\theta_0 \mid y) - \ln M(\theta_0)$.

If x is infinitesimal, an approximation of $I(x)$ and $\theta_0(x)$ can be given through the covariance matrix V (see [5, p142], example 1.4) :

$$I(x) = (\frac{1}{2} x V^{-1} x^t)(1 + \emptyset) \tag{5}$$

$$\theta_0(x) = -x V^{-1}(1 + \emptyset) \tag{6}$$

2.3 A technical lemma

In the following, $B(a, r)$ is the open ball with center a and radius r. The proof of theorem 1 will use the standard following lemma:

Lemma 2 *Let $(u_n)_{n\in\mathbb{N}}$ be a real sequence such that $\lim_{n\to\infty} u_n = +\infty$, let A_0 be bounded borelian subset of $\mathbb{R}^k$, and let $a_0 \in \overline{A_0^0}$. Then there exist a sequence $(a_n)_{n\geq 1}$ of inner points of A_0 and a sequence $(r_n)_{n\geq 1}$ of positive real numbers such that for each $n \geq 1, B(a_n, r_n) \subset A_0$ and such that $\lim_{n\to\infty} a_n = a_0$ and $\lim_{n\to\infty} u_n r_n = +\infty$.*

proof. As $a_0 \in \overline{A_0^0}$, there exist a sequence $(b_n)_{n\geq 1}$ of inner points of A_0 and a sequence $(\rho_n)_{n\geq 1}$ of positive real numbers such that for each $n \geq 1, B(b_n, \rho_n) \subset A_0$ and such that $\lim_{n\to\infty} b_n = a_0$.
We define an increasing function $\varphi : \mathbb{N}^* \to \mathbb{N}^*$ in the following way:

$\varphi(1) := \min \left\{ m \in \mathbb{N}^* : p \geq m \Rightarrow u_p \geq 1.\rho_1^{-1} \right\}$,

$\varphi(2) := \min \left\{ m \in \mathbb{N}^* : (p \geq m \Rightarrow u_p \geq 2.\rho_2^{-1}) \wedge m > \varphi(1) \right\}$,

and more generaly, for $n \geq 2$,

$\varphi(n) := \min \left\{ m \in \mathbb{N}^* : (p \geq m \Rightarrow u_p \geq n.\rho_n^{-1}) \wedge m > \varphi(n-1) \right\}$.

So, for each $n \in \mathbb{N}^*, p \geq \varphi(n) \Rightarrow u_p \rho_n \geq n$.
We now define a nondecreasing function $\psi : \mathbb{N}^* \to \mathbb{N}^*$ in the following way:

for $1 \leq m \leq \varphi(1), \psi(m) := 1$

and for $m > \varphi(1), \psi(m) := \max \{n \in \mathbb{N}^* : \varphi(n) \leq m\}$.

So for each $n \in \mathbb{N}^*, \psi(\varphi(n)) = n$ and for each $m \in \mathbb{N}^*, m \geq \varphi(\psi(m))$; futhermore $\lim_{m\to\infty} \psi(m) = +\infty$ because $\lim_{n\to\infty} \varphi(n) = +\infty$ and ψ is nondecreasing.

At last, we define, for each $m \in \mathbb{N}^*, a_m := b_{\psi(m)}$ and $r_m := \rho_{\psi(m)}$. If $n := \psi(m)$, $u_m r_m = u_m \rho_n \geq n$ because $m \geq \varphi(n)$. Therefore $u_m r_m \geq \psi(m)$ and consequently, as $\lim_{m\to\infty} \psi(m) = +\infty$, $\lim_{m\to\infty} u_m r_m = +\infty$ and $\lim_{m\to\infty} a_m = a_0$ □

2.4 The nonstandard theorem

The following theorem gives a nonstandard result which implies (by transfer) the classical result. If the first part of the proof is based on the law of large numbers as in the classical literature, the second part, based on infinitesimal pavings, seems to be original. This method had been used in [2] and [3]. In this theorem the set A_0 is not necessary standard.

Theorem 3 *Let X be a standard $\mathbb{R}^k$-valued random variable, such that $\mathbb{E}e^{(\theta|X)}$ is finite in a neighbourhood of the origin, $\mathbb{E}X = \mathbf{0}$ and V the covariance matrix is non singular. Let n be un unlimited integer and $X_1, ..., X_n$ be i.i.d. duplicates of X, $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Let A_0 be a limited borelian subset of $\mathbb{R}^k$, such that $d(\mathbf{0}, A_0)$ is appreciable, and let*

$$J_0 = \inf_{x\in\partial A_0} J(x) \text{ where } J(x) = \frac{1}{2}xV^{-1}x^t$$

Let $\alpha \in \left]\frac{\mathcal{L}}{\sqrt{n}}, \emptyset\right]$. We suppose that there exists a point $a_0 \in \partial A_0 \cap \overline{A^0}$ such that:

- $J(a_0) = J_0$,
- $\exists (b_0, r_0) \in A_0 \times \emptyset, \|a_0 - b_0\| \simeq 0 \wedge B(b_0, r_0) \subset A_0 \wedge \alpha r_0 = \frac{\infty}{\sqrt{n}}$

Then we have

$$\frac{1}{n}\ln P^n(\overline{X} \in \alpha A_0) = -\alpha^2 J_0(1 + \emptyset) \tag{7}$$

proof. Let $A := \alpha A_0$, $a := \alpha a_0$, $b := \alpha b_0$ and $r := \alpha r_0 = \alpha\emptyset$. So $B(b, r) \subset A$, $r = \frac{\infty}{\sqrt{n}}$, and $\|b - a\| = \|a\|\,\emptyset$.

• Following the notations of §2.2, let $\theta_0 := \theta_0(b)$. As $\|b\| = \alpha@$, then $\|\theta_0\| = \alpha@$ (cf. 2.2) and let

$$Q := \frac{1}{M(\theta_0)}e^{(\theta_0|.)}P$$

Q is a probability on $(\Omega, \mathcal{B})$ and $\mathbb{E}_Q X = b$ (see[7, section 4]). We can write

$$P = M(\theta_0)e^{-(\theta_0|.)}Q$$

and so

$$P^n(dx_1, .., dx_n) = M(\theta_0)^n e^{-n(\theta_0|\overline{x})}Q^n(dx_1, .., dx_n)$$

where $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$. Then,

$$\begin{aligned} P^n(\overline{X} \in A) &\geq P^n(\overline{X} \in B(b, r)) \geq \\ &\geq \int_{B(b,r)} M(\theta_0)^n e^{-n(\theta_0|\overline{x})}Q^n(dx_1, .., dx_n)) \geq \\ &\geq M(\theta_0)^n \exp(-n\int_{B(b,r)} (\theta_0 \mid \overline{x})Q^n(dx_1, .., dx_n)) \end{aligned}$$

by Jensen's inequality. If $(x_1, .., x_n) \in B(b, r)$ then $\overline{x} = b + r\mathcal{L}^k$ where $\mathcal{L}^k$ is the external set of limited vectors of $\mathbb{R}^k$.

Denoting $\overline{X} = (\overline{X}_j)_{1\leq j\leq k}$ and $b = (b_j)_{1\leq j\leq k}$, according to (8.4) in [1], as $r = \frac{\infty}{\sqrt{n}}$, $Q^n(\overline{X}_j \in [b_j - r/2, b_j + r/2]) = 1 + \emptyset$ for each $j = 1..k$. So the nonstandard law of large numbers yields $Q^n(\overline{X} \in B(b, r)) = 1 + \emptyset$ and we can write

$$\begin{aligned} P^n(\overline{X} \in A) &\geq M(\theta_0)^n e^{-n(\theta_0|b+r\mathcal{L}^k)} \int_{B(b,r)} Q^n(dx_1, .., dx_n) \\ &\geq M(\theta_0)^n e^{-n(\theta_0|b+r\mathcal{L}^k)}(1 + \emptyset). \end{aligned}$$

Thus,

$$\begin{aligned}\frac{1}{n}\ln(P^n(\overline{X} &\in A)) \geq \\ &\geq \ln M(\theta_0) - (\theta_0 \mid b) + (\theta_0 \mid r\mathcal{L}^k) + \frac{1}{n}\ln(1+\emptyset) \\ &\geq -I(b) + \alpha r\mathcal{L} + \frac{\emptyset}{n}\end{aligned}$$

since $\theta_0 = \alpha\mathcal{L}^k$. Using (2.1), $I(b) = \frac{1}{2}bV^{-1}b^t(1+\emptyset)$, and as $\|b-a\| = \|a\|\,\emptyset$,

$$\frac{1}{2}bV^{-1}b^t = \frac{1}{2}aV^{-1}a^t(1+\emptyset) = \alpha^2 J_0(1+\emptyset)$$

Finally,

$$\frac{1}{n}\ln(P^n(\overline{X}\in A)) \geq -\alpha^2 J_0(1+\emptyset) + \alpha r\mathcal{L} + \frac{\emptyset}{n}$$

We know that J_0 is appreciable, that $\alpha^2 = \frac{\infty}{n}$, $r = \alpha\emptyset$; then

$$\frac{1}{n}\ln(P^n(\overline{X}\in A)) \geq -\alpha^2 J_0(1+\emptyset). \tag{8}$$

• Conversly, as the limited set A_0 is included in an hypercube $[-p,p]^k$ where p is a standard integer, the infinitesimal set $A = \alpha A_0$ is included in the hypercube $[-\alpha p, \alpha p]^k$. It is possible to pave it with $(2p\omega)^k$ hypercubes (T_l) with side $\delta = \frac{\alpha}{\omega}$ where ω is an unlimited integer choosed such that $\ln\omega = n\alpha^2\emptyset$. Among these hypercubes, eliminate the ones which do not intersect A and for the others choose one $x_l \in \overline{A}\cap T_l$. In the aim of simplicity, we shall denote again the selected hypercubes by $(T_l)_{1\leq l\leq N}$. We denote $\theta_l := \theta_0(x_l)$ and $Q_l := \frac{1}{M(\theta_l)}e^{(\theta_l|.)}P$. From the relation $N < (2p\omega)^k$ we deduce $\frac{\ln N}{n} = \frac{\ln\omega}{n}\mathcal{L} = \alpha^2\emptyset$. Then, we can write

$$\begin{aligned}P^n(\overline{X} &\in A) \leq \\ &\leq \sum_{l=1}^{N}\int_{T_l} dP^n \\ &= \sum_{l=1}^{N}\int_{T_l} M(\theta_l)^n e^{-n(\theta_l|\overline{x})}Q_l^n(dx_1,..,dx_n) \\ &\leq \sum_{l=1}^{N} M(\theta_l)^n e^{-n(\theta_l|\overline{x})} = \sum_{l=1}^{N} M(\theta_l)^n e^{-n(\theta_l|x_l+\delta\mathcal{L}^k)} \\ &\leq N\max_{l=1..N}\left\{M(\theta_l)^n e^{-n(\theta_l|x_l+\delta\mathcal{L}^k)}\right\}\end{aligned}$$

Thus

$$\begin{aligned}\frac{1}{n}\ln P^n(\overline{X} &\in A) \leq \\ &\leq \frac{1}{n}\ln N + \max_{l=1..N}\left\{\ln M(\theta_l) - (\theta_l \mid x_l + \delta\mathcal{L}^k)\right\} \\ &\leq \frac{1}{n}\ln N + \ln M(\theta_{l_0}) - (\theta_{l_0} \mid x_{l_0} + \delta\mathcal{L}^k)\end{aligned}$$

where l_0 is an index for which the maximum of $\ln M(\theta_{l_0}) - (\theta_{l_0} \mid x_{l_0})$ is obtained. In the aim of simplicity, we denote $x_0 := x_{l_0}$ and $\theta_0 := \theta_{l_0}$; $x_0 \in \overline{A} = \alpha\overline{A_0}$, then $\|x_0\| = \alpha@$ because $d(\mathbf{0}, A_0) = @$ and A_0 is limited; $\|\theta_0\| = \alpha@$ according to (2.2). Then:

$$\begin{aligned}\frac{1}{n}\ln P^n(\overline{X} \in A) \leq \\ &\leq \frac{1}{n}\ln N + \ln M(\theta_0) - (\theta_0 \mid x_0) + \alpha\delta\mathcal{L} \leq \\ &\leq \frac{1}{n}\ln N - \frac{1}{2}x_0 V^{-1} x_0^t(1+\emptyset) + \alpha\delta\mathcal{L} \leq \\ &\leq \frac{1}{n}\ln N - \frac{1}{2}\widehat{x_0} V^{-1}\widehat{x_0}^t(1+\emptyset) + \alpha\delta\mathcal{L}\end{aligned}$$

where $\widehat{x_0} \in \partial A \cap [\mathbf{0}, x_0]$ (the segment between the origin $\mathbf{0}$ and x_0). Indeed $x_0 V^{-1} x_0^t \geq 0$ because V is a positive matrix and $\widehat{x_0} = \lambda x_0$ with $\lambda \in]0, 1]$.

As $\alpha\delta\mathcal{L} = \alpha^2\emptyset$ (because $\delta = \alpha\emptyset$) and $\frac{\ln N}{n} = \alpha^2\emptyset$, we finally can write

$$\frac{1}{n}\ln P^n(\overline{X} \in A) \leq -\alpha^2 J_0(1+\emptyset) + \alpha^2\emptyset$$

and then since $J_0 = \frac{1}{2}a_0 V^{-1} a_0^t = @$ (because $d(\mathbf{0}, A_0) = @$, and A_0 limited),

$$\frac{1}{n}\ln P^n(\overline{X} \in A) \leq -\alpha^2 J_0(1+\emptyset) \tag{9}$$

>From (2.4) and (2.5), we can conclude :

$$\frac{1}{n}\ln P^n(\overline{X} \in A) = -\alpha^2 J_0(1+\emptyset)$$

□

Proof of theorem 1. Theorem 1 is proved by transfer of theorem 2. If n is an unlimited integer, we put $\alpha := \frac{u_n}{\sqrt{n}}$and then $\alpha \in \left]\frac{\mathcal{L}}{\sqrt{n}}, \emptyset\right]$. Following the notations of lemma 1, we put $b_0 := a_n$ and $r_0 := r_n$. So $B(b_0, r_0) \subset A_0$ and $\alpha\ r_0 := \frac{u_n r_n}{\sqrt{n}} = \frac{\mathscr{O}}{\sqrt{n}}$. □

References

[1] I. van den Berg, An external probability order theorem with applications, in F. & M. Diener editors, *Non Standard Analysis in Practice*, Springer-Verlag, Universitext, 1995

[2] J. Bosgiraud, "Nonstandard chi-squared test", *J.* Information & Optimization Sciences, **26** (2005) 443–470.

[3] J. Bosgiraud, Nonstandard likelihood ratio test in exponential families, in I. van den Berg & V. Neves editors, *The strength of Nonstandard Analysis,* Springer-Verlag, 2007.

[4] A. Dembo, O. Zeitouni, *Large deviations Techniques and Applications,* Springer-Verlag, 1998.

[5] M.I. Freidlin, A.D. Wentzell, *Random Pertubations of Dynamical Systems,* Springer-Verlag, 1998.

[6] F. Koudjeti and I. van den Berg, Neutrices, external numbers and external calculus, in F. et M. Diener editors*, Non Standard Analysis in Practice*, Springer-Verlag, Universitext, 1995.

[7] S.R.S. Varadhan , *Large Deviations and Applications*, S.I.A.M., Philadelphia, 1984.

Adresse de l'auteur:

Laboratoire d'Analyse, Géométrie et Applications (U.M.R. 7539)
Université de Paris 8
93526 Saint-Denis cedex 02, FRANCE

Email: jacques.bosgiraud@univ-paris8.fr

Des lois log-normales presque normales

Jacques Bosgiraud

Abstract

If a normal distribution and a log-normal distribution have same appreciable expectation and same infinitesimal variance, then these distributions are asymptotic on their bodies.

KEYWORDS: normal distribution, log-normal distribution, external calculus.

1 Introduction.

Quand on modélise une grandeur aléatoire positive X par une loi normale, on est face au paradoxe qui fait que dans le modèle X prend des valeurs négatives, alors que dans la réalité X est strictement positive (et même comprise entre des valeurs connues). Il y a bien sûr des réponses à ce paradoxe. Mais il n'est pas possible, dans la modélisation, d'écrire des expressions comme $\log_b X$ ou X^α (qui interviennent dans la formulation des fonctions d'utilité en économie, par exemple). Par contre, ce genre d'expressions est tout à fait licite si X suit une loi log-normale. Nous allons voir que si l'écart type σ de X peut être considéré comme négligeable devant son espérance μ, les modélisations par une loi normale ou une loi log-normale sont pratiquement identiques.

2 Densités normales et log-normales.

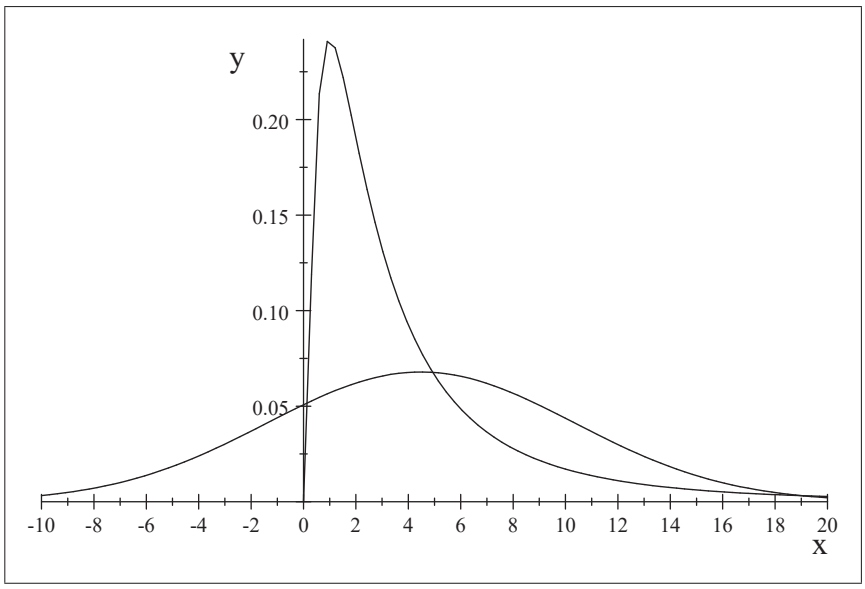
On dit qu'une variable aléatoire X suit une loi log-normale de paramètres a et b si la variable aléatoire $\ln X$ suit la loi normale de paramètres a et b. Ainsi $a = E(\ln X)$ et $b^2 = V(\ln X)$. La densité f de la loi de X vérifie $f(x) = 0$ si $x \leq 0$; et pour $x > 0$,

$$f(x) = \frac{1}{bx\sqrt{2\pi}} \exp(-\frac{(\ln(x) - a)^2}{2b^2}).$$

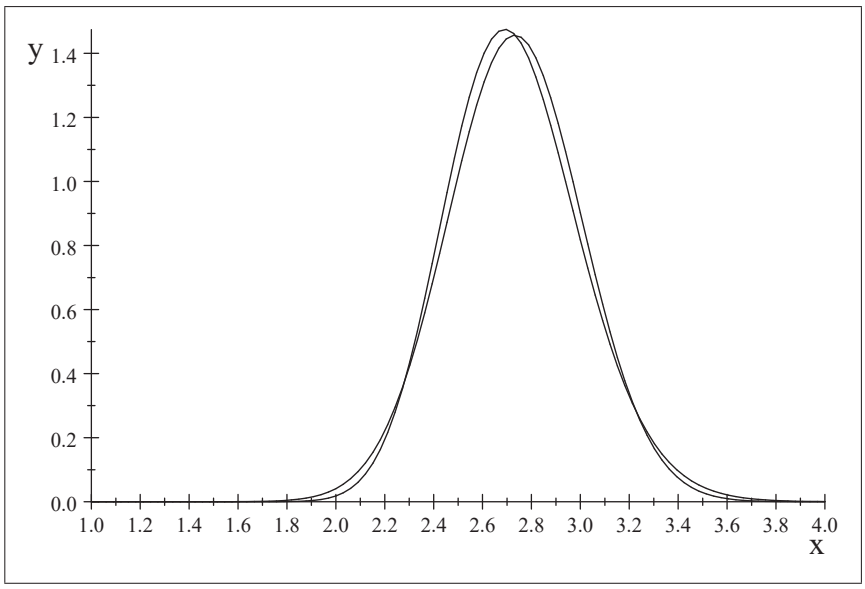
Notons $\mu := E(X) = e^{a+\frac{b^2}{2}}$ et $\sigma^2 := V(X) = e^{2a+b^2}(e^{b^2} - 1)$. Soit g la densité de la loi normale de paramètres μ et σ : pour tout réel x,

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x - \mu)^2}{2\sigma^2}).$$

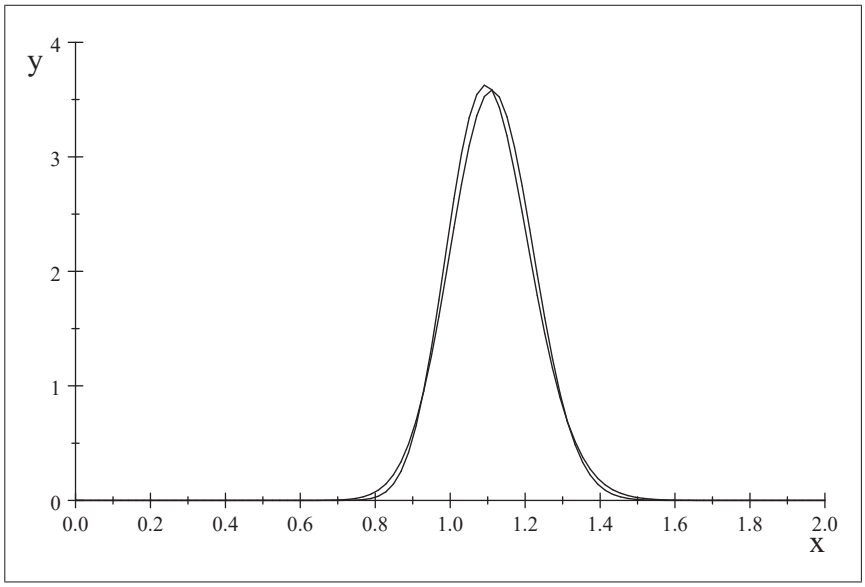
Ci dessous les graphes de f et g pour $a = b = 1$ (alors $\mu = 4.481689070$ et $\sigma = 5.874743662$):



Pour $a = 1$ et $b = 0.1$, les graphes sont les suivants (alors $\mu = 2.731907273$ et $\sigma = 0.2738751280$):



Enfin, pour $a = 0.1$ et $b = 0.1$, les graphes sont les suivants (alors $\mu = 1.110710610$ et $\sigma = 0.1113493176$):



3 Un résultat non standard.

Nous nous proposons de démontrer le résultat suivant (nous utilisons les notations du calcul externe [2]) :

Proposition. *L'espérance μ est appréciable et l'écart-type σ infinitésimal si et seulement si a est limité et b infinitésimal. Dans ce cas, pour tout réel x élément de $\mu + \mathcal{L}\sigma$, $f(x) = (1 + \emptyset)g(x)$.*

Nous travaillons sur la galaxie $\mu + \mathcal{L}\sigma$, car $PR(X = \mu + \mathcal{L}\sigma) = 1 + \emptyset$ (relation (8.4) dans [1]).C'est ce que l'on nomme le corps des distributions.

Démonstration. Les expressions de μ et σ en fonction de a et b établissent immédiatement le premier résultat. Etablissons le second.

Puisque b est infinitésimal, on peut écrire: $e^{b^2} - 1 = (1 + \emptyset)b^2$, $e^{a+\frac{b^2}{2}} = (1 + \emptyset)e^a$, $e^{2a+b^2} = (1 + \emptyset)e^{2a}$, et donc $\sigma = e^{a+\frac{b^2}{2}}(e^{b^2} - 1)^{1/2} = (1 + \emptyset)e^a b$ et $\mu = (1 + \emptyset)e^a$. Ainsi

$$g(x) = \frac{1 + \emptyset}{be^a\sqrt{2\pi}} \exp(-\frac{(x - \mu)^2}{2\sigma^2}).$$

Soit donc x élément de $\mu + \mathcal{L}\sigma$, noté $x = \mu + l\sigma$ ou encore $x = e^a(1 + \lambda b)$. La relation

$$e^{a+\frac{b^2}{2}} + le^{a+\frac{b^2}{2}}(e^{b^2} - 1)^{1/2} = e^a(1 + \lambda b)$$

implique (puisque b est infinitésimal)

$$1+\lambda b = e^{\frac{b^2}{2}}(1+l(e^{b^2}-1)^{1/2}) = (1+(1+\emptyset)b^2)(1+(1+\emptyset)lb) = 1+(1+\emptyset)lb+(1+\emptyset)b^2.$$

Ainsi $\lambda = (1+\emptyset)(l+b)$, et donc (puisque b est infinitésimal):

l est infinitésimal si et seulement si λ est infinitésimal,

l est appréciable si et seulement si λ est appréciable, et alors $\lambda = (1+\emptyset)l$.

Etablissons maintenant $f(x) = (1+\emptyset)g(x)$ pour ces deux cas. D'une part,

$$g(x) = \frac{1+\emptyset}{be^a\sqrt{2\pi}}\exp(-\frac{l^2}{2}).$$

D' autre part, puisque $\mu = (1+\emptyset)e^a$ est appréciable et que σ est infinitésimal, alors $x = (1+\emptyset)e^a$ et donc

$$f(x) = \frac{1+\emptyset}{be^a\sqrt{2\pi}}\exp(-\frac{(\ln(e^a(1+\lambda b))-a)^2}{2b^2}) =$$

$$\frac{1+\emptyset}{be^a\sqrt{2\pi}}\exp(-\frac{(\ln(1+\lambda b))^2}{2b^2}) = \frac{(1+\emptyset)}{be^a\sqrt{2\pi}}\exp(-(1+\emptyset)\frac{\lambda^2}{2}).$$

Dans le cas où $l = \emptyset$, alors $\lambda = \emptyset$ et il est clair que

$$f(x) = g(x)(1+\emptyset) = \frac{1+\emptyset}{be^a\sqrt{2\pi}}.$$

Et dans le cas où l est appréciable, la relation $\lambda = (1+\emptyset)l$ implique bien $f(x) = (1+\emptyset)g(x)$.

On a bien prouvé que si $x = \mu + \mathcal{L}\sigma$, alors $f(x) = (1+\emptyset)g(x)$.□

Références

[1] I. van den Berg, An external probability order theorem with applications, *Nonstandard Analysis in Practice*, F. and M. Diener Eds., Springer-Verlag, Universitext (1995) 171–183.

[2] F. Koudjeti and I. van den Berg, Neutrices, external numbers and external calculus, *Nonstandard Analysis in Practice*, F. and M. Diener Eds., Springer-Verlag, Universitext (1995) 145–170.

Adresse de l'auteur :

Laboratoire d'Analyse, Géométrie et applications (U.M.R. 7539)
Université de Paris 8
93526 Saint-Denis cedex 02, FRANCE

Email : jacques.bosgiraud@univ-paris8.fr

On the implicite interest rate in the Yunus equation

Marc Diener, Pheakdei Mauk

1 Introduction : microcredit

Microcredit is a set of contracts taylored to provide very small loans to very poor people to help develop small businesses or activities generating income. The basic idea came from the finding that a large part of humanity has no access to traditional credit because banks require their borrowers to meet a range of criteria, such as being able to read and write, bears some identification documents, or to have already secured a minimum deposit. The first experiments date back to the 70s in Bangladesh as an initiative of Muhammad Yunus, then a professor of economics at Chittagong University. In 1974, he watched helplessly as a terrible famine in the little village Joha near to his University. He then with his students asks the craftsmen and peasants of the village in order to try to understand their needs and lists a demand for small loans for 42 women to whom he finally decides to pay himself a total of about 27 Euros. Then he spends nearly 10 years trying to persuade banks to take on these loans before finally deciding to start his own bank, the Grameen Bank in 1983. This bank and himself receive the Nobel Prize for Peace in 2006. Currently microcredit activity has spread to most countries in the world, it is ensured by close 10 000 Micro Finance Institutes (MFIs) who lend 50 billions euros to almost 500 millions beneficiaries. The main characteristics of microcredit are

- Very small loans over short periods (10 Euros a year) with frequent (weekly) reimbursements.
- Beneficiaries are mostly women.
- Borrowers can't provide personal wealth to secure the loan.
- Usually loans are with joint liability of a group of borrowers (5 to 30) each borrower receiving her loan individually but all are interdependent in that they must assume all or part of the failure (called the "default") of any member of the group.
- Interest rates are high, around 20%, some of them up to 30%.
- Possibility of a new loan granted automatically in case of timely refunds (dynamic incentive mechanism).
- repayment rate close to 100%.

2 The Yunus polynomial and equation

Exemple : The following example has been given by Muhammad Yunus [1][2].

Grameen lends 1000 BDT (Bangladesh Taka) to borrowers that pay back 22 BDT[1] each week during 50 weeks. Let's denote by r the annual continuously compound interest rate. The present value of the 22 BDT refunded after one week is $22e^{\frac{-r}{52}}$ this value of those of the next

[1]The value of 100 Bangladesh Taka (BDT) is about 1 Euro.

payment is $22e^{\frac{-2r}{52}}$ ...and so on. So, letting $q = e^{-\frac{r}{52}}$, as the 50 refundings balance the 1000 BDT received, we get following equation for q :

$$1000 = 22\sum_{k=1}^{50} q^k = 22\frac{q - q^{51}}{1-q} \tag{1}$$

that reduces to (the degree 51 polynomial equation) $\mathcal{Y}(q) = 0$ where $\mathcal{Y}$ denotes what we shall call que *Yunus polynomial*

$$\mathcal{Y}(q) := 22q^{51} - 1022q + 1000. \tag{2}$$

We observe that $\mathcal{Y}$ has obviously $q = 1$ as zero, has two other zeros $q_- < 0 < q_+ < 1$, and all other zeros are complex conjugate. An approximation of q_+ gives $q_+ = 0.9962107\ldots$ which leads to $r = 19,74\ldots$, so nearly 20%.

But some borrowers don't pay in time, so the n-th payment takes place at some random time $T_n = T_{n-1} + \frac{1}{52}X_n = \frac{1}{52}(X_1 + X_2 + \ldots + X_n)$

Lets assume $(X_i)_{i=1..50}$ i.i.d, $X_i \rightsquigarrow \mathcal{G}(p)$, the geometric distribution, $p = \mathbb{P}(X_i = 1)$, close to 1 ; in other words each week the borrower has probability p to be able to pay the 22 BDT she should pay, weekly refunding accidents being assumed to be independent. So r becomes a random variable, $R = r(X_1, \ldots, X_{50})$, satisfying the "Yunus equation" :

$$1000 = \sum_{n=1}^{50} 22e^{-\frac{R}{52}(X_1+\ldots+X_n)} \left(= 22\sum_{n=1}^{50} v^{R(X_1+\ldots+X_n)} \quad , \quad v = e^{-\frac{1}{52}} \right). \tag{3}$$

For the sake of getting a better understanding of the risks faced by the lender under these new asumptions we wish to have informations on the probability law of the random variable R. The sequel of this paper is devoted to the results we got so far.

3 Actuarial expected rate

Let us call *actuarial expected rate* the positive real number $\overline{r}$ such that, replacing R by $\overline{r}$, it satisfies the expectation of the Yunus equation :

$$\begin{aligned} 1000 &= \mathbb{E}\left(\sum_{n=1}^{50} 22v^{-\overline{r}(X_1+\ldots+X_n)}\right) \quad , \quad v = e^{-\frac{1}{52}} \\ &= 22\sum_{n=1}^{50} \mathbb{E}\left(v^{-\overline{r}X_1}\right)\ldots\mathbb{E}\left(v^{-\overline{r}X_n}\right) \text{ , as } X_1\ldots X_n \text{ are independent} \\ &= 22\sum_{n=1}^{50} \overline{q}^n = 22\ \frac{\overline{q} - \overline{q}^{51}}{1-\overline{q}} \text{ , with } \overline{q} = \mathbb{E}(e^{-\frac{\overline{r}}{52}X_1}) = M_{X_1}\left(-\frac{\overline{r}}{52}\right), \end{aligned}$$

where $M_{X_1}(t) = \frac{pe^t}{1-(1-p)e^t}$ is the moment generating function of $\mathcal{G}(p)$. So $M_{X_1}\left(-\frac{\overline{r}}{52}\right) = \overline{q} = q_+$, the positive non trivial zero of the Yunus polynomial, which leads to

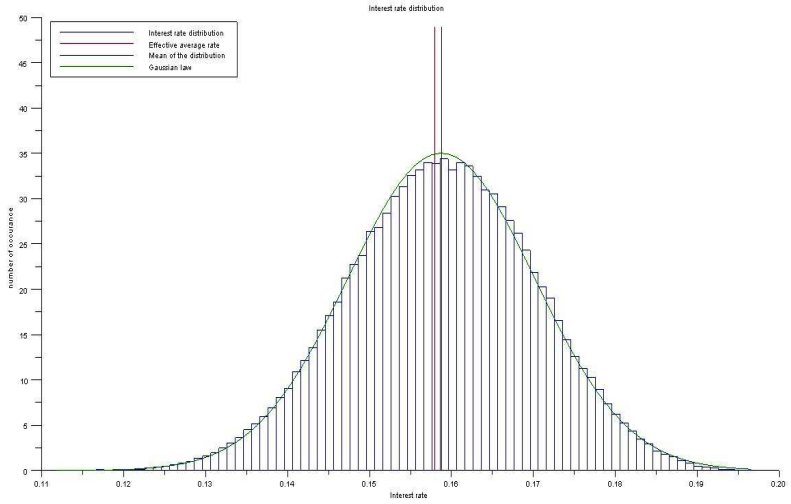
$$e^{-\frac{\overline{r}}{52}} = \frac{q_+}{q_+ + p(1-q_+)} = \frac{1}{1 + p\left(\frac{1}{q_+} - 1\right)}.$$

So $\overline{r} = 52\ln\left(1 + p\left(\frac{1}{q_+} - 1\right)\right)$.

4 Some experimental results

We have the chance to have with us three good students[2] from Polytech'Nice with skills in Scilab that did some numerical experiments. It turns out that the law for R is impressively similar to a Gaussian $\mathcal{N}(\mu, \sigma)$, with $\mu = \overline{r}$

[2] Léo Augé, Aurore Lebrun, and Anaïs Pozin



On the other hand, obviously $0 \leq R \leq 20\%$, so R can't be Gaussian, and indeed, even if the skewness of R if very small, its kurtosis is close to 1, not 3.

So the question is (and stays, up to here) : what is the law of R ?

5 Where could infinitesimals enter the model ?

Here some remarks related to the idea that "50=N is large". Observe that, for $N = 50$ and $a = 10\%$ the equation $\mathcal{Y}(q) = 0$ is equivalent (dividing both members by 20) to

$$\varphi(q) \ := (1+a)q^{N+1} - (N+1+a)q + N = 0 \tag{4}$$

Now assume N is infinitely large and let $q = 1 + \frac{x}{N}$, so x is a blow-up of q around $q = 1$. Let $\psi(x) := \varphi(1 + \frac{x}{N})$, so (4) reduces to $\psi(x) = 0$. But, as N is infinitely large, and denoting by ϕ any infinitesimal, we have

$$\begin{aligned}
\psi(x) &= \varphi\left(1+\frac{x}{N}\right) \\
&= (1+a)\exp\left((N+1)\ln\left(1+\frac{x}{N}\right)\right) - (N+1+a)\left(1+\frac{x}{N}\right) + N \\
&= (1+a)\exp\left((N+1)\frac{x}{N}(1+\phi)\right) - x - (1+a)\left(1+\frac{x}{N}\right) \\
&= (1+a)\exp(x+\phi) - x - (1+a)(1+\phi) \\
&\simeq (1+a)(e^x - 1) - x =: \psi_a(x).
\end{aligned}$$

Actually, the equation $\psi_0(x) = 0$ has two solutions : $x = 0$ and $-x_+ < 0$ so we get the approximation $q_+ = 1 - \frac{x_+}{N}(1+\phi)$ for the non-trivial positive solution of $\mathcal{Y}(q) = 0$. For $a = 10\%$ for instance we have $-x_+ = -0.1937476\ldots$, and $1 - \frac{x_+}{50} = 0.9961250\ldots$.

So, denoting by $-x_+(a)$ the solution of $\psi_a(x) = 0$ different of 1 we have

Proposition 1 *Assume that the number N of refunds is infinitely large. Then the actuarial expected rate is*

$$\overline{r}(a) = \frac{1}{1 + p\left(\frac{1}{1-\frac{x_+(a)}{N}(1+\phi)}\right)}. \tag{5}$$

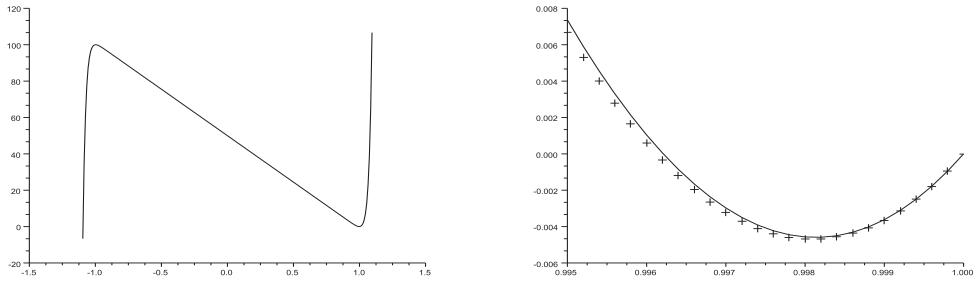


Figure 1: The graph of function $q \mapsto \varphi(q)$, and its blow-up near $q = 1$, for $N = 50$ and $a = 10\%$. The "+" signs are on the graph of $x \mapsto \psi_0(x)$ rescaled so that $q = 1 + \frac{x}{N}$.

References

[1] M. Yunus avec Alan Jolis. *Vers un monde sans pauvreté.* JC Lattès, 1997.

[2] M. Yunus with Alan Jolis. *Banker to the Poor : micro-lending and the battle against world poverty.* Public Affairs, 1999.

Address of the authors:

Université de Nice Sophia-Antipolis
Laboratoire de Mathématiques Jean Dieudonné
Parc Valrose
06108 Nice cedex 2, France

E-mail: `diener@unice.fr, pheak@unice.fr`

The Blasius equation

Bernard Brighi, Augustin Fruchard, Tewfik Sari

Abstract. The Blasius problem $f''' + ff'' = 0$, $f(0) = -a$, $f'(0) = b$, $f'(+\infty) = \lambda$ is investigated, in particular in the difficult and scarcely studied case $b < 0 \leqslant \lambda$. The shape and the number of solutions are determined. The method is first to reduce to the Crocco equation $uu'' + s = 0$ and then to use an associated autonomous planar vector field. The most useful properties of Crocco solutions appear to be related to canard solutions of a slow fast vector field.

KEYWORDS : Blasius equation, Crocco equation, boundary value problem on infinite interval, canard solution.

1 Introduction

In this article we present a selection of results of [6]. The reader is referred to [6] for complete proofs, additional and intermediate results. We take the occasion to completely change the order of presentation: in [6] we first give the results on the Blasius equation with a sketch of proof, then we introduce the Crocco equation and the vector field, we establish results and proofs on these intermediate equations and then we return to the proof of the initial result. Here we choose a different order and we postpone the main result at the end of the article. We hope that this article may be a first approach before a thorough study of [6].

The article is organized as follows. In Section 2, we state the main problem of the paper which is the investigation of the following *Blasius Boundary Value Problem* (BBVP for short)

$$f''' + ff'' = 0 \quad \text{on} \quad [0, +\infty[, \tag{1}$$

$$f(0) = -a, \quad f'(0) = b, \quad \lim_{t \to +\infty} f'(t) = \lambda. \tag{2}$$

We list some former results, according to the relative values of b and λ, and we focus our attention on the case $b < 0 \leqslant \lambda$, which is our case of interest. In Section 3, we show that, in the latter case, this boundary value problem is equivalent to the *Crocco Boundary Value Problem* (CBVP)

$$\begin{cases} uu'' + s = 0 \quad \text{on } [b, \lambda[, \\ u'(b) = a, \quad \lim\limits_{s \to \lambda} u(s) = 0. \end{cases} \tag{3}$$

where $[b, \lambda[$ appears as the maximal right-interval of definition of the solution. In Section 4, we show that the similarity properties of the Blasius or the Crocco solutions permit to reduce the non autonomous second order differential equation of Crocco to an autonomous planar vector and we notice that the maximal right-interval of definition of the solutions of the Crocco equation presents a discontinuity with respect to the initial condition. It is well known that the maximal right-interval of definition of the solution of a differential equation is not continuous in general with respect to the initial conditions. It is simply lower semicontinuous. Actually, in Section 6, we see that the solutions of the Crocco differential equation which are close to 0 for s close to 0 are *canard* solutions of a slow-fast vector field. These solutions play an important role in the

description of the discontinuity of the maximal right-interval of definition of the solution of the Crocco equation. In Section 7, we analyze this discontinuity which occurs along a particular orbit of the planar vector field considered in Section 4. In Section 8, we give a lower bound of the number of solutions of the boundary value problem associated to the Blasius equation, in the case $b < 0 \leqslant \lambda$. Section 9 describes a difficulty encountered in numerical simulations. Indeed, due to the canard solutions phenomenon, some solutions of the Crocco equation become exponentially small for $s < 0$ and the numerical scheme cannot give the right solution. We show how to use the theoretical study in Section 5 to overcome this difficulty.

2 The Blasius Boundary Value Problem

The Blasius Boundary Value Problem (1-2) arises for the first time, with $a = b = 0$ and $\lambda = 2$, in 1907 in the thesis of Blasius [3, 4]. In the case $a = b = 0$, Hermann Weyl [16] proves that the BBVP has one and only one solution. The proof is elementary but strongly uses the fact that $a = b = 0$, see also [5, 8]. The BBVP plays a central role in fluid mechanics [12]: The Blasius equation (1) was obtained using a similarity transform and enabled successful treatment of the laminar boundary layer on a flat plate. By considering equation (1) as a first order linear differential equation for f'', we obtain

$$f''(t) = f''(0) \exp\left\{ - \int_0^t f(\tau) d\tau \right\}.$$

Hence, the BBVP splits into three cases, called respectively linear, concave and convex:

- If $\lambda = b$, then the BBVP has a unique solution, given by $f(t) = bt - a$.
- If $\lambda > b$, then any possible solution must satisfy $f''(t) > 0$ for all $t \geqslant 0$, i.e. has to be convex.
- If $\lambda < b$, then possible solutions are concave.

The concave case is completely solved and well-known [1].

Proposition 1 — *In the case $\lambda < b$, the BBVP (1 - 2) has exactly one solution if $0 \leqslant \lambda < b$, and no solution if $\lambda < 0$.*

When $b \geqslant 0$, the convex case $\lambda > b$ is also well-known, see [8].

Proposition 2 ([6] Corollary 3.6) — *The BBVP (1 - 2), where $b \geqslant 0$ and $\lambda > b$, has exactly one solution when $a \leqslant 0$ or $b > 0$. When $a > 0$ and $b = 0$, the BBVP has exactly one solution for all $\lambda > a^2\lambda_+$ and no solution if $0 < \lambda \leqslant a^2\lambda_+$, where $\lambda_+ \simeq 1.304$ is defined in Proposition 5.*

It is known that every solution of the Blasius equation (1) such that $f''(0) > 0$ is defined for all t and its derivative has a finite and non-negative limit as $t \to +\infty$ ([6] Proposition 3.1). Thus

Proposition 3 — *The BBVP (1 - 2) has no solution if $b < \lambda < 0$.*

In this article, we focus on the remaining case $b < 0 \leqslant \lambda$, which is much richer and trickier. Non uniqueness for the BBVP is mentioned in the literature, but either only supported by numerical investigations [9], or with incomplete proofs [10, 13].

The Blasius equation (1) has the following similarity property:

$$\text{If } t \mapsto f(t) \text{ is a solution of (1), so is } t \mapsto \sigma f(\sigma t), \text{ for all } \sigma \in \mathbb{R}. \tag{4}$$

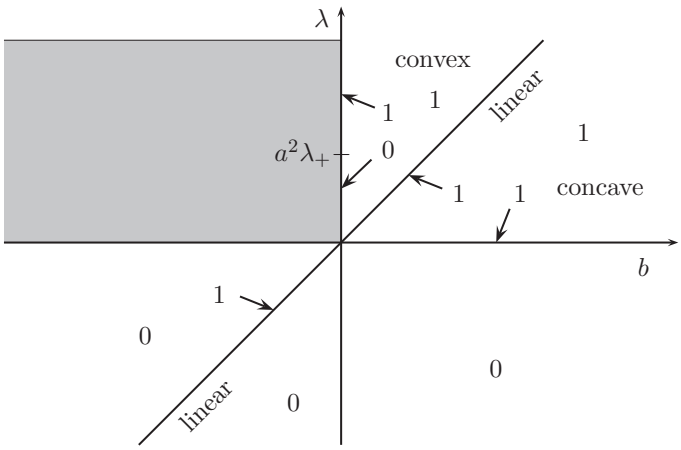


Figure 1: *In the plane* (b, λ)*, the number of solutions of the BBVP* (5) *in each region and on their border. In gray, the remaining region to investigate, purpose of this article.*

This allows us to restrict our attention without loss of generality to the case $b = -1$, i.e. to the BBVP

$$\begin{cases} f''' + ff'' = 0 \ \text{ on } [0, +\infty[, \\ f(0) = -a, \quad f'(0) = -1, \quad \lim\limits_{t\to+\infty} f'(t) = \lambda \geqslant 0. \end{cases} \tag{5}$$

The purpose of this article is to count the number of solutions of (5). The main result, at the end of Section 8, gives a minimum number of solutions of (5) depending on the values of a and λ. We conjecture that this minimum number is the exact number of solutions.

3 The Crocco Boundary Value Problem

Let f be a convex solution of (1). Since $f'' > 0$ it follows that $t \mapsto f'(t)$ is a diffeomorphism. Hence we can use f' as an independent variable and express f'' as a function of f'. This is the so-called *Crocco transformation* [7]

$$s = f', \qquad f'' = u(s)$$

Differentiating $f'' = u(f')$ we obtain $u'(s) = -f$. Differentiating once again we obtain that u satisfies the following so-called *Crocco differential equation*

$$u''u + s = 0. \tag{6}$$

As we will see, the BBVP (5) is equivalent to the Crocco Boundary Value Problem (3) for $b = -1$, rewritten below for convenience

$$\begin{cases} uu'' + s = 0 \ \text{ on } [-1, \lambda[, \\ u'(-1) = a, \quad \lim\limits_{s\to\lambda} u(s) = 0. \end{cases} \tag{7}$$

The equivalence between (5) and (7) will become clear after the following remarks. In order to solve (5) we use the *shooting method*. Let $f(\,\cdot\,; a, c)$ denote the solution of the Blasius Initial

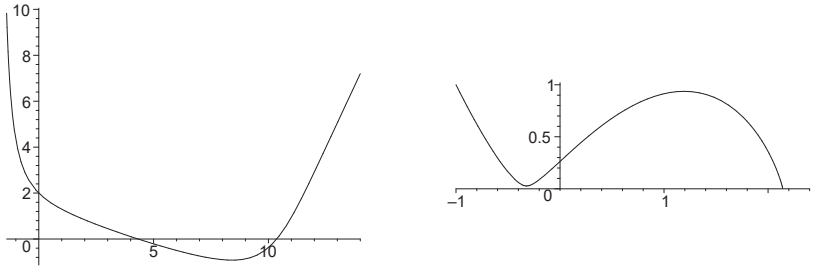


Figure 2: *On the left, the Blasius solution* $t \mapsto f(t;-2,1)$*; on the right, the corresponding Crocco solution* $s \mapsto u(s;-2,1)$.

Value Problem (BIVP)

$$f''' + ff'' = 0, \quad f(0) = -a,\ f'(0) = -1,\ f''(0) = c > 0. \tag{8}$$

The solution $f(\,\cdot\,;a,c)$ is defined for all $t \geqslant 0$ and its derivative has a finite and non-negative limit as $t \to +\infty$ ([6] Proposition 3.1). Let $\widetilde{\Lambda}(a,c)$ denote the limit

$$\widetilde{\Lambda}(a,c) := \lim_{t\to+\infty} f'(t;a,c) \geqslant 0. \tag{9}$$

Then ([6] Proposition 2.1), $[-1,\widetilde{\Lambda}(a,c)[$ is the maximal right interval of existence of the solution $u(\,\cdot\,;a,c)$ of the Crocco Initial Value Problem (CIVP)

$$uu'' + s = 0, \qquad u(-1) = c > 0,\ u'(-1) = a. \tag{10}$$

See Figure 2 for a comparison between a Blasius solution and the corresponding Crocco solution.

Moreover, we have

$$\lim_{s\to\widetilde{\Lambda}(a,c)} u(s) = 0 \quad \text{and} \quad \left(\widetilde{\Lambda}(a,c) > 0 \Rightarrow \lim_{s\to\widetilde{\Lambda}(a,c)} u'(s) = -\infty\right)$$

This shows that (5) is equivalent to (7).

4 Symmetries

The similarity property (4) is rewritten as follows for the Crocco equation (6)

$$\text{If } \sigma > 0 \text{ and } s \mapsto u(s) \text{ is a solution of (6), so is } u^\sigma : s \mapsto \sigma^3 u(\sigma^{-2}s). \tag{11}$$

This similarity property reduces the Crocco equation (6) to a system of autonomous differential equations. Actually, the change of variables

$$x(s) = (-s)^{-1/2}u'(s), \qquad y(s) = (-s)^{-3/2}u(s)$$

leads to the system

$$x' = \frac{\frac{1}{2}\,x + \frac{1}{y}}{-s}, \qquad y' = \frac{x + \frac{3}{2}\,y}{-s}.$$

Then, using the change of independent variables $s = -e^{-\tau}$, we obtain the planar vector field

$$\dot{x} = \frac{1}{2}\,x + \frac{1}{y}, \qquad \dot{y} = x + \frac{3}{2}\,y, \tag{12}$$

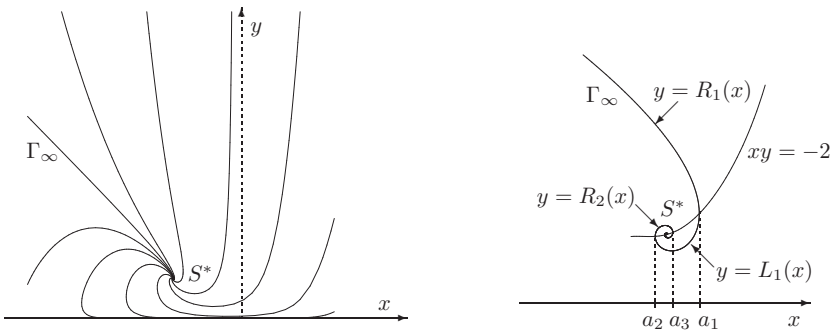


Figure 3: *On the left: the phase portrait of* (12). *On the right: sketch of enlargement of* Γ_∞ *near* S^*. *The functions* L_n *and* R_n *are defined in Section 7.*

where the dot is for differentiating with respect to the new independent variable τ. The initial conditions $u(-1) = c,\ u'(-1) = a$ in the CIVP (10) correspond to

$$x(0) = a, \qquad y(0) = c.$$

Notice that this vector field describes the Crocco equation (6) only for $s < 0$, since τ tends to $+\infty$ as s tends to 0.

Because the transformation $u \mapsto u^\sigma$ given by (11) corresponds to the shift $\tau \mapsto \tau + 2\ln\sigma$, to each orbit $\big\{\big(x(\tau), y(\tau)\big); \tau \in \mathbb{R}\big\}$ of a solution of (12) corresponds a whole family $(u^\sigma)_{\sigma>0}$ of solutions of (6) connected by the similarity (11). In particular, the unique stationary point $S^* = \big(-\sqrt{3}, \frac{2}{\sqrt{3}}\big)$ of (12) corresponds to the unique self-similar positive solution u_* of (6), *i.e.* satisfying $u_*(s) = \sigma^3 u_*(\sigma^{-2}s)$ for $s < 0 < \sigma$, namely

$$u_*(s) = \tfrac{2}{\sqrt{3}}(-s)^{3/2}. \tag{13}$$

A study of this vector field, detailed in [6], shows the following.

- All nonstationary solutions of (12) are defined on $\mathbb{R}$; they tend to S^* as $\tau \to -\infty$ and to infinity as $\tau \to +\infty$.
- There is one and only one orbit, denoted by Γ_∞, such that any solution (x, y) parametrizing Γ_∞ satisfies that $\frac{x(\tau)}{y(\tau)}$ tends to -1 as $\tau \to +\infty$, see Figure 3.
- For all solutions $\big(x(\tau), y(\tau)\big)$ except those on $\Gamma_\infty \cup \{S^*\}$, the quotient $\frac{x(\tau)}{y(\tau)}$ tends to 0 as $\tau \to +\infty$.
- ([6] Theorem 2.4) For all solutions $\big(x(\tau), y(\tau)\big)$ except those on $\Gamma_\infty \cup \{S^*\}$, $\frac{x(\tau)^3}{y(\tau)}$ has a limit $k \in \mathbb{R}$ as $\tau \to +\infty$. The number k parametrizes the orbit of (12) denoted by Γ_k. In terms of positive Crocco solutions, we have the following properties:
  1. if $(a, c) \in \Gamma_\infty$ then $\lim\limits_{s\to 0^-} u(s; a, c) = 0$ and $\lim\limits_{s\to 0^-} u'(s; a, c) < 0$,
  2. if $(a, c) = S^*$ then $\lim\limits_{s\to 0^-} u(s; a, c) = 0$ and $\lim\limits_{s\to 0^-} u'(s; a, c) = 0$,
  3. Otherwise $(a, c) \in \Gamma_k$ for some $k \in \mathbb{R}$. In that case, $u(0; a, c) > 0$ and $u'(0; a, c)$ is of the same sign as k.

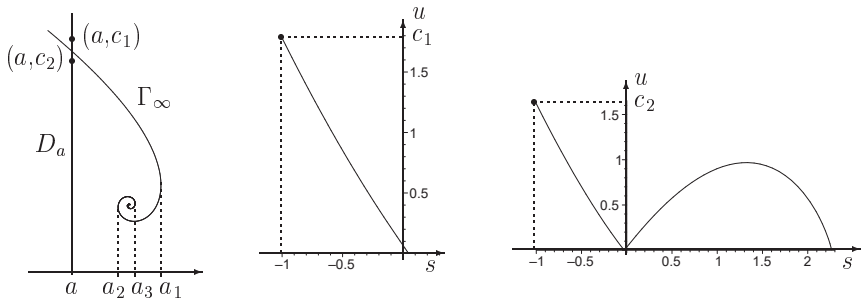


Figure 4: *A sketch of the spiral* Γ_∞ *and two numerical Crocco solutions with initial conditions* $u_1(-1) = c_1 = 1.78$, $u_1'(-1) = a = -2$ *and* $u_2(-1) = c_2 = 1.62$, $u_2'(-1) = a = -2$, *where* (a, c_1) *and* (a, c_2) *are on the convex and on the concave sides of* Γ_∞ *respectively.*

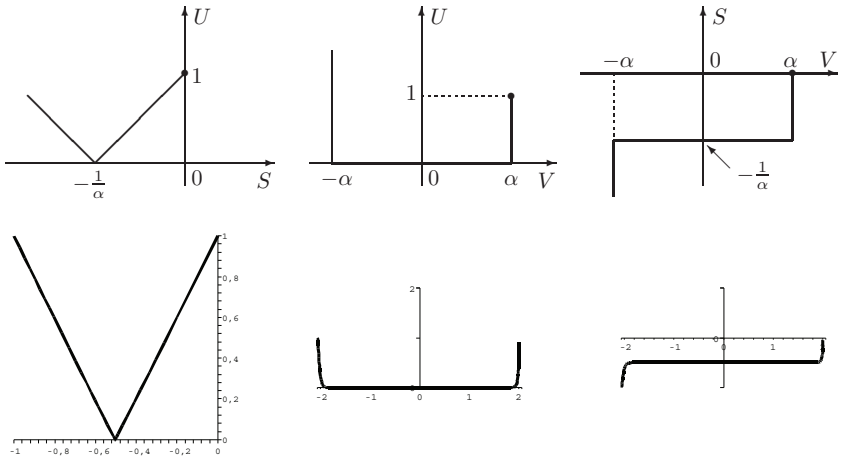


Figure 5: *Above: schematic graphs of the solution of* (14) *in the limit* $\varepsilon \to 0$, *respectively in the variables* S, U, *the variables* U, V *and* S, V. *Below: the numerical solution corresponding to* $\varepsilon = 0.1$ *and* $\alpha = 2$.

Let (a, c) be an initial condition which is close to Γ_∞. If (a, c) lies on the convex side of Γ_∞, then $u(0; a, c)$ is close to 0 and $u'(0; a, c) < 0$. Since $u(s; a, c)$ becomes concave for $s > 0$, it follows that $\widetilde{\Lambda}(a, c)$ is close to 0. Otherwise, if (a, c) lies on the concave side of Γ_∞, then $u(0; a, c)$ is small and $u'(0; a, c) > 0$. In fact, this implies that $\widetilde{\Lambda}(a, c)$ is not close to 0, see Figure 4.

This shows that the function $\widetilde{\Lambda}(a, c)$ is discontinuous on Γ_∞. The precise description of this discontinuity needs the knowledge of the behavior of the solutions $u(s)$ of the Crocco equation (6) for which $u(0)$ is close to 0. This behavior is described in the following section.

5 Crocco solutions near $u = 0$ ...

The following statement describes Crocco solutions close to 0 and with positive slope for $s = 0$: they take an exponentially small value at some small negative abscissa of s and then go far from the s axis.

Proposition 4 — *Fix* $\alpha > 0$ *and let* $0 < \varepsilon \to 0$. *Let* $u = u(s, \varepsilon)$ *denote the solution of* (6) *such that* $u(0, \varepsilon) = \varepsilon$ *and* $u'(0, \varepsilon) = \alpha$. *Then* $u(s, \varepsilon)$ *reaches its minimum at some abscissa*

$s = \kappa(\varepsilon) < 0$ *satisfying* $\kappa(\varepsilon) = -\frac{\varepsilon}{\alpha}\big(1 + o(1)\big)$ *and*

$$u(\kappa(\varepsilon), \varepsilon) = \exp\Big(-\frac{\alpha^3}{2\varepsilon}\big(1 + o(1)\big)\Big) \quad \text{as} \quad \varepsilon \to 0.$$

Moreover, for all $B < 0$ *fixed, we have* $u(\varepsilon S, \varepsilon) = \varepsilon\big(|\alpha S + 1| + o(1)\big)$ *as* $\varepsilon \to 0$, *uniformly for* $S \in [B, 0]$.

Proof. The reference [6] contains two proofs: one in Section 5 and an alternative one in Section 6. We give here an overview of the second one. The solution $u(s, \varepsilon)$ is defined for all $s \leqslant 0$ and is positive. The function $U(S, \varepsilon)$, defined by

$$U(S, \varepsilon) = \frac{1}{\varepsilon} u\left(\varepsilon S, \varepsilon\right),$$

is the solution of the initial value problem

$$U\frac{d^2U}{dS^2} + \varepsilon S = 0, \qquad U(0) = 1, \quad \frac{dU}{dS}(0) = \alpha. \tag{14}$$

Except near the axis $U = 0$, and for bounded values of S, U'' is close to 0, *i.e.* the solutions are almost affine. Precisely, one has for all fixed $S_0 \in \left]-\frac{1}{\alpha}, 0\right]$

$$U(S, \varepsilon) = \alpha S + 1 + o(1) \quad \text{as} \quad \varepsilon \to 0, \quad \text{uniformly for } S \in [S_0, 0] \tag{15}$$

What is less obvious is that this approximation is still valid up to $-\frac{1}{\alpha}$ and that, for any fixed $B \leqslant -\frac{1}{\alpha}$, the solution satisfies the approximation $U(S, \varepsilon) = -\alpha S - 1 + o(1)$ uniformly for $B \leqslant S \leqslant -\frac{1}{\alpha}$. In other words, after its passage near the axis, the solution $U(S, \varepsilon)$ behaves like a light ray reflecting on a mirror, see Figure 5. To see this, we use the new variable $W = \varepsilon \ln U$ and we choose $V = \frac{dU}{dS}$ as an independent variable; we obtain

$$\frac{dS}{dV} = -\frac{\mathrm{e}^{W/\varepsilon}}{\varepsilon S}, \quad \frac{dW}{dV} = -\frac{V}{S}. \tag{16}$$

In a interval where $W < 0$ and $S < 0$, we have $\lim\limits_{\varepsilon \to 0} \frac{\mathrm{e}^{W/\varepsilon}}{\varepsilon S} = 0$. Thus (16) is a regular perturbation of

$$\frac{dS}{dV} = 0, \quad \frac{dW}{dV} = -\frac{V}{S}.$$

Since S is close to $-\frac{1}{\alpha}$ when U is close to 0, we deduce that

$$S(V, \varepsilon) = -\frac{1}{\alpha} + o(1), \quad W(V, \varepsilon) = \frac{\alpha V^2}{2} + W_0 + o(1) \quad \text{as} \quad \varepsilon \to 0, \tag{17}$$

uniformly for $V \in [-V_0, V_0]$, where $W_0 < 0$ and $V_0 < \sqrt{-2W_0/\alpha}$, see Figure 6, left. With the condition $W(\alpha, \varepsilon) = o(1)$, we obtain $W_0 = -\frac{\alpha^3}{2} + o(1)$. Thus we have

$$S(V, \varepsilon) = -\frac{1}{\alpha} + o(1), \quad W(V, \varepsilon) = \alpha\,\frac{V^2 - \alpha^2}{2} + o(1) \quad \text{as} \quad \varepsilon \to 0,$$

uniformly for $V \in [-A_0, A_0]$, where A_0 can be chosen as close to α as we want. Hence we have

$$U(V, \varepsilon) = o(1) \text{ uniformly for } V \in [-A_0, A_0]. \tag{18}$$

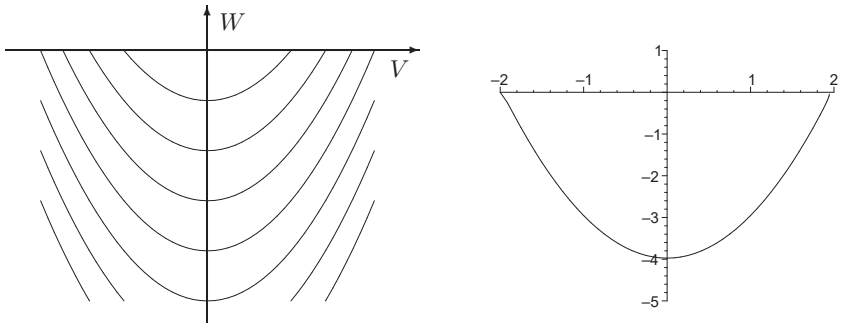


Figure 6: *On the left, a scheme of the vector field in the variables V, W. On the right, the numerical solution corresponding to $\varepsilon = 0.1$ and $\alpha = -2$.*

The minimum of $U(S,\varepsilon)$ is reached for $S = K(\varepsilon)$ which corresponds to $V = 0$. Hence

$$K(\varepsilon) = -\frac{1}{\alpha} + o(1), \qquad U(K(\varepsilon),\varepsilon) = \exp\left(\frac{W(0,\varepsilon)}{\varepsilon}\right) = \exp\left(-\frac{\alpha^3 + o(1)}{2\varepsilon}\right).$$

Thus $\kappa(\varepsilon) = \varepsilon K(\varepsilon) = \varepsilon\big(\frac{1}{\alpha} + o(1)\big)$ and, using $\varepsilon = \exp\frac{\varepsilon \ln \varepsilon}{\varepsilon} = \exp\frac{o(1)}{\varepsilon}$, we have

$$u(\kappa(\varepsilon),\varepsilon) = \varepsilon\, U(K(\varepsilon),\varepsilon) = \varepsilon\, \exp\left(-\frac{\alpha^3 + o(1)}{2\varepsilon}\right) = \exp\left(-\frac{\alpha^3 + o(1)}{2\varepsilon}\right).$$

Using again the differential equation in (14), we have $V(S,\varepsilon) = -\alpha + o(1)$ uniformly for $S \in [B, S_1]$, where $B < S_1$ and S_1 is as close to $-\frac{1}{\alpha}$ as we want. Thus

$$U(S,\varepsilon) = -\alpha\left(S + \frac{1}{\alpha}\right) + o(1), \text{ uniformly for } S \in [S_2, S_1]. \tag{19}$$

Using (15) and (19), together with (18) we conclude that

$$U(S,\varepsilon) = |\alpha S + 1| + o(1) \quad \text{uniformly for } S \in [S_2, 0].$$

Hence $u(\varepsilon S,\varepsilon) = \varepsilon\big(|\alpha S + 1| + o(1)\big)$ uniformly for $S \in [B, 0]$, as $\varepsilon \to 0$. □

6 ... are canard solutions!

The solution $U(V,\varepsilon)$ considered in the proof of Proposition 4 is a canard solution. Indeed, $\big(S(V,\varepsilon), U(V,\varepsilon)\big)$ is a solution of the slow fast system

$$\varepsilon\frac{dS}{dV} = -\frac{U}{S}, \quad \varepsilon\frac{dU}{dV} = -\frac{VU}{S}. \tag{20}$$

whose slow manifold $U = 0$ is attractive when $V < 0$ and repulsive when $V > 0$. Notice that

$$S(V) = \text{constant} < 0, \qquad U(V) = 0, \tag{21}$$

are canard solutions of (20) since they are on the attractive part of the slow manifold when $V < 0$ and on its repulsive part when $V > 0$. These solutions do not correspond to actual solutions of the differential equation in (14) since the latter are for $U \neq 0$.

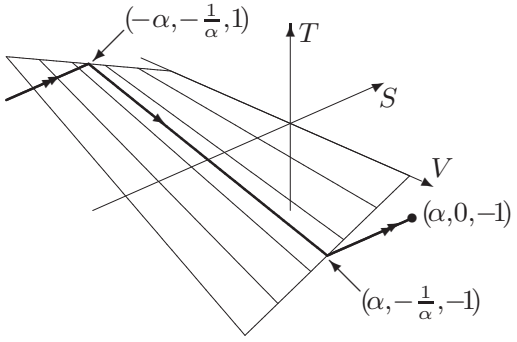


Figure 7: *The canard of (22).*

Considered as a system in $\mathbb{R}^3$, (20) is a slow-fast system with two fast variables S and U and one slow variable V. However, it is possible to rewrite it as a system with two slow variables and only one fast. Actually $T = VS - U$ is a slow variable. With this variable, (20) becomes

$$\varepsilon \frac{dS}{dV} = \frac{T - VS}{S}, \quad \frac{dT}{dV} = S. \tag{22}$$

This is a singularly perturbed system whose slow manifold is the surface $T = VS$. This slow manifold is attractive for $V < 0$ and repulsive for $V > 0$. The Tikhonov theorem (see [14, 11] and [15] Section 39) describes the behavior of the solution $\big(S(V,\varepsilon), T(V,\varepsilon)\big)$ of (22) when $V > 0$. There is a fast transition (see Figure 7) taking the trajectory $\big(V, S(V,\varepsilon), T(V,\varepsilon)\big)$, from its initial point $(\alpha, 0, -1)$, to a $o(1)$ neighborhood of the point $\big(\alpha, -\frac{1}{\alpha}, -1\big)$ of the slow manifold, preceded by a slow transition near a solution $\big(-\frac{1}{\alpha}, -\frac{V}{\alpha}\big)$ of the reduced problem

$$S = \frac{T}{V}, \quad \frac{dT}{dV} = S.$$

More precisely, for any A_0 and A_1, such that $0 < A_1 < A_0 < \alpha$, we have

$$\begin{aligned} S(V,\varepsilon) &= -\frac{1}{\alpha} + o(1) \quad \text{uniformly for } V \in [A_1, A_0], \\ T(V,\varepsilon) &= -\frac{V}{\alpha} + o(1) \quad \text{uniformly for } V \in [A_1, \alpha]. \end{aligned} \tag{23}$$

Notice that A_0 (resp. A_1) is fixed but may be chosen as close to α (resp. 0), as we want. The approximation for S does not hold near $V = \alpha$ since there is a boundary layer (fast transition) from $S = 0$ at $V = \alpha$ to $S = -\frac{1}{\alpha}$ for V close to α. We deduce that

$$U(V,\varepsilon) = VS(V,\varepsilon) - T(V,\varepsilon) = o(1) \quad \text{uniformly for } V \in [A_1, \alpha]. \tag{24}$$

A priori, Tikhonov theorem does not apply for $V \leqslant 0$, because for $V = 0$ the slow manifold becomes repulsive, but it turns out that (24) still holds for negative values of V. This is the so-called bifurcation delay [2]. The slow manifold is foliated by the explicit solutions $S(V) = S_0 =$ constant, $T(V) = VS_0$, corresponding to the solutions (21). These solutions are canard solutions since they follow the attractive part and then the repulsive part of the slow manifold, see Figure 7. Knowing the "exit" value $V = \alpha$ of the solution $T(V,\varepsilon)$ in a small neighborhood of the slow manifold, we can compute the "entry" value for which the solution was far from the slow manifold. Since $U = VS - T > 0$, we use the change of variable $W = \varepsilon \ln U$. In this variable, it appears that the "entry" of the solution in the neighborhood of the slow manifold holds asymptotically for $V = -\alpha$, as shown in the proof of Proposition 4. For details and complements see [6] Section 6.

7 The discontinuity of the function $\widetilde{\Lambda}$

Two particular solutions of the Crocco equation (6) play an important role in our study.

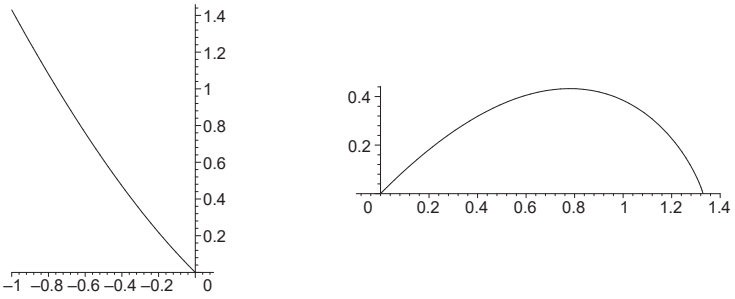


Figure 8: *The graphs of u_- on the left and of u_+ on the right.*

Proposition 5 ([6] Theorem 2.2) — *The Crocco equation* (6) *has two solutions, denoted by u_- and u_+ such that u_- is the unique solution of* (6) *satisfying*

$$\lim_{s\to 0^-} u_-(s) = 0, \qquad \lim_{s\to 0^-} u'_-(s) = -1$$

and u_+ is the unique solution of (6) *satisfying*

$$\lim_{s\to 0^+} u_+(s) = 0, \qquad \lim_{s\to 0^+} u'_+(s) = 1.$$

The solution u_- is defined on $]-\infty, 0[$ and the solution u_+ is defined on $]0, \lambda_+[$ for some $\lambda_+ > 0$.

Numerical computations give $\lambda_+ \approx 1.303918$. See Figure 8 for the graphs of u_- and u_+.

The orbit Γ_∞ on which $\widetilde{\Lambda}$ is discontinuous (See Figure 4 for an illustration of this discontinuity) is given by

$$\Gamma_\infty = \left\{ m(s) = \left((-s)^{-1/2} u'_-(s), (-s)^{-3/2} u_-(s) \right) \ ; \ s < 0 \right\}.$$

A consequence of Proposition 4 is the following; see [6] Section 5.2 for the proofs.

Proposition 6 — *Let $\big((\alpha_n, \gamma_n)\big)_{n\in\mathbb{N}}$ be some sequence of points tending to $m(-1)$.*
Then the sequence $\big(u'(0; \alpha_n, \gamma_n)\big)_{n\in\mathbb{N}}$ is bounded and has at most two cluster points: 1 *and* -1.
More precisely, if (α_n, γ_n) tends to $m(-1)$ on the convex side then $u'(0; \alpha_n, \gamma_n)$ tends to -1, and if (α_n, γ_n) tends to $m(-1)$ on the concave side then $u'(0; \alpha_n, \gamma_n)$ tends to 1.

Remark — This statement seems to contradict the well-known property of continuity with respect to initial conditions: if (a_1, c_1) and (a_2, c_2) are two points close to $m(-1)$ such that $u'(0; a_1, c_1)$ is close to -1 and $u'(0; a_2, c_2)$ close to 1, then this continuity property seems to imply that, for any fixed $d \in]-1, 1[$ there would exist (a, c) between (a_1, c_1) and (a_2, c_2) with $u'(0; a, c) = d$. In fact there is no contradiction: any small path joining (a_1, c_1) and (a_2, c_2) has to cross the "singular line" $\left\{ \big(u_-(s), u'_-(s)\big) \ ; \ s < 0 \right\}$ at some point (a_0, c_0) and the solution with this initial condition is no longer defined at 0. This could explain an error in [13] Lemma 2 p. 257, which asserts the continuity of $\widetilde{\Lambda}$.

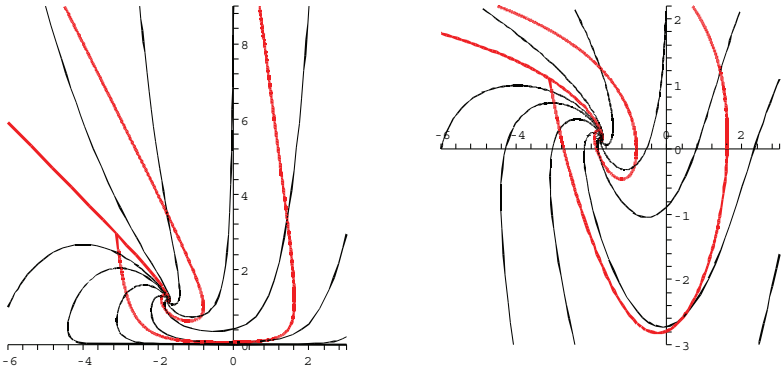


Figure 9: *Numerical graphs of some orbits* Γ_k *and of the level set curves of the function* $\widetilde{\Lambda}$. *On the left nine curves* Γ_k *for various values of* $k \in \mathbb{R} \cup \{\infty\}$ *and* $\widetilde{\Lambda}(a,c) = \lambda$ *for* $\lambda = 0, 1$ *and* 10. *On the right the sames curves in the plane* $(a, \ln c)$, *showing the details for small values of* c. *The flow of* (12) *transforms a level curve of* $\widetilde{\Lambda}(a,c)$ *into another level curve. Notice that the level set curve* $\widetilde{\Lambda}(a,c) = 0$ *is the orbit* Γ_∞.

Proposition 6 shows that the discontinuity of $\widetilde{\Lambda}$ at the point $m(-1)$ of Γ_∞ is equal to λ_+, see [6] Theorem 2.5. The discontinuity of $\widetilde{\Lambda}$ at any point $m(s)$ of Γ_∞ can be obtained using the similarity property (11). To see this, consider $f = f_{a,b,c}$ the solution of the BIVP

$$f''' + ff'' = 0, \quad f(0) = -a,\ f'(0) = b,\ f''(0) = c > 0$$

and set

$$\Lambda(a,b,c) = \lim_{t \to +\infty} f'_{a,b,c}(t).$$

This limit is finite and non negative ([6] Proposition 3.1). The function $\widetilde{\Lambda}$ is thus given by

$$\widetilde{\Lambda}(a,c) = \Lambda(a,-1,c).$$

Then ([6] Proposition 2.1), $[-1, \Lambda(a,b,c)[$ is the maximal right interval of existence of the solution $u(s)$ of the CIVP

$$uu'' + s = 0, \qquad u(b) = c > 0,\ u'(b) = a.$$

The similarity property (11) implies

$$\forall \sigma > 0, \qquad \Lambda(\sigma a, \sigma^2 b, \sigma^3 c) = \sigma^2 \Lambda(a,b,c). \tag{25}$$

This formula justifies, as said in the introduction, that the properties of Λ for $b < 0$ can be deduced from the case $b = -1$, i.e. from the properties of the function $\widetilde{\Lambda} : (a,c) \mapsto \Lambda(a,-1,c)$.

Notice that, for any positive solution u of the Crocco equation defined on some interval I, we obviously have

$$\forall s \in I, \quad \widetilde{\Lambda}\big(u'(-1), u(-1)\big) = \Lambda\big(u'(-1), -1, u(-1)\big) = \Lambda\big(u'(s), s, u(s)\big).$$

As a consequence, (25) gives

$$\forall s < 0, \qquad \widetilde{\Lambda}\big(u'(-1), u(-1)\big) = -s\widetilde{\Lambda}\big((-s)^{-1/2}u'(s), (-s)^{-3/2}u(s)\big).$$

In terms of the associated vector field, we deduce that for all (x, y) solution of (12)

$$\forall \tau \in \mathbb{R}, \qquad \widetilde{\Lambda}\big(x(\tau), y(\tau)\big) = e^{\tau}\widetilde{\Lambda}\big(x(0), y(0)\big). \tag{26}$$

This formula shows how the τ-map flow of (12) transforms the level curve $\widetilde{\Lambda}(a, c) = \lambda$ into the level curve $\widetilde{\Lambda}(a, c) = e^{\tau}\lambda$, see Figure 9. Hence the similarity property (11) yields the discontinuity at any point of Γ_∞.

Corollary 7 — *The discontinuity of $\widetilde{\Lambda}$ at a point $m(s)$ of Γ_∞ is as follows: on the convex side of Γ_∞, $\widetilde{\Lambda}$ tends to 0, whereas on the concave side, $\widetilde{\Lambda}$ tends to $-\frac{\lambda_+}{s}$.*

8 The number of solutions of the BBVP

To count the number of solutions of (5) we adopt the following strategy: for any values of $a \in \mathbb{R}$ and $\lambda \geqslant 0$, we count the number of values of c for which $\widetilde{\Lambda}(a, c) = \lambda$ where $\widetilde{\Lambda} : \mathbb{R}\times]0, +\infty[\to [0, +\infty[$ is the limit defined by (9). Let $A(\lambda)$ denote the abscissa of the point of Γ_∞ where $\widetilde{\Lambda}$ takes the values 0 and λ on each side of Γ_∞ respectively. Corollary 7 yields $-s = \frac{\lambda_+}{\lambda}$, from which we deduce that

$$A :]0, +\infty[\to]-\infty, 0[\,, \ \lambda \mapsto \sqrt{\tfrac{\lambda}{\lambda_+}}\, u'_-\left(-\tfrac{\lambda_+}{\lambda}\right).$$

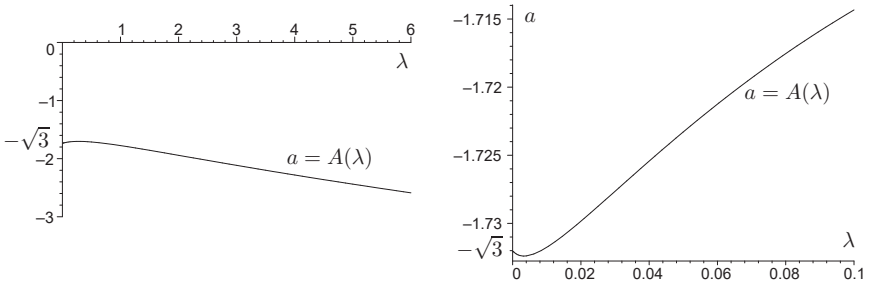


Figure 10: *On the left: the graph of A. On the right: an enlargement near $(0, -\sqrt{3})$.*

See Figure 10 for a numerical graph of A and Figure 12 for a sketch showing the oscillations near $\lambda = 0$. A careful study of the vector field shows that Γ_∞ has no inflexion point and S^* is a focus. Therefore for all $n \geqslant 1$, with the convention $a_0 = -\infty$, there exist functions

$$L_n : [a_{2n}, a_{2n-1}] \to \mathbb{R}, \qquad R_n : [a_{2n-2}, a_{2n-1}] \to \mathbb{R},$$

L_n convex and R_n concave, such that Γ_∞ is the union of the graphs of the mappings $x \mapsto L_n(x)$ and $x \mapsto R_n(x)$; see Figure 3 right for the graphs of R_1, R_2 and L_1. As a consequence, the function A has the following properties.

Proposition 8 ([6] Proposition 1.1) — *The function A is C^∞ and has an infinite sequence of extremal points $(\lambda_n)_{n\geqslant 1}$ decreasing to 0: local minima at λ_{2n} and local maxima at λ_{2n+1}, and no other extremum. Let $A(\lambda_n) = a_n$ denote these extremal values. Sequences (a_{2n}) and (a_{2n+1}) are adjacent and*

$$\lim_{n\to+\infty} \frac{\lambda_{n+1}}{\lambda_n} = e^{-\pi\sqrt{2}}, \qquad \lim_{n\to+\infty} \frac{a_{n+1} + \sqrt{3}}{a_n + \sqrt{3}} = -e^{-\pi\sqrt{2}}. \tag{27}$$

The map $\lambda \mapsto A(\lambda)$ is increasing on each interval $[\lambda_{2n}, \lambda_{2n-1}]$ and decreasing on each $[\lambda_{2n-1}, \lambda_{2n-2}]$.

Hence for all $n \geqslant 1$, with the convention $\lambda_0 = +\infty$, there exist one-to-one mappings

$$l_n : [a_{2n}, a_{2n-1}] \to [\lambda_{2n}, \lambda_{2n-1}], \quad r_n : [a_{2n-2}, a_{2n-1}] \to [\lambda_{2n-1}, \lambda_{2n-2}],$$

such that the graph of $\lambda \mapsto A(\lambda)$ is the union of the graphs of $a \mapsto l_n(a)$ and $a \mapsto r_n(a)$, see Figure 12 left for the graphs of r_1, r_2 and l_1.

Given $a \in \mathbb{R}$ and $\lambda \geqslant 0$, counting the number of solutions of the Blasius Problem (1 - 5) amounts to counting the number of times the function $\widetilde{\Lambda}$ takes the value λ on a vertical ray

$$D_a := \{a\} \times]0, +\infty[. \tag{28}$$

For that purpose, we introduce the function

$$\widetilde{\Lambda}_a :]0, +\infty[\to [0, +\infty[, \quad c \mapsto \widetilde{\Lambda}(a, c).$$

The description below is succinct. We refer to [6] Section 2.4 for proofs, additional details and explanatory figures.

Let $n \geqslant 1$ be such that a is between a_{n-2} and a_n, possibly $a = a_n$ (with the convention $a_{-1} = +\infty,\ a_0 = -\infty$). Then the ray D_a crosses $n-1$ times the spiral Γ_∞ (if $a = a_n$, there is an n-th point of contact but without crossing, hence without creating any discontinuity for $\widetilde{\Lambda}_a$). To fix ideas, assume that n is odd. A similar description can be done for n even. Then the

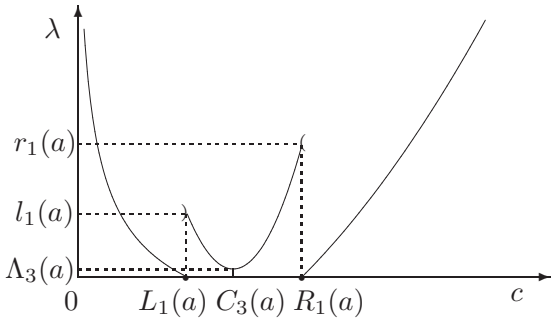


Figure 11: *A sketch of graph of $\widetilde{\Lambda}_a$ in the case $a_3 < a < a_1$, a close to a_3.*

graph of $\widetilde{\Lambda}_a$ consists of n branches: $\frac{n-1}{2}$ on the left, one central and $\frac{n-1}{2}$ on the right. On the central part, by continuity, if a is close to a_n then $\widetilde{\Lambda}_a$ has a minimum close to 0. Therefore we consider $d_n \in]a_n, a_{n-2}]$ as close to a_{n-2} as possible such that, for any $a \in [a_n, d_n[$, the central branch of $\widetilde{\Lambda}_a$ attains its infimum, at some (possibly non unique) abscissa $c = C_n(a)$. We already know that $d_1 = +\infty$. For $n \geqslant 2$, let $\mu_n \in]\lambda_{n-1}, \lambda_{n-2}]$ be such that $d_n = A(\mu_n)$. For $a \in [a_n, d_n[$, we define $\Lambda_n(a)$ as the minimum of $\widetilde{\Lambda}_a$ on its central branch. With the convention $\mu_1 = +\infty$, this yields a continuous map $\Lambda_n : [a_n, d_n[\to [0, \mu_n[$, satisfying $\Lambda_n(a_n) = 0$ and $\Lambda_n(a) \to \mu_n$ as $a \to d_n^-$. See Figure 11 for a sketch of graph of $\widetilde{\Lambda}_a$, Figure 13 for some graphs of $\widetilde{\Lambda}_a$ when $a \approx a_1$ and Figure 14 for the graph of $\widetilde{\Lambda}_{-\sqrt{3}}$. We now present our main result.

Theorem 9 — *The BBVP* (5) *has*

- *no solution if and only if $a > a_1$ and $0 \leqslant \lambda < \Lambda_1(a)$,*
- *at least n solutions (where $n > 0$) if (λ, a) belongs to one of the regions marked n in Figure 12, right, in other words, if:*

- *either* $a = A(\lambda)$ *with* $\mu_{n+1} \leqslant \lambda < \mu_n$,
- *or* $\lambda = \Lambda_n(a)$ *with* $a \in [a_n, d_n[$ *if* n *is odd,* $a \in]d_n, a_n]$ *if* n *is even, including the end-point* $(0, a_n)$,
- *or* (λ, a) *is in the open region below the graphs of* Λ_2 *and* A *in the case* $n = 1$, *and in the open region between the graphs of* $\Lambda_{n-1}, \Lambda_{n+1}$ *and* A *in the case* $n \geqslant 2$,
- *or* $\lambda = 0$ *and* $a_{n-1} < a < a_{n+1}$ *if* n *is odd,* $a_{n+1} < a < a_{n-1}$ *if* n *is even.*

- *infinitely many solutions if* $\lambda = 0$ *and* $a = -\sqrt{3}$.

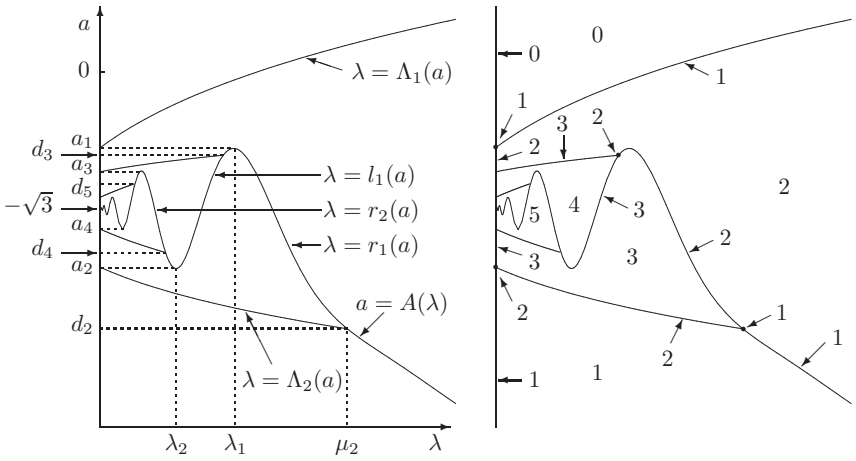


Figure 12: *In the* (λ, a) *plane. On the left, a sketch of the graphs of the functions* A *and* Λ_n*; on the right, a lower bound of the number of solutions of* (5). *We conjecture that this number is exact. We stress that the distances are not respected: due to* (27) *with* $e^{\pi\sqrt{2}} \approx 85$, *on the true graph of* A *no more than one extremal point is visible, see Figure 10.*

Remark — We conjecture that each branch of $\widetilde{\Lambda}_a$ is monotonous, except possibly the central one, which can be either monotonous or first decreasing then increasing, depending on the place of a with respect to the a_k and d_l. A consequence of this conjecture would be that this lower bound is tight.

9 Numerical simulations

When (a, c) is close to Γ_∞ and on its concave side, $u(s; a, c)$ is exponentially small for the small negative value of s at which u reaches its minimum on $[-1, 0]$. This phenomenon can lead to bad numerical simulations. As we will see, the passage to the variables $s(v), w(v)$, corresponding to the variables $S(V), W(V)$ described in Section 5, is appropriate, not only for theoretical but also for numerical reasons. As an illustration, let us solve numerically, with the use of Maple, the CIVP (in the phase plane, i.e. with $v = u'$) with the initial conditions

$$\big\{(u(-1) = c_1 = 2.94,\ v(-1) = a = -3.12\big\}$$

and

$$\big\{u(-1) = c_2 = 2.95,\ v(-1) = a\big\}.$$

The first initial condition lies on the concave side of Γ_∞ and the second one on its concave side. For the convenience of the reader we give hereafter the Maple instructions and the resulting

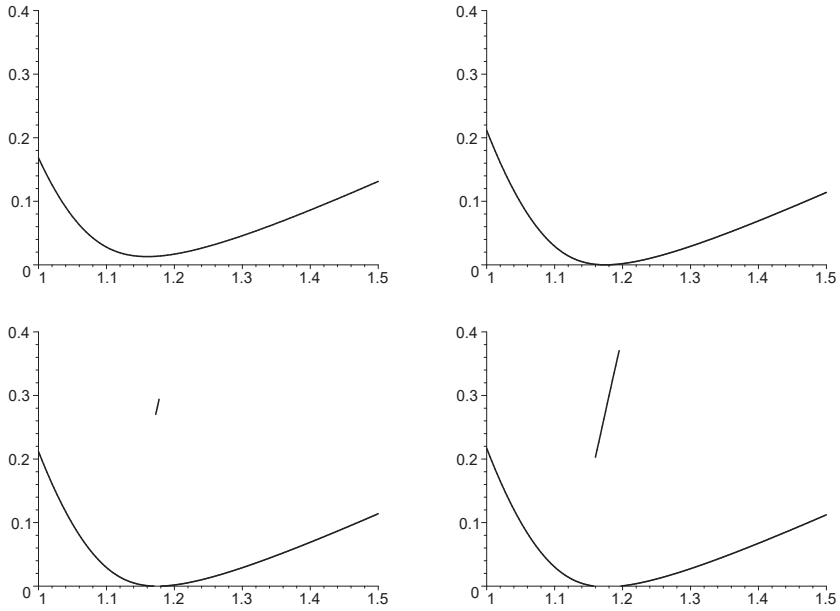


Figure 13: *Scenario of bifurcation of the graphs of* $\widetilde{\Lambda}_a$ *near* $a_1 \approx -1.702704$. *Top left:* $a = -1.68$, *top right:* $a = -1.7027$, *bottom left:* $a = -1.7028$, *bottom right:* $a = -1.705$.

output. Because the aim is to compute bounds of existence intervals of solutions, the output is an error message, with a numerical value of a possible singularity.

```
> restart:
> a:=-3.12: c1:=2.94: c2:=2.95:
> EqCroccoUV:=diff(u(s),s)=v(s),diff(v(s),s)=-s/u(s):
```

$$\frac{du}{ds} = v, \qquad \frac{dv}{ds} = -\frac{s}{u} \tag{29}$$

```
> SolCroccoUV:=proc(c)
>    Sol:=dsolve({EqCroccoUV,u(-1)=c,v(-1)=a},{u(s),v(s)},
>                numeric,output=listprocedure):
>    SolU:=eval(u(s),Sol):
>    SolU(50):
> end proc:
> SolCroccoUV(c1);
Error, (in SolU) cannot evaluate the solution further right of
-0.21850903e-2, probably a singularity
> SolCroccoUV(c2);
Error, (in SolU) cannot evaluate the solution further right of
0.99652635e-3, probably a singularity
```

The result for c_1 is not correct since the solution must be defined for all $s \leqslant 0$. The second result c_2 is correct and predicts that $\widetilde{\Lambda}(a, c_2) \approx 0.001$. It is a fact, not completely elucidated, that the use of the logarithmic change of variable $w = \ln u$ does not yield better numerical results: in the variables (w, v) the numerical solutions are still incorrect for c_1 (and correct for c_2).

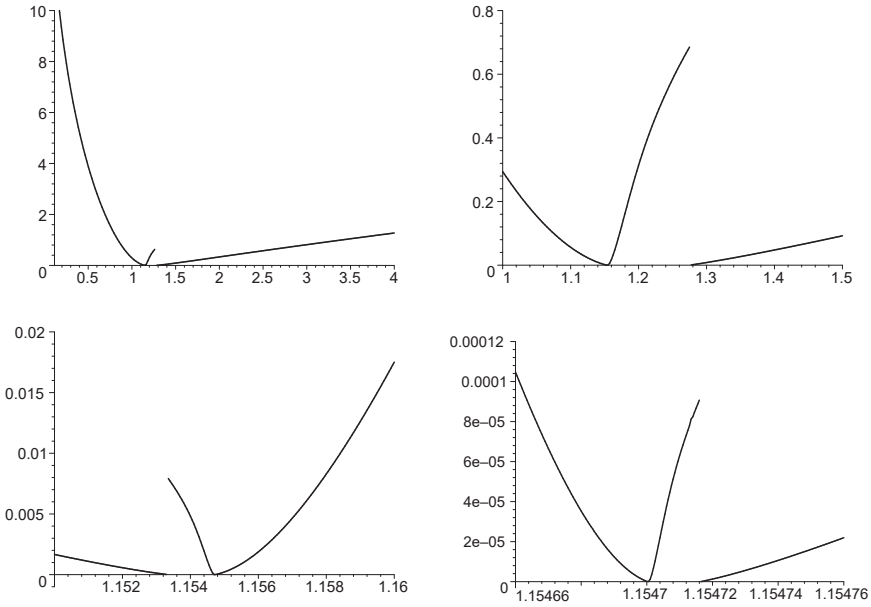


Figure 14: *Numerical graph of* $\widetilde{\Lambda}_a$ *for* $a = -\sqrt{3}$*, with successive enlargements.*

```
> EqCroccoVW:=diff(w(s),s)=v(s)*exp(-w(s)),
>             diff(v(s),s)=-s*exp(-w(s));
```

$$\frac{dw}{ds} = ve^{-w}, \qquad \frac{dv}{ds} = -se^{-w} \tag{30}$$

```
> SolCroccoVW:=proc(c)
>    Sol:=dsolve({EqCroccoVW,w(-1)=ln(c),v(-1)=a},{w(s),v(s)},
>                numeric,output=listprocedure):
>    SolW:=eval(w(s),Sol):
>    SolW(50):
> end proc:
> SolCroccoVW(c1);
Error, (in SolW) cannot evaluate the solution further right of
-0.21862961e-2, probably a singularity
> SolCroccoVW(c2);
Error, (in SolW) cannot evaluate the solution further right of
0.99531941e-3, probably a singularity
```

Following Section 5, we now consider the change of variable $w = \ln u$ and use the variable v as an independent variable. This gives correct numerical results.

```
> EqCroccoSW:=diff(s(v),v)=-exp(w(v))/s(v),diff(w(v),v)=-v/s(v);
```

$$\frac{ds}{dv} = -\frac{e^w}{s}, \qquad \frac{dw}{dv} = -\frac{v}{s} \tag{31}$$

```
> SolCroccoSW:=proc(c)
>    Sol:=dsolve({EqCroccoSW,w(a)=ln(c),s(a)=-1},{w(v),s(v)},
>                numeric,output=listprocedure):
>    SolS:=eval(s(v),Sol):
>    SolS(10):
> end proc:
> SolCroccoSW(c1);
Error, (in SolS) cannot evaluate the solution further right of
2.7651840, probably a singularity
> SolCroccoSW(c2);
Error, (in SolS) cannot evaluate the solution further right of
-2.7726621, probably a singularity
```

We see that for c_1, the solution $w(v)$ is computed until the value $v = 2.7651840$. Hence the numerical solution succeeded to pass exponentially close to the axis $u = 0$ and to reflect on this axis and get a positive derivative. The singularity encountered now at $v = 2.7651840$ is caused by the fact that system (31) is defined only in the half space $s < 0$. For $s \geqslant 0$ we must return to the original variables $u(s)$ and $v(s)$. Hence we use the following procedure to evaluate $\widetilde{\Lambda}(a, c)$.

```
> Lambda:= proc(a,c)
>   erreur:=0.0000000001:
>   fSW:=dsolve({EqCroccoSW,w(a)=ln(c),s(a)=-1},{s(v),w(v)},
>               numeric,stop_cond=[s(v)+erreur]):
>   fSW(10);
>   IC:=subs(%,[v,s(v),w(v)]):
>   v0:=IC[1]: s0:=IC[2]: w0:=IC[3]:
>   f:=dsolve({EqCroccoUV,u(s0)=exp(w0),v(s0)=v0},{u(s),v(s)},
>             numeric,stop_cond=[u(s)-erreur]):
>   f(50):
> subs(%,s):
> end proc:
```

This procedure is easy to understand. First system (31) is solved with initial conditions $w(a) = \ln(c)$, $s(a) = -1$ as far as $s \leqslant -\texttt{erreur}$. Next, one computes the values v_0, $s_0 = s(v_0)$ and $w_0 = w(v_0)$ such that the stopping condition $s_0 = -\texttt{erreur}$ is reached. Then system (29) is solved with initial conditions $u(s_0) = e^{w_0}$, $v(s_0) = v_0$, as far as $u \geqslant \texttt{erreur}$. The procedure gives the value s_1 such that the stopping condition $u(s_1) = \texttt{erreur}$ is attained. This value is a very good approximation of $\widetilde{\Lambda}(a, c)$.

```
> Lambda(a,c1);
Warning, cannot evaluate the solution further right of
2.7651840, stop condition #1 violated
Warning, cannot evaluate the solution further right of
10.012058, stop condition #1 violated
                      10.0120586420604791
> Lambda(a,c2);
Warning, cannot evaluate the solution further right of
-2.7726621, stop condition #1 violated
Warning, cannot evaluate the solution further right of
.99606109e-3, stop condition #1 violated
```

```
0.000996061092810659674
```

Hence, $\widetilde{\Lambda}(a, c_1) \approx 10.012$ and $\widetilde{\Lambda}(a, c_2) \approx 0.001$. The following instructions which use the procedure `Lambda` produce the numerical graph of $\widetilde{\Lambda}_a$ for $a = -\sqrt{3}$ and $1.15466 \leqslant c \leqslant 1.15476$, see Figure 14 bottom right.

```
> a:=-sqrt(3): c1:=1.15466: c2:=1.15476: N:=200;
> for n from 0 to N do LLambda[n]:=c1+n*(c2-c1)/N,
>                       Lambda(a,c1+n*(c2-c1)/N) end do:
> L1:=[[LLambda[i]]$i=0..112]: L2:=[[LLambda[i]]$i=113..N]:
> plot([L1,L2],c1..c2,-.0000031..0.00012,thickness =4,
>      color=black);
```

References

[1] Z. Belhachmi, B. Brighi K. Taous, On the concave solutions of the Blasius equation, *Acta Math. Univ. Comenianae* 69:2 (2000) 199-214.

[2] E. Benoît Ed., *Dynamic Bifurcations*, Proceedings Luminy 1990, Lect. Notes Math. 1493 Springer-Verlag, 1991.

[3] H. Blasius, Thesis, Göttingen (1907).

[4] H. Blasius, Grenzschichten in Flüssigkeiten mit kleiner Reibung, *Zeitschr. Math. Phys.* 56 (1908) 1-37.

[5] B. Brighi, Deux problèmes aux limites pour l'équation de Blasius, *Revue Math. Ens. Sup.* 8 (2001) 833-842.

[6] B. Brighi, A. Fruchard T. Sari, On the Blasius Problem, *Adv. Differential Equations* 13, no. 5-6 (2008) 509–600.

[7] L. Crocco, Sull strato limite laminare nei gas lungo una lamina plana, *Rend. Math. Appl. Ser. 5* 21 (1941) 138-152.

[8] P. Hartman, *Ordinary Differential Equations*, Wiley, 1964.

[9] M. Y. Hussaini W. D. Laikin, Existence and non-uniqueness of similarity solutions of a boundary layer problem, *Quart. J. Mech. Appl. Math.* 39:1 (1986) 15-24.

[10] M. Y. Hussaini, W. D. Laikin A. Nachman, On similarity solutions of a boundary layer problem with an upstream moving wall, *SIAM J. Appl. Math.* 47:4 (1987) 699-709.

[11] C. Lobry, T. Sari S. Touhami, On Tykhonov's theorem for convergence of solutions of slow and fast systems, *Electron. J. Differential Equations* 19 (1998) 1-22.

[12] O. A. Oleinik, V. N. Samokhin, *Mathematical Models in Boundary Layer Theory*, Applied Mathematics and Mathematical Computations 15, Chapman& Hall/CRC, Washington, 1999.

[13] E. Soewono, K. Vajravelu R.N. Mohapatra, Existence and nonuniqueness of solutions of a singular non linear boundary layer problem, *J. Math. Anal. Appl.* 159 (1991) 251-270.

[14] A. N. Tikhonov, Systems of differential equations containing small parameters multiplying the derivatives, *Mat. Sb.* 31 (1952) 575-586.

[15] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Krieger, New York, 1976.

[16] H. Weyl, On the differential equations of the simplest boundary-layer problems, *Ann. Math.* 43 (1942) 381-407.

Address of the authors:

Laboratoire de Mathématiques, Informatique et Applications, EA3993
Faculté des Sciences et Techniques, Université de Haute Alsace
4, rue des Frères Lumière, F-68093 Mulhouse cedex, France

Current address for T. Sari:

UMR ITAP, CEMAGREF, Domaine de Lavalette
361, rue J.-F. Breton, BP 5095
F-34196 Montpellier Cedex 5, France

E-addresses:

Bernard.Brighi@uha.fr
Augustin.Fruchard@uha.fr
Tewfik.Sari@uha.fr

De nouveaux développements asymptotiques combinés pour la perturbation singulière

Augustin Fruchard, Reinhard Schäfke

Résumé. On présente une théorie de développements asymptotiques pour des fonctions de deux variables, combinant à la fois des fonctions d'une des variables et des fonctions du quotient de ces deux variables. Ces développements asymptotiques combinés (DAC) sont bien adaptés à la description des solutions d'équations différentielles ordinaires singulièrement perturbées au voisinage de points tournants. Le lien et les différences avec les méthodes de matching et les développements combinés classiques sont décrits. Cette théorie est appliquée à des problèmes de solutions canards.

Mots-clés : . point tournant, DAC, série Gevrey, canard, perturbation singulière, équation différentielle complexe.

Classification MSC : 34E.

1 Introduction.

Cet article reprend de larges extraits du mémoire [14], auquel nous renvoyons le lecteur pour des preuves complètes et des définitions et résultats complémentaires. La principale différence avec [14] est que le point de vue mis en avant ici est celui de l'analyse réelle, plus familier à bon nombre de lecteurs, alors que dans [14] les résultats sont présentés dans le cadre complexe.

Nous nous intéressons ici aux équations différentielles singulièrement perturbées de la forme

$$\varepsilon y' = f(x, y, \varepsilon) \tag{1.1}$$

où x et y sont des variables réelles et $\varepsilon > 0$ un petit paramètre. Nous nous intéressons en particulier au comportement des solutions lorsque ε tend vers 0 (ou lorsque ε est i-petit dans le contexte de l'analyse non standard). En un point (x, y) du plan où $f(x, y, 0) \neq 0$ la solution est quasi verticale, vers le haut ou vers le bas suivant le signe de $f(x, y, 0)$. La situation est plus compliquée au voisinage de la *variété lente* V, qui est l'ensemble d'équation $f(x, y, 0) = 0$. L'attractivité de V se mesure à l'aide de la dérivée de f par rapport à y : un point $(x, y) \in V$ est *attractif* (resp. *répulsif*, resp. *tournant*) si $\frac{\partial f}{\partial y}(x, y, 0)$ est négatif (resp. positif, resp. nul). Si $(x^*, y^*) \in V$ n'est pas tournant (un tel point est dit *régulier*) alors le théorème des fonctions implicites implique que localement V est le graphe d'une fonction continue y_0, dite *lente*, vérifiant $y_0(x^*) = y^*$. Lorsqu'on cherche à prolonger une telle fonction y_0, deux situations peuvent se produire. Ou bien on arrive au bord du domaine de définition de y_0 (le cas "générique"), ou bien on arrive à un point tournant, où y_0 est encore définie mais où la courbe lente n'a plus d'attractivité ni de répulsivité.

Le comportement des solutions est bien compris au voisinage des points réguliers. Par exemple, si la fonction f est de classe $\mathcal{C}^\infty$, il est facile de montrer que l'équation (1.1) a une unique solution formelle $\widehat{y} = \sum_{n\geq 0} y_n(x)\varepsilon^n$ en puissances de ε ; il est connu aussi qu'au voisinage d'un point régulier il existe des solutions ayant cette solution formelle comme développement asymptotique. Par ailleurs, la théorie des développements asymptotiques combinés classiques [26, 2]

permet de décrire la *couche limite* (appelée aussi la couche intérieure) des solutions commençant à longer une courbe lente $y = y_0(x)$ en un point (x^*, y^*) régulier. Par exemple si ce point est attractif, on peut donner une approximation d'une solution $y = y(x, \varepsilon)$, uniforme sur un intervalle $[x^*, x^* + \delta]$, sous la forme

$$\sum_{n\geq 0} \Big(y_n(x) + z_n\big(\tfrac{x-x^*}{\varepsilon}\big)\Big)\varepsilon^n,$$

à l'aide d'une part de la solution formelle $\widehat{y}$ (la partie dite *lente* du développement combiné) et d'autre part de fonctions z_n à décroissance exponentielle à l'infini (la partie *rapide*).

En un point tournant, la méthode la plus répandue pour obtenir une approximation des solutions est le recollement de deux développements dits *intérieur* et *extérieur*. C'est ce qu'on appelle le *matching* dans la littérature anglo-saxonne. Plus qu'une méthode, le matching est une idée très générale, qui regroupe des méthodes diverses dans de nombreuses situations où apparaissent à des équations fonctionnelles (différentielles ordinaires, aux différences, aux dérivées partielles, etc.)

Le but de cet article est de présenter de nouveaux développements asymptotiques combinés (pour faire court, nous écrirons DAC), particulièrement adaptés aux points tournants des équations singulièrement perturbées de la forme (1.1). L'avantage de notre approche est de donner une approximation uniforme des solutions dans un intervalle qui contient à la fois des points loin du point tournant et des points dans un petit voisinage de ce point tournant. En dépit du caractère naturel, presque familier, de ces développements, nous n'avons trouvé presque aucune trace de ces développements dans la littérature existante. En particulier leur similitude apparente avec les DAC classiques cache des différences profondes. Les seuls travaux déjà existants ayant une relation avec nos DAC sont ceux de Lindsay A. Skinner [23, 24, 25], de L. E. Fraenkel [11], de Wiktor Eckhaus [9] et surtout de Thomas Forget [10].

Contenu de l'article.

Dans une première partie, nous présentons des exemples, les plus simples possibles, qui montrent que les solutions d'équations singulièrement perturbées ont naturellement des DAC près des points tournants.

La définition générale des DAC et leurs premières propriétés sont présentées dans la partie 3.1. La partie 3.3 établit une comparaison de ces DAC avec le matching. Dans la partie 3.4, nous présentons la notion de DAC Gevrey, qui est une notion incontournable pour une étude approfondie des équations singulièrement perturbées. Cette étude est présentée dans la partie 4. Enfin, en application, nous mettons en œuvre notre théorie pour la résolution d'un problème de canard en un point tournant dégénéré (partie 5.1) et de canard non lisse (partie 5.2). Ce dernier problème avait déjà été étudié par Emmanuel Isambert et Marc Diener il y a une quinzaine d'années et suscite toujours de l'intérêt de nos jours.

Dans cet article, dans un soucis de clarté et de simplicité, les résultats sont présentés dans le cadre réel. Cependant nous insistons sur le fait qu'une part importante de la théorie ne peut pas se passer du cadre complexe. Les preuves des résultats énoncés ici sont détaillées dans le mémoire [14]. Pour ces preuves, il est indispensable de considérer x et ε comme des variables du champ complexe. C'est la raison pour laquelle nous avons été amenés à ne pas utiliser l'analyse non standard, contrairement à beaucoup de travaux sur la perturbation singulière. Nous mentionnerons le plan de la variable complexe de manière sporadique dans certains commentaires, par exemple dans l'idée de preuve du théorème 4.2 qui est notre résultat principal, mais une méconnaissance de l'analyse complexe ne nuit pas du tout à la lecture de l'article.

2 Trois exemples introductifs.

Nous présentons ici des exemples, les plus simples possibles, qui montrent que les solutions d'équations singulièrement perturbées ont naturellement des DAC près des points tournants. Les trois exemples sont des équations linéaires. Le premier exemple est parmi les plus simples présentant un point tournant. Le deuxième contient un paramètre de contrôle, permettant de « chasser le canard ». Le troisième exemple contient aussi un paramètre de contrôle mais le point tournant n'est plus simple, ce qui a pour conséquence que les canards ne sont pas des solutions surstables au sens de Guy Wallet [27].

2.1 Premier exemple.

Commençons par l'équation

$$\varepsilon \frac{dy}{dx} = 2xy + \varepsilon g(x), \tag{2.1}$$

où $\varepsilon > 0$ est le petit paramètre et x, y sont des variables réelles. On suppose que la fonction $g : \mathbb{R} \to \mathbb{R}$ est de classe $\mathcal{C}^\infty$ et bornée, ainsi que toutes ses dérivées. Ces hypothèses sont là pour simplifier la présentation, mais beaucoup des résultats qui suivent sont valables avec des hypothèses plus faibles. Par exemple l'existence de la solution y^- ci-dessous n'utilise que la continuité de g et pour montrer que cette solution a localement un développement asymptotique à l'ordre de N, il suffit que g soit de classe $\mathcal{C}^N$.

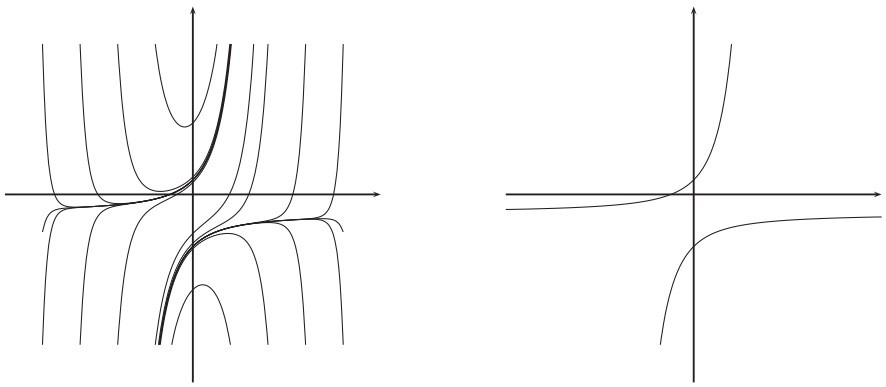


FIG. 1 – À gauche, quelques solutions de (2.1) avec $\varepsilon = 1$, $|x| \leq 4$, $|y| \leq 4$ et $g(x) = x + 1$. À droite, les solutions y^- et y^+.

Avec des modifications mineures, on peut aussi traiter le cas d'un intervalle borné ou de fonctions non bornées.

Puisque l'équation (2.1) est linéaire, ses solutions sont définies sur tout $\mathbb{R}$. La courbe lente mentionnée dans l'introduction est ici $y = 0$; elle est attractive pour $x < 0$ et répulsive pour $x > 0$. L'équation (2.1) présente un point tournant simple en $x = 0$. Pour chaque $\varepsilon > 0$ fixé, on peut voir qu'il existe une unique fonction $y^-(\,\cdot\,, \varepsilon)$ bornée sur $\mathbb{R}^-$ qui est solution de (2.1) pour la valeur ε du petit paramètre. Cette solution est donnée par la formule de variation de la constante

$$y^-(x, \varepsilon) = e^{x^2/\varepsilon} \int_{-\infty}^{x} e^{-t^2/\varepsilon} g(t) dt. \tag{2.2}$$

Puisque dans cet article ε est une variable, ce que nous appelons "solution" est en fait une famille de solutions dépendant de ε. La formule (2.2) définit une famille de solutions de (2.1) qui sont non seulement bornées sur $\mathbb{R}^-$ prises isolément, mais de plus bornées uniformément par rapport

à ε (ou encore, c'est une fonction des deux variables x et ε qui est bornée sur $\mathbb{R}^- \times]0, \varepsilon_0]$ pour $\varepsilon_0 > 0$ fixé). Dans toute la suite, nous utiliserons parfois l'expression "bornée" pour "bornée uniformément par rapport à ε". Dans le contexte de l'analyse non standard, cela correspondrait à une fonction de la seule variable x (puisque ε serait alors une constante i-petite) limitée sur $\mathbb{R}^-$.

Nous voulons avoir une description quantitative de cette solution, non seulement pour $x < 0$ fixé mais aussi pour x proche de 0 et pour $x > 0$. Nous commençons avec le cas $x < 0$.

Par une succession d'intégrations par parties, on montre aisément que, pour tout $\delta > 0$ fixé, la solution y^- admet aussi un développement asymptotique au sens de Poincaré uniforme sur $]-\infty, -\delta[$, de la forme $\widehat{y} = \sum_{n\geq 0} y_n(x)\varepsilon^n$: pour tout entier $N > 0$, il existe une constante C_N telle que pour tout $x \in]-\infty, -\delta[$ et tout $\varepsilon \in]0, \varepsilon_0]$

$$\left| y^-(x,\varepsilon) - \sum_{n=0}^{N-1} y_n(x)\varepsilon^n \right| \leq C_N \varepsilon^N.$$

De fait, ce développement est l'unique solution formelle de (2.1) ; elle est donnée par les premiers termes

$$y_0(x) = 0, \quad y_1(x) = -\frac{1}{2x}\, g(x), \tag{2.3}$$

puis récursivement par

$$y_{n+1}(x) = \frac{1}{2x}\, y_n'(x), \quad n \geq 1. \tag{2.4}$$

Pour voir que $\widehat{y}$ est bien un développement asymptotique de y^-, on peut par exemple écrire $y = y^{(N)} + z\varepsilon^N$ avec $y^{(N)} = \sum_{n=0}^{N-1} y_n \varepsilon^n$ et vérifier que z satisfait une équation du même genre que (2.1), donc est bornée sur la demi-droite $]-\infty, -\delta[$.

Si on remplace $-\infty$ par $+\infty$, la même formule (2.2) fournit aussi une unique solution y^+ bornée sur $\mathbb{R}^+$, qui admet un développement asymptotique sur $]\delta, +\infty[$ pour tout $\delta > 0$. Puisque ce développement est l'unique solution formelle de (2.1), c'est le même que celui de y^-.

• Dans le cas très particulier où g est impaire, alors d'une part on a $y^- = y^+$ et d'autre part les formules (2.3) et (2.4) impliquent que pour tout $n \in \mathbb{N}$ la fonction y_n est paire et sans pôle en $x = 0$; ainsi la solution formelle $\widehat{y}$ reste définie en $x = 0$. Il est donc naturel de se demander si le développement commun $\sum_{n\geq 0} y_n(x)\varepsilon^n$, valide pour x loin de 0, reste valide près de 0. Dans cet exemple c'est le cas et on peut le montrer en utilisant $y^{(N)}$ comme précédemment.

Pour des équations analogues, par exemple en changeant $2x$ par $4x^3$ dans (2.1), *c.f.* 2.2.2, on a toujours $y^- = y^+$ et donc une solution bornée sur tout $\mathbb{R}$, mais les coefficients de la solution formelle admettent en général des pôles en $x = 0$ et les sommes partielles $y^{(N)}$ de $\widehat{y}$ ne peuvent plus être des approximations uniformes de y. L'équation (2.1) est l'un des exemples les plus simples où la théorie de la surstabilité peut s'appliquer. Nous ne poursuivons pas plus loin la discussion dans cette direction car nous voulons présenter les DAC et non la surstabilité.

• Lorsque g n'est pas impaire, le développement de y^+ permet toutefois d'avoir aussi une approximation de y^- sur $]\delta, +\infty[$. En effet, on a $y^-(x,\varepsilon) = y^+(x,\varepsilon) + I(\varepsilon)e^{x^2/\varepsilon}$ avec $I(\varepsilon) = \displaystyle\int_{-\infty}^{+\infty} e^{-t^2/\varepsilon} g(t)dt$. Si g n'est pas impaire, alors la fonction I est non nulle. Pour voir ceci, on peut écrire $I(\varepsilon) = \int_0^{+\infty} e^{-s/\varepsilon}\big(g(\sqrt{s}) + g(-\sqrt{s})\big)\frac{1}{2\sqrt{s}} ds$ et utiliser l'injectivité de la transformation de Laplace. Plus concrètement, si la partie paire de g — donnée par $g^+(x) = \frac{1}{2}\big(g(x)+g(-x)\big)$ — n'est pas plate, alors elle satisfait $g^+(x) \sim Cx^{2N}$ avec $C \neq 0$ et $N \in \mathbb{N}$, et obtient $I(\varepsilon) \sim C'\varepsilon^{N+1/2}$. Par conséquent, en un point fixé $x > 0$, $y^-(x,\varepsilon)$ prend une valeur exponentiellement grande

par rapport à ε : il existe $c, a, \varepsilon_0 > 0$ (qui dépendent de x) tels que pour tout $\varepsilon \in]0, \varepsilon_0[$, $|y^-(x,\varepsilon)| \geq c \exp\left(\frac{a}{\varepsilon}\right)$.

Il est possible de décrire précisément en fonction de N le domaine où y^- reste bornée et le domaine où y^- tend vers l'infini, mais nous ne faisons pas une étude exhaustive ici. La formule (2.2) montre que y^- est bornée, et même de l'ordre de $\eta = \sqrt{\varepsilon}$, pour $x < 0$. Cette formule (2.2) suggère aussi le changement de variable $x = \eta X$ (avec $\eta = \sqrt{\varepsilon}$). On obtient pour tout K réel

$$y^-(\eta X, \varepsilon) = \eta \int_{-\infty}^{X} e^{X^2 - T^2} g(\eta T) dT = \mathcal{O}(\eta) \tag{2.5}$$

quand η tend vers 0, uniformément pour $X \leq K$.

Nous allons voir que y^- admet un développement en puissances de η, mettant en jeu à la fois des fonctions de la variable lente x et de la variable rapide $X = \frac{x}{\eta}$. Ceci peut se voir par une succession d'intégrations par parties. En effet, notons Sg la fonction définie par

$$g(x) = g(0) + xSg(x). \tag{2.6}$$

Puisque g est $\mathcal{C}^\infty$ et bornée sur $\mathbb{R}$, Sg l'est aussi (et même tend vers 0 à l'infini). Une première intégration par parties donne

$$\begin{aligned} y^-(x,\varepsilon) &= e^{x^2/\varepsilon}\left(g(0) \int_{-\infty}^{x} e^{-t^2/\varepsilon} dt + \int_{-\infty}^{x} t e^{-t^2/\varepsilon} Sg(t) dt \right) \\ &= g(0)\eta\, U^-\left(\tfrac{x}{\eta}\right) - \tfrac{\varepsilon}{2}\, Sg(x) + \tfrac{\varepsilon}{2}\, e^{x^2/\varepsilon} \int_{-\infty}^{x} e^{-t^2/\varepsilon} (Sg)'(t) dt. \end{aligned}$$

avec $U^-(X) = e^{X^2} \int_{-\infty}^{X} e^{-T^2}\, dT$. En appliquant (2.5) à $(Sg)'$ au lieu de g, on a ainsi pour tout K réel

$$y^-(x,\varepsilon) = g(0)\eta U^-\left(\tfrac{x}{\eta}\right) - \tfrac{\varepsilon}{2} Sg(x) + \mathcal{O}\left(\eta^3\right)$$

quand $\eta \to 0$ uniformément sur l'ensemble des x avec $\frac{x}{\eta} \leq K$. En itérant l'intégration par parties, on obtient, avec l'opérateur S donné par (2.6) et l'opérateur $D = \frac{d}{dx}$

$$\begin{aligned} y^-(x,\varepsilon) \;=\; & \sum_{n=0}^{N-1} \left(\left(\tfrac{1}{2} DS\right)^n g \right)(0) \eta^{2n+1} U^-\left(\tfrac{x}{\eta}\right) - \\ & \frac{1}{2} \sum_{n=0}^{N-1} S\left(\left(\tfrac{1}{2} DS\right)^n g \right)(x) \eta^{2n+2} + \mathcal{O}\left(\eta^{2N+1}\right) \end{aligned} \tag{2.7}$$

quand $\eta \to 0$ uniformément sur l'ensemble des x avec $\frac{x}{\eta} \leq K$. Il s'agit d'un exemple de développement combiné, de la forme $\sum_{n\geq 0} \left(a_n(x) + g_n^-\left(\frac{x}{\eta}\right) \right) \eta^n$, avec ici

$$a_0 = 0,\ a_{2n} = -\tfrac{1}{2}\, S\left(\left(\tfrac{1}{2} DS\right)^{n-1} g \right),\ a_{2n+1} = 0 \tag{2.8}$$

et

$$g_n^- = c_n U^- \text{ avec } c_{2n} = 0,\ c_{2n+1} = \left(\left(\tfrac{1}{2} DS\right)^n g \right)(0).$$

De plus, la fonction U^- admet un développement asymptotique à l'infini, donné par

$$\begin{aligned} U^-(X) \;&\sim\; \sum_{n\geq 0} (-1)^{n+1}\, 1.3...(2n-1) 2^{-n-1} X^{-2n-1} \\ &=\; -\frac{1}{2X} + \frac{1}{4X^3} - \frac{3}{8X^5} + ..., \quad X \to -\infty. \end{aligned} \tag{2.9}$$

On a aussi une formule analogue pour la solution y^+ bornée sur $\mathbb{R}^+$:

$$\begin{aligned} y^+(x,\varepsilon) &= \sum_{n=0}^{N-1} \Big(\big(\tfrac{1}{2}DS\big)^n g \Big)(0)\eta^{2n+1} U^+\big(\tfrac{x}{\eta}\big) - \\ &\quad \frac{1}{2}\sum_{n=0}^{N-1} S\Big(\big(\tfrac{1}{2}DS\big)^n g \Big)(x)\eta^{2n+2} + \mathcal{O}\big(\eta^{2N+1}\big) \end{aligned}$$

avec $U^+(X) = e^{X^2}\displaystyle\int_{+\infty}^{X} e^{-T^2}\,dT = -U^-(-X)$. Par ailleurs, dans le cas où g est impaire, $(DS)^n g$ est impaire pour tout $n \in \mathbb{N}$, donc la première partie du développement (2.7) est identiquement nulle. On retrouve ainsi le fait que y^- a un développement asymptotique classique en puissances de $\eta^2 = \varepsilon$, avec des coefficients de la variable x uniquement.

Nous sommes bien dans le cadre de la définition 3.1. À l'aide de normes de Nagumo [4], on montre facilement que le développement (2.7) est Gevrey d'ordre $1/2$ en η. Observons d'ailleurs que le développement (2.9) est aussi un développement Gevrey d'ordre $1/2$.

Notons au passage que les termes a_n et g_n^- sont nuls pour la moitié d'entre eux ; il serait donc possible de réécrire (2.7) sous forme de séries en puissances de ε, mais cela est très particulier à cet exemple. Dans les applications, les termes de la partie lente d'un DAC sont effectivement souvent nuls sauf pour les puissances qui sont des multiples de p, mais les termes de la partie rapide n'ont pas de raison *a priori* de s'annuler.

2.2 Extensions.

2.2.1. Avant de présenter le deuxième exemple, nous voudrions explorer des généralisations et extensions du premier exemple. La première généralisation concerne la nature purement locale du résultat. Tout d'abord, toute autre solution de (2.1), de condition initiale $y(x_0,\varepsilon)$ bornée (pour $\varepsilon \in]0,\varepsilon_0]$) en un point fixé $x_0 < 0$, admet un DAC sur $[x_0+\delta, 0]$ pour tout $\delta \in]0, |x_0|[$; de plus ce DAC est le même que celui de y^- puisque les deux solutions sont exponentiellement proches l'une de l'autre sur $[x_0+\delta, 0]$. Pour les mêmes raisons, le résultat reste valide avec une hypothèse seulement locale sur g : si g est de classe $\mathcal{C}^\infty$ sur un intervalle $]-r, r[$, alors pour tout $x_0 \in]-r, 0[$, tout $\delta \in]0, |x_0|[$ et toute fonction $c = c(\varepsilon)$ bornée sur $]0, \varepsilon_0]$, la solution de (2.1) de condition initiale $y(x_0,\varepsilon) = c(\varepsilon)$ admet un DAC de la forme (2.7) sur $[x_0+\delta, 0]$; il en est de même pour les solutions à droite, *i.e.* avec $x_0 \in]0, r[$. Ces DAC ne dépendent pas des conditions initiales ; ils sont *a priori* différents à gauche et à droite mais ils ont la même partie lente $\sum_{n\geq 0} a_n(x)\eta^n$ avec $a_n(x)$ données par (2.8).

2.2.2. La deuxième généralisation est de remplacer le terme $2x$ dans (2.1) par $p\,x^{p-1}$, où p est un entier pair. Nous avons toujours une unique solution y^- bornée sur $\mathbb{R}^-$ et une unique solution y^+ bornée sur $\mathbb{R}^+$. Elles sont données à présent par $y^{\pm}(x,\varepsilon) = e^{x^p/\varepsilon}\displaystyle\int_{\pm\infty}^{x} e^{-t^p/\varepsilon} g(t)dt$. La condition $y^- = y^+$ est toujours équivalente à g impaire. La recherche d'une solution formelle $\widehat{y} = \sum_{n\geq 0} y_n(x)\varepsilon^n$ aboutit à

$$y_0(x) = 0, \quad y_1(x) = -\frac{1}{p\,x^{p-1}}\,g(x), \quad \text{puis} \quad y_{n+1}(x) = \frac{1}{p\,x^{p-1}}\,y_n'(x).$$

En général cette solution formelle n'est pas définie en $x = 0$, même lorsque g est impaire. Dans le cas où g est impaire, le développement de y^- est valide aussi bien pour les x positifs que

pour les x négatifs, mais ne peut pas être valide au voisinage de 0. Cependant la même méthode d'intégrations par parties successives permet de montrer (que g soit impaire on non) que y^- possède un DAC, mêlant à la fois des fonctions de x et des fonctions de la variable rapide $X = \frac{x}{\eta}$, avec $\eta = \varepsilon^{1/p}$. Les calculs sont plus longs et compliqués sans être plus difficiles ; ils font apparaître les $p-1$ « fonctions spéciales » suivantes

$$U_k^-(X) = e^{X^p} \int_{-\infty}^{X} e^{-T^p} T^{k-1} dT, \quad k = 1, ..., p-1.$$

On obtient finalement pour y^- un DAC de la forme $\sum_{n\geq 0} \Big(a_n(x) + g_n\big(\frac{x}{\eta}\big)\Big)\eta^n$ avec $\eta = \varepsilon^{1/p}$ et g_n de la forme $g_n = c_{n1}U_1^- + \cdots + c_{n,p-1}U_{p-1}^-$. De même que précédemment pour U^-, ces fonctions U_{kp}^- ont aussi un développement asymptotique lorsque X tend vers $-\infty$.

2.2.3. Une troisième extension concerne les équations où la fonction g dépend de ε. Si g a un développement asymptotique en puissances de ε, ainsi que toutes ses dérivées par rapport à x, alors on peut montrer que les fonctions $y^\pm$ ont encore des DAC avec des fonctions $g_n^\pm$ proportionnelles à $U^\pm$; le facteur est le même pour les deux signes. Précisément ces DAC sont donnés comme avant par

$$\begin{aligned} y^\pm(x,\varepsilon) \;=\; & \sum_{n=0}^{N-1} A_{N-n,\,n}(0,\varepsilon)\eta^{2n+1} U^\pm\big(\tfrac{x}{\eta}\big) - \\ & \frac{1}{2}\sum_{n=0}^{N-1} B_{N-n,\,n}(x,\varepsilon)\eta^{2n+2} + \mathcal{O}\big(\eta^{2N+1}\big) \end{aligned} \tag{2.10}$$

où $A_{mn} : (x,\varepsilon) \mapsto \sum_{k=0}^{m} A_{mnk}(x)\varepsilon^k$ est le jet d'ordre m par rapport à ε de la fonction $\big(\frac{1}{2}DS\big)^n g$ et B_{mn} est le jet d'ordre m par rapport à ε de la fonction $S\left(\big(\frac{1}{2}DS\big)^n g\right)$.

La seule modification à apporter est la condition nécessaire et suffisante pour avoir $y^- = y^+$. A la place de g impaire, cette condition devient

$$\int_{-\infty}^{\infty} e^{-t^2/\varepsilon}\, g(t,\varepsilon)\, dt = 0.$$

Les résultats sont similaires avec p au lieu de 2. Lorsque $y^- = y^+$, on obtient à nouveau que la solution bornée sur $\mathbb{R}$ a un développement classique dans le cas $p = 2$ car les facteurs de $U^\pm$ dans (2.10) s'annulent. Par contre lorsque $p \geq 4$, ce n'est plus forcément le cas.

2.2.4. Notre dernière extension est d'avoir $f(x)$ à la place de $p\,x^{p-1}$, où f est une fonction de classe $\mathcal{C}^\infty$ vérifiant $xf(x) > 0$ si $x \neq 0$ et $f(x) \sim ax^{p-1}$, $x \to 0$ avec $a \neq 0$. La solution y^- s'écrit alors $y^-(x,\varepsilon) = e^{F(x)/\varepsilon} \displaystyle\int_{-\infty}^{x} e^{-F(t)/\varepsilon} g(t)dt$ avec $F(x) = \int_0^x f(t)dt$, si on ajoute l'hypothèse que $\displaystyle\int_{-\infty}^{0} e^{-F(t)/\varepsilon}\, dt$ converge pour $\varepsilon > 0$ assez petit.

Un difféomorphisme $x = \varphi(\xi)$ permet de se ramener à une équation de la forme

$$\varepsilon \frac{dz}{d\xi} = p\,\xi^{p-1} z + \varepsilon h(\xi),$$

ce qui permet d'obtenir pour y^- un DAC de la forme

$$\sum_{n\geq 0}\Big(a_n(x)+g_n\big(\tfrac{\varphi^{-1}(x)}{\eta}\big)\Big)\eta^n.$$

Notre théorie générale des DAC permet de montrer qu'un tel développement peut aussi se transformer en un DAC de la variable $X=\frac{x}{\eta}$, *i.e.* de la forme $\sum_{n\geq 0}\Big(b_n(x)+h_n\big(\frac{x}{\eta}\big)\Big)\eta^n$, *c.f.* proposition 3.2 (b). Remarquons que, pour cette extension, on a besoin d'autres fonctions "rapides" que U^- et U^+.

2.3 Deuxième exemple.

Considérons à présent une équation déjà considérée sous une forme voisine par Claude Lobry dans le chapitre introductif [18]. Il s'agit de l'équation

$$\varepsilon\frac{dy}{dx}=2xy+\varepsilon g(x)+\varepsilon\alpha, \tag{2.11}$$

où $\varepsilon>0$ est le petit paramètre, $\alpha\in\mathbb{R}$ un paramètre de contrôle, et où la fonction $g:\mathbb{R}\to\mathbb{R}$ est de classe $\mathcal{C}^\infty$, et bornée ainsi que toutes ses dérivées. La question est la suivante.

Existe-t-il des valeurs de α pour lesquelles il existe une solution bornée[1] *sur tout $\mathbb{R}$?*

La réponse est « oui ». Pour le voir, on procède ainsi : étant donné α arbitraire, il existe toujours l'unique solution bornée sur $\mathbb{R}^-$, notée y^- et donnée par

$$y^-(x,\varepsilon)=e^{x^2/\varepsilon}\int_{-\infty}^{x}e^{-t^2/\varepsilon}\big(\alpha+g(t)\big)dt. \tag{2.12}$$

En remplaçant $-\infty$ par $+\infty$, la même formule fournit aussi une unique solution y^+ bornée sur $\mathbb{R}^+$. On en déduit qu'il existe une solution y bornée sur tout $\mathbb{R}$ si et seulement si $y^+=y^-$, ce qui donne une équation pour le paramètre α, dont la solution est

$$\alpha(\varepsilon)=-\Big(\int_{-\infty}^{+\infty}e^{-t^2/\varepsilon}g(t)dt\Big)\Big/\Big(\int_{-\infty}^{+\infty}e^{-t^2/\varepsilon}dt\Big). \tag{2.13}$$

Comme nous avons supposé g de classe $\mathcal{C}^\infty$, on en déduit que α admet un développement asymptotique quand ε tend vers 0. Pour l'étude des solutions $y^\pm$ correspondant à cette valeur de α, on est dans la situation de la deuxième extension (avec pour fonction g la fonction $(x,\varepsilon)\mapsto g(x)+\alpha(\varepsilon)$) et on a $y^-=y^+$ par le choix de α. La solution y admet donc aussi un développement asymptotique, dont les coefficients sont des fonctions $\mathcal{C}^\infty$, y compris en $x=0$.

On calcule directement les premiers termes

$$y_0(x)=0,\quad \alpha_0=-g(0),\quad y_1(x)=-\frac{1}{2x}\big(g(x)+\alpha_0\big). \tag{2.14}$$

Pour la suite, il suffit de remarquer que les α_n et les $y_n(x)$ sont déterminées uniquement par le fait que y_n n'a pas de pôle en $x=0$. On les calcule donc récursivement par

$$y_{n+1}(x)=\frac{1}{2x}\big(y_n'(x)-\alpha_n\big),\quad \alpha_n=y_n'(0). \tag{2.15}$$

[1] Rappelons que " bornée " signifie uniformément par rapport à ε dans un intervalle $]0,\varepsilon_0]$. Dans le présent contexte, il se trouve que, pour tout ε fixé, il existe une unique valeur $\alpha=\alpha(\varepsilon)$ pour laquelle l'équation (2.11) a une solution $y=y(x,\varepsilon)$ bornée sur $\mathbb{R}$ au sens classique, et que la fonction y ainsi définie est aussi bornée sur $\mathbb{R}\times\,]0,\varepsilon_0]$.

Dans le champ complexe, si g est analytique et à croissance au plus exponentielle dans une bande horizontale S centrée sur $\mathbb{R}$, on montre qu'il existe deux solutions y^- et y^+ sur des domaines dont la réunion contient la bande privée d'un petit disque $D(0,\delta)$. Dans le cas où α est donné par (2.13), ces deux solutions coïncident et ont un développement asymptotique en puissances de ε pour $x \in S \setminus D(0,\delta)$. Le principe du maximum permet alors de montrer que ce développement est valide aussi lorsque $|x| < \delta$, *c.f.* [14]. En résumé, du fait que le point tournant $x = 0$ est simple, le relief ne présente que deux montagnes, si bien qu'une solution définie et bornée sur des parties de ces deux montagnes est automatiquement bornée sur tout un voisinage du point tournant. On dit d'une telle solution qu'elle est *surstable*, suivant la terminologie adoptée par Guy Wallet [27].

2.4 Troisième exemple.

Remplaçons le terme $2x$ par $4x^3$; autrement dit, considérons l'équation

$$\varepsilon \frac{dy}{dx} = 4x^3 y + \varepsilon g(x) + \varepsilon\alpha, \tag{2.16}$$

Pour tout $a = a(\varepsilon)$, il existe encore une unique solution y^+ bornée sur $\mathbb{R}^+$ et une unique solution y^- bornée sur $\mathbb{R}^-$. Il existe aussi une unique valeur de α pour laquelle ces deux solutions coïncident pour former une solution y bornée sur $\mathbb{R}$. Cette solution est donnée par

$$\begin{aligned} y(x,\varepsilon) &= e^{x^4/\varepsilon} \int_{-\infty}^{x} e^{-t^4/\varepsilon}\big(\alpha + g(t)\big)dt \\ \text{avec} \quad \alpha &= \alpha(\varepsilon) = -\frac{\int_{-\infty}^{+\infty} e^{-t^4/\varepsilon} g(t)dt}{\int_{-\infty}^{+\infty} e^{-t^4/\varepsilon}dt}. \end{aligned} \tag{2.17}$$

À nouveau on peut montrer que $\alpha(\varepsilon)$ admet un développement asymptotique de la forme

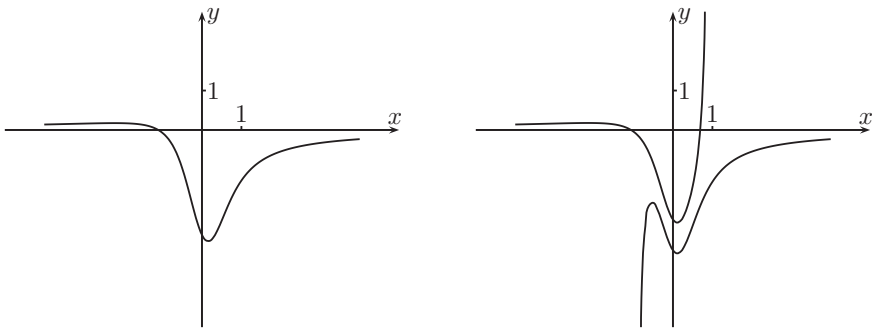


FIG. 2 – Les solutions y^+ et y^- pour $g(x) = 3x^2 + 3x$. À gauche pour $\alpha = -0.507$, à droite pour $\alpha = -0.35$. Ici $|x| \le 5$, $-5 \le y \le 3$ et $\varepsilon = \frac{1}{4}$.

$\alpha(\varepsilon) \sim \sum_{n=0}^{\infty} \alpha_n \varepsilon^{n/2}$. En revanche, la solution y n'a en général pas de développement asymptotique en puissances de $\varepsilon^{1/2}$ dans un voisinage réel de 0. En effet, un tel développement serait une solution formelle $\widehat{y} = \sum_{n\ge 0} y_n(x)\varepsilon^{n/2}$, $\widehat{\alpha} = \sum_{n\ge 0} \alpha_n \varepsilon^{n/2}$ devant vérifier $y_0(x) = 0$, $y_2(x) = -\frac{1}{4x^3}\big(g(x) + \alpha_0\big)$ et $y_{2n+2}(x) = \frac{1}{4x^3}\big(y'_{2n}(x) - \alpha_{2n}\big)$. Dès qu'un coefficient y_{2n} a une dérivée dont le développement de Taylor contient un terme non nul en x ou x^2, le coefficient suivant présente un pôle en $x = 0$, quelque soit le choix de α_{2n}.

Nous allons voir qu'il est cependant possible de donner une approximation de y valide dans tout un voisinage réel de 0, à l'aide de fonctions de $\frac{x}{\eta}$, où $\eta = \varepsilon^{1/4}$. Cette idée était à la base de la thèse de Thomas Forget [10].

En effet, isolons les premiers termes du développement de g en écrivant $g(x) = g_0 + g_1 x + g_2 x^2 + 4x^3 h(x)$. La première formule de (2.17) devient

$$\begin{aligned} y(x,\varepsilon) &= (\alpha(\varepsilon) + g_0) U_0^-\left(\tfrac{x}{\eta}\right) + \eta g_1 U_1^-\left(\tfrac{x}{\eta}\right) + \\ &\qquad \eta^2 g_2 U_2^-\left(\tfrac{x}{\eta}\right) + e^{x^4/\varepsilon} \int_{-\infty}^{x} e^{-t^4/\varepsilon} 4t^3 h(t) dt \end{aligned} \tag{2.18}$$

où on a posé

$$U_j^-(X) = -e^{X^4} \int_{-\infty}^{X} e^{-T^4} T^j dT.$$

Une intégration par parties donne

$$e^{x^4/\varepsilon} \int_{-\infty}^{x} e^{-t^4/\varepsilon} 4t^3 h(t) dt = \varepsilon h(x) - \varepsilon e^{x^4/\varepsilon} \int_{-\infty}^{x} e^{-t^4/\varepsilon} h'(t) dt$$

qui est de la même forme que dans la formule (2.18) avec un facteur ε en plus, ce qui permet de réitérer. De la même manière que dans 2.2.2, on obtient à la fin un DAC. Le fait que y^+ soit égal à y^- entraîne qu'il s'agit d'un DAC sur *tout l'axe réel*. Précisément, il existe des fonctions $a_n \in \mathcal{C}^\infty(\mathbb{R})$ et des combinaisons linéaires g_n des fonctions U_0^-, U_1^-, U_2^- telles que

$$y^\pm(x,\varepsilon) = \sum_{n=0}^{N-1} \left(a_n(x) + g_n\left(\tfrac{x}{\eta}\right) \right) + \mathcal{O}(\eta^N)$$

uniformément sur $\mathbb{R}$.

Dans le champ complexe, le relief associé à (2.16), qui est donné par la partie réelle de x^4, comprend quatre montagnes. La solution étudiée est proche de la courbe lente sur deux montagnes, à l'est et à l'ouest, mais *a priori* pas sur les deux autres montagnes nord et sud ; elle n'est donc pas surstable.

3 Les développements combinés.

Si l'on veut généraliser la méthode de la partie précédente à des équations non linéaires, cela nécessite d'élargir la famille de fonctions dans lesquelles s'écrivent les solutions. En particulier, il est nécessaire de prendre en compte les produits de fonctions $U_j^\pm U_k^\pm$, les solutions d'équations différentielles $\varepsilon y' = p x^{p-1} y + U^\pm\left(\frac{x}{\eta}\right)$, ainsi que des produits de fonctions de x et de fonctions de $\frac{x}{\eta}$. Une stratégie est de construire une algèbre contenant les fonctions $x \mapsto x^n$ et $(x,\varepsilon) \mapsto U_j^\pm\left(\frac{x}{\eta}\right)$ et stable par les opérateurs $\mathcal{J}^\pm$ suivants. Pour chaque signe + et −, $\mathcal{J}^\pm$ associe, à une fonction v à croissance polynomiale (*i.e.* vérifiant $\exists N \in \mathbb{N}\ \exists C > 0\ \forall x \in \mathbb{R},\ |v(x)| \leq C\,|x|^N$), l'unique solution à croissance polynomiale sur $\mathbb{R}^\pm$ de l'équation $\dfrac{dU}{dX} = 4X^3 U + v(X)$ i.e. $\mathcal{J}^\pm$ est donné par

$$\mathcal{J}^\pm v(X) = e^{X^4} \int_{\pm\infty}^{X} e^{-T^4} v(T) dT.$$

La construction de la plus petite algèbre avec ces propriétés conduit à définir un grand nombre de "fonctions spéciales" ; ceci était la stratégie adoptée par Thomas Forget dans sa thèse pour

l'approximation de solutions canard comme dans 2.4. Une complication additionnelle provient du caractère non unique de l'écriture. Par exemple, le terme x peut être considéré aussi bien comme un terme fonction de x d'ordre 0 en η qu'un terme fonction de $\frac{x}{\eta}$ d'ordre 1 en η, s'il est écrit $\frac{x}{\eta}\eta$.

La stratégie que nous avons adoptée est de considérer d'emblée une algèbre plus grosse. Un avantage de cette stratégie est la simplicité ; un inconvénient est de donner moins de renseignements sur les coefficients. Pour simplifier la présentation, dans les parties 3 et 4.1 nous ne considérons des DAC que pour des fonctions définies sur $\mathbb{R}^+$. A partir de la section 4.2, nous aurons à nouveau besoin de regarder des DAC pour des fonctions définies sur $\mathbb{R}^-$.

3.1 Notations et définition.

On se donne $\eta_0, r > 0$, $\mu \geq 0$ et on note $\mathcal{H}$ l'espace vectoriel des fonctions a de classe $\mathcal{C}^\infty$ sur $[0, r]$ et $\mathcal{G}$ l'espace vectoriel des fonctions g de classe $\mathcal{C}^\infty$ et bornées sur $]\mu, +\infty[$ et ayant un développement asymptotique au sens de Poincaré à l'infini sans terme constant $g(X) \sim \sum_{\nu \geq 1} g_\nu X^{-\nu}$, $X \to +\infty$, *i.e.*

$$\forall N \in \mathbb{N} \quad \exists C_N > 0 \quad \forall X > \mu, \qquad \left| g(X) - \sum_{1 \leq \nu \leq N-1} g_\nu X^{-\nu} \right| \leq C_N X^{-N}.$$

Pour définir le développement combiné d'une fonction de deux variables x et η, le plus simple et le plus naturel serait de considérer des fonctions définies sur le produit cartésien $[0, r] \times]0, \eta_0]$. Cependant, pour les applications, il sera commode que l'intervalle par rapport à x évite un voisinage de 0 de taille proportionnelle à η. Pour définir l'intervalle en x, nous avons donc ajouté le paramètre $\mu \geq 0$. Dans la partie 3.4 nous considérerons aussi le cas où μ est négatif, mais pour l'instant nous supposons μ positif.

Définition 3.1 . Soit y une fonction définie et de classe $\mathcal{C}^\infty$ pour $\eta \in]0, \eta_0]$ et $x \in]\mu\eta, r]$. Nous disons que y *admet un* DAC s'il existe des fonctions $a_n \in \mathcal{H}$ et $g_n \in \mathcal{G}$, telles que

$$y(x, \eta) \sim \sum_{n \geq 0} \Big(a_n(x) + g_n\big(\tfrac{x}{\eta}\big) \Big) \eta^n, \quad \eta \to 0$$

autrement dit si, pour tout entier N, il existe une constante $K_N > 0$ telle que, pour tout $\eta \in]0, \eta_0]$ et tout $x \in]\mu\eta, r]$

$$\left| y(x, \eta) - \sum_{n=0}^{N-1} \Big(a_n(x) + g_n\big(\tfrac{x}{\eta}\big) \Big) \eta^n \right| \leq K_N \eta^N. \tag{3.1}$$

Les fonctions a_n forment la *partie lente du* DAC, et les g_n la *partie rapide*.

REMARQUES . 1. Le fait que les fonctions g_n tendent vers 0 quand $X \to +\infty$, implique qu'une fonction $y = y(x, \eta)$ ne peut pas avoir deux DAC différents. En effet, on a $\lim_{\eta \to 0} y(x, \eta) = a_0(x)$ pour $x > 0$, donc $a_0(0)$ est lui aussi déterminé de manière unique ; puis il vient $\lim_{\eta \to 0} y(\eta X, \eta) = a_0(0) + g_0(X)$, d'où g_0 et ainsi de suite. A priori, pour l'unicité du développement, il suffirait donc de demander aux fonctions g_n de tendre vers 0 à l'infini. Cependant, si l'on veut la stabilité pour la multiplication, nous allons voir qu'il est nécessaire que les fonctions de $\frac{x}{\eta}$ aient un développement asymptotique complet à l'infini.

2. Il est aussi utile d'avoir des DAC définis avec un paramètre $\mu < 0$. Ces DAC seront donc des

approximations sur des intervalles contenant 0 à l'intérieur. Malheureusement, dans le cadre $\mathcal{C}^\infty$, on n'a pas unicité de tels développements ; les fonctions a_n ne sont déterminées que pour $x \geq 0$. Dans nos applications, les fonctions a_n, g_n seront réelles analytiques et les valeurs de a_n pour x positifs déterminent la fonction uniquement. Comme ces DAC avec $\mu < 0$ seront, de plus, Gevrey, on en parlera dans la partie 3.4.

3.2 Propriétés.

Multiplication.

Le produit de deux développements combinés se fait en développant le produit terme à terme. Les produits de deux termes lents, *i.e.* de la forme $a(x)b(x)$, ainsi que de deux termes rapides $g\left(\frac{x}{\eta}\right)h\left(\frac{x}{\eta}\right)$, sont les produits de fonctions usuels. Concernant les produits "mixtes" de $a \in \mathcal{H}$ et $g \in \mathcal{G}$, on commence par traiter séparément les premiers termes des développements de a et g : on écrit $a(x) = a_0 + xb(x)$ et $Xg(X) = g_1 + h(X)$ avec $a_0 = a(0)$, $b \in \mathcal{H}$ et $h \in \mathcal{G}$ (où g_1 est le premier terme du développement asymptotique de g), ce qui donne $a(x)g\left(\frac{x}{\eta}\right) = \left(a_0 + xb(x)\right) g\left(\frac{x}{\eta}\right)$ et $x\, g\left(\frac{x}{\eta}\right) = \left(g_1 + h\left(\frac{x}{\eta}\right)\right)\eta$. En combinant ces deux formules, on obtient

$$a(x)\, g\left(\tfrac{x}{\eta}\right) = a_0\, g\left(\tfrac{x}{\eta}\right) + g_1\, b(x)\eta + b(x)\, h\left(\tfrac{x}{\eta}\right)\eta. \tag{3.2}$$

En itérant, on obtient ainsi tout un développement en puissances de η pour ce produit. Pour écrire une formule explicite de ce produit, nous introduisons les opérateurs

$$\mathbf{S} : \mathcal{H} \to \mathcal{H} \ \text{ tel que } \ a(x) = a(0) + x\mathbf{S}a(x)$$

et

$$\mathbf{T} : \mathcal{G} \to \mathcal{G} \ \text{ tel que } \ g(X) = \frac{g_1}{X} + \frac{\mathbf{T}g(X)}{X}.$$

Sur les développements asymptotiques, ces opérateurs ont pour action de décaler vers la gauche et d'effacer le premier terme : si $a(x) \sim \sum_{\nu=0}^{\infty} a_\nu x^\nu$, alors $\mathbf{S}a(x) \sim \sum_{\nu=0}^{\infty} a_{\nu+1} x^\nu$ et si $g(X) \sim \sum_{\nu\geq 1} g_\nu X^{-\nu}$, alors $\mathbf{T}g(X) \sim \sum_{\nu\geq 1} g_{\nu+1} X^{-\nu}$.

La formule de multiplication d'une fonction de $\mathcal{H}$ par une fonction de $\mathcal{G}$ s'écrit alors (avec la convention $g_0 = 0$)

$$a(x)\, g\left(\tfrac{x}{\eta}\right) \sim \sum_{\nu\geq 0} \left(a_\nu\, (\mathbf{T}^\nu g)\left(\tfrac{x}{\eta}\right) + g_\nu\, (\mathbf{S}^\nu a)(x)\right)\eta^\nu. \tag{3.3}$$

Remarques. 1. Les développements combinés classiques [26, 2] entrent dans le cadre des nôtres : il s'agit du cas où les fonctions g_n sont à décroissance exponentielle. Rappelons qu'une fonction $g : J =]\mu, +\infty[\to \mathbb{R}$ est *à décroissance exponentielle* s'il existe $C, A > 0$ vérifiant

$$\forall X \in J,\ |g(X)| \leq C \exp(-AX).$$

Une fonction g à décroissance exponentielle satisfait en particulier $g(X) = \mathcal{O}(X^{-N}),\ X \to +\infty$ pour tout entier N, donc est *plate* : elle admet la série nulle pour développement asymptotique. Une fonction $y = y(x, \eta)$, définie et de classe $\mathcal{C}^\infty$ pour $\eta \in]0, \eta_0]$ et $x \in]\mu\eta, r]$, a un *développement combiné au sens classique* s'il existe des fonctions $a_0, a_1, ...$ de classe $\mathcal{C}^\infty$ sur $[0, r]$ et $g_0, g_1, ...$ de classe $\mathcal{C}^\infty$ et à décroissance exponentielle sur J vérifiant (3.1).

2. On peut vérifier que, dans le cas des développements combinés classiques, la partie lente d'un produit ne dépend que des parties lentes des facteurs, *c.f.* par exemple entre les formules (2.5) et (2.6) de [2]. En revanche, dans le cas des développements combinés du présent article, la formule (3.2) montre que le produit d'un terme lent avec un terme rapide fait aussi apparaître des termes lents, si bien que tout est imbriqué.

3. Le quotient de DAC est un peu plus délicat. La proposition 3.14 de [14] donne des conditions nécessaires et suffisantes, portant sur les développements intérieurs et extérieurs, sous lesquelles le quotient de deux fonctions ayant des DAC a un DAC (les développements intérieurs et extérieurs sont présentés dans la proposition 3.3 ci-dessous et dans la remarque 1 après cette proposition).

Composition.

Les DAC sont aussi compatibles avec la composition à gauche et à droite avec une fonction $\mathcal{C}^\infty$, comme l'exprime l'énoncé suivant.

Proposition 3.2 . (a) *Soit $P(x,z,\eta)$ une fonction de classe $\mathcal{C}^\infty$ définie pour $z \in [-R,R]$, $\eta \in]0,\eta_0]$ et $x \in]\mu\eta, r]$ telle que tous les coefficients P_n du développement asymptotique $P(x,z,\eta) \sim \sum_{n\geq 0} P_n(x,\eta)z^n$ admettent un* DAC *$P_n(x,\eta) \sim \widehat{P}_n(x,\eta)$. Soit $y(x,\eta) = \mathcal{O}(\eta)$ une fonction $\mathcal{C}^\infty$ admettant un* DAC *$\widehat{y}(x,\eta)$ quand $\eta \to 0$ et $x \in]\mu\eta, r]$ sans termes en η^0. On suppose que $y(x,\eta)$ est bornée par R. Alors la fonction $u : (x,\eta) \mapsto P(x, y(x,\eta), \eta)$ admet le* DAC

$$\widehat{\mathbf{Q}}(\widehat{y})(x,\eta) = \sum_{n\geq 0} \widehat{P}_n(x,\eta)\widehat{y}(x,\eta)^n.$$

(b) *Soit φ une fonction numérique de classe $\mathcal{C}^\infty$ au voisinage de 0 telle que $\varphi(0) = 0$ et $\varphi'(0) = 1$ et soit $z = z(u,\eta)$ une fonction ayant un* DAC $\sum_{n\geq 0} \Big(a_n(u) + g_n\big(\frac{u}{\eta}\big)\Big)\eta^n$ *quand $]0,\eta_0] \ni \eta \to 0$ et $u \in]\mu\eta, r]$, avec $a_n \in \mathcal{H}$ et $g_n \in \mathcal{G}$. On suppose que, pour tout $n,k \in \mathbb{N}$, la dérivée $g_n^{(k)}$ a un développement asymptotique quand $X \to \infty$. Alors il existe $\widetilde{\mu}, \widetilde{r}, \widetilde{\eta}_0 > 0$ tels que la fonction $y : (x,\eta) \mapsto z(\varphi(x),\eta)$ admet un* DAC *quand $]0,\widetilde{\eta}_0] \ni \eta \to 0$ et $x \in]\widetilde{\mu}\eta, \widetilde{r}]$.*

REMARQUE . Dans le (a), l'hypothèse "y bornée par R" n'est pas essentielle : elle est satisfaite dès que η_0 est assez petit.

Preuve . (a) Pour tout $N \in \mathbb{N}^*$, la somme finie $\sum_{0\leq n\leq N-1} P_n(x,\eta)y(x,\eta)^n$ admet un DAC (compatibilité avec produit et somme). Il reste à vérifier qu'il existe une constante $L = L(N)$ telle que le reste est borné par $L\eta^N$. Ceci est évident d'après les hypothèses.

(b) Il suffit de montrer que $b\big(\frac{\varphi(x)}{\eta}\big)$ admet un DAC, si $b :]\mu,\infty] \to \mathbb{R}$ est dans $\mathcal{G}$ et si toutes les dérivées de b admettent des développements asymptotiques quand $X \to +\infty$.

Il convient d'introduire les fonctions ψ et h définies par $\frac{1}{\varphi(x)} - \frac{1}{x} = h(x)$ et $\psi(x,t) = x/\big(1 + txh(x)\big)$. La fonction h se prolonge en une fonction $\mathcal{C}^\infty$ sur $]-x_1, x_1[$, notée encore h par abus, et $\psi(x,0) = x, \psi(x,1) = \varphi(x)$. Le développement de Taylor de $b\big(\frac{\varphi(x)}{\eta}\big) = b\big(\frac{\psi(x,1)}{\eta}\big)$ par rapport à t donne pour tout $N \in \mathbb{N}$

$$b\big(\tfrac{\varphi(x)}{\eta}\big) = \sum_{n=0}^{N-1} \tfrac{1}{n!}\, \tfrac{\partial^n}{\partial t^n} b\big(\tfrac{\psi(x,t)}{\eta}\big)\ |_{t=0} + \tfrac{1}{N!}\, \tfrac{\partial^N}{\partial t^N} b\big(\tfrac{\psi(x,t)}{\eta}\big)\ |_{t=\tau}$$

avec un certain $\tau \in]0,1[$. En utilisant le fait que

$$\frac{\partial}{\partial t}\Big[f\big(\tfrac{\psi(x,t)}{\eta}\big)\Big] = (\Delta f)\big(\tfrac{\psi(x,t)}{\eta}\big)\eta h(x)$$

avec l'opérateur Δ défini par $(\Delta f)(X) = -X^2 f'(X)$, on obtient

$$b\big(\tfrac{\varphi(x)}{\eta}\big) = \sum_{n=0}^{N-1} \tfrac{\eta^n}{n!}\,(\Delta^n b)\big(\tfrac{x}{\eta}\big)h(x)^n + \tfrac{\eta^N}{N!}\,(\Delta^N b)\big(\tfrac{\psi(x,\tau)}{\eta}\big)h(x)^N, \tag{3.4}$$

et on peut vérifier que le dernier terme est $\mathcal{O}(\eta^N)$. La compatibilité des DAC avec l'addition et la multiplication entraîne alors l'existence d'un DAC pour $b\big(\frac{\varphi(x)}{\eta}\big)$. □

Intégration et dérivation.

Une difficulté pour la compatibilité avec l'intégration provient du fait que les fonctions de $\mathcal{G}$ ayant un terme avec $\frac{1}{X}$ dans leur développement asymptotique à l'infini ne possèdent pas de primitive dans $\mathcal{G}$. Lorsque ces termes sont tous nuls, l'intégration ne pose pas de problème. Précisément, considèrons un DAC $y(x,\eta) \sim \sum_{n=0}^{\infty} \Big(a_n(x) + g_n\big(\frac{x}{\eta}\big)\Big)\eta^n$ vérifiant l'hypothèse :

toutes les fonctions g_n satisfont $g_n(X) = \mathcal{O}(X^{-2})$ quand $X \to \infty$.

Alors il est facile de montrer que la fonction $(x,\eta) \mapsto \displaystyle\int_r^x y(t,\eta)\,dt$ admet un DAC : on a $\displaystyle\int_r^x y(t,\eta)\,dt \sim \widehat{Y}(x,\eta) - \widehat{Y}(r,\eta)$ avec

$$\widehat{Y}(x,\eta) = A_0(x) + \sum_{n=1}^{\infty} \Big(A_n(x) + G_{n-1}\big(\tfrac{x}{\eta}\big)\Big)\eta^n \tag{3.5}$$

où $A_n(x) = \displaystyle\int_r^x a_n(t)\,dt$ et $G_n(X) = -\int_X^{+\infty} g_n(T)\,dT$. On a $A_n \in \mathcal{H}$ et, d'après l'hypothèse, $G_n \in \mathcal{G}$. Ici on identifie $\widehat{Y}(r,\eta)$ avec la série formelle obtenue en remplaçant $G\big(\frac{r}{\eta}\big)$ par son développement asymptotique quand $\eta \to 0$.

Dans le cas général pour un DAC $y(x,\eta) \sim \displaystyle\sum_{n=0}^{\infty} \Big(a_n(x) + g_n\big(\tfrac{x}{\eta}\big)\Big)\eta^n$, on considère la série formelle $\widehat{R}(\eta) = \sum_{n=0}^{\infty} g_{n1}\eta^n$ des résidus des $g_n(X)$ et, avec une fonction arbitraire $R(\eta) \sim \widehat{R}(\eta)$, la différence $y(x,\eta) - R(\eta)\eta x(x^2+\eta^2)^{-1}$. Comme cette fonction satisfait la condition précédente, son intégrale admet un DAC $\widehat{Z}(x,\eta)$. Ainsi on a

$$\int_r^x y(t,\eta)dt \sim \widehat{Z}(x,\eta) - \widehat{Z}(r,\eta) + \frac{\eta}{2}R(\eta)\big(\ln(x^2+\eta^2) - \ln(r^2+\eta^2)\big).$$

La fonction $x(x^2+\eta^2)^{-1}$ est arbitraire et pourrait être remplacée par $x(x^2+L^2\eta^2)^{-1}$ avec tout autre réel L, ou par une autre fonction sans singularité réelle et ayant un développement asymptotique à l'infini commençant par $\frac{1}{x}$, *c.f.* [14] proposition 3.8.

Dans notre cadre réel, on ne peut pas conclure que la dérivée d'une fonction ayant un DAC admet de nouveau un DAC ; il est bien connu qu'il existe de petites fonctions avec des dérivées non bornées. Par contre, *si* la dérivée d'une fonction ayant un DAC admet elle aussi un DAC, alors la formule (3.5) implique que le DAC de la dérivée peut être obtenu en dérivant terme à terme le DAC de la fonction. Dans le cadre complexe, il est possible de montrer que la dérivée d'une fonction ayant un DAC admet un DAC dans un domaine légèrement réduit, *c.f.* [14] lemme 3.6.

On a aussi un énoncé analogue au théorème classique de Borel-Ritt. Soit $(a_n)_{n\in\mathbb{N}}$ une suite de $\mathcal{H}$ et $(g_n)_{n\in\mathbb{N}}$ une suite de $\mathcal{G}$. Alors, il existe une fonction $y(x,\eta)$ définie pour $\eta \in]0,\eta_0]$ et $x \in]\mu\eta, r]$, telle que $y(x,\eta) \sim \sum_{n=0}^{\infty} \left(a_n(x) + g_n\left(\frac{x}{\eta}\right)\right)\eta^n$. Pour la démonstration, il suffit d'utiliser le théorème de Borel-Ritt pour des développements asymptotiques uniformes classiques deux fois : une fois pour $\sum a_n(x)\eta^n$, une fois pour $\sum g_n(X)\eta^n$.

3.3 Développements combinés et "matching".

Notre notion de développement combiné mélange la notion classique de développement asymptotique au sens de Poincaré et celle de développement dit *intérieur* de la forme $y(\eta X,\eta) \sim \sum h_n(X)\eta^n$. Ces développements intérieurs occupent une place centrale dans la méthode de recollement des développements asymptotiques (méthode des "matched asymptotic expansions" en anglais). Notre approche permet de donner un fondement solide à cette méthode, en montrant que, dans le cas où il existe un développement combiné, la méthode de recollement est applicable. Précisement, on a le résultat suivant.

Proposition 3.3 . *Soit $(a_n)_{n\in\mathbb{N}}$ une famille de fonctions de $\mathcal{H}$ et $(g_n)_{n\in\mathbb{N}}$ une famille de fonctions de $\mathcal{G}$. On note leurs développements asymptotiques $a_n(x) \sim \sum_{m=0}^{\infty} a_{nm}x^m$, $x \to 0$ et $g_n(X) \sim \sum_{m>0} g_{nm}X^{-m}$, $X \to +\infty$. Supposons que*

$$y(x,\eta) \sim \sum_{n\geq 0} \left(a_n(x) + g_n\left(\tfrac{x}{\eta}\right)\right)\eta^n$$

quand η tend vers 0 et $x \in]\mu\eta, r]$ au sens de la définition 3.1.

Alors, pour $x \in]0,r]$ fixé, on a

$$y(x,\eta) \sim \sum_{n\geq 0} c_n(x)\eta^n, \quad \eta \to 0, \tag{3.6}$$

où $c_n(x) = a_n(x) + \sum_{0\leq l\leq n-1} g_{l,n-l}x^{l-n}$. De plus ce développement est uniforme par rapport à $x \in [\rho, r]$ pour tout $\rho > 0$.

De même, pour $X \in]\mu, +\infty[$ fixé, on a

$$y(\eta X,\eta) \sim \sum_{n=0}^{\infty} h_n(X)\eta^n, \quad \eta \to 0, \tag{3.7}$$

où $h_n(X) = \sum_{0\leq l\leq n} a_{n-l,l}X^l + g_n(X)$. Ce développement est uniforme par rapport à X sur toute partie compacte de $]\mu, +\infty[$.

Remarques. 1. Conformément à la littérature, nous appellerons le premier développement (3.6) le *développement extérieur* et le second (3.7) le *développement intérieur*. Chaque fonction c_n du développement extérieur peut avoir une singularité en $x = 0$ mais seulement un pôle d'ordre au plus n ; de même chaque fonction h_n du développement intérieur a une croissance polynomiale d'ordre au plus n lorsque $X \to \infty$. Le *restraint index* au sens de Wasow [28], chapitre VIII est

donc égal à 1.

2. On peut montrer que, pour tout $\kappa \in]0,1[$, le premier développement est uniforme sur $x > \eta^\kappa$, et que le deuxième est uniforme sur $X < \eta^{-\kappa}$, ce qui justifie la méthode de "matched asymptotic expansions" lorsqu'un DAC existe. Cependant, il est souvent préférable d'avoir des approximations uniformes à sa disposition sur tout le domaine d'étude. Ceci est même indispensable si on veut étudier des DAC de type Gevrey.

3. Dans les cas où on peut démontrer indirectement l'existence d'un développement combiné pour une fonction $y(x,\eta)$, mais qu'on ne connaît pas encore les fonctions a_n, g_n, la meilleure méthode pour les déterminer est d'appliquer la proposition 3.3. Pour x loin de 0, on calcule le développement extérieur $y(x,\eta) \sim \sum_{n\geq 0} c_n(x)\eta^n$, puis on rejette les termes avec des puissances négatives. On obtient ainsi les $a_n(x)$. Ensuite, on calcule le développement intérieur $y(\eta X,\eta) \sim \sum_{n\geq 0} h_n(X)\eta^n$ et on rejette les termes de puissances positives de X, ce qui donne les $g_n(X)$. Dans des situations concrètes, le calcul des développements extérieur et intérieur mène souvent à des équations de récurrence pour leurs coefficients. Ceci permet donc le calcul des a_n, g_n sans avoir à utiliser les formules techniques pour la multiplication de DAC.

Dans le cas des équations différentielles singulièrement perturbées, comme le remarquent Emmanuel Isambert et Véronique Gautheron dans [16], le calcul du développement extérieur ne fait intervenir que des opérations algébriques (une fois donnés les développement de Taylor des coefficients de l'équation) alors que celui du développement intérieur nécessite d'intégrer des équations différentielles linéaires, puis de choisir la constante d'intégration pour que la solution ait un comportement asymptotique bien déterminé, ce qui introduit de la transcendance. Dans [17], Emmanuel Isambert appelle ainsi ces développements extérieur et intérieur respectivement *algebraic* et *transcendental expansions*.

Concernant la réciproque, l'énoncé est le suivant.

Proposition 3.4 . *Soit y une fonction définie pour $\eta \in]0,\eta_0]$ et $x \in]\mu\eta, r]$. On suppose qu'il existe des nombres réels a, b, κ avec $0 < a < b$ et $0 < \kappa < 1$, et pour chaque $n \in \mathbb{N}$ une fonction c_n, $c_n(x) = P_n\left(\frac{1}{x}\right) + a_n(x)$, P_n polynomial sans terme constant, $a_n \in \mathcal{H}$ et une fonction $h_n = Q_n + g_n$, Q_n polynomial et $g_n \in \mathcal{G}$, avec les propriétés suivantes.*

Hypothèse 1. *Pour tout $N \in \mathbb{N}$, il existe une constante $C > 0$ telle que*

$$\left| y(x,\eta) - \sum_{n=0}^{N-1} c_n(x)\eta^n \right| \leq C\eta^{N(1-\kappa)} \tag{3.8}$$

pour tout $\eta \in]0,\eta_0]$ et tout $x \in]a\eta^\kappa, r]$, et

$$\left| y(\eta X,\eta) - \sum_{n=0}^{N-1} h_n(X)\eta^n \right| \leq C\eta^{N\kappa} \tag{3.9}$$

pour tout $\eta \in]0,\eta_0]$ et tout $X \in]\mu, b\eta^{\kappa-1}]$.

Hypothèse 2. *Pour tout $n \in \mathbb{N}$, les polynômes P_n et Q_n sont de degré inférieur ou égal à n.*

Alors y admet un DAC *pour $\eta \in]0,\eta_0]$ et $x \in]\mu\eta, r]$; précisément*

$$y(x,\eta) \sim \sum_{n=0}^{\infty} \left(a_n(x) + g_n\left(\tfrac{x}{\eta}\right) \right) \eta^n.$$

REMARQUES. 1. Comme il y a une région commune aux développements (3.8) et (3.9), les développements doivent être compatibles, ce que montre la preuve.

2. On ne peut pas avoir mieux que $\eta^{N(1-\kappa)}$ dans le reste de (3.8) et $\eta^{N\kappa}$ dans celui de (3.9) en général, car les premiers termes négligés ont cette taille lorsque P_N et Q_N sont de degré N.

3. Cet énoncé est un cas particulier d'un théorème général du livre [9] de W. Eckhaus. Dans le procédé classique, on établit d'abord des développements intérieurs et extérieurs sur des domaines qui croissent quand $\eta \to 0$ et qui ont une intersection non vide. Ensuite, on construit des développements dits « composites » dont nos DAC sont donc un exemple.

4. Il se trouve que la proposition 3.3 a un analogue Gevrey, mais pas la proposition 3.4.

Le mémoire [14] présente enfin des résultats de prolongement de DAC. En bref, si une fonction a un dac pour x dans un intervalle et si le développement extérieur, resp. intérieur, existe dans un intervalle plus grand, alors le DAC est valide dans le grand intervalle. Le caractère Gevrey est aussi conservé. Ceci est bien utile pour démontrer les corollaires 4.5 et 4.6 dans la suite.

3.4 Développements asymptotiques combinés : étude Gevrey.

Dans la suite de l'article (partie 4) nous appliquerons les DAC à des problèmes d'équations différentielles singulièrement perturbées. Nous verrons à cette occasion que la notion de DAC Gevrey joue un rôle clé. Cette notion Gevrey a déjà joué un rôle essentiel dans la théorie classique des séries formelles et des développements asymptotiques dans les applications à la perturbation singulière [4, 3]. Pour la théorie Gevrey, il semble indispensable de se placer dans le cadre analytique complexe. Cependant nous voulons dans cet article rester dans le cadre réel.

À partir d'ici, la partie lente de nos DAC sera constituée de fonctions analytiques réelles et la partie rapide de fonctions définies sur une demi-droite ouverte pouvant contenir 0. Précisément, on note I un intervalle ouvert contenant $[0, r]$, $\mathcal{A}$ l'ensemble des fonctions analytiques réelles et bornées sur I, et $\mathcal{G}(\mu)$ l'ensemble $\mathcal{G}$ précédent, c'est-à-dire l'ensemble des fonctions $\mathcal{C}^\infty$ bornées sur $]\mu, +\infty[$, ayant un développement asymptotique à l'infini sans terme constant, avec μ positif ou négatif.

Définition 3.5 . Soit $\eta_0 > 0$ et $\mu \in \mathbb{R}$. On dira qu'une fonction y définie et de classe $\mathcal{C}^\infty$ pour $\eta \in]0, \eta_0]$ et $x \in]\mu\eta, r]$ *admet un* DAC *Gevrey d'ordre* $\frac{1}{p}$ *et de type* (L_1, L_2) s'il existe une constante $C > 0$ et des fonctions $a_n \in \mathcal{A}$ et $g_n \in \mathcal{G}(\mu)$ vérifiant :

(i) pour tout $n \in \mathbb{N}$

$$\sup_{x\in I} |a_n(x)| \leq C L_1^n \Gamma\big(\tfrac{n}{p} + 1\big),$$

(ii) pour tout $n, M \in \mathbb{N}$ et tout $X \in]\mu, +\infty[$

$$\left| g_n(X) - \sum_{m=1}^{M-1} g_{nm} X^{-m} \right| \leq C L_1^n L_2^M \Gamma\big(\tfrac{M+n}{p} + 1\big)\, |X|^{-M}\,, \tag{3.10}$$

(iii) pour tout $N \in \mathbb{N}$, tout $\eta \in]0, \eta_0]$ et tout $x \in]\mu\eta, r]$

$$\left| y(x,\eta) - \sum_{n=0}^{N-1} \Big(a_n(x) + g_n\big(\tfrac{x}{\eta}\big) \Big)\, \eta^n \right| \leq C L_1^N \Gamma\big(\tfrac{N}{p} + 1\big) \eta^N. \tag{3.11}$$

Dans ce cas, nous écrivons $y(x,\eta) \sim_{1/p} \sum_{n\geq 0} \Big(a_n(x) + g_n\big(\tfrac{x}{\eta}\big) \Big)\, \eta^n$, $\eta \to 0$, $x \in]\mu\eta, r]$.

On déduit facilement de (3.10) que $|g_{nm}| \leq CL_1^n L_2^m \Gamma\big(\frac{m+n}{p}+1\big)$ pour tout $n, m \in \mathbb{N}$.

Les inégalités (3.10) sont également indispensables pour la compatibilité de la nouvelle notion avec les opérations élémentaires d'addition, d'intégration et de multiplication. Ceci peut se montrer directement à partir de la définition. On a aussi une compatibilité avec la composition à gauche ou à droite par une fonction analytique et avec la dérivation. Concernant la dérivation, il est nécessaire que les fonctions a_n et g_n soient analytiques complexes. Pour la composition, la compatibilité peut se montrer en utilisant un théorème de type Ramis-Sibuya [20, 21, 22] pour les DAC Gevrey, que nous présentons succintement à la fin de la partie 3.5. L'applicabilité de ce théorème aussi nécessite de se placer dans le cadre complexe. Pour les preuves et une discussion plus approfondie, voir [14].

REMARQUES . 1. Comme on considère des fonctions a_n analytiques, on a l'unicité d'un DAC Gevrey, même si par exemple la limite $\lim_{\eta\to 0} y(x,\eta)$ ne détermine que des valeurs de $a_0(x)$ pour $x > 0$. Ceci permet aussi l'utilisation de μ négatifs dans la définition 3.5.

2. Comme nous le mentionnions dans l'introduction, la théorie classique des développements combinés permet de décrire les couches limites de solutions d'équations singulièrement perturbées en un point régulier à l'aide de fonctions lentes y_n et de fonctions z_n à décroissance exponentielle. Ces développements entrent dans le cadre de notre théorie, *c.f.* [14] pour les détails. Il s'agit du cas où $p = 1$. Or dans ce cas le développement extérieur est régulier, donc coïncide avec la partie lente du DAC. D'après la proposition 3.3, cela signifie que les fonctions g_n sont plates. Puisqu'on montre que ces DAC sont de plus Gevrey, nous retrouvons le fait que les termes rapides d'un développement combiné classique sont exponentiellement décroissants. On a en fait un peu mieux : le (ii) dit un peu plus que simplement les g_n exponentiellement décroissants.

3.5 Fonctions plates Gevrey.

De même que pour les développements asymptotiques classiques, les fonctions *plates* ont un rôle important. Ici, on peut définir deux notions de platitude, suivant que l'on demande aux fonctions a_n et g_n d'être identiquement nulles, ou seulement à leurs coefficients a_{nm} et g_{nm}. Une fonction analytique étant déterminée par les coefficients de sa série de Taylor, la situation pour les g_n et pour les a_n n'est pas symétrique.

Définition 3.6 . Avec les notations de la définition 3.5, on dit que la fonction $y(x,\eta)$ est *plate au sens fort*, si toutes les fonctions a_n et g_n de la série formelle correspondante sont identiquement nulles. On dit qu'elle est *plate au sens faible*, si toutes les fonctions a_n sont nulles et tous les coefficients g_{nm} des développements asymptotiques (3.10) des fonctions g_n s'annulent.

Un des points clés de la théorie classique des développements asymptotiques Gevrey est la relation entre les fonctions plates Gevrey et les fonctions exponentiellement petites. Nous présentons ci-dessous l'analogue pour les DAC.

Proposition 3.7 . *On suppose que y admet un* DAC *Gevrey d'ordre $\frac{1}{p}$ au sens de la définition 3.5 et on utilise les notations de cette définition.*
(a) *Si y est plate au sens fort, alors il existe deux constantes $A, C > 0$ telles que*

$$|y(x,\eta)| \leq C \exp\big(-A/\eta^p\big) \text{ pour } \eta \in]0,\eta_0], x \in]\mu\eta, r]. \tag{3.12}$$

Réciproquement, si une fonction y est définie pour $\eta \in]0,\eta_0]$ et $x \in]\mu\eta, r]$ et satisfait (3.12), alors y admet un DAC *Gevrey d'ordre $\frac{1}{p}$ plat au sens fort.*

(b) *Si y est plate au sens faible, alors il existe deux constantes $B, C > 0$ telles que*

$$|y(x,\eta)| \leq C \exp\big(-B\,|x/\eta|^p\big) \text{ pour } \eta \in]0,\eta_0], x \in]\mu\eta, r]. \tag{3.13}$$

Réciproquement, si une fonction y satisfait (3.13), et si de plus y admet un DAC*, alors y est plate au sens faible.*

c.f. [14] proposition 4.8.

Un ingrédient essentiel dans la théorie complexe des DAC est un théorème du type Ramis-Sibuya ([14] théorème 5.1). Celui-ci peut s'énoncer en gros comme suit : soit (f_l^j) une famille de fonctions définies sur des secteurs (S_l^j) formant de bons recouvrements à la fois pour la variable η et pour la variable x. Si les différences $f_{l+1}^j - f_l^j$ sont plates au sens fort et les différences $f_l^{j+1} - f_l^j$ plates au sens faible, alors les fonctions f_l^j ont toutes un DAC $\sum_{n\geq 1} \Big(a_n(x) + g_n^j\big(\frac{x}{\eta}\big)\Big)\,\eta^n$.

4 DAC de solutions d'équations différentielles d'ordre 1.

4.1 Enoncé du résultat principal.

On considère l'équation

$$\varepsilon y' = f(x)y + \varepsilon P(x,y,\varepsilon) \tag{4.1}$$

où f est analytique (réelle) au voisinage d'un intervalle $[a,b]$ avec $a < 0 < b$ et vérifie $xf(x) > 0$ si $x \neq 0$, et où P est analytique au voisinage de $[a,b] \times \{0\} \times \{0\} \subset \mathbb{R}^3$. Soit $p = 1 + \mathrm{val}(f;0)$, où val désigne la valuation. On a donc $p \geq 2$ et $f(x) \sim c\,x^{p-1}$ lorsque x tend vers 0 pour un certain $c > 0$. Ici l'entier p est le même qu'avant : on pose $\eta = \varepsilon^{1/p}$.

La condition $xf(x) > 0$ si $x \neq 0$ entraîne que l'entier p est pair et que la courbe lente $y = 0$ est attractive sur $[a,0[$ et répulsive sur $]0,b]$. Dans cette situation, il est connu que, d'une part les solutions se prolongent sur la partie attractive de la courbe lente et d'autre part les solutions sont exponentiellement proches les unes des autres. On peut résumer ces deux résultats classiques dans l'énoncé suivant, qui est énoncé pour $]0,b]$ seulement ; pour $[a,0[$ il y a des résultats analogues.

Proposition 4.1 . 1. *Soit $c \in \mathbb{R}$ et $x_0 \in]0,b]$ et soit y la solution de (4.1) de condition initiale $y(x_0,\varepsilon) = c$. Alors y est définie et proche de la courbe lente pour $x \in]0,x_0[$ dans le sens suivant : pour tout $\delta > 0$ il existe $\varepsilon_0 > 0$ tel que la solution $y = y(x,\varepsilon)$ est définie pour tout $x \in [\delta, x_0]$ et tout $\varepsilon \in]0,\varepsilon_0]$ et satisfait de plus $|y(x,\varepsilon)| < \delta$ pour tout $x \in [\delta, x_0 - \delta]$ et tout $\varepsilon \in]0,\varepsilon_0]$.*

2. *Notons F la primitive de f s'annulant en 0. Soit $x_0 \in]0,b]$, soit $c_1 \neq c_2 \in \mathbb{R}$ et soit y_1, y_2 les solutions de (4.1) de conditions initiales respectivement $y_1(x_0) = c_1$ et $y_2(x_0) = c_2$. Enfin, soit $x_1 \in [a,x_0]$, positif ou négatif. Si y_1 est définie et proche de la courbe lente pour $x \in]x_1,x_0[$ (au sens précédent) et si $F(x_1) < F(x_0)$, alors y_2 est, elle aussi, définie et proche de la courbe lente pour $x \in]x_1,x_0[$. De plus on a*

$$|y_2(x,\varepsilon) - y_1(x,\varepsilon)| = \exp\big\{\tfrac{1}{\varepsilon}\big(F(x) - F(x_0) + o(1)\big)\big\}.$$

Le choix de $\eta = \varepsilon^{1/p}$ est naturel au vu de cette dernière formule : les valeurs de x de l'ordre de η sont celles pour lesquelles $F(x)/\varepsilon$ est bornée (uniformément en ε).

Le résultat présenté ci-dessous exprime le fait que, non seulement une solution venant de la droite se prolonge sur un intervalle contenant un petit voisinage de 0 de taille proportionnelle à η, mais de plus elle admet un DAC sur cet intervalle.

Comme précédemment, $\mathcal{G}(\nu)$ désigne l'ensemble des fonctions de classe $\mathcal{C}^\infty$ bornées sur $]\nu, +\infty[$, ayant un développement asymptotique à l'infini sans terme constant, avec ν positif ou négatif et $\mathcal{A}$ désigne l'ensemble des fonctions analytiques réelles sur un intervalle I, mais à présent I est un intervalle ouvert contenant $[a, b]$ où la fonction f est analytique.

Théorème 4.2 . *Soit $x_1 \in]0, b]$ et $y_1 \in \mathbb{R}$ assez petit. Soit $y = y(x, \varepsilon)$ la solution de (4.1) de condition initiale $y(x_1, \varepsilon) = y_1$.*

Alors pour tout $\mu > 0$ il existe $\varepsilon_0 > 0$ tel que y est définie au moins pour $0 < \varepsilon \leq \varepsilon_0$ et $-\mu\eta \leq x \leq x_1$, où $\eta = \varepsilon^{1/p}$.

De plus y possède un DAC *Gevrey d'ordre $\frac{1}{p}$ par rapport à η : il existe des fonctions $a_n \in \mathcal{A}$ et $g_n \in \mathcal{G}(-\mu)$, telles que, pour tout $\delta > 0$,*

$$y(x, \varepsilon) \sim_{1/p} \sum_{n \geq 1} \left(a_n(x) + g_n\left(\tfrac{x}{\eta}\right) \right) \eta^n, \quad \eta \to 0$$

uniformément pour $x \in]-\mu\eta, x_1 - \delta]$.

4.2 Commentaires.

1. Plus précisément, c'est la fonction $(x, \eta) \mapsto y(x, \eta^p)$ qui a un DAC Gevrey d'ordre $\frac{1}{p}$ uniforme pour $0 < \eta \leq \varepsilon_0^{1/p}$ et $-\mu\eta \leq x \leq x_1 - \delta$. Notons qu'une autre solution $\widetilde{y}$ de condition initiale $\widetilde{y}(x_2, \varepsilon) = y_2$ avec $x_2 \in]0, b]$ et $y_2 \in \mathbb{R}$ a le même DAC que la solution y de l'énoncé. En effet ces deux solutions sont exponentiellement proches l'un de l'autre sur tout intervalle de la forme $]0, c]$ avec $c < \min(x_1, x_2)$.

2. On a un résultat identique pour les solutions venant de la gauche. Précisément, notons $\mathcal{G}^-(\mu)$ l'espace vectoriel des fonctions $\mathcal{C}^\infty$ bornées sur $]-\infty, \mu[$, ayant un développement asymptotique à l'infini sans terme constant. Alors une solution $y = y(x, \varepsilon)$ de (4.1) de condition initiale $y(x_1, \varepsilon) = y_1$ avec $x_1 \in [a, 0[$ et y_1 assez petit admet un DAC Gevrey d'ordre $\frac{1}{p}$ par rapport à η : il existe des fonctions $a_n \in \mathcal{A}$ et $g_n^- \in \mathcal{G}^-(\mu)$, telles que, pour tout $\delta > 0$,

$$y(x, \varepsilon) \sim_{1/p} \sum_{n \geq 1} \left(a_n(x) + g_n^-\left(\tfrac{x}{\eta}\right) \right) \eta^n, \quad \eta \to 0$$

uniformément pour $x \in [x_1 + \delta, \mu\eta[$. Les fonction g_n^- sont a priori différentes des fonctions g_n des DAC à droite. Par contre, comme nous allons le voir dans le commentaire 5 ci-dessous, les fonctions a_n sont les mêmes à gauche et à droite. Pour éviter les confusions nous noterons dans la suite les termes rapides des développements à droite par g_n^+ au lieu de g_n.

3. Dans le cadre complexe, l'hypothèse "P analytique" peut être affaiblie en ce qui concerne la dépendance par rapport à ε : il suffit que P soit analytique et Gevrey d'ordre 1 pour ε dans un petit secteur $S(-\delta, \delta, \varepsilon_0)$. Cet affaiblissement de l'hypothèse peut sembler artificiel dans le cas d'une équation sans paramètre, mais cela permet d'utiliser le résultat dans le cas d'une équation avec paramètre de contrôle α, par exemple lorsque α a été ajusté pour que l'équation présente des canards en un autre point tournant.

4. **A propos de la preuve.** Les détails de la preuve de ce résultat se trouvent dans [14]. Les grandes lignes sont les suivantes. Nous montrons d'abord qu'il existe des solutions pour ε et pour x dans des secteurs formant des recouvrements de l'origine, puis que ces solutions sont exponentiellement proches les unes des autres. Précisément, lorsqu'on change d'un secteur en ε à l'autre, les deux solutions sont sur une même montagne, donc leur différence est exponentiellement petite

de la forme $\exp(-\alpha/|\varepsilon|)$. En revanche, lorsqu'on change de secteur en x, les solutions sont définies sur deux montagnes adjacentes et il faut descendre dans la vallée les séparant pour qu'elle deviennent exponentiellement proches. Leur différence est alors de la forme $\exp(-\alpha|x^p/\varepsilon|)$. Il se trouve que ces estimations correspondent exactement aux conditions d'application du théorème de type Ramis-Sibuya pour les DAC évoqué en fin de partie 3. Nous ne donnons pas plus de détails sur cette preuve car, de fait, ce ne sont pas ces détails qui apportent des informations sur les développements.

5. **Pour le calcul** de ce DAC, on utilise la proposition 3.3, comme indiqué à la fin de la partie 3.3 : on commence par déterminer les développements extérieur et intérieur, puis on rejette la partie polaire du développement extérieur pour obtenir le développement lent et la partie polynomiale du développement intérieur pour obtenir le développement rapide.

Détaillons la procédure. Le développement extérieur est la solution formelle $\sum_{n\geq 1} y_n(x)\varepsilon^n$ de (4.1). Cette solution est à coefficients réguliers en dehors du point tournant 0 ; elle est donnée récursivement par

$$y_0 = 0, \qquad y_{n+1} = \frac{1}{f}(y_n' - q_n)$$

où q_n est le coefficient (dépendant de $y_1, \ldots, y_n$) du terme d'ordre n en ε du développement de Taylor par rapport à ε de la fonction

$$\widetilde{Q} : (x, \varepsilon) \mapsto \varepsilon Q\big(x, \sum_{1\leq\nu\leq n} y_\nu(x)\varepsilon^\nu, \varepsilon\big).$$

Notons que, vu comme développement en puissances de η, ce développement extérieur comporte de nombreux termes nuls : en comparaison avec le développement (3.6), on a $c_{np} = y_n$ et $c_k = 0$ si $k \not\equiv 0 \mod p$. Par ailleurs, ces fonctions y_n sont bien entendu les mêmes à gauche et à droite. Puisque les fonctions a_n sont les parties régulières des fonctions c_n, nous avons aussi $a_n = 0$ si $n \not\equiv 0 \mod p$ et ces fonctions a_n sont les mêmes à gauche et à droite.

Pour déterminer le développement intérieur, on pose $x = \eta X, y(x) = Y(X)$ (avec $\eta^p = \varepsilon$) et on aboutit à l'*équation intérieure*

$$\frac{dY}{dX} = c\, X^{p-1} Y + \eta G(X, Y, \eta) \tag{4.2}$$

avec $G(X, Y, \eta) = X^p f_1(\eta X)Y + P(\eta X, Y, \eta^p)$, où f_1 est la fonction analytique déterminée par $f(x) = c\, x^{p-1} + x^p f_1(x)$.

On montre qu'il existe une unique solution formelle $\widehat{Y}^+ = \sum_{n\geq 1} Y_n^+(X)\eta^n$ telle que $Y_n^+(X)$ est à croissance polynomiale lorsque $X \to \infty$, *i.e.* satisfait

$$\forall n \in \mathbb{N}^* \; \exists C, A > 0 \; \forall X \in]0, \infty], \qquad |Y_n^+(X)| \leq C\,|X|^A.$$

Cette solution formelle est déterminée récursivement en calculant l'unique solution à croissance polynomiale sur $\mathbb{R}^+$ de l'équation différentielle

$$\frac{dY_n^+}{dX} = c\, X^{p-1} Y_n^+ + G_n^+(X) \tag{4.3}$$

où G_n^+ est le coefficient (dépendant de $Y_1^+, \ldots, Y_{n-1}^+$) du terme d'ordre $n-1$ en η dans le développement de Taylor de la fonction $\eta \mapsto G\big(X, \sum_{1\leq\nu<n} Y_\nu^+(X)\eta^\nu, \eta\big)$. On trouve

$$Y_n^+(X) = \int_\infty^X \exp(X^p - s^p)\, G_n^+(s) ds.$$

De même, en changeant $+\infty$ par $-\infty$, on montre qu'il existe une unique solution formelle $\widehat{Y}^- = \sum_{n\geq 1} Y_n^-(X)\eta^n$ avec des solutions Y_n^- à croissance polynomiale sur $\mathbb{R}^-$.

Remarquons que les développements asymptotiques des $Y^{\pm}(X)$ quand $X \to +\infty$, resp. $X \to -\infty$, sont les mêmes. On peut le voir de deux manières : ils sont déterminés par la même récurrence que les $Y^{\pm}$ (dans l'algèbre des séries formelles) ou encore ils sont liés aux parties singulières des coefficients y_ν de la solution formelle extérieure.

Pour chaque $n \in \mathbb{N}^*$, la fonction $Y_n^{\pm}$ s'écrit $Y_n^{\pm} = P_n + g_n^{\pm}$, où P_n est un polynôme qui correspond aux parties régulières des fonctions y_ν. Ce polynôme est le même à gauche et à droite ; par conséquent on a $Y_n^+ - g_n^+ = Y_n^- - g_n^-$. Nous retrouvons ainsi le fait que chacun des DAC à gauche et à droite est déterminé de manière unique — en particulier ne dépend pas du choix de la condition initiale — comme nous l'avions écrit dans le commentaire 4.2.1.

6. Dans le cadre complexe, le *relief associé à (4.1)* est le graphe de la fonction

$$R : \mathbb{C} \simeq \mathbb{R}^2 \to \mathbb{R},\ x \mapsto \operatorname{Re} F(x), \quad \text{avec } F(x) = \int_0^x f(\xi)d\xi.$$

Il consiste en une succession de p montagnes M_k, $k = 0, ..., p-1$ où $R > 0$ et p vallées V_k où $R < 0$, délimitées par les séparatrices de col $R = 0$. A chaque montagne M_k dans la variable x pour le relief R est associé le "secteur-montagne" $S_k = S\big(\frac{2k\pi}{p} - \frac{\pi}{2p}, \frac{2k\pi}{p} + \frac{\pi}{2p}, \infty\big)$ dans la variable X, et on montre qu'il existe une unique solution formelle $\widehat{Y}^k = \sum_{n\geq 1} Y_n^k(X)\eta^n$ telle que $Y_n^k(X) = 0$ est à croissance polynomiale lorsque $X \to \infty$ dans S_k. Cette solution formelle est déterminée récursivement par $Y_0^k = 0$ et en calculant l'unique solution à croissance polynomiale sur S_k de l'équation (4.3). On trouve $Y_n^k(X) = \displaystyle\int_{\infty e^{2k\pi i/p}}^X \exp(X^p - s^p)\, G_n^k(s)ds$ où G_n^k est le coefficient du terme d'ordre $n-1$ en η dans le développement de Taylor de $(X, \eta) \mapsto G\big(X, \sum_{1\leq\nu<n} Y_\nu^k(X)\eta^\nu, \eta\big)$. L'analogue complexe du théorème 4.2 montre aussi qu'une solution venant d'une montagne a un DAC sur cette montagne et dans un voisinage de 0 de taille d'ordre η.

Deux de ces montagnes, l'une à l'ouest, l'autre à l'est, contiennent $[a, 0[$ et $]0, b]$. Dans la situation présente, la solution y donnée par le théorème 4.2, *i.e.* venant de la montagne est, ne se prolonge pas *a priori* aux autres montagnes, mais elle se prolonge à toutes les vallées ; elle admet aussi un DAC dans ces vallées, qui est le même que celui que nous venons de calculer. En particulier la solution Y_n^+ à croissance polynomiale sur $\mathbb{R}^+$ se prolonge dans tous les "secteurs-vallées" $S\big(\frac{2k\pi}{p} + \frac{\pi}{2p}, \frac{2k\pi}{p} + \frac{3\pi}{2p}, \infty\big)$, $k = 1, ..., p$. Ce ne sera plus le cas dans la généralisation qui suit.

4.3 Généralisation.

Dans beaucoup de situations, une équation singulièrement perturbée ne peut pas se ramener à une équation de la forme (4.1). En effet, *a priori* cette mise sous cette forme "quasi-linéaire" ne peut se faire localement qu'en-dehors d'un point tournant. Le résultat qui suit permet de traiter de telles situations. Des résultats dans cette direction ont été obtenus par Eric Matzinger, notamment dans [19], sur des équations très voisines.

On considère l'équation

$$\varepsilon y' = \big(f(x) + \varepsilon g(x, \varepsilon)\big)y + \varepsilon h(x, \varepsilon) + y^2 P(x, y, \varepsilon) \tag{4.4}$$

avec, comme précédemment, f analytique réelle dans un voisinage d'un intervalle $[a, b]$ avec $a < 0 < b$, vérifiant $xf(x) > 0$ si $x \neq 0$, g et h analytiques dans un voisinage de $[a, b] \times \{0\}$, et P analytique dans un voisinage de $[a, b] \times \{0\} \times \{0\}$.

Etant donné $r \in \{1, ..., p-1\}$, on dira que l'équation (4.4) *satisfait la condition* (C_r) si, d'une part $h(x,0) = \mathcal{O}(x^{r-1})$, $x \to 0$ et d'autre part le développement $P(x,y,0) = \sum_{k,l\geq 0} p_{k,l} x^k y^l$ satisfait $p_{k,l} = 0$ pour tous les k, l tels que $k + rl < p - r - 1$. Cette condition a pour conséquence que, pour X, Y fixés, $P(\eta X, \eta^r Y, \varepsilon) = \mathcal{O}(\eta^{p-r-1})$.

Théorème 4.3 . *On suppose que (4.4) satisfait la condition* (C_r) *pour un certain* $r \in \{1, ..., p-1\}$. *Soit* $x_1 \in]0, b]$ *et* $y_1 \in \mathbb{R}$ *suffisamment petit. Soit* $y = y(x, \varepsilon)$ *la solution de (4.4) de condition initiale* $y(x_1, \varepsilon) = y_1$.

Alors il existe $\varepsilon_0, \mu > 0$ *tels que* y *est définie pour* $0 < \varepsilon \leq \varepsilon_0$ *et* $\mu\eta \leq x \leq x_1$, *avec* $\eta = \varepsilon^{1/p}$. *De plus* y *possède un* DAC *Gevrey d'ordre* $\frac{1}{p}$ *par rapport à* $\eta = \varepsilon^{1/p}$ *: il existe des fonctions* $a_n \in \mathcal{H}$ *et* $g_n \in \mathcal{G}(\mu)$, *telles que, pour tout* $\delta > 0$, $y(x,\varepsilon) \sim_{1/p} \sum_{n\geq r} \Big(a_n(x) + g_n\big(\frac{x}{\eta}\big)\Big)\eta^n$ *quand* $\eta \to 0$ *uniformément pour* $x \in [\mu\eta, x_1 - \delta]$.

Remarques . 1. La différence principale par rapport au théorème 4.2 est qu'a priori la solution y n'est pas définie dans un voisinage de 0 (de taille proportionnelle à η). La comparaison avec le "matching" (*c.f.* proposition 3.3) montre que les termes lents a_n sont nuls lorsque $n \not\equiv 0 \mod p$. En particulier le coefficient de η^r n'est constitué que du terme rapide $g_r \in \mathcal{G}(\mu)$. De plus, pour $x > 0$ fixé on a $y(x,\varepsilon) = \mathcal{O}(\varepsilon)$, donc ce terme g_r satisfait $g_r(X) \sim \sum_{m\geq p-r} g_{r,m} X^{-m}$ quand X tend vers $-\infty$.

2. L'exemple ci-dessous montre que la condition (C_r) est nécessaire et naturelle. Il s'agit de l'équation

$$\varepsilon y' = 4x^3 y - \varepsilon - xy^2. \tag{4.5}$$

On a donc $p = 4$, $r = 1$ et $p_{1,0} \neq 0$: la condition (C_4) n'est pas satisfaite. On peut montrer facilement que les solutions de (4.5) proches de la courbe lente 0 ne peuvent pas avoir des DAC. Si elles en avaient, alors la proposition 3.3 impliquerait que le coefficient de $\varepsilon^n = \eta^{4n}$ dans le développement extérieur admet au maximum un pôle d'ordre $4n$. On constate, par contre, que la solution formelle de (4.5) présente des pôles d'ordre trop élevé. La recherche d'une solution formelle $\widehat{y} = \sum_{n\geq 1} y_n \varepsilon^n$ conduit à la récurrence

$$y_1(x) = \frac{1}{4x^3}, \qquad y_n(x) = \frac{1}{4x^3}\Big(y'_{n-1}(x) + x\sum_{k=1}^{n-1} y_k(x) y_{n-k}(x)\Big).$$

On montre alors par récurrence que y_n est de la forme $y_n(x) = x^{-5n+2}\big(a_n + xP_n(x)\big)$ où P_n est polynomial et où les nombres a_n satisfont $a_1 = 1$ et pour $n \geq 2$, $a_n = \frac{1}{4}\sum_{k=1}^{n-1} a_k a_{n-k}$ donc sont strictement positifs. Il s'ensuit que y_n a un pôle d'ordre $5n-2$ en $x = 0$.

Par ailleurs, on pourrait étudier cette équation en utilisant le changement de variables et inconnues $x = \mu X, y = \mu^2 Y, \varepsilon = \mu^5$, qui conduit à

$$\mu\frac{dY}{dX} = 4X^3 Y - 1 - XY^2.$$

Une étude de cette équation singulièrement perturbée (avec des moyens en dehors de notre propos) montre que ses solutions restant bornées sur des intervalles $[L, K/\mu]$ avec certains $L, K > 0$ ont des singularités à droite de $X = 0$. Ceci signifie que les solutions correspondantes de l'équation originale (4.5) admettent des pôles à l'échelle $\mu = \varepsilon^{1/5}$, ce qui est incompatible avec l'existence de DAC.

3. La preuve de theorème 4.3 est basée sur le lemme suivant (ou plutôt sa généralisation dans le champ complexe) :

Lemme 4.4 . *Soit $n \in \{0, 1, .., p-1\}$ et B une fonction bornée par 1. On pose*

$$z_n(x,\varepsilon) = \int_{x_1}^{x} e^{\left(F(x)-F(u)\right)/\varepsilon} u^n B(u) du.$$

Alors il existe une fonction $\delta :]0, +\infty[\to \mathbb{R},\ L \mapsto \delta(L)$ tendant vers 0 quand $L \to +\infty$ telle que pour tout $\eta \in]0, \eta_0]$ et tout $x \in [L\eta, x_1]$ on a $|z_n(x,\varepsilon)| \leq \delta(L)\eta^{n+1}$.

Dans le cas où la condition initiale est $y(x_1, \varepsilon) = 0$, on réécrit alors (4.4) en une équation de point fixe par la variation de la constante :

$$y(x,\varepsilon) = \int_{x_1}^{x} e^{\left(F(x)-F(u)\right)/\varepsilon} \Big(g(u,\varepsilon)y(u,\varepsilon) + h(u,\varepsilon) + \frac{1}{\varepsilon} y(u,\varepsilon)^2 P\big(u, y(u,\varepsilon), \varepsilon\big) \Big) du.$$

Si μ est assez grand, le lemme précédent et les conditions du théorème sur P permettent de montrer que le côté droit définit un opérateur contractant dans l'ensemble des fonctions continues et bornées par η^r quand $\eta \in]0, \eta_0]$ et $x \in]\mu\eta, x_1]$. Comme pour la preuve du théorème 4.2, il faut généraliser ceci dans des secteurs formant des recouvrements de l'origine dans $\mathbb{C}^2$, puis montrer que les solutions ainsi construites sont exponentiellement proches et appliquer le théorème de type Ramis-Sibuya mentionné à la fin de la partie 3.

4. Ici encore, on a un résultat identique concernant les solutions venant de la gauche. De même que dans le cas "quasi-linéaire", les DAC peuvent être calculés à partir du développement formel de (4.4) et des solutions formelles à droite et à gauche de l'équation intérieure analogue à (4.2). En particulier, ils ne dépendent pas du choix des conditions initiales.

5. L'équation intérieure qui généralise (4.2), obtenue en posant $x = \eta X, y(x) = \eta^r Y(X)$, a pour limite l'*équation intérieure réduite* lorsque $\eta \to 0$

$$\frac{dY}{dX} = pX^{p-1}Y + cX^{r-1} + Y^2 Q(X,Y). \tag{4.6}$$

où c est donné par $h(x,0) = cx^{r-1} + \mathcal{O}(x^r)$ et où Q est la partie quasi-homogène de plus bas degré de P, donnée par

$$Q(X,Y) = \sum_{k+rl=p-r-1} p_{k,l} X^k Y^l \tag{4.7}$$

Le théorème 4.3 a le défaut de ne pas contenir d'information sur le domaine de validité du DAC, en particulier sur le nombre μ. L'énoncé suivant permet d'avoir cette information à partir d'informations sur la solution Y_0 de (4.6).

Corollaire 4.5 . *Sous les conditions du théorème 4.3, on suppose que la solution Y_0 de l'équation intérieure réduite (4.6) satisfaisant $Y_0(X) \sim -\frac{c}{p}X^{r-p}$ quand $X \to +\infty$ peut être prolongée sur un intervalle ouvert contenant $[\widetilde{\mu}, +\infty[$ avec un certain $\widetilde{\mu} \in \mathbb{R}$. Alors la solution y du théorème 4.3 peut être prolongée et admet un DAC Gevrey pour l'ensemble des (x,η) tels que $\eta \in]0, \eta_1]$ et $x \in]\widetilde{\mu}\eta, x_1 - \delta]$.*

Nous proposons un cas particulier séparément.

Corollaire 4.6 . *Sous les conditions du théorème 4.3, on suppose que $p_{kl} = 0$ si $k+rl = p-r-1$. Alors, pour tout $\widetilde{\mu} \in \mathbb{R}$, il existe $\eta_1 > 0$ et une solution $y(x,\eta)$ de (4.4) définie pour $\eta \in]0, \eta_1]$ et $x \in]\widetilde{\mu}\eta, x_1 - \delta]$.*
De plus y a un DAC Gevrey d'ordre $\frac{1}{p}$ quand $\eta \to 0$ et $x \in]\widetilde{\mu}\eta, x_1 - \delta]$.

La preuve du corollaire 4.5 utilise d'une part le théorème de dépendance de solutions d'équations différentielles par rapport aux paramètre et aux conditions initiales et d'autre part les propriétés de prolongement de DAC évoquées à la fin de la partie 3.3. Nous renvoyons à [14] pour les détails. La preuve du corollaire 4.6 est immédiate : dans ce cas, l'équation (4.6) est linéaire et la condition du corollaire 4.5 est donc trivialement satisfaite. □

5 Applications.

5.1 Canard en un point tournant multiple.

Notre théorie des DAC permet de donner une condition nécessaire et suffisante pour l'existence de canards. Ici, l'aspect Gevrey est indispensable. C'est pourquoi nous sommes amenés à faire une hypothèse d'analyticité dans un voisinage complexe d'un intervalle réel $[a, b]$. On considère l'équation

$$\varepsilon y' = f(x)y + \varepsilon P(x, y, \varepsilon) \tag{5.1}$$

où f est analytique dans un voisinage complexe d'un intervalle réel $[a, b]$ avec $a < 0 < b$, f réelle sur $\mathbb{R}$, $xf(x) > 0$ si $x \neq 0$, et P analytique au voisinage de $[a, b] \times \{0\} \times \{0\} \subset \mathbb{C}^3$, $P(x, y, \varepsilon)$ réel si x, y, ε le sont. On suppose de plus que $x = 0$ est un point tournant *multiple*, *i.e.* $f(x) = c\,x^{p-1} + \mathcal{O}(x^p)$ si $x \to 0$ avec p pair, $p \geq 4$ et $c > 0$.

Un *canard local* est une solution de (5.1) bornée sur un intervalle ouvert contenant 0 ("bornée" sous-entend uniformément par rapport à ε).

Un *canard global* est une solution de (5.1) bornée sur tout $[a, b]$.

Dans [13], nous avions établi une équivalence entre l'existence de solutions formelles $\widehat{y} = \sum_{n\geq 0} y_n \varepsilon^n$ avec y_n sans pôle en $x = 0$, l'existence de solutions surstables et l'existence de ce que nous avions appelé *canards-$\mathcal{C}^\infty$*. Ce sont des canards dont toutes les dérivées sont bornées (uniformément par rapport à ε) dans un voisinage de 0. En particulier, l'existence d'une solution formelle sans pôles implique l'existence de canards locaux. On peut vérifier que ceci correspond exactement à la situation où les fonctions g_n sont identiquement nulles pour tous les n, si bien que le DAC devient un développement asymptotique au sens classique du terme.

Revenons à la situation générale. Nous sommes dans les conditions d'application du théorème 4.2. Une solution y^- de condition initiale $y^-(x_1, \varepsilon) = y_1$ avec $x_1 \in [a, 0[$ admet un DAC Gevrey

$$y^-(x, \eta) \sim_{\frac{1}{p}} \sum_{n\geq 1} \left(a_n(x) + g_n^-\left(\tfrac{x}{\eta}\right)\right) \eta^n$$

pour $\eta \in]0, \eta_0]$ et $x \in [x_1 + \delta, \mu\eta[$ pour certains $\eta_0, \delta, \mu > 0$. De même, une solution y^+ de condition "initiale" $y^+(x_2, \varepsilon) = y_2$ avec $x_2 \in]0, b]$ admet un DAC

$$y^+(x, \eta) \sim_{\frac{1}{p}} \sum_{n\geq 0} \left(a_n(x) + g_n^+\left(\tfrac{x}{\eta}\right)\right) \eta^n.$$

pour $\eta \in]0, \eta_0]$ et $x \in]-\mu\eta, x_2 - \delta]$. De plus les fonctions a_n (qui sont les mêmes à gauche et à droite), g_n^- et g_n^+ ne dépendent pas des conditions initiales.

Supposons à présent qu'il existe un canard local sur un intervalle $[-c, c]$. Alors ce canard admet des DAC à droite et à gauche sur un intervalle $[-c + \delta, c - \delta]$ pour $\delta > 0$ arbitrairement petit. En particulier les fonctions g_n^- et g_n^+ sont de prolongements les unes des autres. Le résultat qui suit montre que cette condition nécessaire $g_n^- \equiv g_n^+$ est aussi une condition suffisante pour l'existence d'un canard local.

Théorème 5.1 . *Les assertions suivantes sont équivalentes.*
(a) *Il existe un canard local.*
(b) *Pour tout* $n \in \mathbb{N}$, *on a* $g_n^- \equiv g_n^+$.
(c) *Pour tout* $n \in \mathbb{N}$, *on a* $g_n^-(0) = g_n^+(0)$.

Preuve . L'implication (a)⇒(b) a été présentée avant l'énoncé ; l'implication (b)⇒(c) est évidente. Il suffit donc de démontrer (c)⇒(a). Si (c) est satisfaite, alors les solutions $y^{\pm}(0,\eta)$ admettent *le même* développement asymptotique quand η tend vers 0. De plus, d'après notre théorie, il s'agit de développements Gevrey d'ordre $\frac{1}{p}$ par rapport à η. La différence $d(\eta) = y^+(0,\eta) - y^-(0,\eta)$ est donc exponentiellement petite, *i.e.* satisfait $|d(\eta)| \leq \exp(-\alpha/\varepsilon)$ avec un certain $\alpha > 0$. Il est bien connu (voir par exemple [4]), que le lemme de Gronwall entraîne que les deux solutions $y^{\pm}(x,\eta)$ existent et que leur différence $y^+(x,\eta) - y^-(x,\eta)$ reste exponentiellement petite sur un voisinage $]-\delta,\delta[$ indépendent de η. On a donc bien un canard local. □

Remarques . 1. Dans [5], Peter De Maesschalck démontre l'équivalence entre l'existence d'un canard local et l'existence d'un canard global. Une technique analogue à celle présentée dans [13] permet aussi de démontrer ce résultat. Puisque ceci ne concerne pas directement les DAC, nous ne mettons pas de détails et renvoyons à [14].

2. Nous insistons sur le fait que le caractère Gevrey des DAC joue un rôle crucial dans cette preuve. C'est pour ce problème et parce que les développements asymptotiques classiques ne sont plus applicables dans ce contexte que nous avons développé la théorie des DAC Gevrey.

4. La condition $g_n^- \equiv g_n^+$ peut être exprimée aussi sur les développements formels de solutions de l'équation intérieure ; ceci permet une autre preuve de l'implication (c)⇒(b). L'équation intérieure pour $x = \eta X\,, y(x) = Y(X)$, $\varepsilon = \eta^p$ est de la forme

$$\frac{dY}{dX} = c\,X^{p-1}Y + \eta G(X,Y,\eta) \tag{5.2}$$

avec $G(X,Y,\eta) = X^p f_1(\eta X)Y + P(\eta X, Y, \eta^p)$, f_1 donnée par $f(x) = c\,x^{p-1} + x^p f_1(x)$. Chacune des deux solutions Y^+ et Y^- correspondant à y^+ et y^- a un développement en puissances de η de la forme $\sum_{n\geq 1} Y_n^{\pm}(X)\eta^n$ où les $Y_n^{\pm}$ sont donnés récursivement par

$$Y_0^{\pm} = 0,\ Y_n^{\pm}(X) = \int_{\pm\infty}^{X} \exp(X^p - s^p)\, G_n^{\pm}(s)ds,$$

où $G_n^{\pm}$ est le coefficient (dépendant de $Y_1^{\pm}, \ldots, Y_{n-1}^{\pm}$) du terme d'ordre $n-1$ en η obtenu en développant $G\big(X, \sum_{1\leq\nu<n} Y_\nu^{\pm}(X)\eta^\nu, \eta\big)$ par Taylor par rapport à η.

Comme nous l'avons vu dans le commentaire 4.2.5, la partie polynomiale de Y_n^+, qui est $Y_n^+ - g_n^+$, correspond aux parties régulières des coefficients y_ν de l'unique solution formelle $\sum_{\nu\geq 1} y_\nu(x)\varepsilon^\nu$ de (5.1), donc est égale à $Y_n^- - g_n^-$. Ainsi la condition $g_n^+ = g_n^-$ est équivalente à la condition $Y_n^+ = Y_n^-$.

Par récurrence, si pour $k < n$ on a $Y_k^+ \equiv Y_k^-$, alors on a $G_n^+ \equiv G_n^-$. La condition (c) de l'énoncé implique de plus $Y_n^-(0) = Y_n^+(0)$; les fonctions Y_n^- et Y_n^+ sont solutions d'une même équation d'ordre 1 avec même condition initiale, donc sont égales, ce qui implique (b). Par ailleurs, la condition $Y_n^+ \equiv Y_n^-$ est aussi équivalente à $\int_{-\infty}^{+\infty} \exp(-s^p)\, G_n^+(s)ds = 0$.

5. On peut affiner l'énoncé du théorème 5.1 dans une direction qui rejoint les canards-$\mathcal{C}^\infty$ de la façon suivante. Pour $m \geq 1$, appelons *canard-*$\mathcal{C}^m$ une solution de (5.1) dont toutes les dérivées sont bornées sur un intervalle $]-\delta,\delta[$ (uniformément par rapport à ε) jusqu'à l'ordre m inclus. Alors nous avons l'énoncé suivant :

Il existe un canard-$\mathcal{C}^m$, $m \geq 1$, si, et seulement si, d'une part une des conditions équivalentes de théorème 5.1 est satisfaite et d'autre part $g_n^+ \equiv g_n^- \equiv 0$ pour tout $n \leq m-1$.

6. Ici nous n'avons parlé que de canards d'équations sans paramètre de contrôle additionnel. L'étude des canards concernant des équations avec paramètre de contrôle est beaucoup plus répandue. Outre l'abondante littérature de l'école non standard française, en particulier des membres du réseau Georges Reeb (dont un grand nombre sont dans les références de [7] par exemple) on peut citer [3, 4, 6].

5.2 Canards non lisses.

Equations de type « Union Jack ».

On considère une équation différentielle de la forme

$$\varepsilon y' = y(y-x)(y+x) + P(x,y,\varepsilon) + \varepsilon c, \tag{5.3}$$

où P est analytique réelle dans un domaine $\mathcal{D} \subset \mathbb{R}^3$ qui sera précisé dans la suite et $c \in \mathbb{R}$ un paramètre additionnel. On fait l'hypothèse que $P(0,0,\varepsilon) = \mathcal{O}(\varepsilon^2)$ et que la valuation homogène de $P(x,y,0)$ est au moins 4, *i.e.* il existe des $p_{kl} \in \mathbb{R}$ tels que

$$P(x,y,0) = \sum_{k+l \geq 4} p_{kl} x^k y^l, \quad \text{pour } |x|, |y| \text{ assez petits.} \tag{5.4}$$

L'ensemble lent de (5.3), d'équation $y(y-x)(y+x) + P(x,y,0) = 0$, peut être désingularisé par un éclatement $y = xz$. On obtient l'équation $z(z-1)(z+1) + xQ(x,z) = 0$ avec $Q(x,z) = x^{-4}P(x,xz,0)$ analytique dans $\mathcal{D}$, y compris pour $x = 0$, à laquelle on peut appliquer le théorème des fonctions implicites aux points $(0,0)$ et $(0,\pm 1)$.

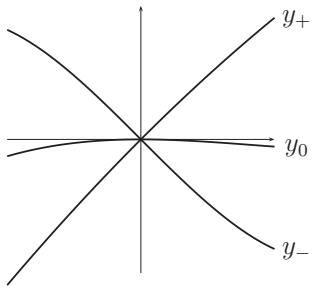


FIG. 3 – Les branches y_-, y_+ et y_0 de l'ensemble lent $y(y-x)(y+x) + P(x,y,0) = 0$.

On a ainsi localement trois courbes lentes analytiques $y_0(x) = \mathcal{O}(x^2)$ et $y_\pm(x) = \pm x + \mathcal{O}(x^2)$. Marc Diener a donné pour nom *Union Jack* à l'équation (5.3) car l'équation « modèle » $\varepsilon y' = y(y-x)(y+x) + \varepsilon c$ a pour ensemble lent la réunion des trois droites $y = 0$ et $y = \pm x$, et ressemble donc au drapeau du Royaume-Uni. Les trois courbes lentes y_0, y_+ et y_- du cas général forment ainsi un « Union Jack modifié ».

On suppose que la courbe lente y_0 peut être prolongée analytiquement sur $[a, \delta]$ et la courbe lente y_+ sur $[-\delta, b]$ avec un certain $\delta > 0$. On note encore y_0 et y_+ ces prolongements. L'hypothèse que doit vérifier le domaine $\mathcal{D}$ d'analyticité de P est

$$\forall x \in [a,0] \quad \big(x, y_0(x), 0\big) \in \mathcal{D} \quad \text{et} \quad \forall x \in [0,b] \quad \big(x, y_+(x), 0\big) \in \mathcal{D}.$$

Par ailleurs, les droites $y = 0$ et $y = \pm x$ sont aussi des solutions particulières de l'équation modèle pour certaines valeurs de c : la solution $y \equiv 0$ pour $c = 0$, $y \equiv x$ pour $c = 1$ et $y \equiv -x$ pour $c = -1$. Il en est de même pour l'équation intérieure réduite de (5.3), obtenue en posant $x = \eta X$, $y = \eta Y$, $\varepsilon = \eta^3$ et en faisant tendre η vers 0, *i.e.*

$$Y' = Y(Y - X)(Y + X) + c. \tag{5.5}$$

Quelque soit la valeur de c, l'équation (5.5) admet une unique solution $Y_g(X, c)$ telle que

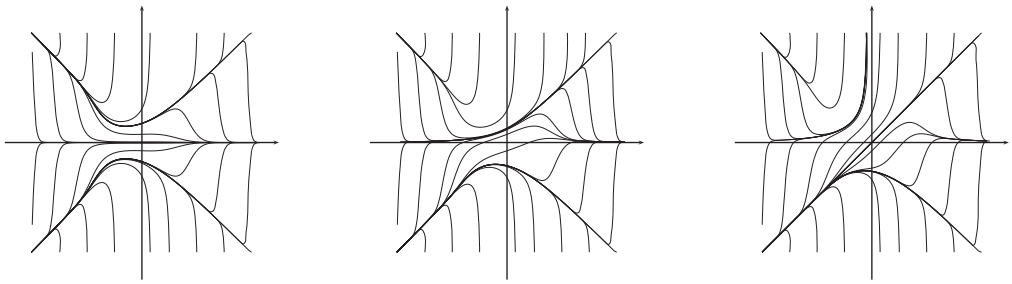


FIG. 4 – Trois portraits de phase de l'équation (5.5), pour $c = 0$, $c = 0.3621759411$ et $c = 1$. Sur les trois figures, $|X|, |Y| \leq 4$.

$Y_g(X, c) \to 0$ quand $X \to -\infty$; en effet, cette équation est de la forme $Y' = -X^2 Y + \mathcal{O}(1 + |X|)$ quand Y reste bornée. Pour une raison analogue, il existe, pour toute valeur de c, deux solutions uniques $Y_d^{\pm}(X, c)$ telles que $Y_d^{\pm}(X, c) \sim \pm X$ quand $X \to +\infty$. Ces trois solutions sont les fleuves répulsifs que l'on voit sur chacun des portraits de phase de la figure 4. On vérifie que $Y_g(X, c) = \mathcal{O}(X^{-2})$ quand $X \to -\infty$ et $Y_d^{\pm}(X, c) = \pm X + \mathcal{O}(X^{-2})$ quand $X \to \infty$; ceci est le cas uniformément pour des compacts en c.

Il a été démontré dans [8] que (5.5) admet une valeur unique $c = c_0 \in]0, 1[$, telle que Y_g et Y_d^+ coïncident, *i.e.* $Y_g(X, c_0) \equiv Y_d^+(X, c_0) \equiv: Y_0(X)$. Cette valeur de c est une valeur à long canard pour l'équation correspondante $\varepsilon y' = y(y - x)(y + x) + \varepsilon c$ car la solution $y(x, \varepsilon) = Y_0\big(\frac{x}{\eta}\big)$ est attractive quand $x \leq \delta < 0$ et répulsive quand $x \geq \delta > 0$; il s'agit d'un canard *non lisse* car la limite uniforme $z(x)$ de $y(x, \varepsilon)$ quand $\varepsilon \to 0$ est $z(x) = 0$ quand $x \leq 0$ et $z(x) = x$ quand $x > 0$: la courbe lente associée n'est pas dérivable. Par ailleurs, la valeur $c = -c_0$ est aussi une valeur à canard non lisse puisque, par symétrie, les solutions Y_g et Y_d^- coïncident pour cette valeur de c. Néanmoins, nous restreignons notre étude aux canards longeant y_0 et y_+.

Le résultat qui suit répond à la question naturelle si ce phénomène persiste pour l'équation complète (5.3) et si on peut décrire les valeurs à canards et les solutions canards correspondantes. Cette question a été résolue par Marc Diener [8] et Emmanuel Isambert [17]. Notre théorie des DAC se révèle particulièrement bien adaptée ici et permet d'apporter des compléments d'information : une approximation uniforme des solutions canards et le caractère Gevrey des développements asymptotiques.

Théorème 5.2 . *Avec les hypothèses et notations précédentes, on suppose que y_0 est attractive sur $[a, 0[$, i.e. $3y_0(x)^2 - x^2 + \frac{\partial P}{\partial y}\big(x, y_0(x), 0\big) < 0$ pour tout $x \in [a, 0[$, tandis que y_+ est répulsive sur $]0, b]$.*

Alors l'équation (5.3) admet une valeur à canard non lisse $c = c(\eta)$ et une solution canard $y(x, \eta)$ correspondante telles que $y(x, \eta) - z(x) = \mathcal{O}(\eta)$, où $z(x) = y_0(x)$ quand $x \leq 0$ et $z(x) = y_+(x)$ quand $x > 0$.

De plus, la fonction $c = c(\eta)$ admet un développement asymptotique Gevrey d'ordre $\frac{1}{3}$ de la forme

$$c(\eta) \sim_{1/3} \sum_{n=0}^{\infty} c_n \eta^n$$

avec pour premier terme la valeur c_0 introduite précédemment. De même, la fonction y admet des DAC *Gevrey d'ordre $\frac{1}{3}$*

$$y(x,\eta) \sim_{1/3} y_0(x) + \sum_{n=1}^{\infty} \Big(a_{gn}(x) + b_{gn}\big(\tfrac{x}{\eta}\big) \Big) \eta^n \tag{5.6}$$

quand $\eta \to 0$ et $x \in [a, L\eta[$ pour $L > 0$ arbitraire et

$$y(x,\eta) \sim_{1/3} y_+(x) + \sum_{n=1}^{\infty} \Big(a_{dn}(x) + b_{dn}\big(\tfrac{x}{\eta}\big) \Big) \eta^n \tag{5.7}$$

quand $\eta \to 0$ et $x \in]-L\eta, b]$, où les a_{gn} sont analytiques sur un voisinage (complexe) de $[a, \delta]$, les a_{dn} sur un voisinage de $[-\delta, b]$ et où les b_{gn} et les b_{dn} sont analytiques sur un voisinage de $\mathbb{R}$. Les fonctions b_{gn} admettent des développements Gevrey d'ordre $\frac{1}{3}$ compatibles au sens de (3.10) quand $X \to -\infty$, les b_{dn} quand $X \to +\infty$.

Enfin, l'énoncé analogue (avec des hypothèses analogues sur y_-) est vrai pour $y_-(x)$ à la place de $y_+(x)$.

REMARQUES. 1. La série asymptotique $\widehat{c}(\eta)$ de $c(\eta)$ et les séries formelles combinées du théorème se calculent comme avant à partir des développements extérieurs et intérieurs. Néanmoins, ici il faut commencer par le développement intérieur pour pouvoir déterminer $\widehat{c}(\eta)$; tout ceci a été fait par Emmanuel Isambert dans [17].

De même que dans la remarque 4 qui suit le théorème 5.1, la procédure de calcul des parties lente et rapide (*c.f.* commentaire 4.2.5) montre que les deux développements intérieurs doivent coïncider, *i.e.* $y(\eta X, \eta) \sim \sum_{n=1}^{\infty} u_n(X)\eta^n =: \eta \widehat{Y}(X,\eta)$ avec une solution formelle de l'équation intérieure $\widehat{Y}(X,\eta)$ avec $c = \widehat{c}(\eta)$. Puisque, d'après la proposition 3.3, ce développement intérieur est constitué de la partie rapide du DAC et d'une partie polynomiale, les coefficients $u_n(X)$ ont une croissance polynomiale quand $X \to +\infty$ et $X \to -\infty$. On montre (comme dans [17]) que ceci détermine de manière unique les valeurs des c_n et les fonctions $u_n(X)$. Par exemple, on a la relation $b_{g1}(X) = u_1(X) = Y_0(X) = X + b_{d1}(X)$.

Les parties lentes des DAC (5.6) et (5.7) sont ensuite déterminées de façon usuelle comme étant les parties non polaires des solutions formelles extérieures de (5.3) avec $c = \widehat{c}(\eta)$ pour les courbes lentes $y_0(x)$, respectivement $y_+(x)$. Ceci implique par exemple $a_{gn} = 0$ et $a_{dn} = 0$ pour $n = 1$ et 2.

2. Comme cela est classique dans les problèmes de canards, il n'y a pas unicité des valeurs à canard $c(\eta)$ ni des solutions canards correspondant à une valeur à canard : une modification de $c(\eta)$ ou de la condition initiale par un terme exponentiellement petit $\mathcal{O}\big(e^{-K/\eta^p}\big)$ avec $K > 0$ suffisamment grand ne change pas la conclusion du théorème. Réciproquement, deux valeurs à canard sont exponentiellement proches. Ceci peut se démontrer avec la variation de la constante, *c.f.* par exemple [1, 3, 4], ou encore [15] et les références qui y sont incluses.

Preuve. Sur un voisinage de 0, disons $|x| < \gamma$, on fait le changement de variable $y = y_0(x) + z$.

L'équation obtenue s'écrit d'abord sous la forme

$$\begin{aligned}\varepsilon z' &= Q(x,z,\varepsilon)+\varepsilon c\\ &:= \big(z+y_0(x)\big)\big(z+y_0(x)-x\big)\big(z+y_0(x)+x\big)+\\ &\quad P\big(x,z+y_0(x),\varepsilon\big)-\varepsilon y_0'(x)+\varepsilon c.\end{aligned}$$

Par construction, on a $Q(x,0,0)\equiv 0$; la fonction $\widetilde{P}:(x,z,0)\mapsto Q(x,z,0)-z(z-x)(z+x)$ satisfait une propriété analogue à (5.4). L'équation précédente peut donc être écrite

$$\varepsilon z' = r(x)z+\varepsilon\big(c+s(x,\varepsilon)\big)+zR(x,z,\varepsilon), \tag{5.8}$$

où on a décomposé $Q(x,z,\varepsilon)=r(x)z+\varepsilon s(x,\varepsilon)+zR(x,z,\varepsilon)$ avec $r(x)=\frac{\partial Q}{\partial z}(x,0,0)=-x^2+\mathcal{O}(x^3)$, $s(x,\varepsilon)=\frac{1}{\varepsilon}Q(x,0,\varepsilon)$ satisfait $s(0,0)=0$ et R satisfait $R(x,0,0)=0$. D'après notre hypothèse sur P et l'observation précédente pour $\widetilde{P}$, la fonction R peut être développée pour $\varepsilon=0$. Précisément, on a

$$R(x,z,0)=z^2+\sum_{k\geq 0,l\geq 1,k+l\geq 3} R_{kl}x^k z^l,$$

quand $|x|<\gamma$, $|z|$ petit.

Après l'homothétie $x\to -\sqrt[3]{3}x$, l'équation (5.8) entre donc dans le cadre du théorème 4.3 avec $p=3$ et $r=1$. De plus, son équation intérieure réduite est $Z'=Z(Z-X)(Z+X)+c$. La solution de cette dernière équation dont le comportement asymptotique est de la forme const. X^{-2} quand X tend vers $-\infty$ est Y_g. Quand $c=c_0$, celle-ci coïncide avec Y_d^+ ; en particulier, elle peut être prolongée analytiquement sur $\mathbb{R}$. Étant donné $M>0$, il existe donc un voisinage $|c-c_0|<\rho$ tel que Y_g peut être prolongée sur $]-\infty,M[$. Maintenant on applique le corollaire 4.5. On obtient que (5.8) admet une solution $z=z(x,c,\eta)=\mathcal{O}(\eta)$ ayant un DAC quand $\eta\to 0$, $x\in[-\gamma,L\eta[$ et $|c-c_0|<\rho$ avec un certain $L>0$. Or il est bien connu que notre hypothèse d'attractivité de $y_0(x)$ sur $[a,0[$ entraîne que la solution de (5.8) avec condition initiale 0 en un point un peu avant a admet un développement asymptotique Gevrey d'ordre 1 en ε sans terme constant sur l'intervalle $[a,-\gamma/2]$, disons. Elle est donc exponentiellement proche de la solution $z=z(x,c,\eta)$ construite ci-dessus, uniformément pour $x\in[-\gamma,-\gamma/2]$ et $|c-c_0|<\rho$, Ainsi, on obtient l'existence d'une solution $y_g(x,c,\eta)$ de (5.3) analytique pour $\eta\in]0,\eta_1]$, $x\in[a,L\eta[$ et $|c-c_0|<\rho$, pour certains $L,\eta_1>0$, ayant un DAC Gevrey d'ordre $\frac{1}{3}$

$$y_g(x,c,\eta)\sim_{1/3} y_0(x)+\sum_{n=1}^{\infty}\Big(a_{gn}(x,c)+b_{gn}\big(\tfrac{x}{\eta},c\big)\Big)\eta^n, \tag{5.9}$$

où $b_{g1}(X,c)=Y_g(X,c)$.

Le changement $y=y_+(x)+z$ dans l'équation (5.3) donne une équation en z qui entre aussi dans le cadre du corollaire 4.5. En utilisant ici la répulsivité de y_+ sur $]0,b]$, on obtient l'existence d'une solution $y_d(x,c,\eta)$ de (5.3) analytique pour $\eta\in]0,\eta_1]$, $x\in]-L\eta,b]$ et $|c-c_0|<\rho$ admettant un DAC Gevrey d'ordre $\frac{1}{3}$

$$y_d(x,c,\eta)\sim_{1/3} y_+(x)+\sum_{n=1}^{\infty}\Big(a_{dn}(x,c)+b_{dn}\big(\tfrac{x}{\eta},c\big)\Big)\eta^n, \tag{5.10}$$

où $b_{d1}(X,c)+X=Y_d^+(X,c)$.

On obtient des valeurs à canards non lisses du théorème en résolvant l'équation

$$y_g(0,c,\eta) = y_d(0,c,\eta). \tag{5.11}$$

Il faut donc montrer que la solution $c = c(\eta)$ existe et qu'elle a les propriétés énoncées, ainsi que les fonctions $y_{g|d}\big(x, c(\eta), \eta\big)$.

On applique le théorème des fonctions implicites à l'équation (5.11) modifiée en

$$f(c,\eta) = 0, \text{ où } f(c,\eta) = \tfrac{1}{\eta}\big(y_g(0,c,\eta) - y_d(0,c,\eta)\big). \tag{5.12}$$

On a d'abord les égalités

$$\lim_{\eta\to 0} \tfrac{1}{\eta} y_g(0,c_0,\eta) = Y_0(0) = \lim_{\eta\to 0} \tfrac{1}{\eta} y_d(0,c_0,\eta)$$

et donc $\lim_{\eta\to 0} f(c_0,\eta) = 0$. En utilisant l'analyse complexe, on peut montrer que

$$\lim_{\eta\to 0} \tfrac{1}{\eta} \frac{\partial y_g}{\partial c}(0,c_0,\eta) \neq \lim_{\eta\to 0} \tfrac{1}{\eta} \frac{\partial y_d}{\partial c}(0,c_0,\eta) \tag{5.13}$$

et donc $\lim_{\eta\to 0} \frac{\partial f}{\partial c}(c_0,\eta) \neq 0$; ceci sera détaillé plus bas. Les conditions du théorème des fonctions implicites sont donc satisfaites et on obtient l'existence d'une solution $c = c(\eta)$ de (5.11) avec $c(0) = c_0$. *A priori*, ce théorème dit seulement que c est une fonction continue, mais l'analyse complexe permet de montrer qu'elle admet un développement Gevrey d'ordre $\frac{1}{3}$. On utilise pour cela la formule suivante, conséquence de la formule des résidus

$$c(\eta) = \tfrac{1}{2\pi i} \int_{|x-c_0|=\rho/2} \frac{x\, \frac{\partial f}{\partial c}(x,\eta)}{f(x,\eta)}\, dx,$$

et la compatibilité des développements Gevrey avec les opérations élémentaires. À l'aide des résultats de composition de DAC avec des fonctions analytiques, on obtient que les compositions $(x,\eta) \mapsto y_g\big(x, c(\eta), \eta\big)$ et $(x,\eta) \mapsto y_d\big(x, c(\eta), \eta\big)$ admettent des DAC Gevrey d'ordre $\frac{1}{3}$. En tant que solutions de la même équation (5.3) avec la même condition initiale $y_g\big(0, c(\eta), \eta\big) = y_d\big(0, c(\eta), \eta\big)$, elles coïncident. Ceci démontre les énoncés du théorème, en particulier (5.6) et (5.7) dont les coefficients peuvent être obtenus en développant $a_{g|d,n}\big(x, \widehat{c}(\eta)\big) = a_{g|d,n}(x, c_0) + \ldots$ respectivement $b_{g|d,n}\big(X, \widehat{c}(\eta)\big)$ par la formule de Taylor dans (5.9) et (5.10).

Pour la démonstration de (5.13), on utilise que les DAC de $y_{g|d}$ sont uniformes par rapport à c dans un voisinage complexe de c_0. On peut donc obtenir les dérivées partielles par rapport à c en utilisant la formule de Cauchy ; par conséquent les dérivées partielles ont aussi des DAC et ces DAC sont obtenus en dérivant ceux de $y_{g|d}$ terme à terme. Une comparaison avec le développement extérieur montre que $a_{g1} = 0$ et $a_{d1} = 0$, donc $\frac{\partial y_g}{\partial c}(0,c_0,\eta) = \eta Z_g(0,c_0) + \mathcal{O}(\eta^2)$ avec $Z_g = \frac{\partial Y_g}{\partial c}$ et $\frac{\partial y_d}{\partial c}(0,c_0,\eta) = \eta Z_d(0,c_0) + \mathcal{O}(\eta^2)$ avec $Z_d = \frac{\partial Y_d^+}{\partial c}$. Or les fonctions $X \mapsto Z_{g|d}(X,c_0)$ sont des solutions de l'équation (5.5) dérivée par rapport à c prise en $c = c_0$ et pour $Y = Y_g(X,c_0) = Y_0(X)$, resp. $Y = Y_d^+(X,c_0) = Y_0(X)$. Autrement dit, ce sont des solutions de

$$Z' = \big(3Y_0(X)^2 - X^2\big)Z + 1.$$

Précisément, $Z_g(X,c_0)$ en est la solution tendant vers 0 quand $X \to -\infty$ et $Z_d(X,c_0)$ celle tendant vers 0 quand $X \to +\infty$. Notons $I_0(X)$ la primitive de $3Y_0(X)^2$ s'annulant en $X = 0$ et $J_0(X)$ celle de $3Y_0(X)^2 - 3X^2$ s'annulant en $X = 0$. Puisque $3Y_0(X)^2 = \mathcal{O}(X^{-1})$ quand

$X \to -\infty$ et $3Y_0(X)^2 - 3X^2 = \mathcal{O}(X^{-1})$ quand $X \to +\infty$, ces primitives ont une croissance au plus logarithmique en $-\infty$, resp. $+\infty$. La formule de variation de la constante donne alors

$$Z_g(0, c_0) = \int_{-\infty}^{0} \exp\big(X^3/3 - I_0(X)\big)\, dX > 0$$

et

$$Z_d(0, c_0) = -\int_{0}^{\infty} \exp\big(-2X^3/3 - J_0(X)\big)\, dX < 0.$$

Ceci démontre (5.13) et la preuve du théorème est complète. □

Bien entendu, la théorie des DAC n'est pas indispensable pour démontrer l'existence de valeurs à canards pour (5.3) ou des équations similaires, ni le fait que le côté droit de l'équation différentielle soit analytique. En effet, on pourrait utiliser le théorème du point fixe pour montrer l'existence de $y_g(x, c, \eta)$ et de $y_d(x, c, \eta)$ pour $x \in [a, -L\eta[$, resp. $x \in]L\eta, b]$, et ensuite les prolonger sur $[a, 0]$ resp. $[0, b]$ en utilisant l'équation différentielle intérieure. Une valeur à canard $c = c(\eta)$ possible est alors la solution de l'équation $y_g(0, c, \eta) = y_d(0, c, \eta)$ dont on assure l'existence par le théorème des fonctions implicites. Le fait crucial qu'une certaine dérivée partielle ne s'annule pas est alors démontré — comme ci-dessus — en se ramenant à une certaine équation différentielle linéaire pour le paramètre $\eta = 0$. De ce point de vue, le problème de l'existence de valeurs à canards traité dans cette partie est plus simple que, par exemple, celui de conditions nécessaires et suffisantes pour l'existence de canards d'équations différentielles *sans paramètres* traité dans la partie 5.1. Toutefois, la théorie des DAC apporte deux nouvelles propriétés des valeurs à canards non lisses : d'une part elle permet de donner une approximation uniforme des solutions canard, précisément notre DAC, et d'autre part les développements asymptotiques de la fonction $c = c(\eta)$ et des solutions canard sont Gevrey. Ceci peut servir par exemple à démontrer qu'une sommation "au plus petit terme" dans l'esprit de ce qui est fait dans [12] fournit une valeur à canard.

Equations non lisses .

Les canards du théorème 5.2 sont appelés non lisses car la courbe lente associée présente un angle en $x = 0$ mais, pour chaque ε, les solutions canards de l'équation analytique (5.3) sont bien sûr analytiques en x. Dans [16, 17], Emmanuel Isambert et Véronique Gautheron ont aussi chassé le canard sur des équations différentielles non lisses. Dans ce qui suit, nous voulons indiquer comment la théorie des DAC s'applique aussi à cette situation.

Il s'agit des équations du type *canard angulaire*

$$\varepsilon y' = \big(y - f(x)\big)\big(y - g(x)\big) + \varepsilon c, \tag{5.14}$$

où $f(x) = \alpha x + \mathcal{O}(x^2)$, $g(x) = \beta x + \mathcal{O}(x^2)$, $\alpha \neq \beta$ et où les restrictions de f et g à des intervalles $[a, 0]$ et $[0, b]$ sont des fonctions réelles analytiques, mais où f et g ne sont pas supposées $\mathcal{C}^\infty$ en 0. On suppose que la courbe lente $y = f(x)$ est attractive pour $x \in [a, 0[$ et répulsive pour $x \in]0, b]$, *i.e.* $x\big(f(x) - g(x)\big) \geq 0$ sur $[a, b]$; en particulier on a $\alpha > \beta$. L'exemple classique de [17] correspond à $f(x) = |x|^3/3$ et $g(x) = -x + |x|^3/3$. Dans un autre exemple de [16, 17], on a $f(x) = x(1 + |x|)$ et $g = -f$.

Traitons d'abord les solutions sur l'intervalle $[0, b]$. On écrit l'équation différentielle sur un voisinage de $[0, b]$ avec les prolongements analytiques f_+, g_+ des restrictions $f\,|_{[0,b]}$ et $g\,|_{[0,b]}$

$$\varepsilon y' = (y - f_+(x))\,(y - g_+(x)) + \varepsilon c. \tag{5.15}$$

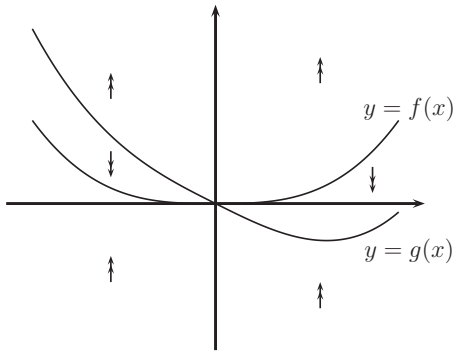


FIG. 5 – Les courbes lentes de (5.14) et l'orientation du champ.

Le changement de variable $y = f_+(x) + z$ mène alors à l'équation

$$\varepsilon z' = \big(f_+(x) - g_+(x)\big)z + z^2 + \varepsilon\big(c - f'_+(x)\big)$$

qui entre, après une homothétie sur la variable x, dans le cadre du théorème 4.3 avec $p = 2$, $r = 1$. Pour un voisinage $|c| < \rho$, ρ assez petit, on peut comme avant utiliser le corollaire 4.5 ; la répulsivité permet comme dans la preuve ci-dessus d'obtenir une solution ayant un DAC jusqu'à b. On obtient qu'il existe une solution holomorphe $y = y_d(x, c, \eta)$ de (5.15) pour $\eta \in]0, \eta_1]$, $x \in]-L\eta, b]$ et $|c| < \rho$ avec certains $\eta_1, L, \rho > 0$ assez petits et qu'elle admet un DAC Gevrey d'ordre $\frac{1}{2}$

$$y_d(x, c, \eta) \sim_{1/2} \widehat{y}_d(x, c, \eta) := f_+(x) + \sum_{n=1}^{\infty} \Big(a_{dn}(x, c) + b_{dn}\big(\tfrac{x}{\eta}, c\big)\Big)\eta^n \tag{5.16}$$

avec $a_{d1} = 0$ et $b_{d1}(X, c) = U_d(X, c)$, où U_d est la solution de l'équation intérieure réduite

$$U' = (\alpha - \beta)XU + U^2 + c \tag{5.17}$$

tendant vers 0 quand $X \to +\infty$. On vérifie, comme à la fin de la preuve du théorème 5.2, que $\frac{\partial U_d}{\partial c}(0, 0) < 0$.

Pour l'intervalle $[a, 0]$, l'équation (5.14) se simplifie aussi en une équation analytique

$$\varepsilon y' = \big(y - f_-(x)\big)\big(y - g_-(x)\big) + \varepsilon c \tag{5.18}$$

avec les prolongements analytiques f_- et g_- des restrictions $f\,|_{[a,0]}$ et $g\,|_{[a,0]}$ sur un voisinage de $[a, 0]$. On obtient ici l'existence d'une solution holomorphe $y = y_g(x, c, \eta)$ de (5.18) pour $\eta \in]0, \eta_1]$, $x \in [a, L\eta[$ et $|c| < \rho$ avec certains $\eta_1, L, \rho > 0$ petits et ayant un DAC Gevrey d'ordre $\frac{1}{2}$

$$y_g(x, c, \eta) \sim_{1/2} \widehat{y}_g(x, c, \eta) := f_-(x) + \sum_{n=1}^{\infty} \Big(a_{gn}(x) + b_{gn}\big(\tfrac{x}{\eta}\big)\Big)\eta^n \tag{5.19}$$

avec $a_{g1} = 0$ et $b_{g1}(X, c) = U_g(X, c)$, où U_g est la solution de la même équation intérieure réduite (5.17) tendant vers 0 quand $X \to -\infty$. On vérifie que $\frac{\partial U_g}{\partial c}(0, 0) > 0$.

En appliquant le théorème des fonctions implicites à l'équation $\frac{1}{\eta}y_g(0, c, \eta) = \frac{1}{\eta}y_d(0, c, \eta)$ au voisinage de $c = 0$ comme dans la preuve du théorème 5.2, on déduit de nouveau l'existence de valeurs à canards non lisses $c = c(\eta)$ ayant un développement asymptotique $\widehat{c}(\eta)$ Gevrey d'ordre

$\frac{1}{2}$ et l'existence de DAC Gevrey pour la solution canard définie par $y(x,\eta) = y_g\big(x, c(\eta), \eta\big)$ quand $x \leq 0$ et par $y(x,\eta) = y_d\big(x, c(\eta), \eta\big)$ quand $x \geq 0$ sur les intervalles $[a, 0]$ resp. $[0, b]$.

A priori, $\widehat{c}(\eta) = \sum_{n=1}^{\infty} c_n \eta^n$ est une série formelle en puissances de η ; la théorie des DAC nous apprend déjà que cette série est Gevrey. En complément à cette étude, nous redémontrons et améliorons ci-dessous un résultat de [17]. C'est l'occasion de présenter un DAC convergent, le seul non nul dans cet article. Ce DAC est en fait obtenu par un développement de Taylor d'une fonction spéciale.

Proposition 5.3 . *Dans le cas du canard angulaire classique, i.e. (5.14) avec* $f(x) = |x|^3/3$ *et* $g(x) = -x + |x|^3/3$, *la série formelle associée aux valeurs à canard*

$$c(\eta) = \sum_{m=1}^{\infty} c_{4m} \eta^{4m} \tag{5.20}$$

ne contient que des puissances multiples de 4 et elle converge.

REMARQUES. 1. Emmanuel Isambert [17] a démontré la forme (5.20) de la série formelle associée aux valeurs à canards. Nous montrons, de plus, sa convergence.

2. Pour cette équation modèle, on peut aussi exprimer les solutions de façon plus simple et on obtient des DAC très particuliers comme nous le verrons dans la preuve et à la fin de cette partie.

Preuve . Dans la suite, nous utilisons seulement ε, car la plupart des fonctions qui entrent en jeu sont des fonctions de cette variable, et non de $\eta = \sqrt{\varepsilon}$. Il s'agit donc de montrer que la valeur à canard $c = c(\varepsilon)$ est analytique dans un voisinage de 0 et paire.

Considérons d'abord l'équation sur $[0, +\infty[$:

$$\varepsilon y' = \left(y - \tfrac{x^3}{3}\right)\left(y + x - \tfrac{x^3}{3}\right) + \varepsilon c. \tag{5.21}$$

Le changement de variables $y = \frac{x^3}{3} + z$ la ramène à une équation simple

$$\varepsilon z' = xz + z^2 + \varepsilon(c - x^2),$$

qu'on peut encore simplifier. Notons $d = d(\varepsilon)$ la solution de $d + d^2 = \varepsilon$ analytique dans un voisinage de $\varepsilon = 0$ et telle que $d(0) = 0$. Alors le changement de variables $z = d(\varepsilon)x + u$ mène à l'équation

$$\varepsilon u' = \big(1 + 2d(\varepsilon)\big)xu + u^2 + \varepsilon\big(c - d(\varepsilon)\big).$$

L'homothétie $x = t/\gamma(\varepsilon)$, $u = \gamma(\varepsilon)v$ avec la fonction analytique $\gamma(\varepsilon)$ vérifiant $\gamma(0) = 1$ et $\gamma(\varepsilon)^2 = 1 + 2d(\varepsilon)$ aboutit à l'équation

$$\varepsilon \frac{dv}{dt} = tv + v^2 + \varepsilon C(\varepsilon), \quad \text{où} \quad C(\varepsilon) = \frac{c - d(\varepsilon)}{\gamma(\varepsilon)^2}.$$

Pour cette équation, on peut éliminer le petit paramètre ε, sauf dans l'argument de C. En effet, le changement de variables $t = \sqrt{\varepsilon}\, T$, $v = \sqrt{\varepsilon}\, V$ mène à

$$\frac{dV}{dT} = TV + V^2 + D \tag{5.22}$$

avec $D = C(\varepsilon)$, qui est l'équation intérieure réduite, hormis le fait que D n'est pas indépendant de ε ; c'est une fonction analytique $D = C(\varepsilon)$, avec $C(0) = c$. Pour D arbitraire dans $\mathbb{R}$, notons

maintenant $V_d(T, D)$ la solution de (5.22) tendant vers 0 quand T tend vers $+\infty$. Elle peut être exprimée par des solutions de l'équation de Weber, mais cela n'apporte rien ici. Constatons simplement que son prolongement complexe est une fonction entière de D, méromorphe de T, qui admet un développement asymptotique Gevrey d'ordre $\frac{1}{2}$ en puissances de T, uniforme par rapport à D dans tout compact de $\mathbb{C}$, de la forme $V_d(T, D) \sim_{1/2} \sum_{m=0}^{\infty} W_m(D)T^{-2m-1}$ avec $W_0(D) = -D$ et enfin que $V_d(T, 0) = 0$ et $\frac{\partial V_d}{\partial D}(0, 0) < 0$.

En résumé, si l'on tient compte de tous les changements de variables, pour $|c|$ petit (5.21) admet une unique solution y_d telle que $y_d(x, c, \varepsilon) - \frac{x^3}{3} - d(\varepsilon)x$ tend vers 0 quand x tend vers $+\infty$. Il s'agit de la fonction

$$y_d(x, c, \varepsilon) = \tfrac{x^3}{3} + d(\varepsilon)x + \sqrt{\varepsilon}\gamma(\varepsilon)V_d\left(\gamma(\varepsilon)\tfrac{x}{\sqrt{\varepsilon}}, \tfrac{c-d(\varepsilon)}{\gamma(\varepsilon)^2}\right).$$

Cette solution est analytique pour $\varepsilon \in]0, \varepsilon_1]$, $x \in [-L\sqrt{\varepsilon}, +\infty[$ et $|c| < \rho$ avec $\varepsilon_1, L, \rho > 0$ peut-être petits.

Sur $]-\infty, 0]$, l'équation du canard angulaire classique est

$$\varepsilon y' = \left(y + \tfrac{x^3}{3}\right)\left(y + x + \tfrac{x^3}{3}\right) + \varepsilon c. \tag{5.23}$$

De manière analogue, et en utilisant que (5.22) ne change pas sous la transformation $T \to -T$, $V \to -V$, on obtient que la fonction

$$y_g(x, c, \varepsilon) = -\tfrac{x^3}{3} + d(-\varepsilon)x - \sqrt{\varepsilon}\gamma(-\varepsilon)V_d\left(-\gamma(-\varepsilon)\tfrac{x}{\sqrt{\varepsilon}}, \tfrac{c-d(-\varepsilon)}{\gamma(-\varepsilon)^2}\right)$$

est son unique solution telle que $y_g(x, c, \varepsilon) + \frac{x^3}{3} - d(-\varepsilon)x$ tende vers 0 quand x tend vers $-\infty$. Elle est analytique pour $\varepsilon \in]0, \varepsilon_1]$, $x \in]-\infty, L\sqrt{\varepsilon}]$ et $|c| < \rho$ avec certains $\varepsilon_1, L, \rho > 0$.

L'équation déterminant la valeur à canard non lisse $c = c(\varepsilon)$ est donc

$$\gamma(\varepsilon)V_d\left(0, \tfrac{c-d(\varepsilon)}{\gamma(\varepsilon)^2}\right) = -\gamma(-\varepsilon)V_d\left(0, \tfrac{c-d(-\varepsilon)}{\gamma(-\varepsilon)^2}\right). \tag{5.24}$$

À cette équation, on peut appliquer le théorème des fonctions implicites pour les fonctions analytiques. On obtient d'abord que $c = c(\varepsilon)$ est une fonction analytique de ε dans un voisinage de 0 et ensuite qu'elle est paire en vertu de la symétrie de l'équation (5.24). Ceci démontre l'énoncé.

□

Remarque. Non seulement la valeur à canard $c = c(\varepsilon)$, mais aussi les solutions canards y_g et y_d, ont des développements convergents en puissances de ε. Pour $x \geq 0$, la solution canard est donnée par

$$y(x, \varepsilon) = y_d(x, c(\varepsilon), \varepsilon) = \tfrac{x^3}{3} + d(\varepsilon)x + \sqrt{\varepsilon}\gamma(\varepsilon)V_d\left(\gamma(\varepsilon)\tfrac{x}{\sqrt{\varepsilon}}, \tfrac{c(\varepsilon)-d(\varepsilon)}{\gamma(\varepsilon)^2}\right)$$

avec les fonctions y_d et V_d de la preuve. Or la fonction $(X, \varepsilon) \mapsto \gamma(\varepsilon)V_d\big(\gamma(\varepsilon)X,\ C(\varepsilon)\big)$ avec $C(\varepsilon) = \big(c(\varepsilon) - d(\varepsilon)\big)/\gamma(\varepsilon)^2$ est analytique bornée dans $V(-\alpha, \alpha, \infty, L) \times D(0, \varepsilon_1)$, si $\alpha, L, \varepsilon_1 > 0$ sont assez petits. Comme $c(0) = d(0) = 0$, on en déduit que la série de Taylor

$$\gamma(\varepsilon)V_d\big(\gamma(\varepsilon)X, C(\varepsilon)\big) = \sum_{n=1}^{\infty} W_n(X)\varepsilon^n$$

converge uniformément sur $V(-\alpha, \alpha, \infty, L)$ et sur tout compact de $D(0, \varepsilon_1)$. Par conséquent, le DAC de y est une série convergente en $\eta = \sqrt{\varepsilon}$. Ce DAC est constitué d'une partie lente ne

contenant que des puissances paires de η, avec pour terme principal $x^3/3$ et les autres termes des multiples scalaires de x, et d'une partie rapide ne contenant que des puissances impaires de η et dont les coefficients sont des fonctions $W_n(X)$ ayant des développements asymptotiques Gevrey d'ordre $\frac{1}{2}$ quand X tend vers $+\infty$. Comme pour la fonction V_d, ces derniers développements ne contiennent que des puissances impaires de X^{-1}. Les propriétés de la solution canard pour $x \leq 0$ sont analogues.

Références

[1] E. Benoît, J.-L. Callot, F. Diener, M. Diener, Chasse au canard, *Collect. Math.* 31, 1–3 (1981) 37–119.

[2] E. Benoît, A. El Hamidi, A. Fruchard, On combined asymptotic expansions in singular perturbations *Electron. J. Diff. Eqns.*, 2002, No. 51 (2002) 1–27.

[3] E. Benoît, A. Fruchard, R. Schäfke, G. Wallet, Solutions surstables des équations différentielles complexes lentes-rapides à point tournant, *Ann. Fac. Sci. Toulouse Math.* Vol. VII, n° 4 (1998) 627–658.

[4] M. Canalis-Durand, J.-P. Ramis, R. Schäfke, Y. Sibuya, Gevrey solutions of singularly perturbed differential equations, *J. Reine Angew. Math.* 518 (2000) 95–129.

[5] P. De Maesschalck, Ackerberg-O'Malley resonance in boundary value problems with a turning point of any order, *Commun. Pure Appl. Anal.* 6, 2 (2007) 311–333.

[6] P. De Maesschalck, F. Dumortier, Canard solutions at non-generic turning points, *Trans. Amer. Math. Soc.* 358 (2006), 2291–2334.

[7] F. Diener, M. Diener, *Nonstandard analysis in practice*, Universitext, Springer, Berlin, 1995

[8] M. Diener, Regularizing microscopes and rivers, *SIAM J. Math. Anal.* 25 (1994) 148–173.

[9] W. Eckhaus, *Asymptotic analysis of singular perturbations,* Studies in Mathematics and its Applications, 9. North-Holland, 1979.

[10] T. Forget, *Points tournants dégénérés,* Thèse de Doctorat, Université de La Rochelle, 2007.

[11] L. E. Fraenkel, On the method of matched asymptotic expansions, *Proc. Cambridge Philos. Soc.* 65 (1969) 209–284

[12] A. Fruchard, R. Schäfke, Exceptional complex solutions of the forced van der Pol equation, *Funkcialaj Ekvacioj* 42, 2 (1999) 201–223.

[13] A. Fruchard, R. Schäfke, Overstability and resonance *Ann. Inst. Fourier*, Grenoble, 53, 1 (2003) 227–264.

[14] A. Fruchard, R. Schäfke, Développements asymptotiques combinés et perturbation singulière, Manuscrit arXiv :1004.5254 (2010). Composite asymptotic expansions and turning points of singularly perturbed ordinary differential equations, soumis (2011).

[15] A. Fruchard, R. Schäfke, A survey of some results on overstability and bifurcation delay, *Discrete Cont. Dyn. Syst. S (DCDS-S)* 2, 4 (2009) 931–965.

[16] V. Gautheron, E. Isambert, Finitely differentiable ducks and finite expansions, in *Dynamic Bifurcations*, E. Benoît Ed., Lect. Notes Math. 1493 (1991) 40–56.

[17] E. Isambert, Nonsmooth Ducks and Regular Perturbations of Rivers, I et II, *J. Math. Anal. Appl.* 200 (1996) 14–33 et 289–306.

[18] C. Lobry, Dynamic Bifurcations, in *Dynamic Bifurcations*, E. Benoît Ed., Lect. Notes Math. 1493 (1991) 1–13.

[19] E. Matzinger, *Etude d'équations différentielles ordinaires singulièrement perturbées au voisinage d'un point tournant,* Thèse, Preprint IRMA 2000/53, Strasbourg, 2000.

[20] J.-P. Ramis, Les séries k-sommables et leurs applications, In *Complex Analysis, Microlocal Calcul and Relativistic Quantum Theory*, Lect. Notes Physics 126 (1980) 178–199.

[21] Y. Sibuya, *Linear differential equations in the complex domain, problems of analytic continuation*, AMS, Providence (RI), 1990.

[22] Y. Sibuya, A theorem concerning uniform simplification at a transition point and a problem of resonance, *SIAM J.Math. Anal.* 12, 5 (1981) 653–668.

[23] L. A. Skinner, Uniform solution of boundary layer problems exhibiting resonance, *SIAM J. Appl. Math.* 47, 2 (1987) 225–231.

[24] L. A. Skinner, Matched expansion solutions of the first-order turning point problem, *SIAM J. Math. Anal.* 25, 5 (1994) 1402–1411.

[25] L. A. Skinner, A class of singularly perturbed singular Volterra integral equations, *Asymptot. Anal.* 22, 2 (2000) 113–127.

[26] A. B. Vasil'eva, V. F. Butuzov, Asymptotic expansions of the solutions of singularly perturbed equations, *Izdat. "Nauka"* (en russe) Moscou, 1973.

[27] G. Wallet, Surstabilité pour une équation différentielle analytique en dimension un, *Ann. Inst. Fourier*, 40, 3 (1990) 557–595.

[28] W.Wasow, *Linear Turning Point Theory*, Springer, New York, 1985.

Adresses des auteurs :

Augustin Fruchard
Laboratoire de Mathématiques, Informatique et Applications, EA 3993
Faculté des Sciences et Techniques, Université de Haute Alsace
4, rue des Frères Lumière, 68093 Mulhouse cedex, France

Courriel : `Augustin.Fruchard@uha.fr`

Reinhard Schäfke
Institut de Recherche Mathématique Avancée, UMR 7501
U.F.R. de Mathématiques et Informatique
Université Louis Pasteur et C.N.R.S.
7, rue René Descartes, 67084 Strasbourg cedex, France

Courriel : `schaefke@math.u-strasbg.fr`

La Modélisation de la Persistance en Écologie

Claude Lobry, Tewfik Sari

1 Introduction

Depuis bientôt un siècle (les premiers articles de Lotka et de Volterra remontent aux années vingt du siècle dernier) on utilise des systèmes différentiels pour représenter la dynamique des populations de différentes espèces dans des écosystèmes. La physique nous a tellement habitués à la représentation de certains phénomènes par des équations différentielles que nous oublions, la plupart du temps, certains présupposés de ces modélisations.

Par exemple, considérons deux "points matériels" assujetis à se déplacer sur une droite et obéissant aux lois de la mécanique classique. Si x_1, m_1 et x_2, m_2 sont la position et la masse de chaque point le mouvement *est décrit* par le système d'équations différentielles :

$$\frac{dx_1}{dt} = v_1, \quad \frac{dv_1}{dt} = km_1m_2\frac{x_2 - x_1}{|x_2 - x_1|^3}, \qquad \frac{dx_2}{dt} = v_2, \quad \frac{dv_2}{dt} = km_1m_2\frac{x_1 - x_2}{|x_1 - x_2|^3}.$$

Ce qui est entendu par *est décrit* est que, au moins pour des vitesses faibles par rapport à celle de la lumière, dans la situation idéale où les masses sont effectivement ponctuelles et où il n'y a pas de frottement, la solution mathématique *exacte* de ce système prédit la valeur de la position et de la vitesse à un instant donné. De petits écarts avec la réalité seront attribués à "l'imperfection" de l'expérimentation. Dans cette théorie la position d'un point dans l'espace est définie par trois nombres réels ce qui veut dire que *l'espace physique est identifié à l'espace vectoriel réel* $\mathbb{R}^3$. Dans un modèle de ce type, si l'unité de distance est l'année lumière et si nous prévoyons que deux particules vont se croiser à une distance de 10^{-22}, c'est-à-dire une distance de l'ordre du micromètre, nous n'avons pas de problème parce qu'à cette distance la représentation de l'espace par des nombres réels garde tout son sens physique. *Mais il n'en serait pas de même avec* 10^{-50}.

Considérons maintenant l'équation de la diffusion dans un milieu à une dimension :

$$\frac{\partial U}{\partial t}(t,x) = k\frac{\partial^2 U}{\partial x^2}(t,x) \tag{1}$$

Dans une telle équation la quantité $U(t,x)$ représente la concentration en un point x, à l'instant t d'une certaine substance. Que se passe-t-il si nous adoptons le même point de vue que précédemment ? Dans ce type de modèle une "concentration" est une *quantité de molécules* par unité de volume, donc, à *priori* un nombre entier mais l'unité choisie est grande (de l'ordre de 10^{22}) ce qui conduit à traîter cette concentration comme un nombre réel[1]. Si notre équations prévoit une valeur de $U(t,x)$ égale à 10^{-22} nous aurons une concentration de quelques dizaines de molécules par unitié de volume, nombre si faible que la notion de concentration n'a plus de

[1]Ce point de vue est relativement récent. Avant que l'hypothèse atomique soit admise des équations telles que (1) étaient déja utilisées pour décrire des milieux jugés "continus".

sens. Donc dans cet exemple $U(t,x)$ perd son sens physique vers la valeur 10^{-20}, bien plus tôt que dans l'exemple précédent.

Ces deux exemples nous rappellent que les nombres réels prédits par le modèle mathématique (ici une équation différentielle) peuvent être si petits qu'ils n'ont pas de sens dans la théorie physique représentée. Dit autrement, le domaine de validité d'une équation différentielle décrivant une loi de la nature est toujours limité. Mais nous n'y pensons généralement pas et nous avons tendance à prendre ces équations différentielles de la physique pour des lois exactes - en tout cas extraordinairement exactes à l'échelle des problèmmes d'ingénierie courants. En d'autres termes le système différentiel est considéré comme "parfait" et l'on demande aux méthodes d'intégration numérique de fournir une solution "approchée" de la solution dont l'existence et l'unicité sont affirmées par un théorème bien connu.

Nous allons voir que dans le cas de la représentation de populations en interaction par des systèmes différentiels il est très souhaitable de se souvenir que les nombres réels que nous manipulons sont en fait des *nombres d'individus* comptés avec une unité assez grande - par exemple le milliard - et que donc une taille de population de 10^{-12} veut alors dire 10^{-3} individu, ce qui, évidemment, n'a pas de sens. La forme la plus générale d'un système décrivant l'évolution de populations en interaction est celle dite de Kolmogorov :

$$\frac{dx_i}{dt} = x_i f_i(x_1, \ldots, x_i, \ldots, x_n), \qquad x_i \geq 0, \quad i = 1 \ldots N, \tag{2}$$

où les x_i sont les tailles des populations, les f_i des taux de croissance. Cette forme a été retenue parce qu'elle laisse l'orthant positif invariant (les tailles des populations doivent rester positives) et les faces $x_i = 0$ également (pas de génération spontannée). En dynamique des populations on s'intéresse, entre autres, à la question de la "persistance". On veut savoir si telle ou telle population sera éternellement présente dans l'écosystème. Il va de soi que dire que la population n'a pas disparu tant que $x_i(t)$ est strictement positif n'est pas satisfaisant dans le cas de notre modèle (2) puisque, pour toute condition initiale telle que $x_i(0) > 0$ on a pour, tout $t \geq 0$, $x_i(t) > 0$. La quatité $x_i(t)$ est, certes, positive, mais elle peut tendre vers 0. C'est pourquoi on a introduit la définition mathématique suivante [10, 23] de la persistance dans laquelle on interdit à $x_i(t)$ de tendre vers 0. :

Définition 1.1 (**Persistance**.) *Le système (2) est persistant si pour toute condition initiale telle* $x_i(0) > 0$ *la trajectoire est bornée et si on a :*

$$\liminf x_i(t) > 0,$$

si de plus il existe une constante $a > 0$*, indépendante de la condition initiale, telle que :*

$$\liminf x_i(t) \geq a$$

on dit que le système est "fortement persistant".

C'est la pertinence de cette notion purement qualitative que nous étudions et critiquons à travers un exemple. Comme notre approche se veut à la fois qualitative et quantitative nous ferons un très large appel à des simulations sur ordinateur que nous utiliserons également pour illustrer notre propos. Notre objectif est de présenter un article d'exposition où toutes les définitions sont explicitées mais les aspects techniques et les démonstrations sont évités. Pour ces derniers nous renvoyons aux articles pertinents.

2 Persistance dans le modèle “ressource-consomateur”

2.1 Mise en place du modèle classique

Le système d'équations ci-dessous représente une relation “ressource-consommateur”[2].

$$\begin{cases} \dfrac{ds}{dt} = f(s) - \mu(s)x \\ \dfrac{dx}{dt} = \varepsilon(\mu(s) - m)x \end{cases} \tag{3}$$

Les quantités s et x représentent la concentration de la ressource et des consommateurs. Dans le domaine de l'écologie marine s pourrait être une concentration de sardines, mesurée, par exemple, en milliard par km^3 et x une quantité de thons mesurée dans la même unité. En écologie microbienne s pourrait être une concentration de bactéries, mesurée, par exemple, en milliard par litre et x une concentration d'un consommateur de ces bactéries. Le unités peuvent varier beaucoup suivant les situations et nous retiendrons simplement que des chiffres de 10^7 à 10^{11} individus par unité de volume sont possibles. La fonction f, que nous supposons ici nulle en 0, positive sur $[0, K]$ négative ensuite[3], est souvent de type logistique et la fonction μ, nulle en 0, est souvent une fonction de Monod. Le paramètre ε est important ; le terme ε évoque une quantité petite et exprime le fait que, lorsque l'unité choisie pour s et pour x reflète la masse, la transformation de la “masse de ressource” en “masse de consommateur” se traduit par une perte ; par exemple il faut $50kg$ d'herbe sèche pour faire $1kg$ de vache.

Comme nous allons beaucoup travailler sur des illustrations obtenues par simulation nous allons préciser ce modèle général de Kolmogorov en le modèle particulier suivant :

$$\begin{cases} \dfrac{ds}{dt} = rs\left(1 - \dfrac{s}{K}\right) - \dfrac{\mu_{max}s}{e + bs}x \\ \dfrac{dx}{dt} = \varepsilon\left(\dfrac{\mu_{max}s}{e + bs} - m\right)x \end{cases}$$

où r, K, μ_{max}, e, b et m sont des paramètres positifs. Nous pouvons faire trois changements d'unité (sur s, sur x et le temps) qui permettront de diminuer de trois le nombre de paramètres indépendants. Ainsi, après avoir renommé les paramètres, nous pouvons travailler sur le modèle :

$$\begin{cases} \dfrac{ds}{dt} = s\left(1 - \dfrac{s}{K}\right) - \dfrac{sx}{e + s} \\ \dfrac{dx}{dt} = \varepsilon\left(\dfrac{s}{e + s} - m\right)x \end{cases}$$

Par changement d'unité de temps nous pouvons diviser les deux équations par ε. Remarquons aussi que s peut être mis en facteur dans la première équation et donc le système devient

[2] On dit aussi relation “proie-prédateur” mais on préfère maintenant relation “ressource-consomateur” ou encore relation “mangeur-mangé”. En effet l'expression “proie-prédateur” évoque mal la relation qui existe entre la vache et l'herbe qu'elle broute !

[3] Une ressource est dite “biotique” lorsqu'elle est constituée d'organismes vivant capables de se reproduire et “abiotique” lorsque c'est une substance chimique inerte. Dans le second cas, pour que le système soit persistant il faut en permanence apporter de la ressource au milieu, alors que dans le premier la ressource se renouvelle en permanence (en consommant des ressources qui ne sont pas prises en compte dans le modèle). Cela conduit à des fonctions “f” différentes, croissantes puis décroissantes dans le cas biotique, monotones décroissantes dans le cas “abiotique” ; dans le cadre de cet article nous considérons des ressources biotiques mais nous n'insistons pas sur ce point car ce n'est pas notre objet.

finalement :

$$\left\{\begin{array}{l} \dfrac{ds}{dt} = \dfrac{s}{\varepsilon}\left(1 - \dfrac{s}{K} - \dfrac{x}{e+s}\right) \\ \dfrac{dx}{dt} = \left(\dfrac{s}{e+s} - m\right)x \end{array}\right. \tag{4}$$

qui est la forme sur laquelle nous allons travailler.

2.2 Persistance théorique

Tout ce qui est dit dans ce paragraphe est classique et se trouve par exemple dans [26, 37] Nous commençons par établir une proposition de persistance.

Proposition 2.1 (Persistance) *Si $f'(0) > 0$ le système (3) est persistant.*

Preuve : Nous commençons par remarquer que les solutions sont bornées. En effet nous avons :

$$\frac{ds}{dt} \leq f(s)$$

et donc, compte tenu des propriétés de f, $\limsup s(t) \leq K$ et donc $s(t)$ est bornée. Considérons la somme $\varepsilon s + x$. Nous avons :

$$\frac{d(\varepsilon s + x)}{dt} = f(s) - mx = f(s) + \varepsilon m s - m(\varepsilon s + x).$$

Puisque s est bornée :

$$\frac{d(\varepsilon s + x)}{dt} \leq A - m(\varepsilon s + x)$$

ce qui montre que $\varepsilon s + x$ est bornée et donc, puisque s et x sont positifs, que x est bornée. Soit maintenant $\gamma(t) = (s(t), x(t))$ une trajectoire telle que $s(0) > 0$ et $x(0) > 0$. Supposons que :

$$\liminf_{t \to +\infty} s(t) = 0$$

Ceci veut dire qu'il existe une suite t_n tendant vers l'infini telle que $s(t_n)$ tende vers 0 et, comme x est bornée, on peut extraire une sous suite $t_{n'}$ de t_n telle que :

$$x(t_{n'}) \ \to x^*$$

et donc le point $(0, x^*)$ est un point ω-limite[4] de de la trajectoire γ. Si x^* est strictement positif on en déduit que tout le demi axe vertical est dans l'ensemble ω-limite de γ ce qui est impossible puisque γ est bornée. Reste le cas où $x^* = 0$. Comme le point $(0, 0)$ est un col (c'est içi que l'hypothèse $f'(0) > 0$ est essentielle) dont le demi axe vertical est une partie de la variété stable, on en déduit encore que ce dernier est dans l'ensemble ω-limite de γ, ce qui est impossible. Le cas où l'on aurait :

$$\liminf_{t \to +\infty} x(t) = 0$$

se traite de manière semblable (en plus du fait que $(0, 0)$ est un col on utilisera aussi le fait que $(K, 0)$ est également un col).

Avec quelques hypothèses supplémentaires nous pouvons montrer la persistance forte.

[4] Soit $t \longrightarrow \gamma(t)$ une trajectoire d'une équation différentielle ; on dit que x^* est un point ω-limite s'il existe une suite t_n tendant vers l'infini telle que $\gamma(t_n)$ tende vers x^*.

Proposition 2.2 (Persistance forte) *Lorsque l'isocline $\mu(s) = m$ est à gauche du maximum de l'isocline $s' = 0$, le système possède un cycle limite unique qui est de plus globalement asymptotiquement stable dans le quadrant strictement posistif.*

Preuve : Les équilibres $(0, 0)$ et $(K, 0)$ situés sur les axes sont instables. Comme l'isocline $\mu(s) = m$ est à gauche du maximum de l'isocline $s' = 0$, l'équilibre persistant est instable aussi. Par conséquent il ne peut pas faire partie de l'ensemble ω-limite d'une solution. On a vu que les solutions sont positivement bornées. Par conséquent l'ensemble ω-limite d'une solution est non vide. D'après le théorème de Poincaré-Bendixon cet ensemble limite est un cycle, car il ne peut contenir aucun des équilibres du système. L'unicité du cycle se démontre en montrant sa stabilité. En effet les éventuels cycles doivent entourer l'équilibre persistant. Or deux cycles adjacents ne peuvent pas être stables tous les deux sur leur faces opposées. La stabilité se démontre en montrant que l'indice de stabilité du cycle, c'est à dire l'intégrale de la divergence du champ le long du cycle, est strictement négatif. Voir [28] pour les détails.

2.3 Persistance réelle : simulations

Nous considérons le modèle (4) pour les valeurs suivantes des paramètres :

$$\varepsilon = 0.05, \qquad K = 2, \qquad e = 0.4, \qquad m = 0.6,$$

c'est à dire le système :

$$\begin{cases} \dfrac{ds}{dt} = \dfrac{s}{0.05}\left[\left(1 - \dfrac{s}{2}\right) - \dfrac{x}{0.4 + s}\right] \\ \dfrac{dx}{dt} = \left(\dfrac{s}{0.4 + s} - 0.6\right) x \end{cases} \tag{5}$$

et nous effectuons une simulation sur ordinateur ce qui donne la Figure 1.

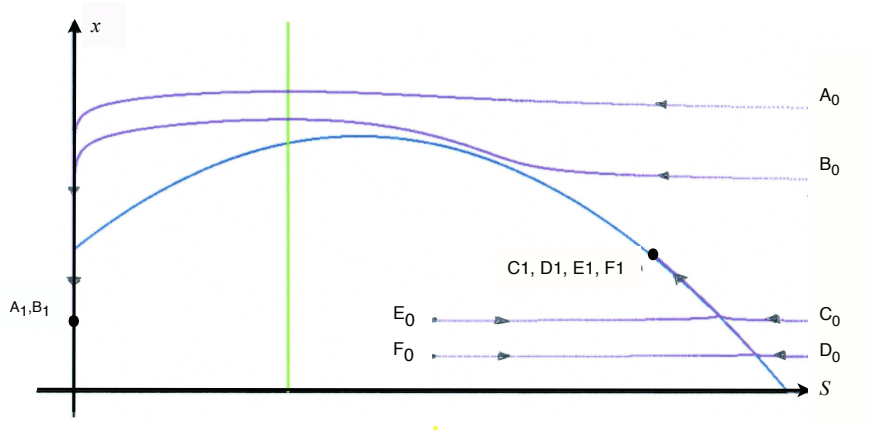


FIG. 1 – Six trajectoires du système (5)

En bleu on a tracé la partie de l'isocline "de la ressource" $(s' = 0)$ constituée par le graphe de :

$$s \to (1 - s/K)(e + s), \qquad s > 0$$

(l'autre partie de l'isocline est constituée du demi axe vertical positif) et, en vert, la partie de l'isocline "du consomateur" ($x' = 0$) constituée de la verticale :

$$s = \frac{em}{1-m}$$

(l'autre partie de l'isocline est le demi-axe horizontal positif). Nous avons simulé six trajectoires, issues des points A_0, B_0, C_0, D_0, E_0, F_0. Les quatre trajectoires issues de C_0, D_0, E_0, F_0 se rendent rapidement dans le voisinage de l'isocline "de la ressource", elles la "longent" (sur notre figure, à la précision du pixel, elles sont confondues avec) jusqu'aux points (confondus sur la figure) C_1, D_1, E_1, F_1 où la simulation est arrêtée. On voit que les deux trajectoires qui sont issues des points A_0 et B_0 se dirigent vers le demi axe vertical puis sont confondues avec lui (à l'épaisseur du pixel près), ce qui veut dire que la ressource s est petite ; ces deux trajectoires "longent" l'axe vertical jusqu'aux points A_1, B_1 où la simulation est arrêtée.

Théoriquement il y a persistance, donc la ressource devrait finir par reprendre des valeurs plus grandes. Intégrons pendant une durée plus longue ce qui donne la Figure 2.

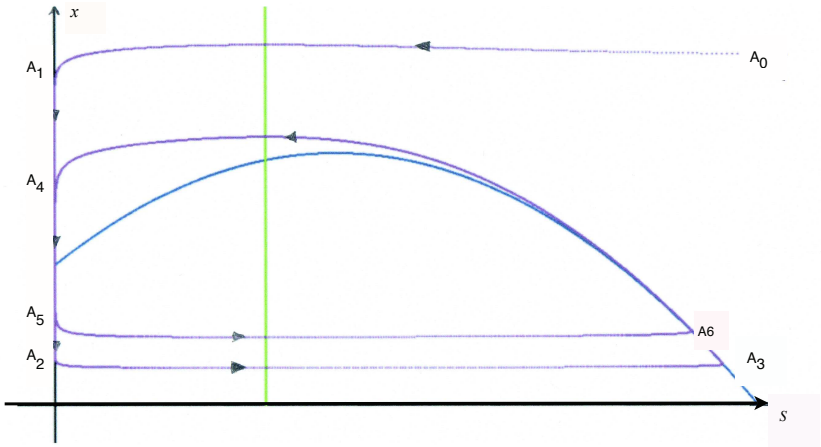


FIG. 2 – Le cycle limite du système (5)

Nous avons pris pour condition initiale le point A_0 de coordonnées $(2.1, 1)$ et nous observons une trajectoire qui se dirige rapidement vers l'axe vertical, le "longe" à partir du point A_1 en descendant, quitte l'axe vers la droite au point A_2 selon la trajectoire quasi horizontale la plus basse, rejoint l'isocline "de la ressource" au point A_3, la longe, puis la quitte vers son sommet pour rejoindre à nouveau l'axe vertical au point A_4, le longe et le quitte au point A_5 selon la trajectoire quasi horizontale au dessus de la précédente et rejoint à nouveau "l'isocline de la ressource" au point A_6. Compte tenu de l'imprécision graphique, à partir de maintenant, on ne distingue plus la trajectoire du cycle limite que nous venons de mettre en évidence. Donc, dorénavant, la ressource $s(t)$ va osciller entre deux valeurs : un maximum proche de la valeur 2 et un minimum qui est une quantité petite. Il y a bien "persistance de s" et on fait le même constat sur x.

Si maintenant nous demandons à l'ordinateur d'afficher la valeur de $s(t)$, nous constatons que pendant la première descente de la trajectoire le long de l'axe vertical la valeur de $s(t)$ **diminue jusqu'à** $2.7\ \ 10^{-10}$, donc, si l'unité correspond à un milliard d'individus, **il y a longtemps que la population a disparu !** Toutefois, à ce stade, nous ne pouvons avoir qu'une confiance limitée dans la valeur affichée par l'ordinateur qui n'est qu'une valeur "proche" de la valeur de la solution exacte de (5). Une petite étude asymptotique permet de conforter ce constat.

2.4 Confortation théorique des simulations

Lorsque nous longeons l'axe vertical la variable $s(t)$ est pratiquement constante, sa dérivée est presque nulle, soit :

$$\frac{s}{\varepsilon}\left(1-\frac{s}{K}-\frac{x}{e+s}\right)\approx 0$$

ce qui n'est possible que si $s(t)$ est de l'ordre de ε de manière à contrer le terme en $\frac{1}{\varepsilon}$. Donc, le long de l'axe vertical, nous négligeons s devant l'unité ce qui donne :

$$\begin{cases} \dfrac{ds}{dt}=\dfrac{s}{\varepsilon}\left(1-\dfrac{x}{e}+O(\varepsilon)\right) \\ \dfrac{dx}{dt}=-mx+O(\varepsilon) \end{cases}$$

(*Par $O(\varepsilon)$ nous entendons un terme qui est de l'ordre de grandeur de ε*). Nous avons :

$$x(t)=x_0\exp(-mt)+O(\varepsilon)$$

Comme $s(t)$ décroit tant que $x(t)$ est au dessus de "l'isocline de la ressource" le minimum de s est atteint au moment T où $x(t)$ croise cette isocline, donc T est approximativement le moment où $x(T)=e+O(\varepsilon)$. Ainsi

$$T=-\frac{1}{m}\ln\left(\frac{e}{x_0}\right)+O(\varepsilon).$$

Donc, tant que t est plus petit que T, nous avons :

$$\frac{ds}{dt}=\frac{s}{\varepsilon}\left(1-\frac{s}{K}-\frac{x}{e+s}\right)=\frac{s}{\varepsilon}\left(1-\frac{x_0\exp(-mt)}{e}+0(\varepsilon)\right)$$

ce qui donne, après intégration :

$$s(T)=s_0\exp\left[\frac{1}{\varepsilon m}\left(1-\frac{x_0}{e}+\ln\left(\frac{x_0}{e}\right)\right)+O(1)\right]$$

soit encore :

$$s(T)=O(1)s_0\exp\left[\frac{1}{\varepsilon m}\left(1-\frac{x_0}{e}+\ln\left(\frac{x_0}{e}\right)\right)\right]. \tag{6}$$

Comme la quantité $1-\frac{x_0}{e}+\ln\left(\frac{x_0}{e}\right)$ est toujours négative nous pouvons prévoir pour $s(T)$ des valeurs très petites. Nous avons comparé les valeurs prédites par la formule (6), aux simulations sur ordinateur. Les résultats sont consignés dans les tableaux suivants où nous faisons varier successivement m, e et x_0 et où on a pris $\varepsilon=0.1$ et $s_0=\varepsilon$ [5] [6].

[5] Notons qu'un ordinateur qui procède en "virgule flottante" peut, au voisinage de 0, manipuler des nombres aussi petits que 10^{-250} sans les assimiler à zéro.

[6] Ce bon accord entre l'approximation asymptotique et les simulations montre que ce que nous observons sur l'ordinateur n'est pas un artefact numérique. Des modifications du pas d'intégration entraînent de petites modifications mais l'ordre de grandeur reste toujours le même.

m	*Simulation*	*Estimation*
0.8	$7.38\ 10^{-5}$	$O(1)\ 10^{-5}$
0.7	$2.47\ 10^{-5}$	$O(1)\ 10^{-5}$
0.6	$5.76\ 10^{-6}$	$O(1)\ 10^{-6}$
0.5	$7.50\ 10^{-7}$	$O(1)\ 10^{-7}$
0.4	$3.53\ 10^{-8}$	$O(1)\ 10^{-8}$
0.3	$2.21\ 10^{-10}$	$O(1)\ 10^{-10}$
0.2	$1.89\ 10^{-13}$	$O(1)\ 10^{-14}$

Tableau 1 : $e = 0.4, \quad x_0 = 1$

e	*Simulation*	*Estimation*
0.6	$4.73\ 10^{-3}$	$O(1)\ 10^{-3}$
0.5	$2.08\ 10^{-4}$	$O(1)\ 10^{-4}$
0.4	$7.50\ 10^{-7}$	$O(1)\ 10^{-7}$
0.35	$8.14\ 10^{-9}$	$O(1)\ 10^{-9}$
0.3	$1.22\ 10^{-11}$	$O(1)\ 10^{-11}$
0.25	$6.91\ 10^{-16}$	$O(1)\ 10^{-16}$
0.2	$1.26\ 10^{-22}$	$O(1)\ 10^{-22}$

Tableau 2 : $m = 0.5, \quad x_0 = 1$

x_0	*Simulation*	*Estimation*
2	$1.52\ 10^{-22}$	$O(1)\ 10^{-22}$
1	$7.50\ 10^{-7}$	$O(1)\ 10^{-7}$
0.9	$1.40\ 10^{-5}$	$O(1)\ 10^{-5}$
0.8	$2.09\ 10^{-4}$	$O(1)\ 10^{-4}$
0.7	$2.43\ 10^{-3}$	$O(1)\ 10^{-3}$
0.6	$2.25\ 10^{-2}$	$O(1)\ 10^{-2}$
0.5	$5.84\ 10^{-2}$	$O(1)\ 10^{-2}$

Tableau 3 : $e = 0.4, \quad m = 0.5$

3 Le modèle classique revisité

3.1 Persistance forcée et extinction automatique

Les considérations qui ont conduit à l'écriture du modèle (3) que nous reproduisons ci-dessous :

$$\begin{cases} \frac{ds}{dt} &= f(s) - \mu(s)x \\ \frac{dx}{dt} &= \varepsilon\left(\mu(s) - m\right)x \end{cases}$$

sont les suivantes : On se donne un petit intervalle de temps dt et on écrit des équations de "bilan" :

$$\begin{cases} s(t+dt) = s(t) + dt\left[f\left(s(t)\right) - \mu\left(s(t)\right)x(t)\right] \\ x(t+dt) = x(t) + dt\left[\varepsilon(\mu(s(t)) - m)x(t)\right] \end{cases}$$

Dans ces équations le terme le plus simple, le terme de "disparition" :

$$dt[-\varepsilon m x(t)]$$

exprime que la quantité de biomasse $x(t)$ qui disparaît pendant un intervalle de temps dt est proportionnelle à dt et à la quantité de biomasse $x(t)$. Cette expression n'a de sens que tant que $x(t)$ est assez grand pour qu'il soit possible de parler de "concentration". Le modèle reste totalement muet sur ce qui se passe lorsque $x(t)$ devient plus petit qu'un certain seuil. Si, par exemple, l'unité correspond à 10^9 individus et que nous acceptons de parler de concentration jusqu'à 10^3 individus, que se passe-t-il lorsque $x(t) = 10^{-6}$?

Il faut alors recourir à d'autres types de modèles comme, par exemple, des modèles probabilistes où la variable exprimant la quantité de biomasse est la variable aléatoire :

$$N(t) = \text{“Nombre d'individus à l'instant” } t$$

et où des hypothèses sont faites sur les probabilités de “naissance” ou de “mort” d'un individu pendant un petit intervalle de temps dt[7]. Toutefois de tels modèles ne peuvent pas être couplés facilement avec le modèle à variables continues et, à notre connaissance, ne le sont pas.

Il existe cependant deux situations particulières (mais extrêmes) où une modélisation mathématique simple est possible : La *“persistance forcée”* et *“l'extinction automatique”*. Par *persistance forcée* d'une variable, notée $u(t)$, nous entendons que si $u(t)$ est inférieur à un certain seuil α alors $u(t)$ est croissant[8] et, à *contrario*, par *extinction automatique* nous entendons le cas où la variable $u(t)$ est décroissante dès qu'elle est inférieure à un seuil α[9]. Bien entendu dans le premier cas la question de la persistance de l'espèce représentée par $u(t)$ ne se pose plus, elle est “forcée” et contenue dans les hypothèses et dans le second cas la question de la persistance se pose d'une autre manière : il n'y a pas persistance et il faut déterminer quel est le *bassin de persistance*, c'est-à dire l'ensemble des conditions initiales pour lesquelles les trajectoires restent toujours supérieures au seuil d'extinction. C'est ce que nous allons examiner de plus près après avoir précisé un modèle mathématique reflétant ces considérations.

3.2 Modèles avec seuils

Nous considérons à nouveau le modèle très général de type Kolmogorov (2) évoqué dans l'introduction :

$$\frac{dx_i}{dt} = x_i f_i(x_1, \ldots, x_i, \ldots, x_n), \qquad x_i \geq 0, \quad i = 1 \ldots N.$$

Pour chaque espèce i nous nous donnons un seuil α_i qui est un nombre positif : C'est la limite inférieure en dessous de laquelle nous estimons que la quantité x_i n'a plus de sens pour représenter la population. Pour chaque i nous introduisons la fonction **discontinue** :

$$f_i(x_1, \ldots, x_i, \ldots, x_n, \alpha_i, \rho_i) = f_i(x, \alpha_i, \rho_i)$$

définie par :

$$f_i(x, \alpha_i, \rho_i) = \begin{cases} f_i(x_1, \ldots, x_n) & \text{si} \quad x_i > \alpha_i \\ \rho_i & \text{si} \quad x_i \leq \alpha_i \end{cases}$$

où ρ_i est un nombre réel non nul, positif ou négatif. Nous considérons maintenant le système différentiel :

$$\begin{cases} \dfrac{dx_i}{dt} = x_i f_i(x, \alpha_i, \rho_i), \qquad x_i \geq 0, \quad i = 1 \ldots N. \end{cases}$$

Ce type de système différentiel n'est pas classique car ses seconds membres ne sont pas des *fonctions continues* des variables x_i qui décrivent l'état du système. On peut utiliser pour ces équations différentielles une notion de solution appelée solution de Filippov ([21, 22]) qu'il n'est pas possible de définir en toute généralité dans le cadre de cet article mais qui dans le cas particulier qui nous intéresse peut être décrite simplement[10]. Plaçons nous en un point où aucun des x_i n'est égal à α_i ; il n'y a pas de problème, on intègre le système différentiel **jusqu'à ce**

[7] On parle de “processus de naissance et de mort” dont on peut trouver une introduction dans [20]

[8] Cela peut être le cas en écologie microbienne lorsque le milieu dans lequel vit l'écosystème est en permanence “contaminé” par une source extérieure en individus constituant la population représentée par $u(t)$.

[9] Ce peut être le cas pour certaines espèces comme les baleines pour lesquelles la reproduction devient insuffisante parce que la probabilité de rencontre de deux individus est trop faible.

[10] Notons que les solutions de Filippov sont également utilisées dans d'autres domaines de modélisation dans les sciences du vivant [2, 14, 25].

qu'à un instant t un des x_i prenne la valeur α_i. Nous supposons, pour simplifier l'exposé, qu'en cet instant un seul des x_i est égal à α_i. L'espace est donc partagé en deux régions par l'hyperplan $\{x\ ;\ \ x_i = \alpha_i\}$ et de chaque côté de cet hyperplan nous avons un système différentiel différent. Trois cas peuvent se produire comme indiqué sur les schémas de la Figure 3. Sur le

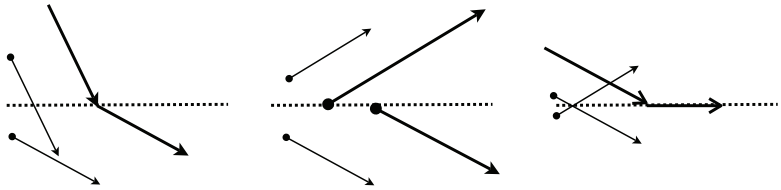


FIG. 3 – Champ discontinu "traversant", "divergeant", "convergeant"

schéma de gauche les deux champs de vecteurs pointent du même côté (du haut vers le bas) au dessus et au dessous de l'hyperplan représenté par la droite horizontale ; dans ce cas la solution de Filippov traverse simplement l'hyperplan, on dira que *le champ est traversant* . Sur le schéma de droite (correspondant à $f_i(x, \alpha_i, \rho_i) < 0$) au dessus de l'hyperplan le champ pointe vers le bas et *vice versa* (ce qui correspond à $\rho_i > 0$) ; la solution de Filippov est contrainte à rester dans l'hyperplan où elle suit la dynamique définie par :

$$\left\{ \begin{array}{l} \dfrac{dx_j}{dt} = x_j f_j(x_1, \ldots, x_n), \qquad j \neq i, \end{array} \right.$$

tant que $f_i(x, \alpha_i, \rho_i)$ reste négatif ; dans ce cas on dira que la *discontinuité est stable*. Enfin, sur le schéma du milieu (qui correspond à $f_i(x, \alpha_i, \rho_i) > 0$ et $\rho_i < 0$), aucune solution ne peut approcher l'hyperplan sauf si la condition initiale est dans l'hyperplan auquel cas on part indifféremment vers le haut ou vers le bas ; on dira que la *discontinuité est instable*. Lorsque nous nous trouvons le long de l'intersection de plusieurs hyperplans la situation est un peu plus complexe à décrire mais nous ne le ferons pas car dans la discussion qui suit nous n'en aurons pas besoin (dans tout l'article nous ne considérons que des modèles en dimension deux).

Remarque. Notons que dans l'ensemble $\{x : x_i \leq \alpha_i\}$ la dynamique ne dépend que du choix du signe des ρ_i, pas de leur module. Comme seul l'ensemble $\{x : x_i \leq \alpha_i\}$ est pertinent en terme de modélisation nous pourrions aussi bien décider que $\rho_i = \pm 1$.

Sous des hypothèses assez générales, qui sont satisfaites ici, il peut être démontré que les solutions de Filippov sont correctement approchées par les solutions numériques définies par le simple schéma d'Euler, ce que l'on peut comprendre facilement en observant le schéma de la Figure (4). Pour plus de détails sur les solutions de Philippov et les schémas d'Euler associés on pourra consulter [32]. Intéressons nous maintenant au sens de ce nouveau modèle.

3.3 L'hypothèse "persistance forcée"

Nous supposons que pour l'espèce i le paramètre ρ_i est **strictement positif**. Nous pouvons alors énoncer le :

Théorème 3.1 *Supposons $\rho_i > 0$. Dès que t est assez grand on a $x_i(t) \geq \alpha_i$*

Preuve : Supposons que $x_i(t_0) \geq \alpha_i$. Si pour tout t on a $x_i(t) \geq \alpha_i$ le résultat est démontré. Sinon il existe t_1 tel que $x_i(t_1) < \alpha_i$. Soit t_1^- le dernier instant avant t_1 pour lequel x_i est supérieur ou égal à α_i.

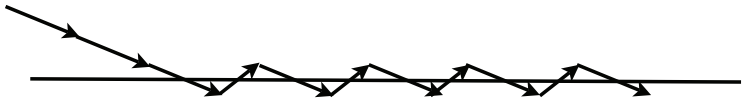


FIG. 4 – Le schéma d'Euler et les solutions de Filippov dans le cas d'une discontinuité attractive

Sur l'intervalle $[t_1^-, t_1]$, par définition de t_1^-, on a $x_i(t) < \alpha_i$ et donc :

$$\frac{dx_i}{dt} = \rho_i x_i > 0$$

ce qui contredit le fait que $x_i(t_1)$ est plus petit que $x_i(t_1^-) = \alpha_i$. Si $x_i(t_0) < \alpha_i$ on a $x_i(t) = x_i(t_0)e^{t-t_0}\rho_i$ et donc $x_i(t) = \alpha_i$ pour $t = \frac{1}{\rho_i} \ln\left(\frac{\alpha_i}{x_i(t_0)}\right)$ ce qui nous ramène au cas précédent et achève la preuve.

3.4 L'hypothèse "extinction automatique"

Nous supposons que pour l'espèce i le paramètre ρ_i est **strictement négatif**. Nous pouvons énoncer le :

Théorème 3.2 *Supposons $\rho_i < 0$. Si $x_i(t_0) < \alpha_i$ alors $x_i(t)$ tend exponentiellement vers 0 avec un taux de décroissance égal à ρ_i.*

Preuve : C'est évident.

Dans ce cas nous interprétons le modèle en disant que en dessous du seuil α_i il y a *extinction automatique* de la population.

Pour conclure cette section nous dirons que les α_i définissent pour chaque population un "seuil" en dessous duquel le comportement de la population n'est plus régi par le modèle général mais obéit à une sorte de loi du **tout ou rien** : hypothèse *persistance forcée* et la population se maintient au seuil α_i tant que des conditions favorable ne la font pas croître à nouveau, hypothèse *extinction automatique* et ce seuil une fois atteint la population disparaît inéluctablement. Notons enfin qu'un modèle peut très bien comporter, pour des raisons de simplicité, des seuils α_i egaux à 0 et le cas où tous les α_i sont nuls est le modèle général (2) de Kolmogorov.

4 Persitance forcée

4.1 Le portrait de phase du modèle classique

Nous reprenons le modèle (4) avec les valeurs numériques pour lesquelles nous avons fait les simulations du paragraphe 2.3, soit les équations (5) que nous reproduisons ici :

$$\begin{cases} \dfrac{ds}{dt} = \dfrac{s}{0.05}\left(1 - \dfrac{s}{2} - \dfrac{x}{0.4+s}\right) \\ \dfrac{dx}{dt} = \left(\dfrac{s}{0.4+s} - 0.6\right)x \end{cases} \tag{7}$$

Nous rappelons que ce système différentiel classique possède un cycle limite globalement asymptotiquement stable dans l'orthant positif, ce qui nous a fait conclure à la persistance forte au sens classique. Sur la Figure 5 nous avons simulé un certain nombre de trajectoires qui convergent toutes vers le cycle limite. Les trajectoires issues de A_0, B_0, C_0, D_0 "pénètrent dans l'axe vertical" aux points A_1, B_1, C_1, D_1 et "ressortent" en ordre inverse aux points D_2, C_2, B_2, A_2. Nous avons vu que pour certaines trajectoires (les plus hautes) les valeurs de $s(t)$ peuvent être très petites.

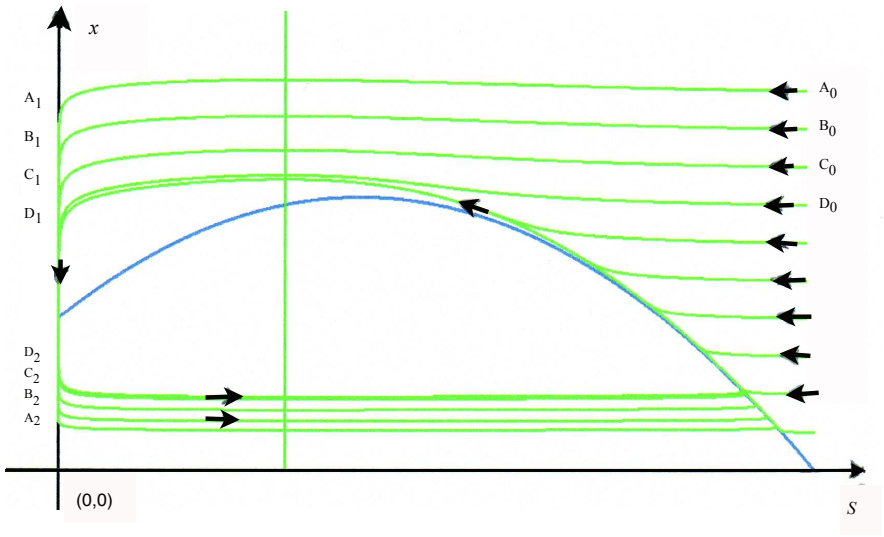


FIG. 5 – Les solutions du système (7) classique

4.2 Le portrait de phase du modèle à persistance forcée

Nous ajoutons au modèle classique (7) un seuil pour la variable s (pour la simplicité nous ne mettons pas de seuil pour la variable x, ce qui sur cet exemple ne change rien). Le modèle avec seuil en s est :

$$\begin{cases} \dfrac{ds}{dt} = \begin{cases} \frac{s}{0.05}\left(1 - \frac{s}{2} - \frac{x}{0.4+s}\right) & \text{si} \quad s > \alpha \\ \rho s & \text{si} \quad s \leq \alpha \end{cases} \\ \dfrac{dx}{dt} = \left(\dfrac{s}{0.4+s} - 0.6\right) x \end{cases} \tag{8}$$

Comme nous nous plaçons dans la cas de la **prersistance forcée** nous supposons que ρ est **positif** et égal à 1 et nous prenons pour valeur du seuil $\alpha = 10^{-6}$. Sur les simulations (Figure 6) nous voyons sans surprise que ce système est persistant et nous avons la garantie que dans ce modèle la taille de la ressource reste supérieure à 10^{-6} (pourvu que la condition initiale soit supérieure à 10^{-6}). Donc, par exemple, nous sommes certains d'avoir toujours au moins 10^3 individus si l'unité choisie est 10^9 individus. Les solutions de ce système avec seuil ressemblent fortement à celles du système classique sans seuil (Figure 5) avec toutefois une petite différence. Les trajectoires issues des points A_0, B_0, C_0, D_0 "ressortent" apparemment toutes au même point de l'axe vertical. Nous rendons plus visible cette différence en représentant sur une même figure (Figure 7) les trajectoires des deux systèmes. On y voit, en rouge, quatre trajectoires du système à seuil issues des quatre conditions initiales A_0, B_0, C_0, D_0 et en vert les quatre

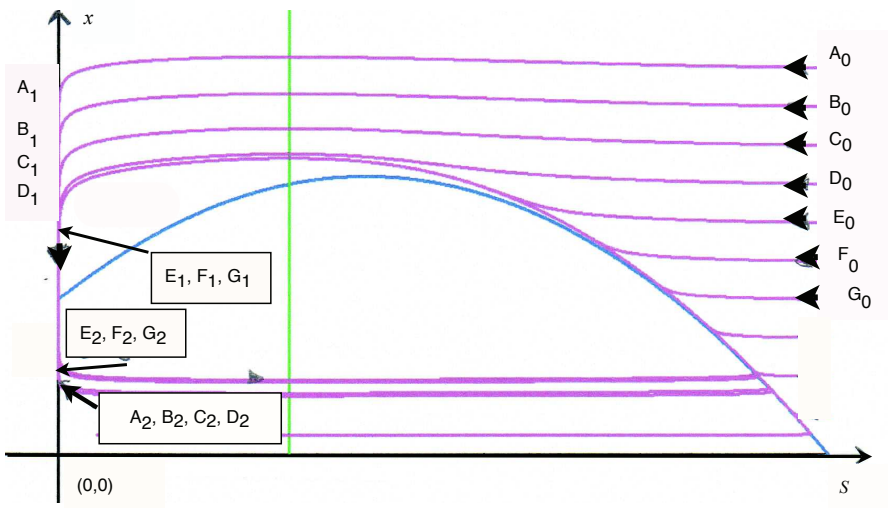


FIG. 6 – Les solutions du système à seuil de persistance (8) avec $\alpha = 10^{-6}$

trajectoires issues des mêmes conditions initiales, pour le système classique. On voit que dans les deux cas les trajectoires des deux systèmes commencent par être identiques, sont confondues pendant un certain temps avec l'axe vertical mais, alors que toutes les trajectoires du système à seuil ressortent à peu près au même point, les trajectoires du système classique ressortent d'autant plus bas qu'elles sont parties de plus haut. On peut dire que quand il y a persistance forcée la croissance de s redémarre beaucoup plus rapidement, ce qui n'est pas surprenant.

Une étude mathématique expliquant ce qui est observé à la Figure 5 a été proposée (dans un cadre différent) pour la première fois dans [5] pour étude les *canards* de l'équation de van der Pol (voir les commentaires bibliographiques à la fin de cet artcile). L'idée de [5] est de faire, dans un système de la forme :

$$\frac{ds}{dt} = \frac{1}{\varepsilon}\phi(s,x)$$

(dans notre cas $\varepsilon = 0.05$) le changement de variable :

$$z = \varepsilon \ln(s)$$

ce qui a pour effet d'étaler l'intervale $]0\ ;\ \varepsilon]$ sur l'intervalle $]-\infty\ ;\ \varepsilon\ln(\varepsilon)]$ et permet de séparer toutes les trajectoires confondues le long de l'axe vertical. Nous ne reprenons pas cette théorie ici mais reprenons l'idée du changement de variable que nous appliquons aux systèmes classique (7) et avec persistance forcée (8) ce qui nous donne les systèmes :

$$\begin{cases} \dfrac{dz}{dt} = 1 - \dfrac{e^{z/\varepsilon}}{2} - \dfrac{x}{0.4 + e^{z/\varepsilon}} \\ \dfrac{dx}{dt} = \left(\dfrac{e^{z/\varepsilon}}{0.4 + e^{z/\varepsilon}} - 0.6 \right) x \end{cases} \tag{9}$$

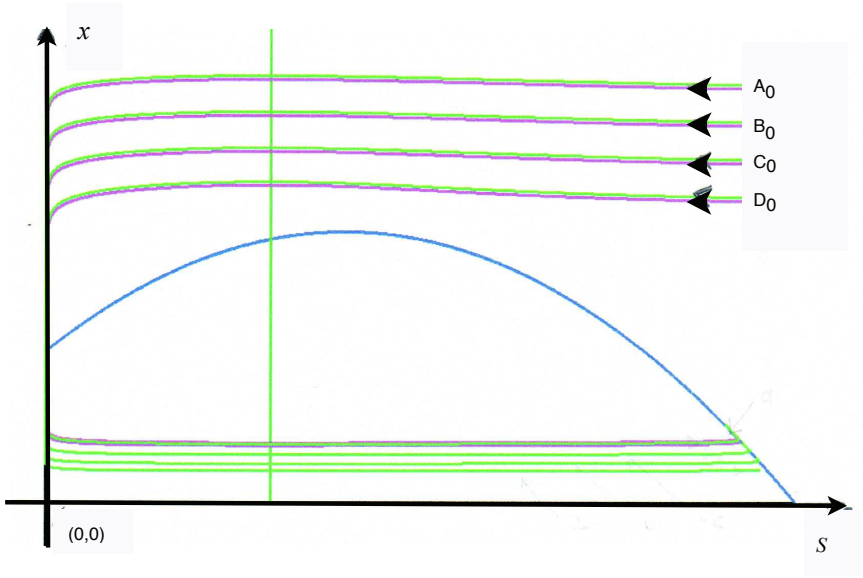


FIG. 7 – En rouge, les solutions du système (8) avec seuil de persistance, en vert, les solutions du système classique (7).

et :

$$\left\{\begin{array}{l} \dfrac{dz}{dt} = \left\{\begin{array}{ll} 1 - \frac{e^{z/\varepsilon}}{2} - \frac{x}{0.4+e^{z/\varepsilon}} & \text{si} \quad s > \alpha \\ \varepsilon\rho & \text{si} \quad s \leq \alpha \end{array}\right. \\ \dfrac{dx}{dt} = \left(\dfrac{e^{z/\varepsilon}}{0.4+e^{z/\varepsilon}} - 0.6\right) x \end{array}\right. \tag{10}$$

dont l'étude du portrait de phase n'est pas difficile. Nous ne la détaillons pas et nous contentons de présenter les résultats d'une simulation (Figure 8) où nous faisons figurer dans les mêmes axes des trajectoires de (9) et (10). Dans les nouvelles variables le seuil est matérialisé par la droite verticale (en rose sur la figure) d'abcisse :

$$\varepsilon \ln(\alpha) = 0.05 \ln(10^{-6})$$

On voit également en bleu l'isocline $z'(t) = 0$. 0n voit, en vert, les trajectoires du système classique issues des points a, b, c, d, e qui vont vers la gauche, traversent le seuil, traversent l'isocline et reviennent vers la droite. Plus une trajectoire part de haut plus elle coupe l'isocline à gauche et, par suite, plus elle recoupe l'axe vertical vers le bas. Le long de la discontinuité le champ (10) est "convergeant" au dessus de l'isocline, "divergent" ensuite. Ce qui explique que les trajectoires issues des point a, b du système (10), en marron, sont identiques à celles du système (9) jusqu'à ce qu'elles rencontrent la droite $z = \varepsilon \ln(\alpha)$; ensuite elles longent la droite $z = \varepsilon \ln(\alpha)$ jusqu'à l'isocline où elles la quittent. En revanche les trajectoires issues des points c, d, e sont identiques pour les deux systèmes. On voit que si la droite $z = \varepsilon \ln(\alpha)$ ne coupe pas le cycle limite du système classique, alors les deux systèmes auront le même cycle limite, seuls les transitoires sont éventuellement modifiés. En revanche, dans le cas contraire, le cycle limite du système avec persistance forcée est modifié.

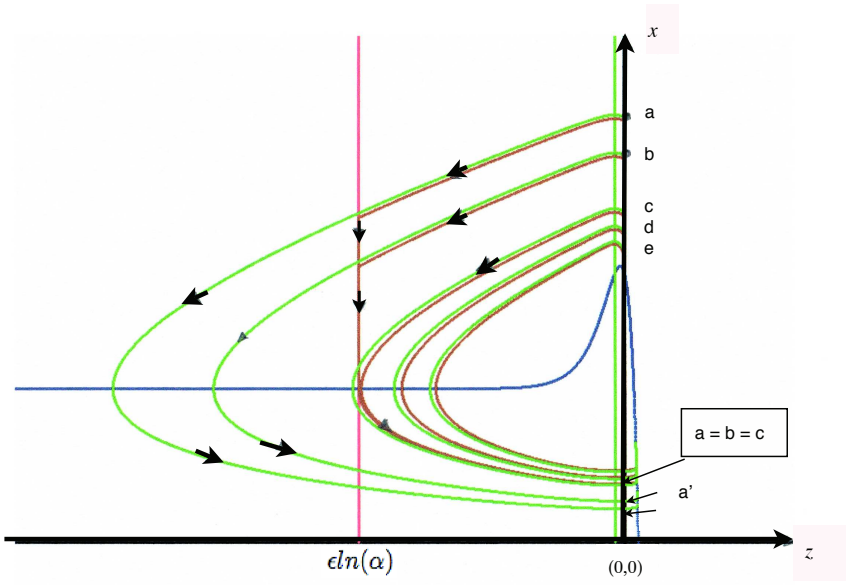


FIG. 8 – Dans les variables $\varepsilon(log(s), x)$: en vert les trajectoire de (9), en marron celles de (10).

5 Extinction automatique

Nous reprenons l'équation classique (4) que nous avons simulée jusqu'à maintenant mais nous ajoutons l'hypothèse d'une "extinction automatique" en dessous du seuil α ce que nous traduisons par le modèle :

$$\begin{cases} \dfrac{ds}{dt} = \begin{cases} \frac{s}{0.05}\left(1 - \frac{s}{2} - \frac{x}{0.4+s}\right) & \text{si} \quad s > \alpha \\ \rho s & \text{si} \quad s \leq \alpha \end{cases} \\ \dfrac{dx}{dt} = \left(\dfrac{s}{0.4+s} - 0.6\right) x \end{cases}$$

où ρ est strictement négatif (et égal à -1 dans nos simulations).

5.1 Description du domaine de persistance

Nous appelons domaine de persistance l'ensemble des conditions initiales telles que $s(t)$ reste toujours supérieur à la valeur α. Nous choisissons les valeurs des paramètres ε, e et m telles que le modèle sans seuil possède un cycle limite. La Figure 9 montre une simulation effectuée avec une valeur relativement élevée ($\alpha = 0.1$) du seuil α pour que l'effet soit bien visible. Le long de la verticale $s = \alpha$ on observe deux parties : au dessus de l'isocline de la ressource le champ est "traversant" et, au dessous, il est "divergeant". Si nous considérons le cas où le système classique possède un cycle limite, nous voyons que ce qui va être décisif c'est le fait que la verticale $\Delta_\alpha = \{(s, x) : \ s = \alpha\}$ coupe ou non ce cycle. Soit δ la distance du cycle limite à l'axe vertical. Nous avons vu dans les précédentes simulations que pour $\varepsilon = 0.05$ la valeur de δ est si petite que le bord gauche du cycle est confondu avec l'axe vertical ce qui rend problématique la visualisation de ce qui va se passer car il n'est pas possible d'agrandir ce qui se passe autour de l'axe vertical tout en continuant à visualiser le cycle sauf à utiliser la variable $\varepsilon \ln(s)$ comme dans le paragraphe précédent, ce qui oblige à réinterpréter le dessin. Nous avons préféré prendre

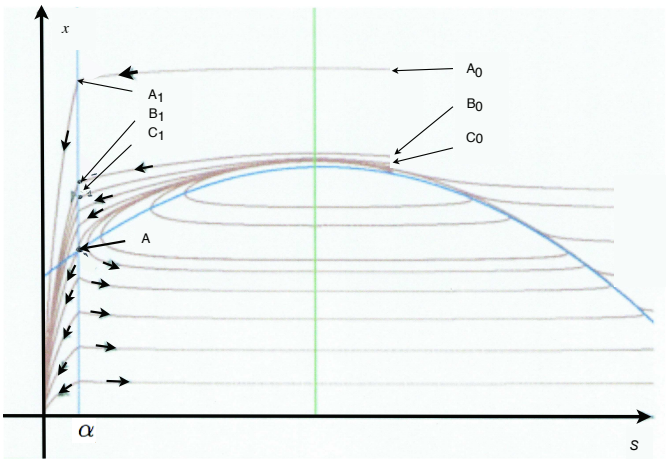


FIG. 9 – Portrait de phase du modèle avec extinction automatique.

pour le modèle classique des valeurs des paramètres pour lesquelles on observe un cycle limite tel que $\delta = 0.37$ ce qui nous éloigne de la problématique de départ mais permet de visualiser le phénomène de bifurcation qui intervient quand on fait varier α au moment où la droite Δ_α correspondante est tangente au cycle limite.

Nous commençons par définir un domaine, noté $\mathcal{A}$ du plan. Pour cela nous considérons l'unique trajectoire γ :

$$t \to \gamma(t) = (s(t), x(t))$$

qui est tangente à Δ_α (c'est la trajectoire qui passe par l'intersection, noté A de l'isocline de la ressource avec la droite Δ_α). Soit γ^- la partie négative de la trajectoire (pour t variant de $-\infty$ à 0) ; Lorsque α est très petit la courbe γ^- ne recoupe pas Δ_α. Dans ce cas le domaine $\mathcal{A}$ est la partie du plan située à droite de Δ_α et en dessous de γ^- (voir la Figure 11, en haut et au milieu, à gauche).

Pour des valeurs plus grandes de α la courbe γ^- recoupe Δ_α. Soit B le premier point où elle recoupe Δ_α. Dans ce cas le domaine $\mathcal{A}$ est la portion bornée de plan délimitée par Δ_α et la portion de γ comprise entre A et B, (voir la Figure 11, au milieu, à droite, et en bas).

5.2 Bifurcations du domaine de persistance

Nous commentons tout d'abord le schéma de la Figure 10. Le point C, intersection de l'isocline $x'(t) = 0$ et de l'axe horizontal (en bleu), est un col. Soit α_0 la valeur de α pour laquelle la trajectoire γ passant par A est un séparatrice du col C. Cette trajectoire instable de ce col pour(laquelle x est positif) vient s'enrouler autour du cycle limite ; deux trajectoires voisines sont représentées en rouge. On voit sur le scéma en quoi α_0 constitue une valeur de bifurcation entre un domaine $\mathcal{A}$ non borné (pour $\alpha < \alpha_0$) et un domaine borné (pour $\alpha > \alpha_0$).

Dans la Figure 11 nous avons représenté plusieurs simulations pour illustrer les bifurcations du domaine de persistance lorsque la valeur du paramètre α augmente et traverses les valeurs α_0 puis δ.

En haut, à gauche, nous voyons le cas $\alpha = 0.001$; dans ce cas la droite Δ_α est confondue avec l'axe vertical. La demi-trajectoire γ^- est également confondue avec le demi-axe vertical pendant une durée importante avant de se diriger vers la droite (en temps rétrograde). On voit, en marron, le cycle limite matérialisé par une trajectoire partant de l'intérieur. Nous avons

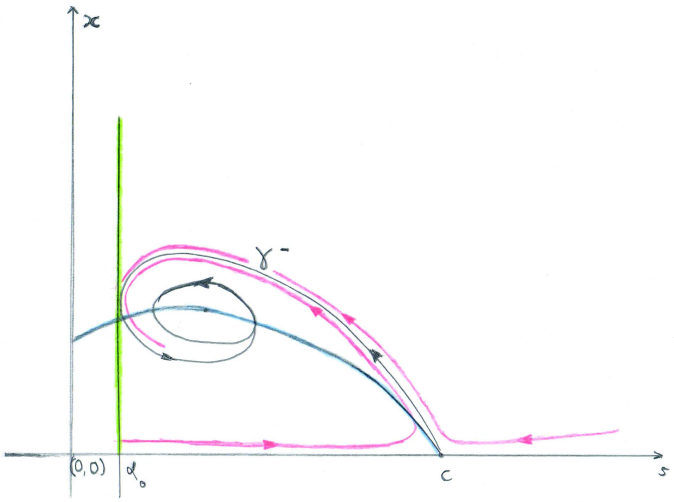


FIG. 10 – Schéma de la bifurcation

hachuré le domaine $\mathcal{A}$. En haut, à droite, nous vovons le cas $\alpha = 0.3$; dans ce cas la droite Δ_α est visible à droite de l'axe vertical. La demi-trajectoire γ^- est plus basse que dans le cas précédent.

Au milieu, à gauche, nous voyons le cas $\alpha = 0.35901091 < \alpha_0$ qui est une valeur pour laquelle Δ_α est proche et à *gauche* de la droite tangente à la trajectoire instable du col. La demi-trajectoire γ^- se rapproche de l'isocline mais continue à se diriger (en temps rétrograde) vers les s infiniment grands. Au milieu, à droite, nous vovons le cas $\alpha = 0.35901093 > \alpha_0$ qui est une valeur pour laquelle Δ_α est encore à gauche du cycle limite. Le domaine $\mathcal{A}$ devient borné et est délimité par les portions de de $\mathcal{A}$ et Δ_α comprises entre A et B .

En bas, à gauche, nous voyons le cas $\alpha = 0.3595 < \delta$ qui est une valeur pour laquelle Δ_α est plus proche du cycle limite. Le domaine $\mathcal{A}$ est toujours délimité par les portions de de $\mathcal{A}$ et Δ_α comprises entre A et B . On voit comment le domaine $\mathcal{A}$ diminue quand α augmente pour disparaître quand la droite Δ_α est tangente au cycle limite. En bas, à droite, nous voyons le cas $\alpha = 0.37 = \delta$; dans ce cas la droite Δ_α est tangente au cycle limite. Le domaine $\mathcal{A}$ est l'intérieur du cycle limite. Pour α un peu plus grand le domaine $\mathcal{A}$ est vide.

Nous avons donc démontré le :

Théorème 5.1 *Soit δ la distance du cycle limite à l'axe vertical. Soit $\underline{\delta}$ la distance du cycle limite à l'axe horizontal. Lorsque $\alpha < \delta$, pour toute condition initiale située dans $\mathcal{A}$ la trajectoire correspondante satisfait :*

$$\liminf s(t) = \delta, \qquad \liminf x(t) = \underline{\delta}.$$

Pour les condition initiales non situées dans $\mathcal{A}$ au bout d'un temps fini on a $s(t) = \alpha$ et, à partir de là $s(t)$ et $x(t)$ disparaissent.
Lorsque $\alpha > \delta$ pour toutes les trajectoires, au bout d'un temps fini on a $s(t) = \alpha$ et par suite disparition des deux espèces.

Le domaine $\mathcal{A}$ est donc le domaine de persistance des deux espèces. Si nous faisons croître le seuil α à partir de la valeur 0 nous avons pour commencer un domaine $\mathcal{A}$ de persistance non borné qui diminue, devient borné pour une première valeur de bifurcation α_0 puis qui diminue

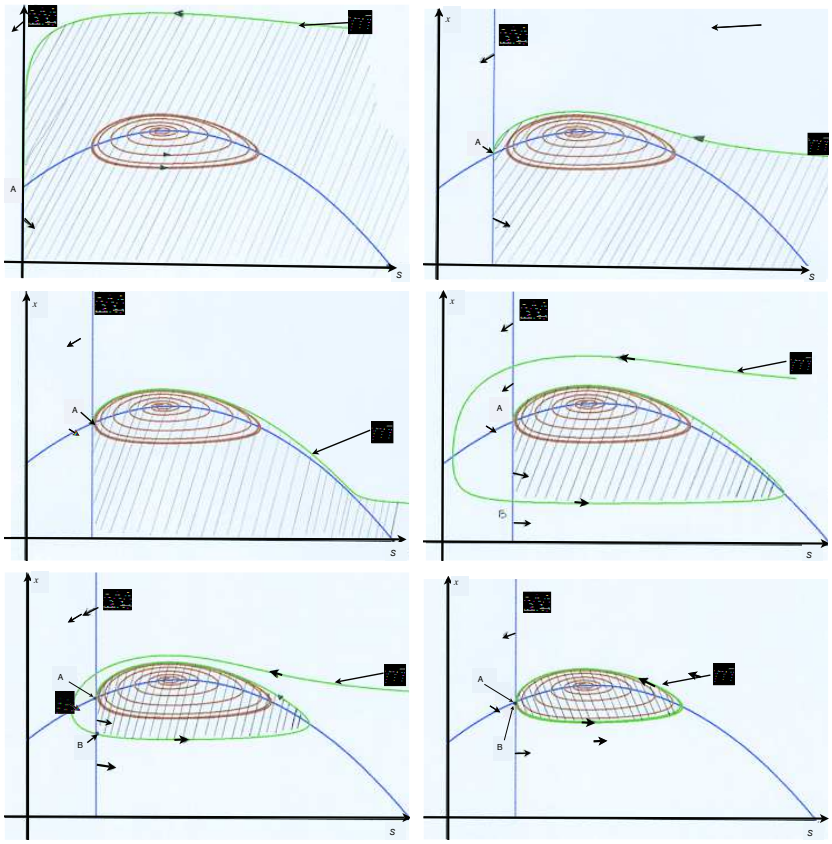


FIG. 11 – Le domaine de persistance $\mathcal{A}$ pour $\alpha = 0.001$ (en haut à gauche), $\alpha = 0.3$ (en haut à droite), $\alpha = 0.35901091$ (au milieu à gauche), $\alpha = 0.35901093$ (au milieu à droite), $\alpha = 0.3595$ (en bas à gauche), et $\alpha = 0.37$ (en bas à droite)

à nouveau jusqu'à la valeur $\alpha = \delta$ où $\mathcal{A}$ est exactement le cycle limite et son l'intérieur, puis $\mathcal{A}$ disparaît dès que α est plus grand que δ.

6 Conclusion

Nous avons considéré un modèle classique de relation *consommateur-ressource* pour lequel il y a *persistance* au sens usuel : Il possède un cycle limite globalement asymptotiquement stable et donc pour toute condition initiale la limite inférieure, lorsque le temps tend vers l'infini, des concentrations de ressource et de consommateur est uniformément bornée inférieurement par un réel strictement positif. Mais nous avons montré sur des exemples que, lors des transitoires, et même le long du cycle limite, les valeurs des concentrations peuvent devenir si petites que le modèle perd toute signification. Nous avons donc proposé d'introduire des "seuils", au delà desquels le modèle n'ayant plus de signification une hypothèse "ad hoc" est nécessaire pour continuer l'analyse ; nous en avons proposé deux, la *persistance forcée* et l'*extinction automatique*. Dans le premier cas la persistance prévue par le modèle classique est bien entendu conservée mais des consitions initiales identiques peuvent conduire à des trajectoires différentes. Dans le second cas il apparaît un domaine de persistance, qui peut être vide et dont la taille dépend de

la valeur du seuil en dessous duquel il y a extinction automatique ; une valeur de bifurcation entre domaine borné et domaine non borné est mise en évidence.

Nous concluons de cette étude sur un cas très particulier que toute simulation d'un système différentiel représentant une situation réelle devrait s'accompagner, entre autres, de la précaution suivante :

- Pour chaque variable d'état x_i décider de façon réaliste de la valeur d'un seuil α_i en dessous duquel la variable perd toute signification physique.
- Introduire dans le programme de simulation la clause :
  Si $x_i(t) \leq \alpha_i$ *tout arrêter et envoyer un message d'alerte.*
- En cas d'alerte reprendre le modèle et aviser.

Il est bien évident que nos deux hypothèses "ad hoc" ne sauraient résoudre tous les problèmes et, à vrai dire, nous pensons qu'elles ne devraient que rarement satisfaire le biologiste. Il pourra préférer utiliser des modèles stochastiques où les variables d'état ne sont plus des concentrations d'individus d'une population mais le "nombre d'individus" et les règles de transition de type probabilité de naissance ou de mort. C'est dans cet esprit que l'article [12] revisite le très classique modèle du chémostat. Il faudra alors envisager le couplage de ce modèle à l'ancien.

A Commentaires bibliographiques

Dans le système (4) les valeurs des paramètres e, m, K étaient relativement grandes par rapport à ε qui a été pris égal à $0,05$. L'étude mathématique de ces systèmes peut se faire soit en faisant tendre ε vers zéro, soit en faisant l'hypothèse que ε est *infiniment petit* au sens de l'Analyse Non Standard (ANS). Pour la méthode classique nous recommandons [27, 38, 44] et pour la méthode non classique [15, 16, 17, 47]. Nos simulations ont mis en évidence des trajectoires qui se confondent pendant un certain temps avec les isoclines. Le théorème de Tychonov rend compte de ce phénomène. On peut en trouver la version classique dans [27, 42, 44] et une version ANS dans [34]. Pour des modèles plus complexes en dimension trois et plus on peut être conduit à utiliser le théorème de Pontryagyn Rodygin [40, 41] qui étudie le cas où la dynamique rapide admet un cycle limite asymptotiquement stable et pas un équilibre asymptotiquement stable, comme cela est le cas pour le théorème de Tikhonov. Pour des applications à des modèles biologiques des théorèmes de Tikhonov et/ou de Pontriagyn-Rodigyn on peut consulter [7, 8, 9, 30, 36, 41].

La première étude, avec des méthodes d'ANS, d'un système du type (3) est due à quatre élèves de Georges Reeb : Eric Benoit, Jean-Louis Callot, Francine Diener et Marc Diener [5]. Ils ont mis en évidence un phénomène nouveau de solutions appelées *solutions canard*. Les solutions canards sont des trajectoires spéciales de champs lents rapides qui sont d'abord proches de la partie stable de la variété lente, ensuite de la partie instable de cette variété. En plus de l'article originel [5], voir [4, 11, 15, 43, 47]. Le phénomène des solutions canard est lié au problème du retard à la bifurcation dans les bifurcations dynamiques (voir [1] p. 179-192 et [3, 29]). L'étude des solutions canards a été faite aussi dans le contexte de l'analyse asymptotique classique [19], de la théorie de la variété centrale et des éclatements [18, 39, 45], et de l'asymptotique complexe Gevrey [6, 13, 24]. Pour plus d'informations on peut consuler l'article *Canards* de Martin Wechselberger paru dans l'encyclopédie en ligne *Scholarpedia* [46]. L'article [33] est un essai de vulgaristion des méthodes de l' ANS dans le domaine de l'automatique. Voir aussi à se sujet [31, 35].

Remerciements

Les auteurs remercient Fabien Campillo, Jérôme Harmand et Alain Rapaport, de l'équipe *Modemic*[11] ainsi que Jean-Luc Gouzé de l'équipe *Biocore*[12] pour de nombreuses et fructueuses discussions sur les questions de méthodologie de la modélisation en dynamique des populations.

Références

[1] V. I. Arnold, *Dynamical Systems V, Bifurcation Theory and Catastrophe Theory*, Encyclopedia Math. Sci., Vol 5, Springer-Verlag, Berlin/New York, 1993.

[2] G. Batt, R. Casey, H. de Jong, J. Geiselmann, J.L. Gouzé, M. Page, D. Ropers, T. Sari, D. Schneider, Analyse qualitative de la dynamique de réseaux de régulation génique par des modèles linéaires par morceaux, in Modélisation et simulation pour la post-génomique, *Revue des sciences et Technologies de l'information, Série Technique et science informatiques*, 26 (2007), 11-45.

[3] E. Benoît (Ed.), *Dynamic Bifurcations*, Proceedings Luminy 1990, Lect. Notes Math. 1493 Springer-Verlag, 1991.

[4] E. Benoît, Perturbation singulière en dimension trois : Canards en un point pseudo-singulier noeud, *Bulletin de la Société Mathématique de France*, 129 (2001), 91-113.

[5] E. Benoît, J.L. Callot, F. Diener, M. Diener, Chasse au canard, *Collect. Math.*, 32 (1981), 37–119.

[6] E. Benoît, A. Fruchard, R. Schaefke, G. Wallet, Solutions surstables des équations différentielles lentes-rapides à point tournant, *Annales de la Faculté des Sciences de Toulouse*, VII (1998), 627–658.

[7] H. Boudjellaba, T. Sari, Oscillations of a prey-predator-superpredator system, *J. Biol. Systems*, 6 (1998), 17–33.

[8] H. Boudjellaba, T. Sari, Stability loss delay in harvesting competing populations. *J. Differential Equations*, 152 (1999), 394–408.

[9] H. Boudjellaba, T. Sari, Dynamic transcritical bifurcations in a class of slow-fast predator-prey models *J. Differential Equations*, 246 (2009), p. 2205–2225.

[10] G. Butler, P. Waltman, Persistence in dynamical systems, *J. Differential Equations*, 63 (1986), 255–263.

[11] JL. Callot, Champs lents-rapides complexes à une dimension lente, *Annales scientifiques de l'Ecole Normale Supérieure*, 4, 26 (1993), 149–173.

[12] F. Campillo, M. Joannides, I. Larramendy-Valverde, Stochastic modeling of the chemostat. *Ecological Modeling*, 22 (2011), 2676–2689.

[13] M. Canalis-Durand, J.P. Ramis, R. Schaefke, Y. Sibuya, Gevrey solutions of singularly perturbed differential equations, *J. Reine Angew. Math.*, 518 (2000), 95–129.

[14] H. de Jong, J.L. Gouzé, C. Hernandez, M. Page, T. Sari, J. Geiselmann, Qualitative simulation of Genetic Regulatory Network using Piecewise Linear Models, *Bulletin Math. Biology*, 66 (2004), 301–340.

[11] `http ://www.inria.fr/equipes/modemic`
[12] `http ://www.inria.fr/equipes/biocore`

[15] F. Diener, M. Diener (Eds.), *Nonstandard Analysis in Practice*. Universitext, Springer-Verlag, 1995.

[16] M. Diener, C. Lobry, (Eds.), *Analyse non standard et représentation du réel*, OPU, Alger, CNRS, Paris, 1985.

[17] M. Diener, G. Wallet (Eds.), *Mathématiques finitaires et analyse non standard*, Publication mathématique de l'Université de Paris 7, Vol. 31-1 et 31-2, 1989.

[18] F. Dumortier, R. Roussarie, Canard Cycles and center manifolds, *Mem. Amer. Math. Soc.* 577, 1996.

[19] W. Eckhaus, Relaxation oscillations including a standard chase on French ducks, in *Asymptotic Analysis II, surveys and new trends*, Lecture Note Math. 985 Springer-Verlag, Berlin/New York (1984), 449–494.

[20] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol 1 and 2. John Wiley & Sons, 2nd edition, 1971.

[21] A.F. Filippov, *Differential equations with discontinuous right-hand sides*, *Mat. Sb.*, 51 (1960), 99–128.

[22] A.F. Filippov,, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic Publishers, 1988.

[23] H.I. Freedmann, P. Moson, Persistence definitions and their connections, *Proc. Amer. Math. Soc.* 109 (1990), 1025–1033.

[24] A. Fruchard, R. Schaefke, Exceptional complex solutions of the forced van der Pol equation, *Funkcialaj Ekvacioj*, 42 : 2 (1999), 201–223.

[25] J.L. Gouzé, T. Sari, A class of piecewise linear differential equations arising in biological models. Special issue : Non-smooth dynamical systems, theory and applications. *Dynamical Systems : An International Journal*, 17 (2002), 299–316.

[26] J. Hofbauer, K. Sigmund, *The Theory of Ecolution and Dynamical Systems*, Cambridge Univ. Press, Cambridge, UK, 1988.

[27] P.V. Kokotovic, H.K. Khalil, and J. O'Reilly, *Singular Perturbations Methods in Control : Analysis and Design.* Academic Press, New York, 1986.

[28] L.P. Liou, K.S. Cheng, On the uniqueness of a limit cycle of a predator-prey system, *SIAM J. Math. Anal.*, 19 (1988), 867–878.

[29] C. Lobry, A propos du sens des textes mathématiques, un exemple : la théorie des "bifurcations dynamiques", *Annales Institut Fourier*, 42 (1992), 327–351.

[30] C. Lobry, A. Rapaport, T. Sari, Stability loss delay in the chemostat with a slowly varying washout rate, *Proceedings MATHMOD 09 Vienna 11-13/02/2009*, Inge Troch, Felix Breitenecker (Editors), ARGESIM Report no. 35 (2009), 1582–1586.

[31] C. Lobry, T. Sari, Singular perturbation methods in control theory, in *Contrôle non linéaire et Applications*, Travaux en Cours no. 64, Hermann, Paris (2005), 155–182.

[32] C. Lobry, T. Sari, Equations différentielles à second membre discontinu, in *Contrôle non linéaire et Applications*, Travaux en Cours no. 64, Hermann, Paris (2005), 255–289.

[33] C. Lobry, T. Sari, Nonstandard Analysis and representation of reality, *International J. Control*, 81, 3 (2008), 519–536.
Traduction française : `http ://hal.inria.fr/inria-00163365/fr/`.

[34] C. Lobry, T. Sari, S. Touhami, On Tykhonov's theorem for convergence of solutions of slow and fast systems, *Electron. J. Diff. Eqns.*, Vol. 1998 (1998), No. 19, 1–22.

[35] C. Lobry, T. Sari, S. Touhami, Fast and slow feedbacks in systems theory, *J. Biol. Systems*, 7 (1999), 1–25.

[36] C. Lobry, T. Sari, K. Yadi, Coexistence of three predators competing for a single biotic resource, in *Advances in the Theory of Control, Signals and Systems with Physical Modeling*, J. Lévine and P. Mullhaupt (Editors) Lectre Notes in Control and Information Sciences, 2010, vol. 407, p. 309–322.

[37] J.D. Murray *Mathematical Biology I. An Introduction*, Series : Interdisciplinary Applied Mathematics , Vol. 17 Springer, Heidelberg, 2004.

[38] R.E. O'Malley, Jr. *Singular Perturbation Methods for Ordinary Differential Equations*, Springer-Verlag, New York, 1991.

[39] D. Panazzolo, On the Existence of Canard Solutions. *Publ. Mat.*, 44, 2 (2000), 503–592.

[40] L.S. Pontryagin, L.V. Rodygin, Approximate solution of a system of ordinary differential equations involving a small parameter in the derivatives, *Soviet. Math. Dokl.*, 1 (1960), 237–240.

[41] T. Sari, K. Yadi, On Pontryagin-Rodygin's theorem for convergence of solutions of slow and fast systems, *Electron. J. Diff. Eqns.*, Vol. 2004 (2004), No. 139, 1–17.

[42] A.N. Tykhonov, Systems of differential equations containing small parameters multiplying the derivatives, *Mat. Sborn.*, 31 (1952), 575–586.

[43] W. Wallet, Entrée-sortie dans un tourbillon, *Annales de l'Institut Fourier*, 36 (1986), 157–184.

[44] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Robert E. Kriger Publishing Company, New York, 1976.

[45] M. Wechselberger, Existence and Bifurcation of Canards in $\mathbb{R}^3$ in the case of a Folded Node, *SIAM J. Applied Dynamical Systems*, 4 (2005), 101–139.

[46] M. Wechselberger, Canards (2007), *Scholarpedia*, 2 (4) : 1356.
`http ://www.scholarpedia.org/article/Canards`

[47] A.K. Zvonkin, M.A. Shubin, Nonstandard Analysis and Singular Perturbations of Ordinary Differential Equations, *Uspekhi Mat. Nauk.*, 39 (1984), 77–127. English transl. : *Russian Math. Surv.*, 39 (1984), 69–131.

Adresses des auteurs :

Claude Lobry
EPI Modemic Inra/Inria,
UMR Mistea, 2 pl. Viala, 34060 Montpellier, France.

Courriel : `Claude.Lobry@inria.fr`

Tewfik Sari
Laboratoire de Mathématiques, Informatique et Applications, EA 3993
Faculté des Sciences et Techniques, Université de Haute Alsace
4, rue des Frères Lumière, 68093 Mulhouse cedex, France
Adresse actuelle
UMR Itap, Cemagref, Domaine de Lavalette
361, rue J.-F. Breton, BP 5095, 34196 Montpellier Cedex 5, France

Courriel : `Tewfik.Sari@uha.fr`

Correspondence between discrete and piecewise linear models of gene regulatory networks

Francine Diener[1], Aparna Das[1], Gilles Bernot[2], Jean-Paul Comet[2], Frédéric Eyssette[1]

Abstract

We know that some proteins can regulate the expression of genes in a living organism. The regulation of gene expression occurs through networks of regulatory interactions in a non linear way between DNA, RNA, proteins and some molecules, called genetic regulatory networks. It is becoming clear that mathematical models and tools are required to analyze these complex systems.

In the course of his study on gene regulatory networks R. Thomas proposed a discrete framework that mimics the qualitative evolution of such systems. Such discrete models are of great importance because kinetic parameters are often non measurable *in vivo* and available data are often of qualitative nature. Then Snoussi proved consistency between the discrete approach of R. Thomas and Piecewise Linear Differential Equation Systems, which are easy to construct from interaction graph and thresholds of interactions.

Our work focuses on the relationships between both approaches: we will prove a result of correspondence between the two models. Finally, we will give some short description of a Maple program which can compute a discrete path, given the ordinary differential equation and starting box.

Supplementary information: The code for computing discrete path and instructions to use it are available on http://math.unice.fr/∼diener/.

1 Introduction

A gene regulatory network is a set of genes coding for proteins (i.e. each gene expresses itself and produces a specific protein) able to activate or inhibit the expression of the other genes of the set. As the number of genes of the interacting networks is usually high, the possible interactions between them build a network of interactions so complex and intricated that it becomes really difficult to predict for example the consequences on the whole network of the over expression of one gene or under expression of another. Building simplified computational models is thus required to understand the dynamic of these networks.

The most obvious method to model such a network consists of a description in term of systems of differential equations. But as the interactions between genes are considered as non linear and as most of the parameters of the differential equations are impossible to identify, it remains difficult, if not impossible, to understand the dynamic and to predict the behavior of the different genes even knowing the form of the differential model, unless one can simplify the description.

In the early seventies, two kinds of simplified models have been introduced on the same idea: the activation or inhibition of one gene on the expression of another gene have a sigmoid profile which means that the regulation is essentially inefficient when the concentration of the active gene is below a threshold value, its effect increases rapidly around this threshold value and is

saturated for higher concentration. With this in mind, Glass and Kauffman [4] have introduced a special class of piecewise linear differential equations, replacing the sigmoids in the differential equations by Heaviside functions. This leads to a discontinuous system of differential equations that is much more tractable than the original smooth ones. Piecewise linear models have been intensively studied (see [11] for example) and produce lots of interesting results. Another class of even more radically simplified models, the so called *logical models* or *discrete models*, introduced by René Thomas [10], is also build on an on/of version of the regulations but in keeping only in the model the description of the dynamic through a boolean interaction graph. This over simplification is shown to be especially useful in allowing automatic explorations of all possible interactions with a computer ([2],[1] for example).

In this paper we will first present, and illustrate with two typical examples, the two approaches, the piecewise linear and the discrete. Then we will prove a result of correspondence between the two models. Doing this we follow a work of Snoussi [8] who first introduced what he called a *discrete mapping* that maps each piecewise linear differential model to a corresponding discrete model in an automatic way. It is thus important to describe precisely what information contained in the piecewise linear differential model is kept by the discrete model and what is lost.

2 Two examples of gene regulatory networks

Before introducing the general form of the piecewise linear differential model we consider here, let us just show first two typical examples, one is an example of what is called a biological switch and the other example of a biological cycle. These both cases, are only toy examples (with arbitrary chosen coefficients), just for illustration.

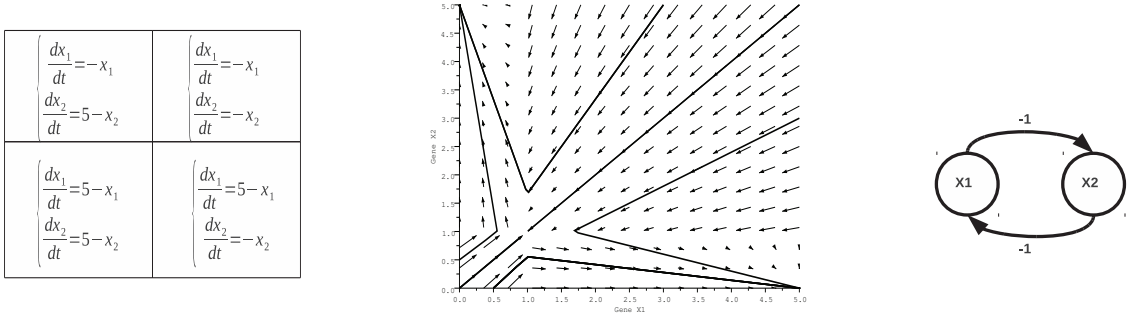


Figure 1: An example of biological switch: the choice of the parameters here is $\gamma_1 = \gamma_2 = \theta_2^1 = \theta_1^2 = 1$ and $k_1^0 = k_2^0 = -k_1^2 = -k_2^1 = 5$. There are 4 boxes in which the differential system is linear, $]0,1[^2$, $]1,5[\times]0,1[$, $]0,1[\times]1,5[$ and $]1,5[^2$. It happens that all trajectories with initial point "below" the diagonal converge to the stable node $(5,0)$ and all the trajectories with initial point above the diagonal converge to the other stable node $(0,5)$. There is a third equilibrium on the diagonal, $(1,1)$, which is a saddle point of the dynamic.

2.1 Example of a biological switch

The first example consists in two genes X_1 and X_2, whose respective level of expression at time t are denoted by $x_1(t)$ and $x_2(t)$ (level of concentration in the two produced proteins). The

model of the dynamic takes into account several effects. First a negative action (inhibition) of the protein produced by the gene X_2 on the expression of the gene X_1, described by

$$\frac{dx_1}{dt} = k_1^0 + k_1^2 I(\theta_2^1, x_2)$$

where k_1^0 is the level of expression of X_1 when X_2 is absent (or inefficient), θ_2^1 is the threshold value of the level of expression of X_2 beyond which the X_2 inhibition of X_1 is assumed to be efficient, k_1^2 is the level of that inhibition and $I(\theta, x)$ is the Heaviside function (equal to 0 for $x < \theta$ and 1 for $x \geq \theta$). A second effect taken into account is the degradation of the protein produced by X_1. The rate of degradation is usually assumed to be proportional to the concentration $x_1(t)$ and then described by

$$\frac{dx_1}{dt} = -\gamma_1 x_1$$

We assume the dynamic of the second gene similar: inhibition by the other gene and decreasing of the level of concentration of its protein due to degradation. This leads to the following differential system:

$$\left\{ \begin{array}{rcl} \frac{dx_1}{dt} & = & k_1^0 + k_1^2 I(\theta_2^1, x_2) - \gamma_1 x_1 \\ \frac{dx_2}{dt} & = & k_2^0 + k_2^1 I(\theta_1^2, x_1) - \gamma_2 x_2 \end{array} \right. \quad (1)$$

It is easy to see that the two thresholds θ_1^2 and θ_2^1 cut the phase space $[0, max_1] \times [0, max_2]$ in 4 domains (or rectangular boxes) in which the differential system is simply linear with constant coefficients and thus easy to solve explicitly. It remains to stick together the trajectories in between the 4 domains to have a good picture of the global behavior (see figure 1). With the chosen set of parameters, the system has two stable equilibrium and an additional equilibria of saddle point type that create for the trajectories a possible switch from one stable equilibrium toward the other one when the initial conditions $(x_1(0), x_2(0))$ change: indeed a small modification of the initial conditions, just crossing the diagonal, is enough to completely modify the evolution of the system. This is the phenomenon of *bistability* or *biological switch*[1]: in one case[2], the system converges to one equilibrium corresponding to a maximal level of expression of X_2 and about no expression of X_1 and in the other case[3], the level of expression of X_1 is maximal and there is about no expression of X_2.

2.2 Example of a biological cycle

The second example is still an example with two genes X_1 and X_2 but we assume now that X_1 activates X_2 and that X_2 activates itself and inhibits X_1. This leads to the following model:

$$\left\{ \begin{array}{l} \frac{dx_1}{dt} = k_1^0 + k_1^2 I(\theta_2^1, x_2) - \gamma_1 x_1 \\ \frac{dx_2}{dt} = k_2^1 I(\theta_1^2, x_1) + k_2^2 I(\theta_2^2, x_2) - \gamma_2 x_2 \end{array} \right. \quad (2)$$

[1]This example is a simplified model of the following situation : the bacteriophage Lambda is a virus able to get into the cell of the Escherichia Coli bacteria and to multiply. The infection of the bacteria by the phage either leads to the destruction of the bacteria (lytic pathways) through a kind of explosion of the cell producing a huge amount of phages able to infect new bacterias, either to a silent integration of the phage genome into the bacteria genome (lysogenic pathway). In the last case, the bacteria will continue to reproduce itself as usual being now resistant to new infection. The choice between lytic and lysogenic pathways is regulated by two antagonist proteins, the protein CI responsible for the lysogenic pathway and the protein CRO responsible for the lytic pathway. These two proteins are encoded by two genes, *ci* and *cro*, whose expression is mutually inhibited by the other.

[2]which corresponds to the lytic pathway

[3]which corresponds to the lysogenic pathway

where the constants have the same meaning as in the first example.
As in the first example, this system is piecewise linear on four boxes delimited by the thresholds.

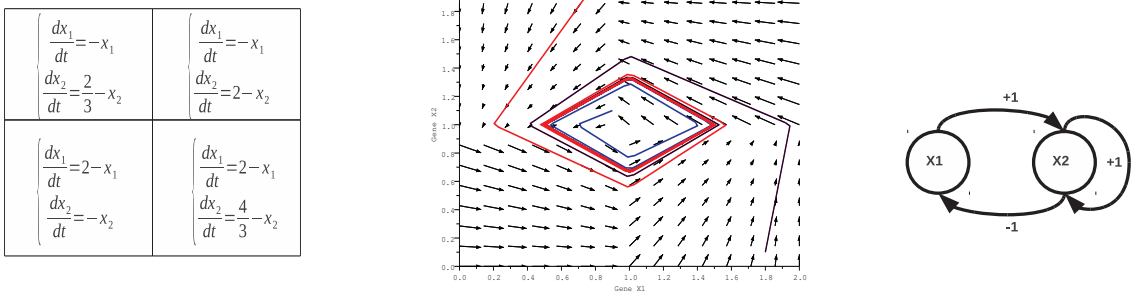


Figure 2: An example of biological cycle: the choice of the parameters here is $\gamma_1 = \gamma_2 = \theta_2^1 = \theta_1^2 = \theta_2^2 = 1$, $k_1^0 = -k_1^2 = 2$, $k_2^1 = \frac{4}{3}$ and $k_2^2 = \frac{2}{3}$. There are 4 boxes in which the system is linear. Whatever its initial condition, all trajectories will spiral toward a unique limit cycle.

It is easy to compute the trajectories in each box and to stick them end to end at the border of the boxes. For a particular choice of the parameters, one can prove that there is a unique attractive cycle (see figure 2), just in computing the first return Poincaré map explicitly.

3 The piecewise linear model

In the paper [3], from which we will adopt here some notations, the piecewise linear model for gene regulatory network we consider is called a *Glass model* because it has been introduced by Glass and Kauffman in [4]. The general form of this model of n genes $X_1, X_2, \ldots, X_n$ is given by

$$\frac{dx}{dt} = f(x) - \Gamma x \tag{3}$$

where $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ represents the concentrations of the proteins produced by the n genes, $f(x)$ is a vector whose i^{th} component $f_i(x_1, x_2, \ldots, x_n)$ represents the rate of synthesis of the i-th gene (which depends on the action of the other genes) and Γ is a diagonal matrix whose diagonal coefficients $\Gamma_i^i = \gamma_i$ are the degradation rates of genes. As in the two examples, the rate of synthesis f_i of each gene X_i is the sum of a constant term which represent a basal synthesis rate of this gene and the aggregated contribution of all the genes (including possibly itself) that activate or inhibit it.

$$f_i(x_1, x_2, \ldots, x_n) = k_i^0 + \sum_{j \in T_i} k_i^j I(\theta_j^i, x_j) \tag{4}$$

The numbers θ_j^i are the threshold values of the action of gene X_j on the expression of gene X_i and the coefficient k_i^j is positive when gene X_j is an activator of gene X_i and negative when it is an inhibitor. The sum is on all indices j such that the gene X_j either activates or inhibits the expression of the gene X_i. We denote by T_i, the set of all these indices. Being a concentration of protein, each quantities $x_i(t)$ ranges in some interval $[0, max_i]$ and thus the differential system will be considered only in the bounded domain $\mathcal{B} = \prod_{i=1}^n [0, max_i]$ called the *phase space* of the continuous model. The different thresholds break down the phase space $\mathcal{B}$ in several boxes where the differential system is simply linear (affine) (and thus easy to solve explicitly).

For each $j = 1, \ldots, n$, the set of all thresholds θ_j^i can be ordered in $\sigma_0^j < \sigma_1^j < \sigma_2^j < \ldots < \sigma_{s_j}^j < \sigma_{s_j+1}^j$ where we add $\sigma_0^j = 0$ and $\sigma_{s_j+1}^j = max_j$ as a boundary of the boxes and the different boxes are defined by the following system of n double inequalities:

$$\left\{ \begin{array}{lcccl} \sigma_{s_1}^1 & < & x_1 & < & \sigma_{s_1+1}^1 \\ \ldots & & \ldots & & \ldots \\ \sigma_{s_n}^n & < & x_n & < & \sigma_{s_n+1}^n \end{array} \right. \tag{5}$$

The multi integer (vector of integers) $s = (s_1, s_2, \ldots, s_n)$ which is an element of the set $S = \prod_{j=1}^n \{0, 1, \ldots, s_j\}$ is a level for the different boxes of the phase space $\mathcal{B}$ that we call a *state* of the network. In the piecewise linear model, there is only a finite number of states for the dynamic. Each of them corresponds to one box $\mathcal{B}_s$ of the phase space, delimited by a lower multi threshold $\sigma(s) = (\sigma_{s_1}^1, \ldots, \sigma_{s_n}^n)$ and an upper multi threshold $\sigma(s+\mathcal{I}_n) = (\sigma_{s_1+1}^1, \ldots, \sigma_{s_n+1}^n)$ where $\mathcal{I}_n$ is the vector having all its components equal to 1. The main point is that, in each box $\mathcal{B}_s$, the model (3) is given by the simple linear system

$$\left\{ \begin{array}{lcccl} \frac{dx_1}{dt} & = & k_1(s) & - & \gamma_1 x_1 \\ \ldots & & \ldots & & \ldots \\ \frac{dx_n}{dt} & = & k_n(s) & - & \gamma_n x_n \end{array} \right. \tag{6}$$

where $k_i(s) = f_i(x_1, \ldots, x_n)$ is the value of the synthesis rate f_i in the box B_s. Thus the solutions in the box are easy to compute

$$x_i(t) = \frac{k_i(s)}{\gamma_i} + e^{-\gamma_i t}(x_i(0) - \frac{k_i(s)}{\gamma_i}) \tag{7}$$

for all $i = 1, \ldots, n$, and it is immediately obvious that, inside each box, all solutions tend to a stable equilibrium called the *focal point* of the box. Let's denote $\phi(s) = (\frac{k_1(s)}{\gamma_1}, \ldots, \frac{k_n(s)}{\gamma_n})$ the focal point of $\mathcal{B}_s$. It can be either inside $\mathcal{B}_s$ (including its boundary), if it satisfy the set of double inequalities (5) and it is then a stable equilibrium of the whole dynamic, either outside (in another box $\mathcal{B}_{s'}$, with $s' \neq s$) and in this case the trajectories follow the dynamic given by (7) as long as they have not reached the boundary of $\mathcal{B}_s$ and will simply switch to the new dynamic of the next box when crossing the boundary.

Remark One difficulty with piecewise linear systems of differential equations like our model here is that they are discontinuous differential equations, and, as such, do not behave like smooth systems of differential equations regarding the question of existence and unicity of the solutions. More precisely, it can happens that the boundary between two boxes is a so called *black wall* when the solutions in both boxes point towards the boundary in such a way that it becomes impossible from each side to exit the box and enter the other. When this happens, one can nevertheless define properly a concept of solutions for such piecewise linear system using the Philipov theory of singular (and set valued) solutions, as explained by Gouzé and Sari in [6].

In this paper nevertheless, we will not consider such singular solutions because we will ever stay away from any black wall. All the solutions we will consider will be build only in putting end to end the boundaries of the boxes, the usual (i.e.well defined) solutions inside the different boxes.

Let us make precise what we consider here as a (regular) solution of (3).

Definition A continuous function $(x(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ defined on an interval $t \in [t^-, t^+]$ is called a *regular solution* of (3) if it is differentiable except for at a finite number of points $t^- = t^0 < t^1 < \ldots < t^L < t^{L+1} = t^+$ and satisfy the equation (3) on each open interval $]t^l, t^{l+1}[$ for $l \in \{0, 1, \ldots, L-1\}$.

In restricting our attention to these solutions only, we loose other kind of (singular) solutions, that could exists also, the ones that spend time inside the boundaries of the boxes. But this is singular behavior that we will not consider here.

4 The discrete model

Before introducing the discrete model, let us remark that it is not necessary to start from the piecewise linear model to introduce the discrete model as we will do here. Indeed, even if it was introduced by R.Thomas as a simplification of the (smooth) systems of differential equations used by the biologists, often much too difficult to study by themself, the discrete model have been develop and used successfully as a model of gene regulatory networks without any connection with a piecewise linear model. Nevertheless we will introduce it here, following Snoussi[8] as a discrete version of a piecewise linear model because we do not need to be more general.

4.1 Snoussi's discrete mapping

The first idea, starting with a piecewise linear model (3), is to introduce a directed graph whose vertices are the states $s \in \mathcal{S}$ (one state per box) and whose edges are the transitions $(s \to s')$, where $s' \in \mathcal{S}$ is the state corresponding to the focal point $\phi(s)$ of the box $\mathcal{B}_s$. This application from $\mathcal{S}$ to $\mathcal{S}$ that map each state s to the state s' of its focal point is what Snoussi called the discrete mapping in [8].

But it is easy to understand that this model does not adequately reflect the piecewise linear dynamic except when $\mathcal{B}_s$ and $\mathcal{B}_{s'}$ are two neighboring boxes. Indeed, as soon as the flow exits $\mathcal{B}_s$ to enter the next box, it is driven by a new dynamic and tends to a focal point which is usually no longer $\phi(s)$. This is why Thomas and Snoussi leave this first description, called synchronous, for an asynchronous one, the *State Transition Graph*, or simply *Transition Graph*.

4.2 The transition graph

The transition graph associated with model (3) is the directed graph defined from the previous one with the same vertices $s \in \mathcal{S}$ and with each edge $(s \to s')$ replaced, when s' is not a neighboring state, by one or several edges from s to neighboring states in the following way. Let denote by e^i for any $i \in \{1, \ldots, n\}$ the state having only zero as component except the i^{th} that is equal to 1 and let define $\tau_i^+(s) = s + e^i$ and $\tau_i^-(s) = s - e^i$. In the transition graph each age $(s \to s')$ of the initial synchronous graph will be replaced, except if $s' = s$ or if s' is already a neighboring state, by one edge $(s \to \tau_i^+(s))$ for each i such that the i^{th} component of $\phi(s) - s$ is strictly greater then 1, and one edge $(s \to \tau_i^-(s))$ for each i such that the i^{th} component of $\phi(s) - s$ is strictly lower than -1.

For example, the transition graphs of the two previous examples are given by :

$$\begin{array}{ccc} 01 & \longleftarrow & 11 \\ \uparrow & & \downarrow \\ 00 & \longrightarrow & 10 \end{array} \qquad\qquad \begin{array}{ccc} 01 & \longleftarrow & 11 \\ \downarrow & & \uparrow \\ 00 & \longrightarrow & 10 \end{array} \quad .$$

The first transition graph contains in fact two additional edges, one is $(01 \to 01)$ and the other $(10 \to 10)$, not drawn in the picture.

It appears that, even very simple, transition graphs are useful for the study of gene regulatory network because they keep the main features of the piecewise linear (or smooth) model but they are much more tractable. They probably contribute to the discovery by R.Thomas of the decisive role of negative and positive circuits that remain one of the main result of the theory

of gene regulatory network. Let us recall here this result as a comment on the choice of our two introductory examples. When a gene exerts an influence on the rate of production of a second one who exerts an influence on the production of a third one, and so on, who finally exerts an influence on the first gene itself, they build a network called a *feedback circuit*. There are in fact two kinds of such circuits that have very contrasting roles in the regulatory networks. The presence of the first one, called *positive* because the number of inhibitions is even (as in our first example with two inhibitions), appears to be a necessary condition for multistationarity while the presence of the second, called *negative* because this number is odd (as in our second example), is a necessary condition for the existence of an attractor (such as a limit cycle). These properties have been proved in the mean time (see for example [9],[5],[7]).

5 A correspondence result

In [8], Snoussi studied two simple situations where the information contained in the transition graph are sufficient to deduce, from the knowledge of it only, the dynamic of the original piecewise linear model. The first is the case of a stable equilibrium (if in the transition graph one has an edge $(s \to s)$, then the former system has a stable equilibrium (the converse is obvious) and the second is the case of a particular negative feedback circuit. In this particular case, the focal point of each box $\mathcal{B}_s$ belongs to the next box (the naïve (synchronous) transition graph is already a complete (asynchronous) transition graph.

On the other hand, it is easy to understand that given any path in the transition graph, it is not always true that the flow of the piecewise linear model follows the corresponding sequence of boxes/states. If the transition graph is for example the following:

$$\begin{array}{ccccc} 01 & \longleftarrow & 11 & & 21 \\ \downarrow & & \uparrow & & \uparrow \\ 00 & \longrightarrow & 10 & \longrightarrow & 20 \end{array} .$$

and if we attach our attention to the sequence of edges from 00 to 10 and from 10 to 20, different dynamics can happens for the piecewise linear model. Either some trajectories starting in the box $\mathcal{B}_{00}$ will enter the box $\mathcal{B}_{20}$ after passing through the box $\mathcal{B}_{10}$ (the others will go up to $\mathcal{B}_{11}$) or none of them will reach $\mathcal{B}_{20}$ (all will go up). Thus it is clear on this example that the path $00 \to 10 \to 20$ that exists in the transition graph do not represent well the dynamic of the model. The same conclusion holds for the path $00 \to 10 \to 11$ for which it is either some or possibly all the solutions that are captured by the dynamic of this path. In both cases, the knowledge of the discrete dynamic only do not allow to understand the original dynamic. This is because the transition graph contains two vertices leaving $\mathcal{B}_{10}$, one toward $\mathcal{B}_{11}$ and one toward $\mathcal{B}_{20}$ (as the focal point of the box $\mathcal{B}_{10}$ belongs to $\mathcal{B}_{21}$). This kind of ambiguity are the origin of the problem.

We will state now a sufficient condition on the paths of the transition graph to avoid such an ambiguity. This is the object of the following correspondence result. The sufficient condition is close to the one introduce in [3] (to prove the existence and unicity of a limit cycle), called alignment of the focal points but it is not exactly the same we consider here.

Theorem 5.1 *Consider a piecewise linear model (3), L an integer and let s^1, s^2, ...s^L be a discrete path of length L belonging to the transition graph associated with the model.*

Assume that

1. *each vertex* s^l, $l = 1, \ldots, L-1$, *is* simple, *in the sens that it is only connected with* s^{l+1} *and there is no other edge* $(s^l \to s)$, *for* $s \neq s^{l+1}$, *in the transition graph*
2. *for any edge* $(s^l \to s^{l+1})$ *in the discrete path, the next edge (if any, i.e. if* $l < L-1$*) is not* $(s^{l+1} \to s^l)$

Then for any point $x_0 \in \mathcal{B}_{s^1}$, *not in the boundary of* $\mathcal{B}_{s^1}$, *there is a unique solution of the piecewise linear dynamic* $x(t)$ *defined for* $t \in [0, t^L]$, *with* $x(0) = x_0$, $x(t^L) \in \mathcal{B}_{s^L}$ *and such that* $x(t)$ *will cross successively* $\mathcal{B}_{s^2}$, $\mathcal{B}_{s^3}$, $\ldots$,$\mathcal{B}_{s^{L-1}}$.

Proof Let us first recall that the dynamic of the piecewise linear model is defined in each box $\mathcal{B}_s$ of the phase space by the linear system (6) which is a smooth differential system with well defined solutions in the whole space. As long as a solution stays inside a box, there is no problem for existence and unicity but we need to say what happens for the solution at the boundaries of the boxes. Notice first that any solution of the linear system starting at a point of the boundary of the box either enter the box when t increases or exits the box or neither enter nor exit (possible behavior for points "in the corners"). If all the points of one face, not belonging to the boundary of the face, are initial points of solutions that enter the box, we will call this face an *entrance face*, and same for *exit faces*. Usually, a face of a box is neither an entrance nor an exit face because it contains together initial points of solutions that enter and that exit.

We will show first the following lemma :

Lemma 5.2 *When a vertex* s *of the transition graph is simple (assumption 1 of the theorem), its associated box* $\mathcal{B}_s$ *has a unique exit face and* $2n-1$ *entrance faces unless it contains its own focal point in which case all the faces of the box are entrance faces.*

Proof Let's denote by $s \to s'$ the unique edge starting from the state s. According to the definition of a transition graph, either $s' = s$ or $s' = \tau_i^{\pm}(s)$ for some $i \in \{1, \ldots, n\}$. The case when $s' = s$ corresponds to the case where the focal point of B_s belongs to B_s itself. It has been already noticed by Snoussi that, in this case the focal point is a stable equilibrium of the global dynamic and this implies that all solutions tend to the focal point, staying in B_s. All faces of B_s are then entrance faces.

Now, assume for example that $s' = \tau_i^+(s)$. Let's call the i^{th} *beam* of the box the set defined by the same set of inequalities then the box (5),

$$\sigma(s) \leq x \leq \sigma(s + I_n)$$

except the i^{th} one, $\sigma^i_{s_i} \leq x_i \leq \sigma^i_{s_i+1}$, replaced by $\sigma^i_{s_i} \leq x_i < +\infty$. This beam is an unbounded domain containing the box that also contains its focal point as $s' = \tau_i^+(s)$ because in this case, s' and s differs by only their i^{th} component. As we know the exact solution inside B_s, it is easy to see that each component tends monotonously to the focal point and thus will satisfy all inequalities that define the beam until it leaves the box. This shows that any solution starting in the box will leave it when crossing the face $x_i = \sigma^i_{s_i+1}$, including the solutions starting on any other faces than this one. Thus this face is indeed the only exit face.

This lemma shows the dynamic inside the boxes. The next lemma explain how to stick together the dynamic of two successive boxes.

Lemma 5.3 *Let* $(s^l \to s^{l+1})$ *an edge of the transition graph between two simple vertices (satisfying assumption* 1*) and satisfying also assumption* 2. *Then for any point* $x^l \in \mathcal{B}_{s^l}$, *not in the boundary of* $\mathcal{B}_{s^l}$, *there is a unique regular solution of the piecewise linear dynamic* $x(t)$ *defined for* $t \in [t^l, t^{l+1}]$, *with* $x(t^l) = x^l$, $x(t^{l+1}) \in \mathcal{B}_{s^{l+1}}$.

Proof The solution starting at x^l, well define and unique as long as it stays in B_{s^l}, will exit B_{s^l} through its unique exit face according to assumption 1 and the previous lemma. This exit face is also a face of the next box, $B_{s^{l+1}}$ and the assumption 2 make impossible that it is the exit face of this box too. Thus it is an entrance face of $B_{s^{l+1}}$ and the exit point from B_{s^l} of the solution we consider is then an initial point of a solution of $B_{s^{l+1}}$, well defined and unique. To build the regular solution we want it suffices to put end to end the two solutions until any point $x^{l+1} \in B_{s^{l+1}}$ in order to build a continuous function that is a regular solution.

To prove the theorem, consider any discrete path s^1, s^2, $...s^L$ of length L belonging to the transition graph of the piecewise linear model that satisfy the two assumptions and take any point in the box B_{s^1}. Follow its solution defined in B_{s^1} up to the unique exit face of this box which exists according to the first lemma and which as to be an entrance face of the next box B_{s^2} according to the second lemma. This allows one to define by continuity the solution in a unique way up to a point belonging to B_{s^2}. Starting again from this point, the same argument allows one to define the solution uniquely up to a point of B_{s^3} and so on. As the same reasoning can be repeated for each vertex of the discrete path, the theorem is proved.

6 Computation (using a MAPLE program) of a discrete path given the ordinary differential equation (3) and a starting box

We are concerned in finding out a set of procedures which can provide us automatically whether there exist a closed path or not and given L, it can find out a discrete path of length L starting from a first box. In the MAPLE program we first input the equation (3), all k_i^j's, θ_j^i's, γ_i's and $\mathcal{B} = \prod_{i=1}^n [0, max_i]$. Here, partition of the $\mathcal{B}$ produces several boxes. If one enter as initial condition of a path, any of these boxes and the length L of a path then the program return the path of the Transition graph beginning at the initial box and of length L except if one of the following conditions is not fulfilled for each box along the path:

1. the path exits the space $\mathcal{B}$
2. the hypothesis of being simple (as in the theorem 5.1) is not satisfied
3. the path reaches a stationary box before it reach the length L

The output of this program is like a numerical integration of a differential equation. The program is less precise but together with the theorem it shows the existence of a true trajectory. Hence, it is a tool for analyzing model like (3) of genetic regulatory network.

7 Conclusion

We have shown a sufficient condition for the existence and unicity of regular solutions of a piecewise linear model of gene regulatory network, expressed in terms of its transition graph. It shows that under easy to check assumptions, the considered path of the transition graph is a faithful representation of the piecewise linear flow. It is easy to find examples where these assumptions are not satisfied and nevertheless the conclusion of the theorem remains true. This gives the idea of possible extensions of this correspondence result to some more general cases especially the ones that needs a more careful study as ambiguities exist in the dynamic.

References

[1] J. Ahmad, O. Roux, G. Bernot, J.-P. Comet, A. Richard, Analysing formal models of genetic regulatory networks with delays, *Int. J. Bioinformatics Research and Applications*, 4(3) (2008) 240-262.

[2] G. Bernot, J.-P. Comet, A. RICHARD, J. GUESPIN, Application of formal methods to biological regulatory networks: extending thomas' asynchronous logical approach with temporal logic, *J. Theor. Biol.*, 229(3) (2004) 339-347.

[3] E. Farcot, J.-L. Gouzé, Periodic solutions of piecewise affine gene network models with non uniform decay rates: The case of negative feedback loop, *Acta Biotheor.*, 57 (2009) 429-455.

[4] L. Glass, S. A. Kauffman, The logical analysis of continuous, non-linear biochemical control networks, *Journal of theoretical Biology*, 39(1) (1973) 103-129.

[5] J.-L. Gouzé, Positive and negative circuits in dynamical systems , *Biol. Syst.*, 6 (1998) 11-15.

[6] J.-L. Gouzé, T. Sari, A class of piecewise linear differential equation arising in biological models, *Dyn. Syst*, 17 (2003) 299-316.

[7] M. Kaufman, C. Soulé, R. Thomas, A new necessary condition on interaction graphs for multistationarity, *J. Theor. Biol.*, 248 (2007) 675-685.

[8] E. H. Snoussi , Qualitative dynamics of piecewise-linear differential equations: a discrete mapping approach, *Dyn. Stab. Syst.*, 4(3) (1989) 189-207.

[9] E. H. Snoussi, Necessary conditions for multistationnarity and stable periodicity, *J. Biol. Syst.*, 6 (1998) 3-9.

[10] R. Thomas, Boolean formalization of genetic control circuits, *J. Biol. Syst.*, 42 (1973) 563-585.

[11] J. Tyson, B. Novak, Regulation of the eukariotic cell cycle: molecular antagonism, hysteresis and irreversible transtions, *J. Theor. Biol.*, 210 (2001) 249-263.

Address of the authors:

[1] Lab. J.A. Dieudonne, Department of Mathematics, University of Nice Sophia-Antipolis, 28 Avenue Valrose, 06108 Nice Cedex-2, France

[2] Lab. I3S, UMR 6070 UNS and CNRS, Algorithmes-Euclide-B, 2000 route des Lucioles, B.P. 121, F-06903, Sophia-Antipolis, France

E-addresses:

Francine.DIENER@unice.fr, Aparna.DAS@unice.fr, bernot@unice.fr, comet@unice.fr, Frederic.Eyssette@unice.fr

Emmanuel Isambert était Professeur à l'université de Paris 13, lorsqu'il est décédé brutalement en septembre 2007.

Logicien et mathématicien, spécialiste d'équations différentielles, il était un membre très actif du réseau Georges Reeb.

La rencontre du réseau de décembre 2007 était dédiée à sa mémoire.

La diversité des thèmes abordés dans ces Actes, philosophie, logique et théorie des ensembles, mathématiques financières, probabilités et équations différentielles, histoire des mathématiques, témoigne de l'ouverture d'esprit d'Emmanuel.