Analysis, control and optimisation of PDEs, application to the biology and therapy of cancer

Camille Pouchol

École doctorale ED386

Analyse, contrôle et optimisation d'EDP, application à la biologie et la thérapie du cancer

THÈSE

présentée et soutenue publiquement le vendredi 29 juin 2018

pour l'obtention du

Doctorat de Sorbonne Université

(spécialité mathématiques)

par

Camille Pouchol

devant le jury composé de

Président :	Benoît PERTHAME
Rapporteurs :	Assia BENABDALLAH Roberto NATALINI
Examinateurs :	Franck BOYER Vincent CALVEZ
Directeurs :	Jean CLAIRAMBAULT Michèle SABBAH Emmanuel TRÉLAT

Laboratoire Jacques-Louis Lions, Laboratoire de Biologie et Thérapeutique des Cancers

Mis en page avec la classe thesul.

Remerciements

Je voudrais commencer par manifester toute ma gratitude à mes directeurs de thèse. La curiosité scientifique de Jean Clairambault restera pour moi une source d'inspiration ; c'est grâce à elle et à l'étendue de ses connaissances que j'ai tant appris en modélisation mathématique. Il ne devait pas être facile d'avoir un thésard si matheux que moi, et je suis donc très reconnaissant envers Michèle Sabbah et l'ouverture d'esprit dont témoigne son choix d'encadrer des thèses interdisciplinaires. Le sens de la pédagogie d'Emmanuel Trélat a été source d'émerveillement mathématique : sa capacité à isoler l'essence - l'idée simple mais puissante - d'une technique mathématique, m'a permis d'en assimiler de nombreuses.

Je retiendrai avant tout leurs qualités humaines, eux qui ont été d'une disponibilité exemplaire, ont su m'encourager et très vite me mettre le pied à l'étrier pour l'écriture d'articles et la participation à des conférences. C'est autant de confiance accumulée, cruciale dans mon apprentissage du métier de chercheur. Je m'estime chanceux et les en remercie.

Je remercie aussi Assia Benabdallah et Roberto Natalini d'avoir accepté de rapporter ma thèse, leur relecture attentive a été très précieuse pour moi et pour la qualité de ce manuscrit. Tout comme eux, leur réputation en contrôle et maths-bio précèdent Franck Boyer et Vincent Calvez, et je suis donc très honoré qu'ils participent au jury.

La joyeuse diversité des thèmes abordés durant ma thèse est intimement liée aux nombreuses personnes avec lesquelles j'ai pu travailler. Je pense à Nathalie Ferrand, qui n'a jamais rechigné à une énième expérience, à Alexander Lorz et ses idées à foison sur les équations intégro-différentielles, Markus Schmidtchen qui sera sans aucun doute un grand chercheur. ll y a bien sûr Tommaso Lorenzi et son incroyable gentillesse, avec qui j'ai adoré échanger, notamment lors de mon passage à St Andrews. J'ai récemment pu profiter de l'étendue des connaissances d'Enrique Zuazua en contrôle, ainsi que le temps qu'il aime passer à les transmettre. Enfin, une grande partie des techniques et des équations que j'ai traitées pendant ma thèse n'existerait pas sans Benoît Perthame. C'est un véritable honneur qu'il dirige mon jury.

J'aimerais aussi avoir un petit mot pour certains professeurs marquants, qui m'ont transmis le goût des mathématiques. M. Espéret est indéniablement celui qui m'a appris à faire des HMM et des maths à l'intuition. Franck Boyer et Pierre Bousquet à Marseille ont été déterminants, et, même si je n'ai pas pu l'avoir comme enseignant, Dominique Barbolosi.

J'en viens maintenant au laboratoire, et je remercie chaleureusement Malika, Salima et Catherine pour leur aide bienveillante dans toutes les démarches administratives et relevant des missions, sans oublier non plus Khashayar et Antoine que j'ai dérangés un nombre incalculable de fois pour AMPL ou FreeFem, le numérique dans ma thèse leur doit beaucoup. La gentillesse et la bonne humeur éternelles de Luis m'ont plus d'une fois donné le sourire, c'était une aubaine d'être son voisin. Merci à Grégoire d'avoir pris le temps de répondre à mes questions de maths-bio.

La thèse a rarement été une expérience solitaire, et ce grâce à la joyeuse ribambelle de

doctorants, Carlo, Dena, Anouk, Olivier, Florian, Hugo, Martin, Ana, Léo, Gabriela, Idriss... et j'en oublie certainement, excusez-moi. Ludie, attends un peu, tu n'es pas au LJLL. Amaury est devenu un indéboulonnable, en atteste son alias Abdelaziz Bouteflika... et je lui dois probablement plus de la moitié des chocolats que j'ai mangés ces deux dernières années. Christophe a le jeu de mots extrêmement prometteur.

Encore plus essentiels sont les camarades de bureau, qui ont tenu pour certains trois ans malgré ma folie (dissimulée au reste du labo, bien sûr). Il y a eu avant nous Pierre qui nous a fournis en canapé, tout de même, et à mon arrivée Pierre, Andrada, Thibault cet énergumène, et enfin Maxime (sans qui il n'y aurait tout simplement pas de simulations de contrôle optimal...). La composition actuelle du bureau défie toute concurrence. Alex, c'est l'homme dans (de ?) la Lune, la gentillesse chevillée au corps, le roi des répliques sorties de nulle part, un fin géomètre des oeufs pseudo-convexes. Federica, c'est l'art du café en toute circonstance, quelqu'un avec qui il a fait bon travailler et qui sort une simulation en un rien de temps, une vilva la revolucion qui rassure sur l'Italie de Salvini. Il y a Ludie, sacrée Ludie, qui est bien sûr trop mon amie, le fleuron des mathématiciens-médecins et de la couleur orange. Il aurait été difficile de ne pas terminer par Antoine, ami ébouriffant, mais qui ne s'est pas contenté du contrôle optimal, un compagnon de footing, de tennis, de fruits, de discussions, tout ça pêle-mêle.

Il est temps de sortir du labo pour aller faire un tour du côté des amis qui ont au moins autant apporté à cette thèse que la lettre ε, et en premier lieu la chorale, une vraie colonne vertébrale de mes semaines, avec laquelle nous alignons les projets fous et franchement rigolards. Avec une mention particulière à DiTi et Matthieu de m'avoir fait partir de rien et m'éclater avec ma voix. Il y a aussi les machines, des amis précieux : Geoffroy qui malgré sa bigorexie reste inénarrable, et continue à rendre les sketchs plus drôles qu'ils ne le sont vraiment, Damien qui est si talentueux et qui refuse d'en prendre conscience, quel saltimbanque. Constance et sa gentillesse légendaire.

Ma vie à Paris doit beaucoup à l'accueil si doux de la famille Devanz, Thien, Christian, Gaby et Lou, merci beaucoup. Une deuxième famille en somme, mais toujours soutenu par ma famille, marseillaise, ou d'ici. Je résiste à l'envie de lister tout le monde, mais aller à Marseille et retrouver Ptitbout, Véro, Fab', Marie, Eric, Nono, c'est une bonne dose de bonheur. Ou rester à Paris avec Laurent, Steph et les minotes, Hélène et Jean-Michel qui m'ont tant et si bien accueilli, les cousins-cousines Antoine, Rafaëlle (que je ne vois pas assez mais ne chéris pas moins), Florence (qui n'est pas à Paris, mais je perds le Nord), et bien sûr Mamy Suzanne.

Je ne me lasse jamais de mon exceptionnelle amitié avec Florian, 20 ans au royaume de l'absurde et ce n'est que le début. Ni de ma relation avec mon frère, Saïm, que j'admire et aime tant. Et encore moins de celle avec mes parents, de leur amour si grand et toujours renouvelé. Je leur dois mon goût pour l'abstraction et la transmission et donc, c'est certain, celui pour les maths et leur partage.

Mon dernier mot est pour la douce et belle, la lumineuse, le dernier mot est pour Clara, son mot préféré : merci.

Je dédie cette thèse à Mamy Suzanne,
à Papy Georges,
et à Mamie Ptitbout.

Contents

General introduction

PhD context

This PhD originates from a joint project on chemotherapy optimisation, bringing together three advisors: Jean Clairambault, medical doctor and mathematician, Michèle Sabbah, cancer biologist, and Emmanuel Trélat, mathematician specialised in optimal control. Most of the work undertaken has thus been motivated by questions from cancer biology or therapy.

Answering them has required using and further developing tools from several different mathematical areas, among them the asymptotic analysis for partial differential equations, and theoretical and numerical optimal control. These developments have in turn posed new mathematical problems, interesting in their own right, with applications in the mathematical fields of adaptive dynamics, population dynamics, optimal control or numerical analysis, a classification which roughly corresponds to Parts I, II, III and IV, respectively.

A chronological overview

More than the first half of the PhD has been devoted to continuing and expanding a work initiated by Jean Clairambault, Alexander Lorz and Emmanuel Trélat on the optimisation of chemotherapy. This has led to the development of Lyapunov functionals for the asymptotic analysis of integro-differential systems (Chapter 1), with an application to a model from population dynamics (Chapter 3). It has also raised related questions in adaptive dynamics about the selection of phenotypes (Chapter 2).

Treating the problem theoretically has combined the previous asymptotic results as well as techniques for the optimal control of ODEs (Chapter 5). Solving it numerically was made possible thanks to classical but expert methods from numerical optimal control, and required introducing a new method based on homotopies, in the setting of a more complicated model (Chapter 6).

The second part of the PhD has consisted in a subproject studying problems of controllability for some other models in population dynamics (Chapter 4), while the central matter was motivated by experiments of cells aggregating in a 3D structure. Trying to understand these patterns thanks to a minimal chemotaxis model, the work ranged from 2D simulations exhibiting similar ones and their analysis (Chapter 7), to the development of appropriate numerical schemes for the 1D equations (Chapter 8).

1 Reaction-diffusion and selection-mutation models

1.1 A general model as a basis

All the partial differential equations (PDEs) analysed in this manuscript have in common that they model the dynamics of a population over time, structured by a continuous variable $x \in \Omega$, where Ω is a bounded domain of $\mathbb{R}^d$. It may represent space, in which case we are interested in the spatial dynamics of the population, or a *trait* or *phenotype* variable, in which case the selection dynamics are the object of study. An example of phenotype could be the size of a giraffe's neck, the colour of a given flower, etc.

Hence, the quantity of interest is $n(t,x)$, the density of individuals at time t, at position (or trait) x. It is assumed to satisfy the PDE

$$\frac{\partial n}{\partial t} - \beta \Delta n = f(x,n), \tag{1}$$

where $t > 0$, $x \in \Omega$, starting from an initial density $n^0 \geqslant 0$. The conditions on the boundary $\partial\Omega$ of Ω are either Neumann or Dirichlet boundary conditions.

There are therefore two main ingredients in these equations, either called reaction-diffusion or selection-mutation depending on the context:

- a *reaction* term, rather called a *selection* term in the phenotype case. It describes the death and proliferation of the individuals, and may depend locally or non-locally upon n. The simplest and most classical modelling for this term is undoubtedly

  $$f(x,n) = n(1-n),$$

  which means that there is exponential growth for $n > 0$ with saturation at $n = 1$. Such a logistic hypothesis of saturation is central and will be shared by most choices we shall be considering for f. It models the competition between the individuals by an additional death rate, proportional to n, which will counterbalance the so-called *intrinsic* reaction rate, here equal to 1.

- a *diffusion* term, which models the random movement of individuals in space. If x is a phenotype, it is pinpointed as a *mutation* term, modelling the random possibility (here local and unbiased) for individuals to change phenotype. In biology, mutations are modifications in the DNA, while *epimutations* are heritable changes in DNA expression (but not in DNA itself), the latter being considered to occur on much shorter time-scales. We shall abusively always speak of mutations or *genetic instability*, grouping mutations and epimutations in their biological sense.

We refer to the monographs [11, 130] for an introduction to those models.

The coefficient $\beta \geqslant 0$ has great importance, because we shall consider the case $\beta = 0$, circumstances under which we will say that equation (1) is *integro-differential*.

Additional advection term. A part of this PhD has been devoted to studying more general models, for which there is also an oriented movement of individuals. They tend to prefer the direction of the gradient of a certain function c, potentially nonlinearly through a *sensitivity* function φ. With $\beta > 0$ normalised to 1, the model reads

$$\frac{\partial n}{\partial t} - \Delta n + \nabla \cdot (\varphi(n)\nabla c) = f(x, n), \tag{2}$$

In that case, c can be given a priori on Ω (Fokker-Planck equation), or be representing the concentration of some signal emitted by the individuals themselves. The equation (2) above is then coupled to an elliptic or parabolic equation for v, of the form

$$\frac{\partial c}{\partial t} - \Delta c = n - c. \tag{3}$$

It is a Keller-Segel model, originally introduced to study collective movement of cells emitting a chemical signal, towards zones of higher concentration [128, 88, 89]. Neumann boundary conditions are enforced for n and c, making of the initial mass for n a formally preserved parameter when $f = 0$.

1.2 Modelling differences in the space and phenotype cases

The logistic hypothesis according to which competition between individuals (for resources, space) induces an additional death rate more generally leads to non-local terms of the form

$$f(x, n) = n \left(r(x) - \int_\Omega K(x, y) n(y)\, dy \right), \tag{4}$$

through a kernel K. The intrinsic reaction rate here depends on the variable x, but the important change is that individuals at x do not compete exclusively with those sharing the same x, but also with those at y, with a weight $K(x, y)$.

The kernel K in the space case. When x stands for space, it is quite natural to assume that K is localised, in order to describe that competition should occur mostly between neighbours. K is thus thought of as a regularised Dirac at $x - y$. Equation (1) with $r = 1$ and normalisation $\int_\Omega K(x, y)\, dy = 1$ for each x then becomes

$$\frac{\partial n}{\partial t} - \beta \Delta n = \left(1 - \int_\Omega K(x, y) n(t, y)\, dy \right) n,$$

namely the *non-local Fisher-KPP equation* [12, 71]. At the limit $K(x, y) \to \delta(x - y)$, it indeed simplifies into the well-known Fisher-KPP equation *i.e.*, that with second term $n(1 - n)$ [92].

In another direction, a usual refinement in population dynamics is to model the necessity for a minimal density of individuals for the population to be viable at a given x (for reproduction or cooperation reasons, for example). The nonlinearity is then called *bistable*

(in contrast with the *monostable* case like $f(n) = n(1-n)$). It is negative below a threshold θ, then positive. The most common example is

$$f(n) = n(1-n)(n-\theta).$$

Such bistable models are commonly used in physics (combustion theory, for instance), but also in electrophysiology, e.g., with the FitzHugh-Nagumo equation.

The kernel K in the phenotype case. At variance with the space case, the phenotype case allows for situations where K is not localised. If K is independent of y for example with $K(x,y) = d(x)$, we obtain

$$\frac{\partial n}{\partial t} - \beta\Delta n = (r(x) - d(x)\rho(t))\, n, \tag{5}$$

where

$$\rho(t) := \int_\Omega n(t,x)\, dx$$

is the total number of individuals at time t. The asymptotic behaviour for these equations in the integro-differential case $\beta = 0$ is particularly interesting and mathematically well understood: under general hypotheses, ρ converges to $\rho^\infty := \max\left(\frac{r}{d}\right)$ and n concentrates on $\arg\max\left(\frac{r}{d}\right)$ [129]. If this set were to be discrete, the limit should be a sum of Dirac masses.

This property of concentration on certain specific phenotypes is generally interpreted as the *selection* of these phenotypes. It justifies the attention these models have attracted in *adaptative dynamics*, the branch of mathematical biology dealing with the modelling of selection processes in ecology [49, 51].

1.3 Mathematical tools and properties

Well-posedness. Let us start by mentioning a few results on (1).

When $\beta > 0$ and if the dependence of f on n is local, the standard theory of semilinear parabolic equations applies, and quite generally there are classical solutions locally in time [56]. If f is furthermore logistic (vanishing at $x = 1$), the solutions are even global and remain between 0 and 1 provided that the initial condition itself is, a fact due to the parabolic maximum principle [11].

When f depends non-locally on n under the form (4), and if $\beta > 0$, solutions are still classical and global in time, by a fixed point argument [46]. If $\beta = 0$, however, the regularity inherited from the initial condition cannot be improved but the exponential structure is enough to prove global existence, for example in a model like (5) [129].

Finally, for the Keller-Segel system (2)-(3), and assuming $f = 0$ for simplicity, solutions are in general classical but the advection term (pushing the individuals towards oneanother) might overcome the diffusion term. A remarkable phenomenon is then the existence of a

critical mass for the initial condition, under which solutions are global, and above which there is finite-time blow-up [54, 16].

An option to get rid of blow-up is to choose a logistic sensitivity, creating a threshold density at which the biased movement in the direction of the gradient of c is shut off [75], and more generally there are sharp results on whether solutions are global or not depending on φ [165].

Asymptotic analysis. For equations (1) with a non-local reaction term, and when solutions are defined globally, the typical behaviour is convergence towards stationary states. A striking example of this general principle is the following result, for $\beta > 0$ and $f(x,n) = f(n)$ analytic (semilinear heat equation): any trajectory must converge to a stationary state [156].

The standard method to arrive at these types of results is the construction of energy functionals. When $\beta = 0$ and if $f(x,n) = \left(r(x) - \int_\Omega K(x,y)n(y)\,dy\right)$ is non-local in n, stationary solutions also attract all trajectories. This result holds for quite general kernels and also rests on some Lyapunov functionals [82], on which we shall come back in much more detail.

If $\beta > 0$ and still in the non-local case, however, a complete understanding is lacking. Only the very peculiar situation when K depends only on y has been tackled: the term $\int_\Omega K(x,y)n(y)\,dy$ does not depend on x and can be temporarily gotten rid of by an exponential change of variable [46, 100]. The result is that any initial datum will converge either to 0 or to a multiple of the first eigenfunction of the operator $\beta\Delta + r(x)$, depending on the sign of the first eigenvalue of this operator.

As far as the Keller-Segel (2)-(3) is concerned, energies can be available, paving the way for proving convergence towards certain stationary states [27, 53].

For these models, being able to build numerical schemes preserving energy dissipation is essential, not only because they are entropies or physical energies and thus have an actual physical meaning, but also because it often guarantees that steady states will also be preserved at the discrete level.

Let us mention the related questions for the non-local Fisher-KPP equation. With Neumann boundary conditions and in the purely local case $n(1-n)$, the local Fisher-KPP equation has 1 as a constant stationary state. It attracts all non-zero initial conditions, and the question is whether this steady state remains attractive in the non-local case, since other stationary states than 0 and 1 might exist. Results on 1 being the only steady state other than 0 if K is localised enough have been obtained in the literature [12, 71].

Control and optimal control. Control of models (1) belongs to the theory of control of PDEs, a whole mathematical branch in itself [103, 41]. The control can act on the whole or a part of the domain Ω, or on the whole or a part of the boundary $\partial\Omega$ (Dirichlet or Neumann controls). Controlling PDEs (or ODEs) roughly means being able to steer from a given state to another one in some fixed or free time, by means of appropriately

chosen controls. The literature on the controllability of parabolic semilinear equations is abundant, at least when f depends locally upon n [103, 41, 176].

Let us consider, for example, the semilinear heat equation controlled with controls acting on ω, non-empty open subset of Ω:

$$\frac{\partial n}{\partial t} - \beta \Delta n = f(n) + u(t,x)\mathbb{1}_\omega.$$

Under fairly general hypotheses with f a small enough nonlinearity depending locally on n, then any initial datum can be brought to 0 in arbitrarily small time $T > 0$, with some control $u \in L^2(\omega \times (0,T))$ [94, 55, 57]. The regularising properties of the heat equation are crucial to prove this result, which therefore requires $\beta > 0$.

The above control ensuring controllability to 0 will of course become larger and larger as we take T smaller [58]. Consequently, these results do not apply to the quite natural context where L^∞ constraints are enforced on the control. They also do not apply, at least directly, if the control does not act additively, although multiplicative controls are ominous in population dynamics for example, such as when the control plays the role of an additional death term. Developing controllability techniques in the presence of constraints is an active research problem [132, 107, 108].

A more flexible alternative to controllability when there are constraints on the controls (or on the state), at least towards numerical experiments, is to set an optimal control problem. In general, it writes as the minimisation of some criterion of the form

$$J(u) = \int_0^T f_0(t, n(t), u(t))\, dt + g(T, n(T))$$

among all controls satisfying the constraints prescribed by the problem. When one aims at targeting $\bar{n}$, a standard example is $J(u) = \|n(T) - \bar{n}\|^2$ in an appropriate norm.

Analysing such an optimal control problem from a theoretical point of view is possible after applying a Pontryagin Maximum Principle (PMP) in a well-chosen space [136, 102], although it is not often the case that precise information will be obtained on the optimal control. From a numerical point of view, all PDE-based optimal control problems can at least formally be solved by direct methods [167]. Their principle is simple: the equation and controls are discretised so that the optimal control problem becomes a high but finite-dimensional optimisation problem. Its resolution is thus not yet granted, even with a good algorithm. The computations can indeed be very heavy due to the high dimension, and a good a priori is needed to initiate the algorithm.

2 Motivations and related mathematical questions

2.1 Optimisation of chemotherapy and modelling

Clinical question and resistance in cancer therapy. Most of the mathematical work presented in this manuscript has originates from the following question: how can one use chemotherapy efficiently to try and eradicate (or control the size of) a given tumour? This is a tall order because any strategy will be faced with the two classical pitfalls of cancer therapy, which are

- the toxicity to the healthy tissue,
- the emergence of drug resistance.

The main chemotherapeutic drugs used fall into two main categories

- the *cytotoxic* drugs, which actively kill cancer cells,
- the *cytostatic* drugs, which lower their proliferation,

the former being believed to be triggering drug resistance, because they subject cancer cells to mortal stress, contrarily to cytostatic drugs.

It is indeed commonly seen in the clinic that using so-called maximum tolerated doses (MTD) for too long will eventually

- damage the healthy tissue to a life-threatening extent,
- lead to regrowth of the tumour [72, 127, 149].

There might be an initial decrease of the tumour burden, because cells that are sensitive to the treatment have been killed, while the resistant ones will take over the whole cancer cell population making the tumour insensitive to further intensive treatment [152, 134]

An ecological paradigm to explain drug resistance relies on the idea that *phenotypic heterogeneity* in cancer cells and the dynamics of cancer cell populations can be understood through the principles of Darwinian evolution [65, 67]. Cancer cells indeed all evolve in the *extracellular matrix* (ECM), which is their structural and biochemical support, while, given a particular tumour micro- and macro-environment (*e.g.*, access to oxygen, nutrients, growth factors, drug exposure), the fittest cells are selected. In the case of resistance, resistant cell subpopulations are assumed to emerge and be selected for their high levels of fitness in the presence of chemotherapeutic agents.

An idea gaining popularity is that some cells, called CSC for cancer stem cells, have a higher *plasticity* because they are phenotypically very close to stem cells [153], and as such are much more resistant to drugs. They are further identified as having undergone the epithelial-to-mesenchymal transition (EMT), *i.e.*, they have passed from an epithelial phenotype with strong cell-cell adhesive properties, to a mesenchymal phenotype characterised by increased migrative properties.

Since the EMT in a given cell can for example be assessed by the E-cadherin protein concentration (necessary for junctions between cells), the level of resistance to a drug is better represented by a *continuous* variable. Resistance to the drug can further be correlated to other continuous biological characteristics, *e.g.*, the intracellular concentration of a detoxication molecule (such as reduced glutathione), the activity of detoxifying enzymes in metabolising the administered drug, or drug efflux transporters eliminating the drug.

Consequently, a relevant modelling alternative to the binary sensitive versus resistant ODE framework (as already proposed long ago in e.g., [43, 44]) consists of studying the cells at the population level using a selection-mutation model as those introduced, and we choose as a basis the model (5)

$$\frac{\partial n}{\partial t} - \beta \Delta n = (r(x) - d(x)\rho(t))\, n,$$

with $\rho(t) = \int_0^1 n(t,x)\, dx$. Here, $x \in [0,1]$ stands for the phenotype for resistance, ranging continuously from sensitiveness ($x = 0$) to full resistance ($x = 1$). Note that the diffusion term can be complemented with an advection term [37, 38]. The idea is to model stress-induced adaptation: individuals actively adapt to their environment and this can be thought of as an appropriate modelling of Lamarckian induction. Such generalisations will not be considered in this thesis.

The model. From the previous equation, we build a model for the problem at hand in the form of a system of two coupled non-local PDEs for healthy and cancer cells densities $n_H(t,x)$ and $n_C(t,x)$, given by

$$\begin{aligned}
\frac{\partial n_H}{\partial t}(t,x) &= \left[\frac{r_H(x)}{1+\alpha_H u_2(t)} - d_H(x) I_H(t) - u_1(t)\mu_H(x)\right] n_H(t,x) + \beta_H \Delta n_H(t,x), \\
\frac{\partial n_C}{\partial t}(t,x) &= \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x) I_C(t) - u_1(t)\mu_C(x)\right] n_C(t,x) + \beta_C \Delta n_C(t,x),
\end{aligned} \tag{6}$$

starting from an initial condition (n_H^0, n_C^0), with Neumann boundary conditions in $x = 0$ and $x = 1$. This model and variants are introduced in [111].

Let us describe in more details the different terms and parameters appearing above, with the functions r_H, r_C, d_H, d_C, μ_H μ_C all continuous and non-negative on $[0,1]$, with r_H, r_C, d_H, d_C positive on $[0,1]$.

- The terms $\frac{r_H(x)}{1+\alpha_H u_2(t)}$, $\frac{r_C(x)}{1+\alpha_C u_2(t)}$ stand for the selection rates lowered by the effect of the cytostatic drugs, with
$$\alpha_H < \alpha_C.$$

- The non-local terms $d_H(x) I_H(t)$, $d_C(x) I_C(t)$ are added death rates to the competition inside and between the two populations, with

$$I_H := a_{HH}\rho_H + a_{HC}\rho_C,\ I_C := a_{CC}\rho_C + a_{CH}\rho_H$$

and as before

$$\rho_i(t) = \int_0^1 n_i(t,x)\,dx,\ i = H, C.$$

We make the important assumption that the competition inside a given population is greater than between the two populations:

$$a_{HC} < a_{HH}, \qquad a_{CH} < a_{CC}.$$

- The terms $\mu_H(x)u_1(t)$, $\mu_C(x)u_1(t)$ are additional death rates due to the cytotoxic drugs. Owing to the meaning of $x = 0$ and $x = 1$, μ_H and μ_C are taken to be decreasing functions of x.
- The terms $\beta_H \Delta n_H(t,x)$ and $\beta_C \Delta n_C(t,x)$ model the random mutations, with their rates β_H, β_C such that

  $$\beta_H < \beta_C,$$

  because cancer cells have higher genetic instability than healthy cells.

We first require the model to be rich enough to recover that high constant doses are deleterious in the long-run, both on the healthy cell count and on the tumour itself. The focus is not on transient behaviours, although numerically, it is also of interest to see whether the model is able to reproduce an initial drop in the tumour size before it starts growing again.

It is therefore natural to tackle this aspect from the point of view of asymptotic analysis for the previous system, with constant doses of drugs. For these systems, the existing techniques for asymptotic analysis (developed for a single equation) do not work, as we shall see.

The optimal control problem. Looking for alternative strategies where the infusion rates are now allowed to vary, we have in mind alternatives that are currently being extensively investigated by oncologists, e.g., metronomic scheduling, which relies on frequent and continuous low doses of chemotherapy [10, 29, 127].

For a fixed final time T, we consider the following optimal control problem, denoted in short by ($\mathbf{OCP_1}$). It consists in minimising the number of cancer cells at the end of the time-frame

$$\inf\ \rho_C(T)$$

as a function of the L^∞ controls u_1, u_2 subject to L^∞ constraints for the controls and two state constraints on (ρ_H, ρ_C), for all $0 \leqslant t \leqslant T$:

- The maximum tolerated doses cannot be exceeded:

  $$0 \leqslant u_1(t) \leqslant u_1^{max},\ \ 0 \leqslant u_2(t) \leqslant u_2^{max}.$$

- The tumour cannot be too big compared to the healthy tissue:

  $$\frac{\rho_H(t)}{\rho_H(t) + \rho_C(t)} \geqslant \theta_{HC}, \tag{7}$$

with $0 < \theta_{HC} < 1$.

- Toxic side-effects must remain controlled:

$$\rho_H(t) \geqslant \theta_H \rho_H(0), \tag{8}$$

with $0 < \theta_H < 1$.

In other words, having fixed some therapeutic time-window $[0, T]$, we consider all possible infusion protocols below the maximum tolerated doses, while we want to keep the tumour size below a critical size (relatively to the healthy cell population) and curb damages to the healthy tissue.

2.2 Transcription into mathematical words and related difficulties

On the asymptotic analysis. For the theoretical study of the model (6), we start by assuming $\beta_H = \beta_C = 0$. Understanding the asymptotic behaviour of the 2x2 resulting integro-differential system for constant doses boils down (after renaming functions) to that of models of the form

$$\frac{\partial n_1}{\partial t}(t,x) = \big(r_1(x) - d_1(x) I_1(t)\big) n_1(t,x),$$
$$\frac{\partial n_2}{\partial t}(t,x) = \big(r_2(x) - d_2(x) I_2(t)\big) n_2(t,x).$$

where the coupling is competitive and appears only in the non-local term

$$I_1 = a_{11}\rho_1 + a_{12}\rho_2, \; I_2 = a_{22}\rho_2 + a_{21}\rho_1$$

with $\rho_i(t) = \int_0^1 n_i(t,x)\, dx$, $i = 1, 2$.

We recall that that the asymptotic behaviour for a single integro-differential equation in the previous form

$$\frac{\partial n}{\partial t}(t,x) = \big(r(x) - d(x)\rho(t)\big)\, n(t,x),$$

is well summed up by the statements that

- ρ converges to $\rho^\infty = \max \frac{r}{d}$,
- n concentrates on the set $\arg\max(\frac{r}{d})$.

There is no difficulty in formally understanding the above formulae: assume for a moment that ρ converges to some $\bar{\rho} > 0$. The asymptotic behaviour of $n(t, x_0)$ for some fixed x_0 is then exponential, with rate arbitrarily close to $r(x_0) - d(x_0)\bar{\rho}$ as time goes by. On the one hand, if it were to be a positive quantity for some x_0, there would be exponential blow-up of n in a neighborhood of x_0. Consequently, ρ would blow-up as well, a contradiction with its convergence. On the other hand, if $r(x) - d(x)\bar{\rho} < 0$ on the whole $[0, 1]$, then n would converge to 0 exponentially, uniformly on $[0, 1]$. This also contradicts the convergence of ρ to a positive limit.

Collecting these results, we have obtained $r(x) - d(x)\bar{\rho} \leqslant 0$ for all x, where equality must hold at least for some phenotypes. For other phenotypes (namely those satisfying $r(x) - d(x)\bar{\rho} < 0$), the previous reasoning shows that all the mass has vanished asymptotically. Thus any mass left must be where equality holds. This proves that $\bar{\rho}$ is uniquely determined as the smallest positive real such that $r(x) - d(x)\bar{\rho} \leqslant 0$, *i.e.*, $\bar{\rho} = \max \frac{r}{d} = \rho^\infty$, but also that n concentrates on the set $\arg\max\left(\frac{r}{d}\right)$.

The difficulty is to show the convergence of ρ. The classical approach is to prove that ρ is BV on $[0, +\infty)$. The intuition is that for the logistic ODE $\rho' = \rho(1-\rho)$, ρ will converge increasingly towards its carrying capacity 1 if the initial condition is below 1. For the integro-differential equation, the monotony is lost but it still holds true that ρ "does not decrease too much": it satisfies that its derivative has integrable negative part $(\rho')_- \in L^1(0, +\infty)$. Since ρ is bounded from above, it is BV and hence must converge. This is detailed in Appendix 1.5 of Chapter 1.

This approach, however, fails for a competitive system and one must thus look for an alternative strategy, extendable to systems. Such results have remained scarce, unless some particular structure is available [25].

On the optimal control. The optimal control problem ($\mathbf{OCP_1}$) is hard to solve for two main reasons: it is infinite-dimensional and has state constraints. Solving it either numerically or theoretically is therefore a challenging task.

On the theoretical side, Pontryagin Maximum Principles (PMP) exist even when there are state constraints [174], but the difficulties in analysing the resulting equations are numerous:

- Lagrange multipliers associated with the constraints are measures, and it is in general difficult to ensure that these measures have no singular continuous part,
- as the system is a PDE, the adjoint equation is a PDE as well and it also makes it more difficult to analyse,
- more generally, even if the constraints were to be neglected in the first place, the complexity of the equations leads to intractable computations.

On the numerical side, indirect methods rely on a PMP, and as stressed above the constraints are not amenable for it to be treated efficiently [166]. Thus, we restrict ourselves to direct methods which consists in discretising the whole problem (in the variable t and in the variable x) so that the optimal control problem is approximated by a finite-dimensional optimisation problem. These problems are in general not easily dealt with because they are non convex. There are, however, some specific difficulties related to the complexity of the equations (of PDE nature and with state constraints):

- discretisation in both variables lead to a very high-dimensional optimisation problem, making the use of expert optimisation routines compulsory.
- the best optimisation algorithms in a nonconvex setting require a good initial guess, which is all the more difficult that the problem itself is.

2.3 Cells forming spheroids in a 3D structure

Biological experiments. In the Laboratoire de Biologie et Thérapeutique des Cancers of INSERM at Saint-Antoine Hospital, Michèle Sabbah and Nathalie Ferrand have performed experiments consisting in putting cells in a 3D hydrogel structure mimicking the actual environment they encounter in vivo, the extracellular matrix, in sharp contrast with usual 2D in vitro experiments in Petri dishes. The results strikingly show that breast cancer cells tend to organise spatially as numerous spheroids in the hydrogel, see Figure 1 below.

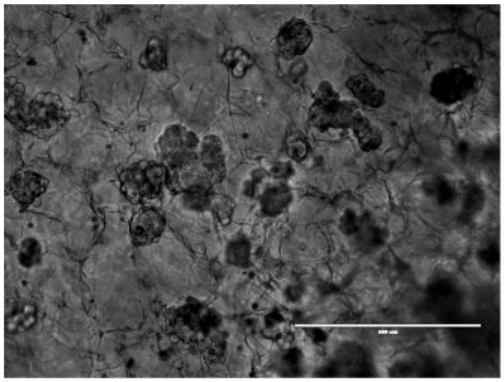


FIGURE 1: 2D image of hydrogel showing spheroids formed by cells from a breast cancer cell line (MCF7) after 8 days of culture, courtesy of N. Ferrand and M. Sabbah.

The interest for these patterns is at least twofold

- the ability of cells to move and aggregate is very much linked to the occurrence of metastases, because it is now very well documented that migrating cells responsible for metastases are not alone, but rather organise as clusters [80],
- in breast cancer, cancer cells invading the surrounding adipose tissue are known to give rise to nontrivial spatial structures.

Turing instabilities. Starting from the seminal work of Turing [168], the so-called *Turing patterns* have emerged as a fundamental tool to explain patterns in biology, such as stripes on the bodies of animals. Mathematically, these are exhibited by reaction-diffusion equations for which, following the definition of Perthame [130], the three following ingredients all coexist regarding some steady state:

- extinction and blow-up are impossible,
- the steady state is linearly unstable,
- the unstable modes have bounded frequencies.

Contrarily to other works [125, 175], we do not enter into the physics of the cell-structure adhesion. Assuming that cells in this hydrogel can move randomly in all directions (provided that the hydrogel is loose enough and isotropic), we choose a simple diffusion term.

Because they also emit chemokines attracting other cells, such as CXCL12 or CXCL8 for breast cancer cells [119], we instead consider chemotaxis to be the main feature and focus on a minimal Keller-Segel model. We also take growth into account since the hydrogel is put in a renewed high glucose medium.

Mathematical model. More precisely, we are interested in Keller-Segel models writing in dimensionless form as

$$\begin{cases} \frac{\partial n}{\partial t} - \Delta n + \nabla \cdot (\varphi(n) \nabla c) = n, \\ \varepsilon \frac{\partial c}{\partial t} - \Delta c = n - c. \end{cases} \tag{9}$$

posed on Ω with Neumann boundary conditions. Here, $\Omega \subset \mathbb{R}^3$ is the hydrogel, a cylinder of small height. Thus, taking $\Omega \subset \mathbb{R}^2$ to be a disk is a reasonable approximation both for simulations and analysis, while we shall also consider an interval of $\mathbb{R}$ for the analysis of numerical schemes.

For the sensitivity function φ, we have in mind the classical case $\varphi(n) = n$ but also $\varphi(n) = n(1-n)$ to prevent chemotaxis from occurring above a critical density 1, or the more flexible nonlinearity $\varphi(n) = ne^{-n}$.

For the model (9), several questions are in order:

- can it exhibit Turing patterns, and if so, what are the key parameters responsible for the emergence of patterns, their size and number?
- do the Turing patterns qualitatively match the structures experimentally obtained?
- how can one explain discrepancies between the model outcomes and the experiments?

Turing instabilities, steady states and numerical schemes.

For these types of models, Turing patterns have been shown to exist but the dynamics can be quite intricate [126, 53]. For example, when $\varphi(n) = n(1-n)$, structures with high density (close to 1) may arise (while n is close to 0 on the rest of the domain), but they tend to merge slowly on very long time-scales [137]. These instabilities thus seem stable on relevant biological scales and are pinpointed as *metastable* in the literature. As a consequence, understanding the possible steady states (which might attract trajectories) is not enough because they would be obtained only on non-realistic time-scales.

There are energies for these Keller-Segel equations (for certain specific functions φ), which sometimes are enough to prove convergence towards stationary states [27, 53]. Mathematically, relying on efficient algorithms preserving energy decrease, pertinent bounds for the problems and steady states is also of great importance, in particular when it comes to the complex question of discriminating between actual steady states, metastable ones, or "wrong" steady states due to poor discretisation.

Such schemes have been developed in [30], where finite-volumes schemes adapted to the gradient flow structure of the equations are designed and shown to preserve energy dissipation and positivity, at the semi-discrete level (*i.e.*, after discretisation of the space variable) but not all the way to the discrete level.

3 Mathematical results and applications

3.1 Asymptotic analysis

From ODEs to integro-differential equations. The approach we have developed for the asymptotic analysis of systems of integro-differential equations is based on some type of Lyapunov functionals, which were introduced in this context by Jabin and Raoul [82]. A good introduction (and probably, how the intuition came to the aforementioned authors) for these functionals is to consider the equation of interest on some set $\Omega \subset \mathbb{R}^d$

$$\frac{\partial n}{\partial t}(t,x) = \left(r(x) - \int_\Omega K(x,y) n(t,y)\, dy \right) n(t,x), \tag{10}$$

and discretise it formally in x through $r_i = r(x_i)$, $a_{ij} = K(x_i, x_j)$, $n(t,x_i) = y_i(t)$, which yields

$$\frac{dy_i}{dt}(t) = \left(r_i - \sum_{j=1}^N a_{ij} y_j(t) \right) y_i(t), \quad i = 1, \dots, N. \tag{11}$$

The equation (11) falls into the category of *Lotka-Volterra ODEs*, where $A = (a_{ij})$ is the so-called interaction matrix [64, 63].

Assume that this equation has a steady state $(y_i^\infty)_{1 \leqslant i \leqslant N} > 0$ and consider

$$V_d := \sum_{j=1}^N \lambda_j \left((y_j - y_j^\infty) - y_j^\infty \ln \left(\frac{y_j}{y_j^\infty} \right) \right) \geqslant 0,$$

where the $\lambda_j > 0$ are to be chosen.

Computing the derivative along trajectories, we find

$$\frac{dV_d}{dt} = -\frac{1}{2} u^T \left(A^T D + D\, A \right) u,$$

where $D = diag(\lambda_i)$, $A = (a_{ij})$, $u = y - y^\infty$. From the classical theorems from the Lyapunov theory, we arrive at a Theorem due to Goh, dating back to 1977.

Theorem ([64])**.** Assume that there is the ODE system has a steady state $(y_i^\infty)_{1 \leqslant i \leqslant N} > 0$. Then if there exists a diagonal matrix D with positive entries such that $A^T D + DA > 0$, y^∞ attracts all non-negative solutions other than 0.

Knowing such a result, it is a quite natural idea to consider its infinite-dimensional counterpart *i.e.*, for a given steady state n^∞ of (10), the following function

$$V(t) := \int_\Omega m(x) \left(n(t,x) - n^\infty(x) - n^\infty(x) \ln \left(\frac{n(t,x)}{n^\infty(x)} \right) \right) dx, \tag{12}$$

with $m \geqslant 0$ a weight to be chosen.

Computing the derivative formally, we find

$$\frac{dV}{dt} = -\frac{1}{2}\int_{\Omega^2} \left(m(x)K(x,y) + m(y)K(y,x)\right)\left(n(t,x) - n^\infty(x)\right)\left(n(t,y) - n^\infty(y)\right)\, dx\, dy.$$

The energy dissipates if there exists a weight m such that $(x,y) \mapsto m(x)K(x,y) + m(y)K(y,x)$ is a positive definite kernel (the infinite-dimensional version of the ODE hypothesis). This should be enough to conclude with some minor technical difficulties due to the infinite dimension.

Dealing with the Dirac masses. As noticed already by Jabin and Raoul, there is actually a hidden difficulty in the well-posedness of the function V itself: recall that for some kernels the asymptotic behaviour should be convergence to Dirac masses! The term $n^\infty \ln(n^\infty)$ does not make sense in these cases and the formula (12) is not even defined. At first glance, this seems like a reasonable hurdle because this term does not depend on time and, as such, can be removed without changing the (possible) decreasing behaviour of V along trajectories.

The main difficulty comes from the fact that V is then no longer a non-negative function, and one has to focus instead on the dissipation $\frac{dV}{dt}$ itself, trying to prove that it tends to 0, at the expense of further computations to estimate the second derivative of V.

In the specific case of $K(x,y) = d(x)$ which we are interested in, this Lyapunov functional happens to be perfectly fitted. Indeed, choosing any measure n^∞ concentrated on $\arg\max\left(\frac{r}{d}\right)$ and of mass $\int_\Omega n^\infty = \rho^\infty = \max\left(\frac{r}{d}\right)$, and with $m = \frac{1}{d}$, we find

$$\frac{dV}{dt} = -(\rho - \rho^\infty)^2 + \int_\Omega m(x)\left(r(x) - d(x)\rho^\infty\right) n(t,x)\, dx =: I_1 + I_2.$$

Both terms are non-positive: the first one is linked to the convergence of ρ, while the second accounts for the concentration of n. The Lyapunov functional naturally decouples the two phenomena.

Estimating V from below and its derivative in the spirit of Jabin and Raoul's work, there holds that $\frac{dV}{dt}$ vanishes asymptotically, and so must the two-terms I_1 and I_2. In particular, ρ converges to ρ^∞ and concentration follows from the reasoning already given earlier.

Extension to systems. With these functionals perfectly suited for our interest, we turn our attention towards systems, motivated by the 2x2 one for cancer and healthy cells, and more generally in any dimension, in the form

$$\frac{\partial}{\partial t} n_i(t,x) = \left(r_i(x) - d_i(x)\sum_{j=1}^N a_{ij}\rho_j(t)\right) n_i(t,x), \quad i = 1,\dots,N. \tag{13}$$

For these equations, if there were to be convergence and concentration, then as for a single equation, the limits for ρ_i and the sets on which n_i concentrate are unambiguously defined

(provided that A is invertible): more precisely, a limit ρ^∞ for ρ must satisfy for each i $r_i(x) - d_i(x)(A\rho^\infty)_i \leqslant 0$ with equalities for some phenotypes, in other words:

$$A\rho^\infty = \left(\max \frac{r_i(x)}{d_i(x)}\right)_{1\leqslant i\leqslant N}$$

from which a formula follows after inversion.

The functional is now a mix between the ODE one and the infinite-dimensional one, and reads

$$V_s(t) := \sum_{i=1}^{N} \lambda_i \int_\Omega m_i(x) \left[n_i^\infty(x) \ln\left(\frac{1}{n_i(t,x)}\right) + (n_i(t,x) - n_i^\infty(x)) \right] dx,$$

with $m_i = \frac{1}{d_i}$ and the $\lambda_i > 0$ to be chosen, and where the n_i are measures with appropriate mass and support. Their use is presented in Chapter 1.

With this functional, we have proved in [140]

Theorem 0.1 (P. and Trélat)**.** *Assume that there exists a diagonal matrix D with positive entries such that DA is symmetric positive definite. Then, convergence and concentration hold for* (13).

The requirement that DA be symmetric is quite restrictive as soon as $N \geqslant 3$ because it translates into some polynomial constraints on the coefficients of A. However, it essentially covers all interesting cases in dimension $N = 2$, which was our initial motivation for the cancer and healthy cells and the first result we proved and used in [139], see Chapter 5.

Finally, we have extended the BV technique to the system (13) in Section 1.4 of Chapter 1, when the system is cooperative (the non-diagonal coefficients of A are non-positive). In this particular setting, comparison principles for ODEs can be used and the BV approach still yields interesting results.

A case with diffusion. To apply such Lyapunov functionals as (12) to a case where there is the Laplacian has one advantage: steady states are smooth and not measures so that the term $n^\infty \ln(n^\infty)$ makes perfect sense. However, a new term appears in the dissipation rate of the functional (12) which has no sign in general and it is an open problem to understand how to compensate it with the first one. This term can however be given a sign if handled properly by choosing the weight in front as $m = n^\infty$.

Such a choice is motivated by the usual entropies for linear parabolic equations, which take the form

$$\int_\Omega n^\infty \phi\, H\left(\frac{n}{n^\infty}\right),$$

where H is convex and ϕ is the solution of the adjoint elliptic problem [130]. In our case, the operator is self-adjoint and $H(z) := z - 1 - \ln(z)$, which gives exactly

$$V(t) = \int_\Omega n^\infty(x) \left(n(t,x) - n^\infty(x) - n^\infty(x) \ln\left(\frac{n(t,x)}{n^\infty(x)}\right)\right) dx.$$

The dissipation of (12) is now

$$\frac{dV}{dt} = -\int_{\Omega^2} n^\infty(x)K(x,y)\left(n(t,x)-n^\infty(x)\right)\left(n(t,y)-n^\infty(y)\right)\,dx\,dy \\ -\beta\int_\Omega \frac{(n^\infty(x))^4}{n^2(t,x)}\left|\nabla\left(\frac{n(t,x)}{n(x)}\right)\right|^2\,dx.$$

Under the assumption that $(x,y)\mapsto n^\infty(x)K(x,y)+n^\infty(y)K(y,x)$ is positive semi-definite, we can conclude. The hypothesis nonetheless remains implicit and therefore not usable as is, because we do not know n^∞. There is an important case, however, where this is explicit, that is $n^\infty = 1$ and then we require that K be positive definite. The equation is then indeed the non-local Fisher-KPP equation (under the additional hypothesis $\int_\Omega K(x,y)\,dy = 1$ for all x).

Difficulties arise if this equation is posed in the whole $\mathbb{R}^d$, but everything works smoothly in bounded domains with Neumann boundary conditions, a fact developed in a note [138] and presented in Chapter 3:

Theorem 0.2 (P.)**.** *If K is positive semi-definite, the state 1 attracts all non-negative non-zero solutions.*

We believe this condition on K to be highly relevant, even on $\mathbb{R}^d$ where the use of this Lyapunov functional is an open problem. This is because when $K(x,y) = \phi(x-y)$ is a convolution, the condition that it should be a non-negative definite kernel on $\mathbb{R}^d$ is essentially equivalent to ϕ having a positive Fourier transform. This condition has already proved to be sufficient for 1 to be the only stationary state other than 0 for the equation, when $d = 1$ [12].

Refining the asymptotic analysis for Dirac masses. A somewhat puzzling feature of integro-differential equations with a kernel $K(x,y) = d(x)$ is that the limit n^∞ is not necessarily unique and there is typically a non countable set of steady states. Assume that the set on which it is supposed to be concentrated is made of two phenotypes x_1 and x_2. Then we have a one-paremeter family of steady states given by

$$\rho^\infty\left(\alpha\delta_{x_1} + (1-\alpha)\delta_{x_2}\right),\ 0\leqslant\alpha\leqslant 1,$$

and the actual limit of n among them depends on the initial condition, so that we cannot say more.

It is then reasonable to wonder whether this kind of degeneracy is due to neglecting the mutation term. In other words, can we recover some uniqueness by instead considering

$$\frac{\partial n_\varepsilon}{\partial t} - \varepsilon\Delta n_\varepsilon = \left(r(x) - d(x)\rho_\varepsilon(t)\right)n_\varepsilon,$$

with some small parameter ε? This question is addressed in Chapter 2.

The goal is then to analyse the asymptotic behaviour of the equation above as t goes to $+\infty$, and assuming that there is a unique limit n_ε^∞, investigate the limit of n_ε^∞ as ε goes

to 0. We are able to do it only if d is constant, and we normalise it to 1. The important quantities are now $\max(r) =: \rho^\infty$ which is the limit of ρ, and $\arg\max(r)$ into which the support of any limit point of $n(t,\cdot)$ must be contained.

When $d = 1$ is constant, there is no particular difficulty in proving that, for ε small enough and $\max(r) > 0$, n_ε converges to a multiple of the first eigenvector ψ_ε of the operator $\varepsilon\Delta + r$, namely $(-\lambda_\varepsilon)\psi_\varepsilon$ where λ_ε is its first eigenvalue and ψ_ε is normalised by $\int_\Omega \psi_\varepsilon = 1$ [46, 100].

Using the Rayleigh quotients, it is possible to prove that the approach is coherent, namely that the limit points of the family $((-\lambda_\varepsilon)\psi_\varepsilon)$ are measures with total mass $\max(r) = \rho^\infty$ and support included in $\arg\max(r)$, and this is totally independent of the initial condition. Thanks to this result, we are thus interested in the behaviour of n_ε^∞ as ε goes to 0 in place of that of n as t goes to $+\infty$. This approach will have brought something if we are able to say more on the support of the limit.

To go further, we must understand more deeply the behaviour of the first eigenfunction (and first eigenvalue) of the operator $\varepsilon\Delta + r$ as ε tends to 0. Of course, there are some cases where symmetry at the level of the data (the function r and the domain Ω) is inherited by the eigenfunction, so that any limit of the first eigenfunction must be symmetric, which will typically yield uniqueness at the limit $\varepsilon \to 0$.

In the absence of symmetry, much more can still be said: interestingly, this question had attracted a lot of attention in semi-classical analysis where $-r = V$ is a potential with several local minima at the same height, and the question is to determine where this particle will be found in the limit of small noise [155]. Translating results obtained in this community [79], the support for limit points of the eigenfunction are narrowed down. A formal and working statement is

Theorem 0.3 (Lorenzi and P.)**.** *Among the set* $\arg\max(r)$*,* n_ε^∞ *must concentrate on the phenotypes for which* r *has the lowest concavity.*

For example, on $[0,1]$ and if r has a two maxima at x_1 and x_2 in $(0,1)$ with $-r''(x_1) < -r''(x_2)$, then n_ε^∞ converges to $\rho^\infty \delta_{x_1}$ as ε tends to 0.

3.2 Optimal control for chemotherapy optimisation

A motivation from the asymptotic analysis. In the particular case of the system for healthy and cancer cells (6) with $\beta_H = \beta_C = 0$ (neglecting the mutations), the result for systems of integro-differential equations applies. It means that, under constant infusion of drugs $\bar{u}_1$, $\bar{u}_2$, the total number of cells ρ_H and ρ_C converge to some ρ_H^∞, ρ_C^∞ while the densities n_H and n_C concentrate on some sets, all of them being computable from the data and $\bar{u}_1$, $\bar{u}_2$.

It these sets are reduced to singletons x_H^∞, x_C^∞ whatever the doses, we have a mapping between $\bar{u}_1$, $\bar{u}_2$, and the resulting asymptotic number of cells ρ_H^∞, ρ_C^∞, as well as the phenotypes x_H^∞, x_C^∞ on which the cell densities have concentrated.

This provides a nice framework showing that our model reproduces the failure of giving high doses: simulations show that the cancer cell count first decreases and then increases to its asymptotic value. This is because the cancer cell density has converged to a weighted Dirac at x_C^∞ which is very close to 1, a very resistant phenotype, meaning that after some time, the treatment becomes inefficient.

This is also a motivation for the analysis of the optimal control problem (**OCP$_1$**) in detail, which we did in [139] for $\beta_H = \beta_C = 0$ (mutations are neglected). This work is presented in Chapter 5.

Theoretical optimal control: a smaller class of controls. As already emphasised, obtaining quantitative results out of a PMP applied to (**OCP$_1$**) is out of reach. We instead provide several heuristics for related optimal control problems on simpler ODE models in Section 5.3 of Chapter 5, all suggesting that, when the final time T is large, the optimal control strategy will first consist in taking constant controls on a long phase to ensure concentration on resistant phenotypes.

Therefore, we proceed to the analysis of the optimal control problem in a (much) smaller but still rich class of controls

$$\mathcal{B}_T := \left\{ (u_1, u_2),\ (u_1(t), u_2(t)) = (\bar{u}_1, \bar{u}_2) \text{ on } (0, T_1),\ T - T_1 \leqslant T_2^M \right\},$$

namely controls that take some constant values on a (long) first phase, and then switch to any controls on a second (short) phase, and we aim at doing so in the limit $T \to +\infty$. Finding the optimal control amounts to determining

- the constant values $(\bar{u}_1, \bar{u}_2)$ for the controls on the long phase $(0, T_1)$,
- the L^∞ controls (u_1, u_2) on the short phase $(T - T_1, T)$,
- the length of second phase $T - T_1 \leqslant T_2^M$.

The results from the asymptotic analysis then come in very handy: at T_1, denoting x_H^∞ and x_C^∞ the phenotypes on which the healthy cells and cancer cell densities have concentrated, there holds $n_H(t, \cdot) \simeq \rho_H(t)\delta_{x_H^\infty}$, $n_H(t, \cdot) \simeq \rho_C(t)\delta_{x_C^\infty}$ at the limit $T \to +\infty$. Furthermore, ρ_H and ρ_C are arbitrarily close to solving the ODE system:

$$\begin{aligned} \frac{d\rho_H}{dt} &= \left[\frac{r_H(x_H^\infty)}{1 + \alpha_H u_2(t)} - d_H(x_H^\infty) I_H(t) - u_1(t)\mu_H(x_H^\infty) \right] \rho_H(t), \\ \frac{d\rho_C}{dt} &= \left[\frac{r_C(x_C^\infty)}{1 + \alpha_C u_2(t)} - d_C(x_C^\infty) I_C(t) - u_1(t)\mu_C(x_C^\infty) \right] \rho_C(t), \end{aligned} \tag{14}$$

on the interval $[T - T_1, T]$.

Theoretical optimal control: the optimal structure in the smaller class. For the previous ODE system (14), minimising $\rho_C(T)$ with the state constraints requires lengthy but manageable computations, if one assumes that the Lagrange multipliers associated with the constraints, which are measures, have no singular continuous part. And indeed

we can prove that for the ODE system, the optimal strategy is such that the last three arcs must be

- a boundary arc along the constraint (7) on $\frac{\rho_H}{\rho_H+\rho_C}$,
- a free arc with controls $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$,
- a boundary arc along the constraint (8) on ρ_H, with $u_2 = u_2^{max}$.

All the previous results put together yield the optimal strategy in the set of controls $\mathcal{B}_T$ as T goes to $+\infty$, which informally reads as follows, and holds under various technical hypotheses.

Theorem 0.4 (Clairambault, Lorz, P. and Trélat)**.** *Asymptotically in T there exists an asymptotically optimal solution to ($\boldsymbol{OCP_1}$) in $\mathcal{B}_T$. More precisely, there exists a control strategy which minimises $\rho_C(T)$ up to an error vanishing as T goes to $+\infty$, and such that the trajectory on $(T-T_1, T)$ is arbitrarily close to the concatenation of at most three arcs:*

- *a quasi-boundary arc along the constraint (7),*
- *a free arc with controls $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$,*
- *a quasi-boundary arc along the constraint (8), with $u_2 = u_2^{max}$.*

Numerical optimal control problem: direct methods and homotopies. The numerical resolution of the optimal control problem is here made through a direct method, thanks to a discretisation both in time and in the variable x, described by respective number of points N_t and N_x. It leads, as explained in detail in Section 6.3 of Chapter 6, to a complex nonlinear constrained optimisation problem with about $2N_tN_x$ variables, which we denote $\mathcal{P}_1$. Even efficient algorithms will fail for N_t and N_x large because they require a good initial guess.

To overcome this, our general approach is to perform a homotopy, the principle of which is simple: we wish to find a much simpler problem $\mathcal{P}_0$ which can be linked to $\mathcal{P}_1$ by a series of optimisation problems $(\mathcal{P}_\lambda)$ where λ ranges from 0 to 1.

Assuming that $\mathcal{P}_0$ is simple enough for the optimisation algorithm to converge regardless of the initial value, this yields the homotopy algorithm

- solve $\mathcal{P}_0$, and set $\lambda = 0$
- while $\lambda \leqslant 1$,

  $\lambda \leftarrow \lambda + d\lambda$.

  Solve $\mathcal{P}_{\lambda+d\lambda}$ with the solution of $\mathcal{P}_\lambda$ as initial guess.

For the specific problem we are dealing with, we use the modelling language AMPL [60] and the optimisation routine IpOpt [173], together with a homotopy on N_t and N_x, starting from low values (*i.e.*, a coarse discretisation). The result of a typical simulation is given below in Figure 2.

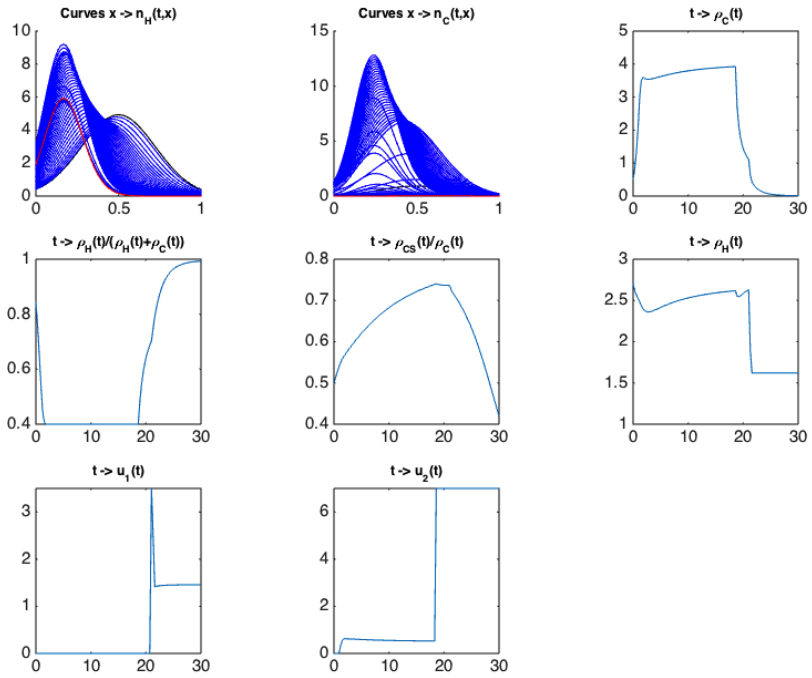


FIGURE 2: Simulation of the solution to ($\mathbf{OCP_1}$) for $T = 30$. Here, $\rho_{CS}(t) := \int_0^1 (1-x) n_C(t,x)\, dx$ accounts for the number of resistant cells.

These simulations clearly indicate that for the chosen numerical data, if T is large enough, then the optimal controls are such that:

- the optimal control u_1 is first equal to 0 on a long arc. Then, on a short-time arc, $u_1 = u_1^{\max}$ and then to a value such that the constraint (8) saturates;
- the optimal control u_2 has a three-part structure, with a long-time starting arc which is a *boundary arc*, that is, an arc along which the state constraint (7) is (very quickly) saturated. It corresponds to an almost constant value for the control u_2. The last short-time arc coincides with that of u_1, and along this arc $u_2 = u_2^{max}$.

In other words, the optimal strategy is made of

- a first long phase with no cytotoxic drugs and low doses of cytostatic drugs (as low as the constraint (7) allows it), at the end of which the cancer cell density has concentrated on a sensitive phenotype.
- a second short phase first with maximal doses (which are efficient on a sensitive tumour), up until the constraint (8) on the side effects has saturated. Then, intermediate cytotoxic drug doses are given (while the cytostatic drugs are still given at full dose), in order to make the tumour still shrink while the constraint (8) remains saturated.

In particular, we numerically recover the three arcs obtained that were obtained theoreti-

cally.

Numerical optimal control problem with mutations. The previous approach to numerically solving the optimal control problem is time-consuming when T becomes large, and even starts failing for fine discretisation parameters. These problems are made worse with the Laplacian terms, which have to be either discretised explicitly or implicitly, leading to a CFL or costly inversions, respectively.

The core idea presented in Chapter 6 which we developed in [124], is also based on homotopies, but first performed on the optimal control problem itself. We simplify the optimal control problem up to a point where a PMP in infinite dimension can provide very precise results on the optimal controls. More precisely, setting $\beta_H = \beta_C = a_{CH} = \theta_H = \theta_{HC} = 0$, we end up with the optimal control problem $(\mathbf{OCP_0})$

$$\min_{(u_1,u_2)\in\mathcal{U}} \rho_C(T)$$

where n_C solves

$$\frac{\partial n_C}{\partial t}(t,x) = \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho_C(t) - \mu_C(x)u_1(t)\right] n_C(t,x),$$

with $0 \leqslant u_1(t) \leqslant u_1^{max}$, $0 \leqslant u_2(t) \leqslant u_2^{max}$, but without state constraints (they are inactive because $\theta_H = \theta_{HC} = 0$).

The main result is the following.

Theorem 0.5 (Olivier and P.)**.** *The optimal controls for problem* $(\boldsymbol{OCP_0})$ *are given by*

$$u_1(t) = u_1^{max}\mathbb{1}_{[t_1,T]}, \qquad u_2(t) = u_2^{max}\mathbb{1}_{[t_2,T]}.$$

for some $t_1 \in [0,T[$, $t_2 \in [0,T[$.

One can therefore reduce the controls (originally in the infinite-dimensional space $L^\infty(0,T)^2$) to their switching times t_1, t_2 (in $\mathbb{R}^2$). Numerically, the problem can be discretised very finely, and the corresponding problem is an optimisation problem from $\mathbb{R}^2$ onto $\mathbb{R}$, which can be solved quickly and efficiently. This serves as a starting point (*i.e.*, as the problem $\mathcal{P}_0$) for the homotopy procedure, at the end of which the problem $\mathcal{P}_1$ (associated with the full optimal control problem $(\mathbf{OCP_1})$) is solved.

Simulations resulting from this technique, shown in Section 6.4 of Chapter 6, indicate that the optimal controls have the same structure as in the integro-differential case, and the computations have been made much quicker (even if $\beta_H = \beta_C = 0$).

It does not only provide a very precise result for the problem with mutations (with a much finer discretisation than with the previous approach) but also a general strategy for numerical optimal control: working at the continuous level by simplifying the problem so that it can be theoretically analysed with a PMP. The discretised counterpart is then an excellent candidate as starting point for a homotopy.

Towards an application to the clinic. In the clinic, we advocate that the optimal strategy obtained here must be thought of as a pattern to be repeated in a kind of periodic way, with three phases:

- (1) a long phase with low doses, as low as the maximal tumour size allows it,
- (2) a short phase with maximal doses up until the side-effects are deemed too great,
- (3) a potential (risky) extension to the previous phase with high cytostatic doses and intermediate cytotoxic ones to let the tumour further decrease while side-effects are at their limit.

Examples of such strategies are given in Section 5.4 of Chapter 5..

An actual implementation of this strategy requires to define it in feedback form, where the decision to switch from one phase to another is made on some biological markers:

- from phase (1) to phase (2): when resistance markers show that the tumour has become sensitive enough for the maximum tolerated doses to be used efficiently or when the tumour is considered to be too big (it becomes a threat to the organ or metastases are too likely to occur),
- from phase (2) to phase (3): when the maximal tolerated doses are no longer tolerated,
- from phase (3) to phase (1): when resistance markers show that the tumour has become resistant again. Alternatively, or equivalently, one could think that if the tumour starts increasing again, this is the sign that resistance is too high and that treatment at high doses must stop.

A key hypothesis (tacit in the model) for these strategies to work is for the tumour to be plastic enough to go back to sensitiveness in the absence of strong drug pressure. This reversibility must also happen quickly enough for phase (1) not to be too long. Finally, it is implicit here that several indicators can be measured for the assessment of several or all of the following: the level of resistance to the drugs, the damage to the healthy tissue and the tumour size.

3.3 Controllability for monostable and bistable 1D equations

Problem statement and link with asymptotic analysis. We are concerned with the control of 1D semilinear parabolic equations for population dynamics in the form

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t,0) = u(t),\ y(t,L) = v(t), \\ y(0) = y^0. \end{cases} \tag{15}$$

with Dirichlet controls on the boundary between 0 and 1, and $0 \leqslant y^0 \leqslant 1$, where f is either of monostable type, e.g., $f(y) = y(1-y)$, or of bistable type, e.g., $f(y) = y(1-y)(y-\theta)$.

We aim at solving the following controllability problem: can we control the equation towards 0 (extinction), 1 (invasion) or even the unstable state θ (in finite time or infinite time)?

For 0 (resp. 1), the question boils down to analysing what happens if we put 0 (resp. 1) on the boundary because of the parabolic comparison principle [142, 93]. This is consequently a problem of asymptotic analysis, for which a key result due to Matano is the following: with static controls, any trajectory must converge to a stationary state. Thus, if we set 0 on the boundary, any trajectory will asymptotically converge to 0 provided that the trivial solution 0 is the only one to the stationary problem

$$\begin{cases} -y_{xx} = f(y), \\ y(0) = y(L) = 0. \end{cases} \tag{16}$$

From phase plane analysis for the ODE $-y'' = f(y)$, it is possible to find a threshold $L^\star$ (computable as a transcendental integral) under which 0 is the only solution to the stationary problem (16), while there is at least another one if $L > L^\star$.

Control towards θ. To control towards θ, the main tool is the staircase method, which relies on the result that, for any two path-connected steady states, there exists a control strategy steering the first one to the second one.

Managing to use this technique properly amounts to finding a steady state y_{init} such that

- a suitable choice of boundary Dirichlet controls will make y_{init} globally attractive,
- there exists a path of steady states linking y_{init} and θ.

Both requirements are ensured thanks to a fine analysis of the phase portrait in the bistable case. The general result obtained in this work presented in Chapter 3, is summed up as follows.

Theorem 0.6 (P., Trélat and Zuazua)**.** (15) *is controllable*

- *in infinite time towards* 0 *if and only if* $L \leqslant L^\star$ *in the monostable case (resp.* $L < L^\star$ *in the bistable case).*
- *in infinite time towards* 1 *independently of* L*.*
- *in finite (or infinite) time towards* θ *if and only if* $L < L^\star$ *in the bistable case.*

3.4 Spheroid formation and Keller-Segel equations

Turing instabilities for Keller-Segel. For the dimensionalised Keller-Segel system without growth (neglected to ease computations)

$$\begin{cases} \frac{\partial n}{\partial t} - D_1 \Delta n + \chi \cdot (\varphi(n) \nabla c) = 0, \\ \frac{\partial c}{\partial t} - D_2 \Delta c = \alpha n - \beta c, \end{cases}$$

we use standard methods in Chapter 7 to determine a necessary and sufficient condition for the steady-state $(M, \frac{\alpha}{\beta}M)$ to be Turing unstable, with M the total and preserved mass, given by $\int_\Omega n^0 = M$. The condition writes simply as

$$\varphi(M) > \frac{\beta D_1}{\alpha\chi}. \tag{17}$$

If φ increases up to some density and then decreases, for example in the logistic ($\varphi(n) = n(1-n)$) and exponential cases ($\varphi(n) = ne^{-n}$), a consequence of the previous condition (17) is the following: all other parameters being fixed, there are no Turing instabilities for M too small or too large, while they can exist for intermediate values for the mass. Interestingly, experiments seem to yield such results, which is a good indicator that these functions are a better choice than the more classical $\varphi(n) = n$.

2D simulations. We provide 2D simulations on a disk for the system without and with growth, either for the logistic or exponential function, obtaining round patterns. These patterns appear abruptly, first at the periphery and then all the way to the center of the disk. A typical example is given in Figure 3.

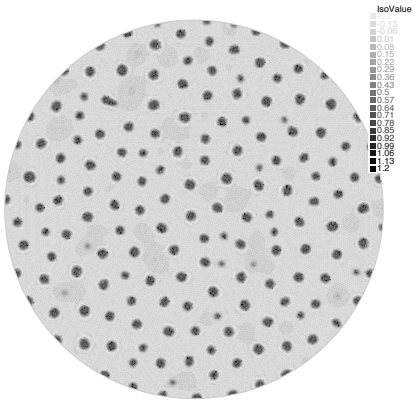


FIGURE 3: An example of a simulation exhibiting patterns, for a logistic sensitivity function and no growth.

Numerical analysis of 1D finite-volume schemes. For models like the Keller-Segel one above, we develop schemes in Chapter 8, in the simplified case where the chemoattractant diffuses much faster than the cells (which leads to neglect the time-derivative for the equation on c). They must preserve important continuous properties such as positivity and energy dissipation. Indeed, for the parabolic-elliptic Keller-Segel system posed on $(0, 1)$,

$$\begin{cases} \frac{\partial n}{\partial t} - \frac{\partial^2 n}{\partial x^2} + \frac{\partial}{\partial x}\left(\varphi(n)\frac{\partial c}{\partial x}\right) = 0, \\ -\frac{\partial^2 c}{\partial x^2} = n - c, \end{cases}$$

the energy

$$\mathcal{E}(t) = \int_0^1 \left[G(n) - \frac{1}{2}nc\right] dx$$

dissipates, where $G' = g$, $g' = \frac{1}{\varphi}$. There are two close ways of writing the equation for n: the first one is the classical *Gradient Flow*

$$\frac{\partial n}{\partial t} - \frac{\partial}{\partial x}\left[\varphi(n)\frac{\partial\left(g(n) - c\right)}{\partial x}\right] = 0,$$

or a *Scharfetter-Gummel* one, inspired by a common technique for the discretisation of PDEs arising in the physics of semi-conductors:

$$\frac{\partial n}{\partial t} - \frac{\partial}{\partial x}\left[e^{c-g(n)}\varphi(n)\frac{\partial e^{g(n)-c}}{\partial x}\right] = 0.$$

Following closely these computations to write the flux terms $F_{i+1/2}$ in a finite-volume approach leads to two semi-discrete schemes which preserve energy dissipation. At this stage, the equations become ODEs for $n_i(t) \simeq n(t, x_i)$, of the form

$$\frac{dn_i(t)}{dt} + \frac{1}{\Delta x}[F_{i+1/2}(t) - F_{i-1/2}(t)] = 0.$$

We use an semi-implicit scheme Euler scheme, which requires choosing appropriately terms that will be discretised explicitly or implicitly in order for the properties preserved at the semi-discrete to be further conserved at the discrete one. This has to be done with care since existence and uniqueness of a solution for the schemes is not granted due to the implicit terms. They actually rely on some appropriate monotonicity of F_{i+1} as a function of n_{i+1}, n_i, see Appendix 8.8.

This whole process yields two different finite-volumes schemes at the discrete level, presented in the article [22]:

Theorem 0.7 (Bubba, Neves de Almeida, P. and Perthame)**.** *For the 1D parabolic-elliptic Keller-Segel system, the Gradient Flow and Scharfetter-Gummel approaches lead to discrete schemes preserving mass conservation, energy dissipation, positivity and steady states.*

4 Some perspectives and open problems

On the asymptotic analysis for a single equation. For the selection-mutation equation with a non-local selection term (4), a complete picture is still lacking. The work of Jabin and Raoul has undoubtedly shed some light in the integro-differential case. Our extension to the PDE case with a weight $m = n^\infty$ has allowed us to treat the very special case of the non-local Fisher-KPP equation for a bounded domain.

Extending this result to the non-local equation on the full space seems trickier due to the fact that asking that $n(t) - 1 - \ln(n(t))$ be in $L^1(\mathbb{R}^d)$ is too restrictive. A natural path is to consider an exponential weight to gain integrability, but it is not clear how to handle the new terms appearing in the dissipation.

Does the implicit condition that $(x, y) \mapsto n^\infty(x)K(x, y)+n^\infty(y)K(y, x)$ be positive definite apply to other interesting cases? For other choices of weight m (not depending on n^∞)

we can always assume that $(x, y) \mapsto m(x)K(x, y) + m(y)K(y, x)$ is positive definite, but the term coming from the diffusion term in the dissipation of V has no sign. It is an open question to see whether it can be balanced by the other term by, for example, appropriate functional analytic inequalities.

On the asymptotic analysis for a system. For systems, the Lyapunov functionals introduced by Jabin and Raoul have been proved to be appropriate for integro-differential systems with an arbitrary number of equations for a specific form of the kernel ($K(x, y) = d(x)$). We have explained how the argument is general for 2x2 systems, but the condition on the interaction matrix becomes much more restrictive in higher dimensions, although simulations do not seem to suggest that the convergence and concentration properties are lost.

When the system is not integro-differential, the situation is essentially as open as for a single equation: only kernels of the form $K(x, y) = d(y)$ are dealt with, as in [100], because they can be handled as some kind of eigenvalue depending on time.

On the optimal control. For the theoretical analysis of ($\mathbf{OCP_1}$) in the integro-differential case, the critical idea was that for a long final time T, it is optimal to try and first reach a very sensitive phenotype. This result, confirmed by numerical simulations and hinted at by several arguments for simpler ODE equations, still remains to be proved rigorously in its full generality.

The numerical approach developed in Section 6 has been proved to successfully deal with ($\mathbf{OCP_1}$), but we believe that the strategy is a relevant option for a wide variety of problems, in particular when it holds that:

- a direct method is appropriate but does not converge for a fine discretisation,
- a priori theoretical knowledge is available on the type of equation: some simplifications lead to optimal control problems for which a PMP significantly reduces the dimension of the space of controls (this is typically the case if they are bang-bang).

An interesting perspective is to illustrate this strategy on some other challenging optimal control problems.

Application to the clinic. Discussions are ongoing with oncologists from the Tenon Hospital of Paris, in order to derive therapeutical protocols in the light of the optimal control strategy obtained in this PhD.

The first step would be to compare the results of infusion strategies already tried on patients. The data would contain the therapeutical strategy over time as well as medical imaging data for the tumour size, and would be used to parametrise the model.

As a second step, the goal would be to implement alternative strategies inspired from the optimal one exhibited for ($\mathbf{OCP_1}$), in vitro and/or in vivo (on mice).

Publications

In preparation

- F. Bubba, N. Ferrand, L. Neves de Almeida, B. Perthame, C.P., M. Sabbah, *A chemotaxis-based explanation of spheroid formation in 3D structures mimicking the extracellular matrix*, 2018.
- T. Lorenzi, C.P., *Finding selected phenotypes among those of equal fitness in the limit of small mutations*, 2018.
- F. Bubba, B. Perthame, C.P., M. Schmidtchen, *Hele-Shaw limit for a 1D system of two interacting populations*, 2018.

Submitted

- C.P., E. Trélat, E. Zuazua, *Phase portrait control for 1D monostable and bistable reaction-diffusion equations*, 2018.
- A. Olivier, C.P., *Combination of direct methods and homotopy in numerical optimal control: application to the optimization of chemotherapy in cancer*, 2017.

Accepted

- F. Bubba, L. Neves de Almeida, B. Perthame, C.P., *Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations*, accepted in Networks and Heterogeneous Media, 2018.
- (Note) C.P., *On the stability of the state 1 in the non-local Fisher-KPP equation in bounded domains*, published in Comptes Rendus Mathématique, 2018.
- C. P., E. Trélat, *Global stability with selection in integro-differential Lotka-Volterra systems modelling trait-structured populations*, accepted in Journal of Biological Dynamics, 2018.
- C.P., J. Clairambault, A. Lorz, E. Trélat, *Asymptotic analysis and optimal control of an integro-differential system modelling healthy and cancer cells exposed to chemotherapy*, published in Journal de Mathématiques Pures et Appliquées, 2017.

Part I

Adaptive dynamics

Chapter 1

Systems of integro-differential selection equations

In this work with Emmanuel Trélat, we investigate the asymptotic properties of Lotka-Volterra integro-differential systems for N species, proving convergence and concentration results for certain types of interaction matrices. A first general result is given using Lyapunov functionals, which happens to be sharp in dimension 2 but more restrictive in higher dimensions. The second result assumes that the interaction between species is mutualistic, setting for which the BV approach yields results, at least when the interaction is small. It is the subject of an article accepted in the Journal of Biological Dynamics, entitled *Global stability with selection in integro-differential Lotka-Volterra systems modelling trait-structured populations* [140].

1.1 Introduction

1.1.1 Biological motivations

We are interested in the evolution of N populations of individuals, each of which is structured by a continuous *phenotypic trait*, also called *trait*. In each species the phenotype models some continuous biological characteristics (such as the size of the individual, the concentration of a protein inside it, etc). We shall consider both interactions inside a given population and between the populations and we assume that mutations can be neglected. Mathematical modelling and analysis of such ecological scenarios is one purpose of the field of adaptive dynamics, a branch of mathematical biology which aims at describing evolution

among a population of individuals, see [49, 117, 129] for an introduction to deterministic models.

The basis for our model stems from the logistic ODE $\frac{dN}{dt} = (r - dN)N$ where r is the intrinsic growth rate, dN the logistic death rate due to competition for nutrients and for space by direct or indirect inhibition of proliferation between individuals. Its natural extension to a density $n(t, x)$ of individuals of phenotype x (say in $[0, 1]$) is a non-local logistic model

$$\frac{\partial}{\partial t} n(t, x) = (r(x) - d(x)\rho(t))\ n(t, x), \tag{1.1}$$

with $\rho(t) := \int_0^1 n(t, x)\, dx$ the total number of individuals.

Note that these models can be derived from stochastic models at the individual level [34, 50, 69], and more generally measure-valued functions n can be considered [28, 70]. The asymptotic behaviour of the previous model (1.1) and variants is analysed in [68, 111, 129], and one important property among others is that solution typically tend to concentrate on a few phenotypes, a convergence to Dirac masses in mathematical terms. These models are thus successful at representing the survival of only a few phenotypes, which we will refer to as *selected*.

The mathematical results available for $N = 1$ naturally call for generalisations on systems of interacting species with such non-local logical terms based on the total number of individuals. For instance, to study resistance in cancer, one may think also of different cancer subpopulations interacting with healthy cells and between them, each one of them being endowed with a specific drug resistance phenotype in a tumour 'bet hedging' strategy [21]. These generalisations, in turn, might help unravel general principles about the underlying ecological processes, and develop new mathematical techniques to analyse them.

1.1.2 The model

We consider N populations structured by respective phenotypes $x \in X_i$, where X_i is some compact subset of $\mathbb{R}^{p_i}$, with $p_i \in \mathbb{N}^*$, for $i = 1, \dots, N$. Although they model distinct quantities, we abusively denote all variables x to improve readability.

The model writes

$$\frac{\partial}{\partial t} n_i(t, x) = \left(r_i(x) + d_i(x) \sum_{j=1}^{N} a_{ij} \rho_j(t) \right) n_i(t, x), \quad i = 1, \dots, N, \tag{1.2}$$

where, for $i = 1, \dots, N$, r_i and $d_i > 0$ are functions in $L^\infty(X_i)$,

$$\rho_i(t) := \int_{X_i} n_i(t, x)\, dx$$

is the total number of individuals in species i, and $a_{ij} \in \mathbb{R}$ are fixed (interaction) coefficients.

The initial conditions are

$$n_i(0,\cdot) = n_i^0 \quad i = 1,\dots,N \tag{1.3}$$

where each initial density n_i^0 is taken to be a non-negative function in $L^1(X_i)$. From now on, we will call these equations *integro-differential Lotka-Volterra equations*.

The matrix $A := (a_{ij})_{1\leqslant i,j\leqslant N}$, called *matrix of interactions*, describes the interactions between the populations: if $a_{ij} > 0$, the species j acts positively on the species i, and negatively if $a_{ij} < 0$. Finally, we will say that the equations are *competitive* (resp. *mutualistic*) if $a_{ij} < 0$ (resp., $a_{ij} > 0$) for all $i \neq j$.

Another interpretation of the equations is to see them as coupled logistic equations of the form

$$\frac{\partial}{\partial t} n_i(t,x) = \big(r_i(x) - d_i(x) I_i(t)\big)\, n_i(t,x), \quad i = 1,\dots,N.$$

In other words, the species i reacts to its environment through the non-local variable I_i defined for $i = 1,\dots,N$ by

$$I_i := -\sum_{j=1}^{N} a_{ij}\rho_j.$$

The terms $r_i(x)$ and $d_i(x)I_i$ respectively stand for the intrinsic proliferation rate and logistic death rate of individuals in species i, of phenotype x.

We will also use the notation $R_i(x,\rho_1,\dots,\rho_N) := r_i(x) + d_i(x)\sum_{j=1}^N a_{ij}\rho_j$, with which the equations rewrite:

$$\frac{\partial}{\partial t} n_i(t,x) = R_i\left(x,\rho_1(t),\dots,\rho_N(t)\right) n_i(t,x), \quad i = 1,\dots,N.$$

These models generalise Lotka-Volterra ordinary differential equation (ODE) models [6]: if the functions r_i, d_i are all constant (say equal to some r_i, and $d_i = 1$), then after integration with respect to $x \in X_i$, the equations boil down to

$$\frac{d}{dt}\rho_i(t) = \left(r_i + \sum_{j=1}^{N} a_{ij}\rho_j(t)\right)\rho_i(t), \quad i = 1,\dots,N, \tag{1.4}$$

which we will from now on refer to as *classical Lotka-Volterra equations*. It can be written in the more compact form $\frac{d\rho}{dt} = (r + A\rho)\rho$. Thus, another advantage of a logistic term directly defined by ρ is that it makes our model more tractable with respect to the corresponding already well understood ODE models. Conversely, the integro-differential model can be seen as a perturbation of the ODE one where the individuals among a given population are allowed to have different proliferation and death rates depending on their phenotype.

Our goal is to understand the asymptotic behaviour of the solutions to these equations, both in terms of convergence at the level of the total number of individuals ρ_i, and in terms of concentration at the level of the densities n_i. The first problem is usual in population dynamics while the second is specific to adaptive dynamics and consists of determining

which traits asymptotically survive, taking over the whole population. These are then called *Evolutionary Attractors*, and the fact that it is the generic situation has been coined *exclusion principle*. Mathematically, this corresponds to a given density n_i converging to a sum of Dirac masses. For one Dirac mass only, concentration writes, for some $x_0 \in X_i$:

$$n_i(t,\cdot) - \rho_i(t)\delta_{x_0} \rightharpoonup 0$$

as $t \to +\infty$, in the weak sense of measures.

The more precise aim of this chapter is to study the global asymptotic stability (GAS) of what we will call *coexistence steady states*, namely of possible ρ^∞ with positive components (all species asymptotically survive) such that ρ converges to ρ^∞, because we will see how it determines on which phenotypes the densities concentrate. When it is possible, we will investigate the speed at which convergence and concentration occur. An interesting question is also to see if a result of that type is sharp, *i.e.*, to compare the assumptions needed to obtain global asymptotic stability in our generalised setting to those known for classical Lotka-Volterra equations.

At this stage, we did not make any restrictive assumptions on the matrix A. However, it will be clear from the results recalled below in the ODE case and the ones presented in Section 1.2, that answers to the previous questions are available when interspecific interactions are low compared to the intraspecific ones. Thus, we are covering the ecological scenario of each species i having its own niche, but inside which competition (if $a_{ii} < 0$) is blind because of the term $a_{ii}\rho_i$.

Notations. In what follows, $\mathbb{R}^N_{>0}$ will stand for the positive orthant in $\mathbb{R}^N$, the set of vectors whose components are all positive, and we will write $x > y$ when $x - y \in \mathbb{R}^N_{>0}$. We will also use the usual ordering on the set of symmetric matrices: for A a real symmetric matrices, we denote $A \geqslant 0$ (resp., $A > 0$) when A is positive semidefinite (resp., positive definite). Finally, $\mathcal{M}^1(X)$ will denote the set of Radon measures supported in X.

1.1.3 State of the art

Classical Lotka-Volterra equations. The ODE system (1.4) has been extensively studied, dating back to the pioneering works of Lotka and Volterra for two populations of preys and predators [113, 171]. Since then, many contributions to the analysis of steady states and their stability have been made, and we refer to [120] for an introduction and to [6] for a review.

Regarding the global asymptotic stability of coexistence steady states, a very well-known result due to Goh [64] states a simple and very general condition on the matrix $A = (a_{ij})_{1\leqslant i,j\leqslant N}$ which ensures convergence to the (unique) coexistence steady state:

Theorem 1.1 ([64])**.** *Assume that the equation $A\rho + r = 0$ (where $r \in \mathbb{R}^N$ and $\rho \in \mathbb{R}^N$ are the vectors $(r_i)_{1\leqslant i\leqslant N}$ and $(\rho_i)_{1\leqslant i\leqslant N}$) has a solution ρ^∞ in $\mathbb{R}^N_{>0}$. If there exists a diagonal matrix $D > 0$ such that $A^T D + DA < 0$, then ρ^∞ is GAS in $\mathbb{R}^N_{>0}$ (and hence is the unique coexistence steady state) for system (1.4).*

A result also worth stating is that the mere existence of a unique coexistence steady state is not enough for it to be GAS. Other steady states on the boundary of $\mathbb{R}^N_{>0}$ can attract trajectories even in dimension $N = 2$. Another possibility is the occurrence of chaotic behaviour even in low dimension as evidenced in [169] for $N = 3$. Finally, we mention the more recent work [46], where the authors tackle the problem of GAS for some type of Lotka-Volterra ODEs with mutations. In particular, they obtain GAS of the coexistence steady state in the case where the logistic variables I_i, $i = 1, \dots, N$ all coincide, that is, when they are equal to some variable $I := \sum_{j=1}^N a_j \rho_j(t)$. In such a case, it is proved that the convergence to the equilibrium is exponential. The result of GAS is also extended to perturbations of this specific case.

Integro-differential Lotka-Volterra equations. The first question for such equations is the existence of a solution for all positive times. This obviously does not hold true in full generality since the ODE $y' = y^2$ is a particular case. Let us first state an existence and uniqueness theorem.

Theorem 1.2. *Assume that for a given $n^0 \in \prod_{i=1}^N L^1(X_i)$, $n^0 \geqslant 0$, there exists $0 < \rho^{sup}$ such that we have an a priori upper bound $\rho(t) \leqslant \rho^{sup}$ for the functions ρ_i whenever they are defined. Then the Cauchy problem (1.2)-(1.3) has a unique solution $n = (n_i)_{1 \leqslant i \leqslant N}$, $n \geqslant 0$, in $C\left([0, +\infty), \prod_{i=1}^N L^1(X_i)\right)$.*

The proof is a straightforward generalisation of that given in [129, Theorem 2.4] for $N = 1$, relying on the Banach Fixed Point Theorem. For completeness, we provide it in Appendix 1.5.

In the case of a single equation, the asymptotic behaviour is well understood. For $N = 1$, assuming $a_{11} < 0$ to avoid blow-up, the equation is simply

$$\frac{\partial}{\partial t} n_1(t, x) = (r_1(x) - d_1(x)\rho_1(t))\ n_1(t, x),$$

where, without loss of generality, we have set $a_{11} = -1$. The first result is that ρ_1 converges.

Theorem 1.3. *Assume some regularity on X_1, r_1, d_1, and $r_1 > 0$, Then, for any positive continuous initial condition n_1^0, ρ_1 the function $t \mapsto \rho_1(t)$ is well defined on $[0, +\infty)$ and converges to $\rho_1^M := \max_{x \in X_1} \dfrac{r_1(x)}{d_1(x)}$ as $t \to +\infty$.*

This, in turn, completely determines where n_1 concentrates.

Corollary 1.1. *Under the previous hypotheses, $n_1(t)$, viewed as a Radon measure on X_1, concentrates on the set*

$$\left\{x \in X_1,\ r_1(x) - d_1(x)\rho_1^M = 0\right\}$$

as $t \to +\infty$. If this set is reduced to some x_1^∞, we obtain in particular

$$n_1(t, \cdot) \rightharpoonup \rho_1^M \delta_{x_1^\infty}$$

weakly in $\mathcal{M}^1(X_1)$, as $t \to +\infty$.

This result is classical and it relies on proving that ρ_1 is a bounded variation (BV) function on $[0, +\infty)$. For completeness, we recall it in Appendix 1.5. Let us stress that when the set on which n_1 concentrates is not reduced to a singleton, the steady state (at the level of n_1) is not unique. For example, if the set is made of two points, the repartition of the limiting density on these two points depends on the initial condition, see for example [45]. This is why for this equation and the general equations considered here, there is no hope in proving general GAS results directly at the level of the densities n_i.

For a general logistic term $\left(\int_X K(x,y)n(t,y)\,dy\right) n(t,x)$ and a single equation, the asymptotic behaviour is also analysed in detail in both [48] and [82]. In the latter, under some suitable assumptions on the kernel K, a Lyapunov functional is used to prove that some measure is GAS, in a very specific sense depending on K. Similar results can be found in [35], where their counterpart for competitive classical Lotka-Volterra equations are also discussed.

In the case of integro-differential systems, however, much less is known about the asymptotic behaviour. We mention [25] where an integro-differential system of two populations is analysed, and whose form does not fit in our framework. A particular triangular coupling structure allows the authors to perform an asymptotic analysis.

The chapter is organised as follows. In Section 1.2, we explain how coexistence steady states can be identified, allowing us to state rigorously what we mean by GAS for system (1.2). We explain why, under the hypothesis of GAS, only some phenotypes are generically selected, and how to compute them. Then, we present the two main results about GAS for such equations. Section 1.3 is devoted to the proof of the first result, which applies for any type of interactions and relies on analysing a suitably designed Lyapunov functional. In the specific case of mutualistic interactions, our second main result gives alternative conditions sufficient for GAS. It is presented in Section 1.4. In Section 1.5, we conclude with several comments and open questions.

1.2 Possible coexistence steady states and main results

For the rest of the chapter, we will work with the following assumptions:

$$r_i, d_i, n_i^0 \in C(X_i),\ n_i^0 > 0 \text{ for } i = 1, \dots, N.$$

This will simplify statements, but we will be more specific below as to which data our results generalise.

1.2.1 Analysis of coexistence steady states

In the context of this system of integro-differential equations, the expression "GAS in $\mathbb{R}^N_{>0}$" must be defined. By that, we mean that there exists $\rho^\infty > 0$ such that, whatever the positive continuous initial conditions n_i^0 are, ρ_i converges to ρ_i^∞ for all i.

First, let us explain how to compute the possible steady states at the level of ρ, *i.e.*, possible limits $\rho^\infty > 0$ for positive continuous initial conditions. We will work with the following topological assumption on the sets X_i:

$$\forall x \in \partial X_i,\ \forall \eta > 0,\ \lambda_{p_i}\left(B(x,\eta) \cap X_i\right) > 0, \tag{1.5}$$

where λ_{p_i} stands for the Lebesgue measure on $\mathbb{R}^{p_i}$ and $B(x,\eta)$ for the open ball of center x and radius η.

Assume that each ρ_i converges to some $\rho_i^\infty > 0$, in which case the exponential behaviour of $n_i(t,x)$ is asymptotically governed by $r_i(x) + d_i(x)\sum_{j=1}^N a_{ij}\rho_j^\infty$, the sign of which we can analyse as follows.

- If this quantity is positive for some x_0, let us prove that $n_i(t,x)$ blows up in its neighbourhood, leading to the explosion of ρ_i.

  If $r_i(x_0) + d_i(x_0)\sum_{j=1}^N a_{ij}\rho_j^\infty > 0$, there exists $\eta > 0$ such that by continuity $r_i(x) + d_i(x)\sum_{j=1}^N a_{ij}\rho_j^\infty > 0$ on $(B(x_0,\eta)\cap X_i)$, and then $\lambda_{p_i}\left(B(x_0,\eta)\cap X_i\right) > 0$ whether $x_0 \in int(X_i)$ or also if $x_0 \in \partial X_i$ thanks to (1.5). For $\varepsilon > 0$ small enough and t large enough (say $t \geqslant t_0$) such that $r_i(x_0) + d_i(x_0)\sum_{j=1}^N a_{ij}\rho_j > \varepsilon$, we can write:

$$\begin{aligned}
\rho_i(t) &\geqslant \int_{B(x_0,\eta)\cap X_i} n_i(t,x)\,dx \\
&\geqslant \int_{B(x_0,\eta)\cap X_i} n_i(t_0,x)\, e^{\int_{t_0}^t R_i(x,\rho_1(s),\dots,\rho_N(s))\,ds}\,dx \\
&\geqslant \lambda_{p_i}\left(B(x_0,\eta)\cap X_i\right)\left(\inf_{B(x_0,\eta)\cap X_i} n_i(t_0,x)\right)\, e^{\varepsilon(t-t_0)},
\end{aligned}$$

  with the right-hand side blowing up as $t \to +\infty$, which cannot be compatible with the convergence of ρ_i.

- If $r_i + d_i\sum_{j=1}^N a_{ij}\rho_j^\infty$ is negative globally on X_i, this clearly implies the extinction of species i, which is also incompatible with the convergence of ρ_i to a positive limit.

Remark 1.1. It is possible to relax the regularity on both the sets X_i and the data r_i and d_i by working only with points which are both Lebesgue points of $\frac{r_i}{d_i}$ and of Lebesgue density 1 for X_i, see [56]. If the functions $\frac{r_i}{d_i}$ are in $L^1(X_i)$, one can indeed check that $r_i + d_i\sum_{j=1}^N a_{ij}\rho_j^\infty \leqslant 0$ *a.e.* on X_i.

The previous results motivate the following definition:

$$I_i^\infty := \max_{x\in X_i} \frac{r_i(x)}{d_i(x)}, \quad i = 1,\dots,N.$$

With this definition, a steady state $\rho^\infty > 0$ exists if and only if the following assumption holds:

$$\text{the equation } A\rho + I^\infty = 0 \text{ has a solution } \rho^\infty \text{ in } \mathbb{R}^N_{>0}, \tag{1.6}$$

which we assume from now on.

The previous computations also show that n_i vanishes where $r_i(x) - d_i(x)I_i^\infty < 0$, which implies the following result:

Proposition 1.1. *Assume that assumption (1.6) holds, and that ρ converges to the coexistence steady state ρ^∞. Then, $n_i(t)$, viewed as a Radon measure, concentrates on the set*

$$K_i := \left\{x \in X_i,\ r_i(x) - d_i(x)I_i^\infty = 0\right\}$$

as $t \to +\infty$, for all $i = 1, \dots, N$.

If, for some i, K_i is reduced to some x^∞, we obtain in particular

$$n_i(t, \cdot) \rightharpoonup \rho_i^\infty \delta_{x_i^\infty}$$

as $t \to +\infty$ in $\mathcal{M}^1(X_i)$.

Densities n_i of total mass ρ_i^∞ and concentrated on K_i are called Evolutionary Stable Distributions (ESD) in [82].

Remark 1.2. In the hypothesis of global existence and convergence of ρ towards ρ^∞, the previous reasoning actually shows that the concentration is ensured as soon as $n_i^0 \in L^1(X_i)$ is bounded by below by a positive constant on a neighbourhood of one of the points of K_i. For more general hypotheses ensuring concentration, we refer to [82].

Remark 1.3. If all the sets K_i are reduces to some singletons x_i^∞, then the dynamics of ρ are asymptotically governed by classical Lotka-Volterra equations concentrated in $(x_1^\infty, \dots, x_N^\infty)$, namely

$$\frac{d}{dt}\rho_i(t) \simeq \left(r_i\left(x_i^\infty\right) + d_i\left(x_i^\infty\right)\sum_{j=1}^{N} a_{ij}\rho_j(t)\right)\rho_i(t), \quad i = 1, \dots, N,$$

as t goes to $+\infty$. For a precise statement, see Chapter 5.

1.2.2 Main results

Our first approach to prove GAS is to mix a Lyapunov functional which is inspired by the one designed in [82] and the Lyapunov functional used for classical Lotka-Volterra equations [64], which is the key tool to obtain Theorem 1.1. With some mild regularity assumptions on the data, we obtain the following result:

Theorem 1.4. *Assume (1.6) and that there exists a diagonal matrix $D > 0$ such that DA is symmetric and $DA < 0$. Then the solution to the Cauchy problem (1.2)-(1.3) is globally defined. Furthermore, the solution ρ^∞ to $A\rho + I^\infty = 0$ is GAS (and hence, unique).*

We emphasise that there is no assumption on the type of interactions, *i.e.*, on the sign of the coefficients of A. However, a necessary condition for the existence of D is that A must be such that $a_{ii}a_{jj} > a_{ij}a_{ji}$ for all i, j. For this result to apply, interactions must therefore be stronger inside species than between them.

We also remark that our hypothesis is exactly the one exhibited in [35] for competitive classical Lotka-Volterra equations. The analysis of the Lyapunov functional allows to determine a speed at which convergence to ρ^∞ and concentration on a given set of phenotypes occur. In dimension 2, we also analyse more deeply the link between this condition and the one for classical Lotka-Volterra equations, which in most interesting cases happen to be equivalent.

Our second main result focuses on the special case of mutualistic interactions, and an informal statement of the theorem is the following.

Theorem 1.5. *Assume (1.6), that for $i = 1, \dots, N$, $r_i > 0$ and that for some explicit $0 < c_i < C_i$, the matrix $\hat{A}$ defined by $\hat{a}_{ii} := c_i a_{ii}$ and $\hat{a}_{ij} = C_i a_{ij}$ for $i \neq j$ is Hurwitz. Then the solution to the Cauchy problem (1.2)-(1.3) is globally defined. Furthermore, the solution ρ^∞ to $A\rho + I^\infty = 0$ is GAS.*

Again, this applies to the case of interspecific interactions being higher that intraspecific ones, because a Hurwitz matrix is a matrix such that all its eigenvalues have negative real part and it has to do with diagonally dominant matrices (see Section 1.4).

Because of the hypothesis of mutualism, the system is cooperative, and sub and supersolution techniques can succeed. More precisely, it is possible to prove that all functions ρ_i are BV on $[0, +\infty)$ and this implies their convergence.

1.3 General interactions

1.3.1 Proof of the main theorem

In this section, we need slightly more regularity for the data, namely that the functions are Lipschitz continuous:

$$\text{for } i = 1, \dots, N,\ r_i, d_i \in C^{0,1}(X_i). \tag{1.7}$$

We can now restate the first theorem:

Theorem 1.6. *Assume (1.6) and (1.7). Assume that there exists a diagonal matrix $D > 0$ such that DA is symmetric and $DA < 0$. Then the solution to the Cauchy problem (1.2)-(1.3) is globally defined.*

Furthermore, the solution ρ^∞ to $A\rho + I^\infty = 0$ is GAS with

$$\rho(t) - \rho^\infty = \mathrm{O}\left(\left(\frac{\ln(t)}{t}\right)^{\frac{1}{2}}\right). \tag{1.8}$$

Concentration of a given n_i occurs at speed $\mathrm{O}\left(\frac{\ln(t)}{t}\right)$, in the following sense:

$$\int_{X_i} m_i(x) R_i\left(x, \rho_1^\infty, \dots, \rho_N^\infty\right) n_i(t, x)\, dx = \mathrm{O}\left(\frac{\ln(t)}{t}\right). \tag{1.9}$$

In particular, if K_i is reduced to a singleton x_i^∞, then

$$\forall \varepsilon > 0, \quad \int_{X_i \setminus B\left(x_i^\infty, \varepsilon\right)} n_i(t,x)\, dx = \mathrm{O}\left(\frac{\ln(t)}{t}\right). \tag{1.10}$$

Proof. First step: definition of the Lyapunov functional. From (1.6), Evolutionary Stable Densities exist and we can pick one of them: for $i = 1, \dots, N$, we choose any measure n_i^∞ in $\mathcal{M}^1(X_i)$ satisfying $n_i^\infty(X_i) = \rho_i^\infty$, which is furthermore concentrated on K_i, *i.e.*,

$$\mathrm{supp}(n_i^\infty) \subset K_i. \tag{1.11}$$

We abusively write integration of functions g against measures μ as $\int_X g(x)\mu(x)\, dx$. We also set $m_i := \frac{1}{d_i}$ and define N functions V_i by

$$V_i(t) := \int_{X_i} m_i(x) \left[n_i^\infty(x) \ln\left(\frac{1}{n_i(t,x)}\right) + (n_i(t,x) - n_i^\infty(x)) \right] dx.$$

In what follows, we consider the following Lyapunov functional:

$$V(t) := \sum_{i=1}^{N} \lambda_i V_i(t)$$

where the positive constants λ_i are to be chosen later. The diagonal matrix of diagonal entries $\lambda_1, \dots, \lambda_N$ is denoted by D.

Second step: computation and sign of the derivative. For any i, we compute

$$\begin{aligned}
\frac{dV_i}{dt} &= \int_{X_i} m_i(x) \left[-n_i^\infty(x) \frac{\partial_t n_i(t,x)}{n_i(t,x)} + \partial_t n_i(t,x) \right] dx \\
&= \int_{X_i} m_i(x)\, R_i\left(x, \rho_1, \dots, \rho_N\right) \left[n_i(t,x) - n_i^\infty(x) \right] dx \\
&= \int_{X_i} m_i(x) \left(R_i\left(x, \rho_1, \dots, \rho_N\right) - R_i\left(x, \rho_1^\infty, \dots, \rho_N^\infty\right) \right) \left[n_i(t,x) - n_i^\infty(x) \right] dx \\
&\qquad + \int_{X_i} m_i(x)\, R_i\left(x, \rho_1^\infty, \dots, \rho_N^\infty\right) \left[n_i(t,x) - n_i^\infty(x) \right] dx.
\end{aligned}$$

The definition of m_i implies that the first term simplifies as follows

$$\int_{X_i} m_i(x) d_i(x) \left[A\left(\rho - \rho^\infty\right) \right]_i \left[n_i(t,x) - n_i^\infty(x) \right] dx = \left[A\left(\rho - \rho^\infty\right) \right]_i \left(\rho_i - \rho_i^\infty \right).$$

For the second term, the choice (1.11) leads to

$$B_i(t) := \int_{X_i} m_i(x)\, R_i\left(x, \rho_1^\infty, \dots, \rho_N^\infty\right) n_i(t,x)\, dx.$$

The functions B_i are all non-positive by definition of ρ^∞.

Defining the vector $u := \rho - \rho^\infty$, we arrive at:

$$\begin{aligned}\frac{dV}{dt} &= \sum_{i=1}^{N} \lambda_i \left[A\left(\rho - \rho^\infty\right)\right]_i \left(\rho_i - \rho_i^\infty\right) + \sum_{i=1}^{N} \lambda_i B_i \\ &= \sum_{i=1}^{N} \lambda_i (Au)_i u_i + \sum_{i=1}^{N} \lambda_i B_i.\end{aligned}$$

Thus, we end up with the expression

$$\frac{dV}{dt} = u^T (DA) u + \sum_{i=1}^{N} \lambda_i B_i.$$

Since the antisymmetric part of DA does not play any role, this can also be expressed

$$\frac{dV}{dt} = \frac{1}{2} u^T (A^T D + DA) u + \sum_{i=1}^{N} \lambda_i B_i.$$

By assumption, $M := A^T D + DA < 0$, from which we deduce $\dfrac{dV}{dt} \leqslant 0$. Furthermore, the convergence of the term $u^T M u$ to 0 is equivalent to that of ρ to ρ^∞. However, we do not have the usual property $V \geqslant 0$ for Lyapunov functions, so that we cannot yet conclude.

Third step: estimates on $\frac{dV}{dt}$. Let

$$G := \frac{1}{2} u^T M u + 2 \sum_{i=1}^{N} \lambda_i B_i.$$

We are going to show that G is non-decreasing.

We denote by $\langle u, v \rangle$ the canonical scalar product of two vectors u and v in $\mathbb{R}^N$.

$$\begin{aligned}\frac{d}{dt}\left(u^T (DA) u\right) &= \frac{d}{dt} \langle (DA)u, u \rangle \\ &= \left\langle (DA)\frac{du}{dt}, u \right\rangle + \left\langle (DA)u, \frac{du}{dt} \right\rangle.\end{aligned}$$

For $i = 1, \ldots, N$, the derivative of B_i is given by

$$\begin{aligned}\frac{dB_i}{dt} &= \int_{X_i} m_i(x) R_i\left(x, \rho_1^\infty, \ldots, \rho_N^\infty\right) R_i\left(x, \rho_1, \ldots, \rho_N\right) n_i(t,x)\, dx \\ &= \int_{X_i} m_i(x) R_i^2\left(x, \rho_1, \ldots, \rho_N\right) n_i(t,x)\, dx \\ &+ \int_{X_i} m_i(x) \left[R_i\left(x, \rho_1^\infty, \ldots, \rho_N^\infty\right) - R_i\left(x, \rho_1, \ldots, \rho_N\right)\right] R_i\left(x, \rho_1, \ldots, \rho_N\right) n_i(t,x)\, dx \\ &\geqslant \left[A\left(\rho^\infty - \rho\right)\right] \int_{X_i} R_i\left(x, \rho_1, \ldots, \rho_N\right) n_i(t,x)\, dx \\ &= -(Au)_i \left(\frac{du}{dt}\right)_i\end{aligned}$$

leading to the bound

$$\frac{d}{dt}\left(\sum_{i=1}^{N}\lambda_i B_i\right) \geqslant -\sum_{i=1}^{N}\lambda_i (Au)_i \left(\frac{du}{dt}\right)_i = -\left\langle (DA)u, \frac{du}{dt}\right\rangle.$$

Put together, these estimates yield:

$$\frac{dG}{dt} \geqslant \left\langle (DA)\frac{du}{dt}, u\right\rangle + \left\langle (DA)u, \frac{du}{dt}\right\rangle - 2\left\langle (DA)u, \frac{du}{dt}\right\rangle = \left\langle (DA)\frac{du}{dt}, u\right\rangle - \left\langle (DA)u, \frac{du}{dt}\right\rangle.$$

The last expression is equal to 0 if DA is symmetric, in which case G is non-decreasing as claimed.

The assumptions that DA is symmetric and that $A^T D + DA < 0$ are equivalent to the assumption made for the theorem: DA is supposed to be a symmetric negative definite matrix.

As a consequence of the monotonicity of G, we get $u^T(-DA)u \leqslant -G(0)$ for all t. The left-hand side is the square of some norm on $\mathbb{R}^N$, which means that ρ has to be bounded: these a priori bounds ensure the global definition of the solution to (1.2)-(1.3).

Fourth step: a lower estimate for V. To estimate V from below, we need a uniform (in x) upper bound on the densities n_i. Because of the regularity assumption (1.7) , there exists $C > 0$ such that:

$$\forall i = 1, \dots, N, \quad \forall (x, y) \in X_i^2, \quad R_i\left(y, \rho_1, \dots, \rho_N\right) \geqslant R_i\left(x, \rho_1, \dots, \rho_N\right) - C|x - y|.$$

The constant C can be chosen to be independent of t since the functions ρ_i are bounded. This implies for a given i

$$\begin{aligned}\int_{X_i} n_i(t, y)\, dy &= \int_{X_i} n_i^0(y) \exp\left(\int_0^t R_i\left(y, \rho_1, \dots, \rho_N\right)\, ds\right) dy \\ &\geqslant \int_{X_i} \frac{n_i^0(y)}{n_i^0(x)} \left(n_i^0(x) \exp\left(\int_0^t R_i\left(x, \rho_1, \dots, \rho_N\right)\right)\right) \exp\left(-Ct|x - y|\right)\, dy \\ &\geqslant \frac{n_i(t, x)}{n_i^0(x)} \int_{X_i} \exp\left(-Ct|x - y|\right)\, dy.\end{aligned}$$

Computing the integral, we write, thanks to the boundedness of ρ_i n_i^0 and (C has changed and is independent of t and x): for t large enough, $n_i(t, x) \leqslant Ct$. The bound on V follows immediately:

$$V(t) \geqslant -C\left(\ln(t) + 1\right).$$

Fifth step: convergence. G bounds $\frac{dV}{dt}$ from above: $\frac{dV}{dt} \leqslant \frac{1}{2}G$. Thus

$$V(t) - V(0) \leqslant \frac{1}{2}\int_0^t G(s)\,ds \leqslant \frac{1}{2}tG(t)$$

thanks to the third step. We can now write $G(t) \geqslant -C\frac{\ln(t)+1}{t}$: $G(t) = \mathrm{O}\left(\frac{\ln(t)}{t}\right)$, consequently, each non-positive term it is composed of also vanishes like $\mathrm{O}\left(\frac{\ln(t)}{t}\right)$ as $t \to +\infty$.

In other words, $\frac{1}{2}u^T M u$ and each B_i onverge to 0 as as well $\mathrm{O}\left(\frac{\ln(t)}{t}\right)$. This is nothing but the two first statements (1.8) and (1.9).

For the last statement, we fix i and $\varepsilon > 0$. We denote $h_i := -m_i R_i\left(\cdot, \rho_1^\infty, \ldots, \rho_N^\infty\right)$, which is non-negative on X_i, and by assumption vanishes at x_i^∞ only. We choose $a > 0$ small enough such that $a\mathbb{1}_{\{X_i \setminus B\left(x_i^\infty, \varepsilon\right)\}} \leqslant h$ on X_i. This enables us to write

$$\int_{X_i \setminus B\left(x_i^\infty, \varepsilon\right)} n_i(t,x)\,dx \leqslant \frac{1}{a}\int_{X_i} m_i(x) R_i\left(x, \rho_1^\infty, \ldots, \rho_N^\infty\right) n_i(t,x)\,dx = \mathrm{O}\left(\frac{\ln(t)}{t}\right).$$

□

Remark 1.4. The first interesting fact is that the convergence rate of G to 0, in $\mathrm{O}\left(\frac{\ln(t)}{t}\right)$, is almost optimal in many cases. Indeed, if the sets K_i are reduced to singletons, there cannot exist any $\alpha > 1$ such that this sum vanishes like $\mathrm{O}\left(\frac{1}{t^\alpha}\right)$. This comes from the fact that if it were true, $\frac{dV}{dt}$ would be integrable on the half-line, which would imply the convergence of V. This is not possible since each V_i has to go to $-\infty$ as t goes to $+\infty$.

This might seem contradictory with the exponential convergence rates obtained in [46] for some classical Lotka-Volterra equations, but the Lyapunov functional gives us information on the speed of both phenomena in the sense defined above (through the function G) and it does not say whether one of the two is faster.

1.3.2 Sharpness in dimension 2

It is clear that if we can find $D > 0$ diagonal such that DA is symmetric and $DA < 0$, then $A^T D + DA < 0$. The condition that DA should be symmetric is constraining, especially if $N \geqslant 3$ in which case it imposes some polynomial equalities on the coefficients of the matrix A. In dimension 2, however, the result is sharp in various contexts, as stated in the following proposition.

Proposition 1.2. *Assume $N = 2$, $a_{11} < 0$, $a_{22} < 0$ and $a_{12}\,a_{21} > 0$. Then the following conditions are equivalent.*

(i) there exists $D > 0$ diagonal such that DA is symmetric and $DA < 0$;

(ii) there exists $D > 0$ diagonal such that $A^T D + DA < 0$;

(iii) $\det(A) > 0$.

Proof. (i) implies (ii) as noticed before. Now, let us assume (ii) and compute $M := A^T D + DA = \begin{pmatrix} 2\lambda_1 a_{11} & \lambda_1 a_{12} + \lambda_2 a_{21} \\ \lambda_1 a_{12} + \lambda_2 a_{21} & 2\lambda_2 a_{22} \end{pmatrix}$, which has positive determinant, *i.e.*, $\det(M) = 4\lambda_1\lambda_2 a_{11}a_{22} - (\lambda_1 a_{12} + \lambda_2 a_{21})^2 > 0$. If $\det(A) \leqslant 0$, $\det(M) \leqslant 4\lambda_1\lambda_2 a_{12}a_{21} - (\lambda_1 a_{12} + \lambda_2 a_{21})^2 = -(\lambda_1 a_{12} - \lambda_2 a_{21})^2 \leqslant 0$, a contradiction.

Now, if (iii) holds, we take $\lambda_1 := \frac{1}{|a_{12}|}$ and $\lambda_1 := \frac{1}{|a_{21}|}$ for which $DA = \begin{pmatrix} \frac{a_{11}}{|a_{12}|} & \operatorname{sgn}(a_{12}) \\ \operatorname{sgn}(a_{21}) & \frac{a_{22}}{|a_{21}|} \end{pmatrix}$ is clearly symmetric negative definite, whence (i). □

1.4 Cooperative case

1.4.1 Some facts about Hurwitz matrices

We now focus on the cooperative case, *i.e.*, on the case where all off-diagonal elements of A are non-negative. We will also assume that the diagonal elements are negative, since otherwise blow-up clearly occurs: there is intra-spectific competition inside any given species. We will say that such a matrix is *cooperative*.

In this case, we can hope for stronger results at the level of the integro-differential system because sub and super-solution techniques work. For our purpose, the following result on ODEs is sufficient.

Lemma 1.1. *For $T > 0$ (possibly $T = +\infty$), let $f : [0,T) \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$ be a continous function on $[0,T) \times \mathbb{R}^N$, locally Lipschitz in $x \in \mathbb{R}^N$ uniformly in $t \in [0,T)$. Denoting $f(t,x) := (f_i(t,x_1,\dots,x_N))_{1\leqslant i\leqslant N}$, further assume that for all $i = 1,\dots,N$, f_i is non-decreasing with x_j for all $j \neq i$.*

Assume that we have a solution z on $[0,T)$ of the following Cauchy problem:

$$\begin{aligned} \frac{dz}{dt} &= f(t,z) \\ z(0) &= z_0, \end{aligned}$$

where $z_0 \in \mathbb{R}^N$, and a function y subsolution to the previous Cauchy problem, i.e.,

$$\begin{aligned} \frac{dy}{dt} &\leqslant f(t,y) \\ y(0) &\leqslant z_0. \end{aligned}$$

Then $y(t) \leqslant z(t)$ on $[0,T)$.

When the matrix A is cooperative, it is possible to give an equivalent condition to the one required in Theorem 1.1 for GAS in classical Lotka-Volterra equations. Let us explain how, starting with the three following lemmas, the two first of which can be found in [6].

Lemma 1.2. *Let A be a cooperative matrix. Then it is Hurwitz if and only if it is negatively diagonally dominant, i.e., if and only if there exists a vector $v > 0$ such that $a_{ii}v_i + \sum_{j\neq i} a_{ij}v_j < 0$ for $i = 1, \dots, N$.*

This first lemma will be useful in its own right in this section. A consequence is that

Lemma 1.3. *If A is cooperative and $r > 0$, $A\rho + r = 0$ has a unique solution in $\mathbb{R}^N_{>0}$ if and only if A is Hurwitz.*

Finally, it comes from the theory of M-matrices (see [135] for a review) that

Lemma 1.4. *Let A be cooperative. Then A is Hurwitz if and only if there exists $D > 0$ diagonal such that $A^T D + DA < 0$.*

An important consequence of these three lemmas is the following rephrasing of Theorem 1.1 for classical Lotka-Volterra equations.

Proposition 1.3. *Assume that A is cooperative, $r > 0$ and that the equations (1.4) have a unique steady state in $\mathbb{R}^N_{>0}$. Then the equations are globally defined and this steady state is GAS.*

Thus, in the context of cooperation between the species, the requirement that A is Hurwitz is somehow optimal to obtain a GAS coexistence steady state, since it is already required to have its mere existence, a fact mentioned in [63]. We will mainly work with this characterisation (rather than the equivalent one given by Lemma 1.4 which we used for a general interaction matrix A) because the next results will lead us to modify the matrix A: analysing whether it is still Hurwitz or not is easier than checking this equivalent condition.

1.4.2 A priori bounds

For the remaining part of this section, we make the assumption that r_i is positive on X_i for $i = 1, \dots, N$, and we define the lower and upper bounds $0 < d_i^m \leqslant d_i(x) \leqslant d_i^M$, $0 < r_i^m \leqslant r_i(x) \leqslant r_i^M$.

Theorem 1.7. *Assume that the matrix $\tilde{A}$ defined by $\tilde{a}_{ii} := d_i^m a_{ii}$ and $\tilde{a}_{ij} := d_i^M a_{ij}$ is Hurwitz. Then the solutions to (1.2) are globally defined and bounded.*

Proof. First remark that since $\tilde{A}$ is Hurwitz, then so is A from Lemma 1.2.

We integrate the equations with respect to x and bound them (through $r_i(x) \leqslant r_i^M$)

$$\frac{d}{dt}\rho_i(t) \leqslant \left(r_i^M + \sum_{j=1}^N a_{ij}\rho_j(t) \int_{X_i} d_i(x) n_i(t,x)\, dx \right) \quad i = 1, \dots, N.$$

Since the diagonal elements are negative, the off-diagonal non-negative, we obtain

$$\frac{d}{dt}\rho_i(t) \leqslant \left(r_i^M + a_{ii} d_i^m \rho_i + \sum_{j\neq i}^N a_{ij} d_i^M \rho_j(t) \right) \rho_i(t), \quad i = 1, \dots, N.$$

Thus, the vector $(\rho_1, \dots, \rho_N)$ is a subsolution of the previous system which is nothing but classical Lotka-Volterra equations with interaction matrix $\tilde{A}$. Thanks to (1.3), the solutions to this system are bounded. Thus, so are those of the integro-differential one. □

Remark 1.5. Note that the assumption that $\tilde{A}$ is Hurwitz reduces to A being Hurwitz in the case of constant coefficients. Thus, this result for boundedness is sharp, since it is exactly the one required for convergence to the coexistence steady state when the equations at hand are classical Lotka-Volterra equations.

Using Theorem 1.7, we can thus define $\rho^M > 0$ as the GAS steady state for the system obtained in the previous proof, that is to the equations

$$\frac{d}{dt}u_i = \left(r_i^M + a_{ii}d_i^m u_i + \sum_{j\neq i}^{N} a_{ij}d_i^M u_j(t)\right)u_i(t), \quad i = 1, \dots, N.$$

In other words, $\rho^M := -\tilde{A}^{-1}r^M$ where r^M is the vector $(r_i^M)_{1\leqslant i\leqslant N}$. This means that we can write

$$\limsup_{t\to+\infty} \rho_i \leqslant \rho_i^M \quad i = 1, \dots, N. \tag{1.12}$$

In a similar fashion to the previous proposition, bounding ρ_i away from 0:

$$\frac{d}{dt}\rho_i \geqslant \left(r_i^m + a_{ii}d_i^M \rho_i + \sum_{j\neq i}^{N} a_{ij}d_i^m \rho_j(t)\right)\rho_i(t), \quad i = 1, \dots, N,$$

leading to

$$\liminf_{t\to+\infty} \rho_i \geqslant \rho_i^m \quad i = 1, \dots, N \tag{1.13}$$

where $\rho^m := -B^{-1}r^m > 0$ with B, a Hurwitz matrix defined by $b_{ii} := d_i^M a_{ii}$, $b_{ij} := d_i^m a_{ij}$ for $i \neq j$ and $r^m := (r_i^m)_{1\leqslant i\leqslant N}$.

1.4.3 GAS in the mutualistic case

We can now state the main result:

Theorem 1.8. *Assume $r_i > 0$ for all $i = 1, \dots, N$, and that the matrix $\hat{A}$ defined by $\hat{a}_{ii} := d_i^m \rho_i^m a_{ii}$ and $\hat{a}_{ij} := d_i^M \rho_i^M a_{ij}$ for $i \neq j$, is Hurwitz. Then $\tilde{A}$, A and B are also Hurwitz. Furthermore, $\rho^\infty := -A^{-1}I^\infty$ lies in $\mathbb{R}^N_{>0}$ and it is GAS.*

Proof. The fact that $\tilde{A}$, A and B are also Hurwitz is a direct consequence of Lemma 1.2. Since $\tilde{A}$ is Hurwitz, the solutions are globally defined with ρ bounded thanks to Theorem 1.7. Since A is Hurwitz, it is invertible and $\rho^\infty := -A^{-1}I^\infty$ is in $\mathbb{R}^N_{>0}$ thanks to Lemma 1.3.

Let us now prove that it is GAS. The idea is to prove that each ρ_i is BV on $[0, +\infty)$. Identifying the limit is straightforward, thanks to the reasoning made in Section 1.2.

For any i, we define $q_i := \dfrac{d\rho_i}{dt}$ and write $R_i = R_i(x, \rho_1, \dots, \rho_N)$ for readability. Since $q_i = \dfrac{d\rho_i}{dt} = \int_{X_i} R_i n_i$, we obtain

$$\frac{dq_i}{dt} = \int_{X_i} R_i^2 n_i + \int_{X_i} \left(\sum_{j=1}^{N} \frac{\partial R_i}{\partial \rho_j} q_j \right) n_i \geqslant \sum_{j=1}^{N} a_{ij} \left(\int_{X_i} d_i(x) n_i(t,x)\, dx \right) q_j.$$

Let $b_i(t) := \int_{X_i} d_i(x) n_i(t,x)\, dx$. The idea is that ρ_i is "mostly" increasing, so we are interested in the negative part of q_i, denoted by $(q_i)_-$. For this quantity we have the (a.e.) bound

$$\frac{d(q_i)_-}{dt} \leqslant b_i \sum_{j=1}^{N} a_{ij} q_j \left(-\mathbb{1}_{\{q_i<0\}} \right)$$

On the one hand,

$$b_i a_{ii} q_i \left(-\mathbb{1}_{\{q_i<0\}} \right) = b_i a_{ii} (q_i)_-.$$

On the other hand, for $i \neq j$,

$$b_i a_{ij} q_j \left(-\mathbb{1}_{\{q_i<0\}} \right) \leqslant b_i a_{ij} (q_j)_-.$$

Combining these two, we get

$$\frac{d(q_i)_-}{dt} \leqslant b_i \left(A(q)_- \right)_i.$$

We fix $\varepsilon > 0$ small enough and t large enough (say $t \geqslant t_0$) such that $\hat{A} + \varepsilon J$ is Hurwitz (where J is the matrix composed of ones only) and such that, for each (i,j), $b_i(t) a_{ij} \leqslant \hat{a}_{ij} + \varepsilon$. The first requirement is easily derived from Lemma 1.2 since $\hat{A} + \varepsilon J$ is clearly cooperative and negatively diagonally dominant for $\varepsilon > 0$ small enough. The second requirement comes from the lower and upper bounds for the functions ρ_i as stated in (1.12) and (1.13).

The resulting inequality is

$$\frac{d(q_i)_-}{dt} \leqslant \left(\left(\hat{A} + \varepsilon J \right) (q)_- \right)_i,$$

so that $((q_1)_-, \dots, (q_N)_-)$ is a subsolution of the system with same initial conditions at t_0, given by

$$\frac{dy}{dt} = \left(\hat{A} + \varepsilon J \right) y.$$

The solutions to this system go exponentially to 0 since $\hat{A} + \varepsilon J$ is Hurwitz.

For any i, we have thus proved that $(q_i)_-$ goes to 0 exponentially. Together with the fact that ρ_i is bounded from above, we conclude that it is BV on $[0, +\infty)$. Indeed it holds true that a function u which is both bounded from above and such that $u_- \in L^1([0, +\infty))$ is BV on $[0, +\infty)$, see [129, Lemma 6.7]. Therefore, ρ converges (to ρ^∞). □

1.5 Conclusion

We have analysed the global asymptotic stability properties for integro-differential systems of N species structured by traits x belonging to different trait spaces X_i. The coupling comes from a non-local logistic term, which is a linear combination of the total number of individuals ρ_i in each species. Theses systems generalise the usual Lotka-Volterra ODEs for which many stability analyses are available in the literature. Our main focus has been on the asymptotic behaviour of the functions $\rho_i(t)$ as $t \to +\infty$, especially towards equilibrium ρ^∞ with positive components, *i.e.*, of persistence of all species. In Section 1.2, we explained how identifying it essentially determines the asymptotic behaviour of the underlying density n_i, namely the phenotypes on which the measures $n_i(t,\cdot)$ concentrate in large time.

In Section 1.3, an adequate Lyapunov functional allowed us to state a general result relying solely on an assumption on the matrix A, regardless of the type of interactions. For $N = 2$, this is essentially a sharp result, but becomes more restrictive for $N \geqslant 3$. This tool also provided us with convergence rates to equilibrium. In Section 1.4, we presented another strategy based on a BV bound which yielded a second result of global asymptotic stability, this time for mutualistic equations.

The result of Theorem 1.8 is partly less general than the one given in Theorem 1.6 because it requires a sign on the coefficients of A. However, the set of matrices which satisfy the hypothesis given in the last theorem is an open subset of the set of real matrices $\mathbb{R}^{N\times N}$ in any dimension. This in sharp contrast with the hypothesis of Theorem 1.6, which, as already mentioned, imposes some polynomial equalities on the coefficients of A as soon as $N \geqslant 3$. In other words, for a small perturbation of a cooperative matrix for which GAS holds, GAS still holds. In particular, if one has weakly (but mutualistically) coupled equations, GAS holds, whereas Theorem 1.6 does not cover the case of any weakly coupled equations for general interactions, unless $N = 2$.

In both cases, the assumptions fall within the class of matrices which cannot have off-diagonal coefficients which are too high compared to the diagonal ones. The present results thus apply to cases where interactions among individuals of a same species are not only blind because of the term $a_{11}\rho_1$, but also stronger than the interactions between species. In other words, each one of them has its own ecological niche inside which interactions are independent of how given phenotypes x and y are distant from another.

Let us remark that the BV method would apply to more general functions $R_i(x, \rho_1, \dots, \rho_N)$, as long as they are increasing in the variables ρ_j, $j \neq i$. However, the Lyapunov functional used in Theorem 1.6 seems to be dependent on the linear coupling chosen here and it is an open problem to generalise our results for other settings. Another open question is about finding whether there are matrices A for which the underlying classical Lotka-Volterra equations converge to the coexistence steady state (for example such that there exists $D > 0$ with $A^T D + DA < 0$), but for which there is no GAS for the integro-differential system. Numerically at least, we could not build any such case.

Appendix A: proof of Theorem 1.2

Proof. The proof is based on the Banach-Picard fixed point theorem. We set $T > 0$, and define the Banach spaces $Z := \prod_{i=1}^{N} L^1(X_i)$ endowed with the max norm, and $E := C([0,T], Z)$ endowed with the norm $\|m\|_E := \sup_{0\leqslant t\leqslant T} \|m(t)\|_Z$. Finally, we consider the following closed subset: $F := \{m \in E \,/\, m \geqslant 0 \text{ and } \|m\|_E \leqslant M\}$ where $M > \rho^{\sup}$.

We now build the application. Let m be a fixed element in F, and let us define for $i = 1, \dots, N$

$$\tilde{\rho}_i(t) = \int_{X_i} m_i(t,x)\, dx.$$

For each $i = 1, \dots, N$ and each fixed $x \in X_i$, we consider the solution $\gamma_{i,x}$ to the following differential equation:

$$\begin{cases} \frac{d\gamma_{i,x}}{dt} = R_i\left(x, \tilde{\rho}_1(t), \dots, \tilde{\rho}_N(t)\right)\, \gamma_{i,x} \\ \gamma_{i,x}(0) = n_i^0(x) \end{cases} \tag{1.14}$$

which is global on $[0,T]$.

We then define for all (t,x) in $[0,T] \times X$ and $i = 1, \dots, N$ the function $n_i(t,x) := \gamma_{i,x}(t)$, thus building an application Φ through $\Phi(m) := n$.

We now show that Φ maps F onto itself.
The equation (1.14) can be solved explicitly by

$$n_i(t,x) = n_i^0(x) e^{\int_0^t R_i(x, \tilde{\rho}_1(s), \dots, \tilde{\rho}_N(s))\, ds},$$

which shows both $n \geqslant 0$ and $n \in E$.

Let us fix some $i = 1, \dots, N$ and bound as follows

$$\frac{\partial}{\partial t} n_i(t,x) \leqslant \left(\|r_i\|_{L^\infty} + \|d_i\|_{L^\infty} \|A\|_\infty \rho^{\sup}\right) n_i(t,x).$$

Integrating in x, we uncover $\frac{d}{dt}\|n(t)\|_Z \leqslant C\|n(t)\|_Z$ for some constant $C > 0$, which leads to

$$\|n(t)\|_Z \leqslant \rho^{\sup} e^{CT}.$$

To obtain $n \in F$, it only remains to choose T small enough so that $\rho^{\sup} e^{CT} \leqslant K$.

The last step is to prove the strong contraction property for Φ. In the following, C will denote various positive constants, which might change from line to line. Let $(m^1, m^2) \in F^2$ and (n^1, n^2) its image by Φ. We define $\tilde{\rho}^k$ as before for $k = 1, 2$. For all i, we write

$$(n_i^1 - n_i^2)(t,x) = n_i^0(x) \left[e^{\int_0^t R_i\left(x, \tilde{\rho}_1^1(s), \dots, \tilde{\rho}_N^1(s)\right) ds} - e^{\int_0^t R_i\left(x, \tilde{\rho}_1^2(s), \dots, \tilde{\rho}_N^2(s)\right) ds} \right].$$

Now, since the argument in the exponentials can be bounded by CT, the mean value theorem yields

$$\begin{aligned}|(n_i^1 - n_i^2)|(t,x) &\leqslant n_i^0(x)e^{CT}\left|\int_0^t \left[R_i\left(x,\tilde{\rho}_1^1(s),\dots,\tilde{\rho}_N^1(s)\right) - R_i\left(x,\tilde{\rho}_1^2(s),\dots,\tilde{\rho}_N^2(s)\right)\right] ds\right| \\ &\leqslant \|d_i\|_{L^\infty}\|A\|_\infty n_i^0(x)e^{CT}\left[\int_0^T \|\tilde{\rho}^1(s) - \tilde{\rho}^2(s)\|_\infty\, ds\right] \\ &\leqslant Cn_i^0(x)Te^{CT}\|m_1 - m_2\|_E.\end{aligned}$$

This implies after integrating in x and taking the supremum both in $t \in [0,T]$ and in $i = 1,\dots,N$:

$$\|n^1 - n^2\|_E \leqslant C\rho^{sup}Te^{CT}\|m^1 - m^2\|_E.$$

It provides us with the contracting property for Φ whenever T is small enough.

We conclude by noticing that T has been chosen small independently of the initial data, so that the argument can be iterated on $[0,T]$, $[T,2T]$, etc. □

Appendix B: proof of convergence in the case $N = 1$

Proof. We are going to prove that ρ is a BV function. To that end, let us prove that ρ is bounded from above, and that it has integrable negative part.

First step: upper bound for ρ. The existence of such a bound comes from integrating the equation with respect to x:

$$\rho'(t) = \int_X (r(x) - d(x)\rho(t))\, n(t,x)\, dx.$$

If ρ is too large, the right hand side is negative, forcing ρ to decrease. It proves the claim on the upper bound for ρ. Similarly, because of assumption (1.3), ρ increases if it is too close to 0: ρ is bounded from below by some $\rho^{min} > 0$.

Second step: estimate on the negative part of ρ_C. We define $q := \rho'$ and wish to prove that $q_- \in L^1(0,+\infty)$ and write in short R for $r(x) - d(x)\rho$. We differentiate $\rho' = \int_X nR$ to obtain:

$$q' = \int_X nR^2 + \left(\int_X n\frac{\partial R}{\partial \rho}\right) q$$

It provides an upper bound for the negative part of q:

$$(q)'_- \leqslant \left(\int_X n\frac{\partial R}{\partial \rho}\right) q_- \leqslant -d^{min}\rho^{min}(q)_-$$

where $0 < d^{min} \leqslant d$ on X. We conclude that q_- vanishes exponentially and is consequently integrable over the half-line. Therefore, ρ converges to some $\rho^\infty > 0$. □

Chapter 2

Selected phenotypes among those of equal fitness for small mutations

---

The goal here is to go further in understanding what happens at the limit for a simple integro-differential equation exhibiting convergence and concentration: on which phenotypes does the latter occur? This cannot be decided at the integro-differential level independently of the initial condition. Together with Tommaso Lorenzi, we propose to introduce a small diffusion term and pass to the limit as time goes to infinity in the non-local PDE, and then to pass to the limit as the diffusion rate vanishes. The analysis then boils down to that of the behaviour of the first eigenfunction of the operator $\varepsilon\Delta + r$ where r is the fitness function. Translating results from the community of semi-classical analysis, we find that uniqueness is recovered, and in the absence of symmetries, that a single phenotype is typically selected. This work is the subject of an upcoming article, under the (yet tentative) name *Finding selected phenotypes among those of equal fitness in the limit of small mutations.*

2.1 Introduction

The purpose of adaptive dynamics is to provide a mathematical framework to study and understand evolution. Following Diekmann [49], a fundamental concept is that of *ecological feedback loop*: the individuals of a given population create the environment they live in. As a consequence, what happens at the individual level is shaped by the population-level.

Assuming on top of it the existence of small mutations, a basic selection-mutation model

like

$$\frac{\partial n}{\partial t}(t,x) - \beta \Delta n(t,x) = \left(r(x) - d(x)\rho(t)\right) n(t,x), \tag{2.1}$$

(with $\beta > 0$ small) and close models have attracted some attention [129, 51]. Here, $\rho(t) = \int_\Omega n(t,x)\,dx$, $\Omega \subset \mathbb{R}^d$ is the set of phenotypes, and we set Neumann boundary conditions on $\partial\Omega$. r is the so-called *fitness function*.

Note that a tacit hypothesis is that individuals of phenotype x are independent of one another, in that they all react the same to the environment created by the whole population through $\rho(t)$.

As explained in the General Introduction, if $\beta = 0$, there holds that $\rho(t) \longrightarrow \max\left(\frac{r}{d}\right)$ while $n(t,\cdot)$ concentrates on the set $\arg\max\left(\frac{r}{d}\right)$

The integro-differential equation thus leads to selection of those phenotypes maximising some function (related to the fitness function), a satisfying property from the applicative point of view. However, a problematic feature remains: if $\arg\max\left(\frac{r}{d}\right)$ is not reduced to a singleton, there are uncountably many possible limits, the one being observed at the limit $t \to +\infty$ depending on the initial condition.

Writing things explicitly, if $\arg\max\left(\frac{r}{d}\right) = \{x_1, \dots, x_N\}$, any limit measure n^∞ is of the form

$$\frac{n^\infty}{\int_\Omega n^\infty} = \sum_{i=1}^{N} \alpha_i \delta_{x_i}.$$

If for example the whole problem is symmetric (r, d and Ω are), then one would expect at the limit to find a repartition n^∞ which preserves this symmetry. This does not hold true at the integro-differential level.

Going back to the model with small mutations, a significant effort has been put in the analysis of the model after rescaling of time [129, 7, 112]. More precisely, inserting $\sqrt{\beta}$ in front of the time derivative and considering instead of (2.1)

$$\sqrt{\beta}\frac{\partial n}{\partial t}(t,x) - \beta \Delta n(t,x) = \left(r(x) - d(x)\rho(t)\right) n(t,x),$$

the goal is to analyse, t being fixed, the behaviour of $n(t,x)$ as β tends to 0.

If $r(x) - d(x)\bar{\rho}$ has a unique maximum point for all $\bar{\rho} > 0$, then the limit is a Dirac located at an appropriate point, and in the long-run $t \to +\infty$ we recover the concentration on $\arg\max\left(\frac{r}{d}\right)$ and convergence of ρ towards $\max\left(\frac{r}{d}\right)$. The proof of these results is involved and relies on the solution of a Hamilton-Jacobi equation.

The situation is somewhat more complicated and not well understood when there are several maximum points [129]. A concentration set of more than one phenotype is consequently a difficulty also with this approach. We here take a different direction to analyse the problem, without rescaling time, and we do so in the simpler case where $d(x)$ is a constant, which we normalise to 1.

To that end, we intend to use the equation with small mutations, of rate now denoted by ε, to obtain uniqueness. From the applicative point of view, this amounts to deciphering how completely neglecting small mutations for these types of models might be inappropriate, and how the problem can be addressed by introducing them back in the model.

We aim at comparing

$$\begin{aligned} &\frac{\partial n_\varepsilon}{\partial t}(t,x) = \big(r(x) - \rho_\varepsilon(t)\big)n_\varepsilon(t,x) + \varepsilon\Delta n_\varepsilon(t,x), \quad x \in \Omega,\ t > 0, \\ &\frac{\partial n_\varepsilon}{\partial \nu}(t,x) = 0, \quad x \in \partial\Omega,\ t > 0, \end{aligned} \tag{2.2}$$

with its integro-differential approximation for $\varepsilon = 0$ given by

$$\frac{\partial n}{\partial t}(t,x) = \big(r(x) - \rho(t)\big)n(t,x), \quad x \in \overline{\Omega},\ t > 0, \tag{2.3}$$

with both equations starting at the same initial condition $n^0 \geqslant 0$, $n^0 \neq 0$ in $L^1(\Omega)$, independent of ε.

The goal is to pass to the limit as t goes to $+\infty$ on (2.2) to obtain some limit n_ε^∞, and then as ε goes to 0 on the (at this stage hypothetical) unique limit n_ε^∞, obtaining some measure n^∞. Since for (2.3), we have convergence of $\rho(t)$ towards $\rho^\infty := \max(r)$ and concentration of $n(t,\cdot)$ on the set $\arg\max(r)$, the first requirement will be for this procedure to yield a measure n^∞ with mass ρ^∞ and support included in $\arg\max(r)$. We shall address these questions in Section 2.2.

In Section 2.3, we will explain how the symmetric case is well dealt with by this method, and then turn our attention towards more general cases without symmetry: using knowledge from semi-classical analysis, we will explain how the typical case is for the limit measure to be concentrated on a single phenotype. Theoretical results with be illustrated by simulations.

2.2 Why total mass and support match

2.2.1 Asymptotic analysis of $t \mapsto n_\varepsilon(t,\cdot)$

We assume

$$n^0 \in L^1(\Omega), \tag{2.4}$$

$$r \in C^{0,1}(\overline{\Omega}), \tag{2.5}$$

Under the previous assumptions (2.4)-(2.5), it holds that (2.2) has a unique classical solution in $C([0,+\infty), L^1(\Omega)) \cap C^1((0,+\infty), C^{2,\alpha}(\Omega))$ [46]. We will also assume $\max(r) > 0$.

In what follows, we will denote (for a given $\varepsilon > 0$) A_ε the elliptic operator defined by

$$A_\varepsilon := \varepsilon\Delta + r.$$

The theory of elliptic operators ensures that A_ε has a principal eigenvalue λ_ε and a unique eigenfunction $\psi_\varepsilon > 0$ (once normalised by $\int_\Omega \psi_\varepsilon = 1$). We also recall that λ_ε is obtained as the infimum of the Rayleigh quotients defined for $\phi \in H^1(\Omega) \setminus \{0\}$, by

$$\mathcal{R}(\phi) := \frac{\varepsilon \int_\Omega |\nabla \phi(x)|^2 \, dx - \int_\Omega r(x)\phi^2(x)\, dx}{\int_\Omega \phi^2(x)\, dx}.$$

In other words,

$$\lambda_\varepsilon = \inf_{\phi \in H^1(\Omega)\setminus\{0\}} \mathcal{R}(\phi),$$

with the infimum being reached by multiples of ψ_ε. The asymptotic behaviour of (2.2) is given by the following alternative, a standard fact [46, 100].

Proposition 2.1. *If $\lambda_\varepsilon \geqslant 0$, then the whole population goes extinct:*

$$\rho_\varepsilon(t) \to 0 \text{ as } t \text{ goes to } +\infty.$$

If $\lambda_\varepsilon < 0$,

$$n_\varepsilon(t,x) \longrightarrow n_\varepsilon^\infty := -\lambda_\varepsilon \psi_\varepsilon,$$

as t goes to $+\infty$, in $L^\infty(\Omega)$ (in particular, ρ_ε converges to $-\lambda_\varepsilon$).

Proof. The idea is to consider the non-local term as an eigenvalue and to introduce a suitable change of variables allowing to momentarily get rid of that term. Let us define $\tilde{n}_\varepsilon(t,x) := a(t) n_\varepsilon(t,x)$ with a solving the Cauchy problem

$$a'(t) = \rho_\varepsilon(t) a(t) + \lambda_\varepsilon a(t), \; a(0) = 1.$$

$\tilde{n}_\varepsilon$ satisfies

$$\begin{aligned}
\frac{\partial \tilde{n}_\varepsilon}{\partial t}(t,x) &= a'(t) n_\varepsilon(t,x) + a(t)\frac{\partial n_\varepsilon}{\partial t}(t,x) \\
&= a'(t) n_\varepsilon(t,x) + \big(r(x) - \rho_\varepsilon(t)\big) a(t) n_\varepsilon(t,x) + a(t)\Delta n_\varepsilon(t,x) \\
&= \big(a'(t) - \rho(t)a(t)\big)\, n(t,x) + A_\varepsilon \tilde{n}(t,x) \\
&= \lambda_\varepsilon \tilde{n}(t,x) + A_\varepsilon \tilde{n}(t,x),
\end{aligned}$$

where the last equality is obtained thanks to the definition of a.

It is a standard fact about parabolic operators that this implies the convergence of $\tilde{n}_\varepsilon(t,x)$ towards $C\psi_\varepsilon(x)$ for some $C > 0$, in $L^\infty(\Omega)$. Integrating the equality $\tilde{n}_\varepsilon(t,x) = a(t) n_\varepsilon(t,x)$ in x, we obtain that $a(t)\rho_\varepsilon(t)$ converges to C.

This allows to recast the ODE on a as $a'(t) = \lambda_\varepsilon a(t) + C + o(1)$, and this is enough to characterise the asymptotic behaviour of a depending on the sign of λ_ε: if $\lambda_\varepsilon > 0$, then $a(t) \to +\infty$ as t goes to $+\infty$, while it converges to $\frac{C}{-\lambda_\varepsilon}$ if $\lambda_\varepsilon < 0$.

From the convergence of $a(t)\rho_\varepsilon(t)$ to C, we get that $\rho_\varepsilon(t)$ goes to 0 for $\lambda_\varepsilon \geqslant 0$, and to $-\lambda_\varepsilon$ in the other case. Owing to $\tilde{n}_\varepsilon = a\, n_\varepsilon$, we get the announced result on the asymptotic behaviour of n_ε. □

2.2.2 Passing to the limit again on n_ε^∞

We now address the question of proving that any limit measure for $(n_\varepsilon^\infty)_{\varepsilon>0}$ must have total mass ρ^∞ and support included in $\arg\max(r)$. Thanks to the normalisation $\int_\Omega \psi_\varepsilon = 1$, we have split these two questions into the analysis of the behaviour of λ_ε and ψ_ε separately.

Remark 2.1. In what follows, we will be led to consider sequences (u_n) of functions in $L^1(\Omega)$ which we will see as elements of the larger space $\mathcal{M}^1\big(\overline{\Omega}\big)$ of Radon measures, in the sense of the dual of $C\big(\overline{\Omega}\big)$. In this setting, such a sequence will be said to *concentrate* on a given set F if for all $g \in C\big(\overline{\Omega}\big)$ such that $supp(g)$ does not intersect F, we have

$$\langle u_n, g\rangle = \int_\Omega u_n g \longrightarrow 0,$$

as n goes to $+\infty$.

A well-known fact from measure theory is that if u_n concentrates on a finite set $\{x_1, \dots, x_N\}$ and converges to a measure $\mu \in \mathcal{M}^1\big(\overline{\Omega}\big)$, then μ must be a linear combination of Dirac masses located at the points x_i, $1 \leqslant i \leqslant N$, because if the previous result holds true, then the support of μ must be contained in F.

The result writes as follows

Proposition 2.2. *Assume that r does not attain its maximum exclusively in $\partial\Omega$. As ε tends to 0, we then have*

$$-\lambda_\varepsilon \longrightarrow \rho^\infty = \max(r),$$

and

$$\psi_\varepsilon \text{ concentrates on the set } \arg\max(r).$$

Proof. We split the proof into two steps, first concentrating on the mass (the first statement), then on the support (the second one).

Step 1: computation of the mass. For the first result, we need to establish that $-\lambda_\varepsilon$ tends to $\max(r)$ as ε tends to 0.

Recall that

$$\lambda_\varepsilon = \inf_{\phi\in H^1(\Omega)\setminus\{0\}} \left\{ \frac{\varepsilon \int_\Omega |\nabla\phi(x)|^2 \, dx - \int_\Omega r(x)\phi^2(x)\, dx}{\int_\Omega \phi^2(x)\, dx} \right\},$$

and because for any $\phi \in H^1(\Omega)$, $\varepsilon \int_\Omega |\nabla\phi(x)|^2 \, dx - \int_\Omega r(x)\phi^2(x)\, dx \geqslant -\max(r)\int_\Omega \phi^2(x)\, dx$, we already have $\lambda_\varepsilon \geqslant -\max(r)$.

For the upper bound, let us build a sequence of functions (ϕ_ε) converging to a Dirac located in $x^0 \in \arg\max(r)$, $x^0 \notin \partial\Omega$, at an appropriate rate depending on ε. Let $G : x \longmapsto Ce^{-|x|^2}$, where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^d$, and C is such that $\int_{\mathbb{R}^d} G = 1$. It is then standard that $\frac{1}{\alpha^d} G\left(\frac{x-x_0}{\alpha}\right)$ converges to the Dirac mass δ_{x_0} located at the point x_0

$$\frac{1}{\alpha^d} G\left(\frac{x-x_0}{\alpha}\right) \rightharpoonup \delta_{x_0}$$

as $\alpha \to 0$ weakly in $\mathcal{M}^1(\overline{\Omega})$[1].

We take $\alpha = \varepsilon^{\frac{1}{4}}$, namely we define the Gaussian $\phi_\varepsilon^2 := x \longmapsto \frac{1}{\varepsilon^{\frac{d}{4}}} G\left(\frac{x-x_0}{\varepsilon^{\frac{1}{4}}}\right)$, which converges to δ_{x_0} as ε tends to 0. Let us study the Rayleigh quotient $\mathcal{R}(\phi_\varepsilon)$ and prove that it converges to $-\max(r)$, which will yield the first statement of the proposition. We already know that $\int_\Omega r(x)\phi_\varepsilon^2(x)\,dx$ and $\int_\Omega \phi_\varepsilon^2(x)\,dx$ converge to $r(x_0) = \max(r)$ and 1, respectively. It remains to show that the term $\varepsilon \int_\Omega |\nabla \phi_\varepsilon(x)|^2\,dx$ vanishes.

We compute $\varepsilon \int_\Omega |\nabla \phi_\varepsilon(x)|^2\,dx = \int_\Omega |x-x_0|^2\,\phi_\varepsilon^2(x)\,dx$, which tends to $|x_0 - x_0|^2 = 0$.

Step 2: computation of the support. As stated in Remark 2.1, our aim is to prove that for all $g \in C(\overline{\Omega})$ such that $supp(g)$ does not intersect $\arg\max(r)$, we have

$$\int_\Omega \psi_\varepsilon g \longrightarrow 0,$$

as ε goes to 0.

Integrating the equation $\varepsilon\Delta\psi_e + r\psi_\varepsilon = -\lambda_\varepsilon\psi_\varepsilon$, we find $\int_\Omega(-\lambda_\varepsilon - r)\psi_\varepsilon = 0$, and from the first step this entails the convergence of $\int_\Omega(\max(r) - r)\psi_\varepsilon$ towards 0.

Fixing now a function $g \in C(\overline{\Omega})$ and defining $\tilde{g} := \frac{g}{\max(r)-r}$ which belongs to $L^\infty(\Omega)$, we write

$$\left|\int_\Omega \psi_\varepsilon g\right| = \left|\int_\Omega (\max(r) - r)\psi_\varepsilon \tilde{g}\right| \leqslant \|\tilde{g}\|_\infty \int_\Omega (\max(r) - r)\psi_\varepsilon,$$

whence the result by letting ε tend to 0. □

Remark 2.2. If r attains its maximum only at the boundary of Ω, it is possible to adapt the previous proof by suitably changing the constant C. This has to be done depending on the geometry of $\partial\Omega$ near the maximum point, but for the sake of simplicity we stick with this simplest case.

Also note that since $\max(r) > 0$, we infer from the previous result that $-\lambda_\varepsilon > 0$ for ε small enough. Owing to Proposition 2.1, a consequence is that there is no extinction for (2.2) when ε is small enough.

2.2.3 Numerical simulations

Let us illustrate this result when $\arg\max(r)$ is reduced to a singleton. In the subsequent simulations (and also in the next subsection), we will always take $\Omega = (0,1)$, $\max(r) = 1$,

[1] Let $\phi \in C(\overline{\Omega})$. Changing variables, we have

$$\int_\Omega \phi(x)\frac{1}{\alpha^d} G\left(\frac{x-x_0}{\alpha}\right) dx = \int_{\mathbb{R}^d} \phi(x_0 + \alpha u)\, G(u) \mathbf{1}_{\{u\in\mathbb{R}^d,\, x_0+\alpha u\in\Omega\}}\, du.$$

Because $x^0 \in \Omega$, there is pointwise convergence of $u \mapsto \phi(x_0 + \alpha u)\, G(u)\mathbf{1}_{\{u\in\mathbb{R}^d,\, x_0+\alpha u\in\Omega\}}$ towards $\phi(x_0)G(u)$ on $\mathbb{R}^d$. Furthermore, this function is bounded by $\|\phi\|_\infty G(u)$. One can thus pass to the limit thanks to Lebesgue's dominated convergence Theorem, obtaining $\phi(x_0)$.

$n^0(x) := Ce^{-\frac{(x-0.5)^2}{2\sigma^2}}$ with $\sigma = 0.1$ and C such that $\int_\Omega n^0 = \frac{1}{2}$, so that n^0 is a (truncated) Gaussian with maximum at 0.5.

We here consider r with a single maximum point at 0.5, given by Figure 2.1 below. We

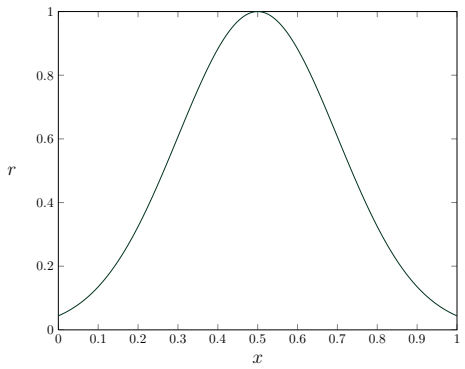


FIGURE 2.1: Plot of the function r with a single maximum point at 0.5.

also compare the time evolution of $\rho(t)$, $n(t,\cdot)$ with $\rho_\varepsilon(t)$, $n(t,\cdot)$ on Figure. At the level of the total mass, we indeed find that ρ converges to $\rho^\infty = \max(r) = 1$, while ρ_ε converges to $-\lambda_\varepsilon < \rho^\infty = 1$. At the level of the density, n concentrates as a Dirac mass on 0.5, while n_ε has stabilised as a smoothed version of this Dirac, also centred at 0.5, which we know is a multiple of the first eigenfunction ψ_ε of A_ε.

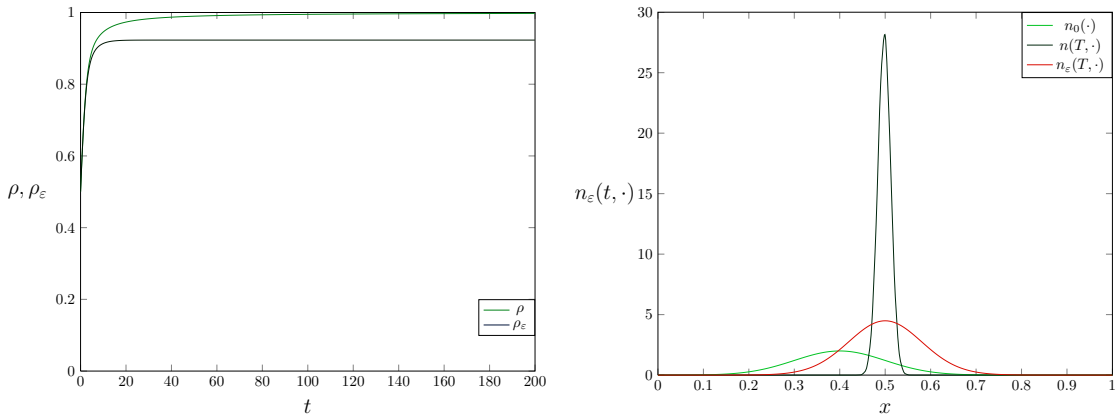


FIGURE 2.2: On the left, plots of functions ρ, ρ_ε, on the right of n and n_ε at the final time, together with the initial condition. Parameter values are $\varepsilon = 5.10^{-4}$ and $T = 200$.

2.3 Inferring the real Dirac masses

At this stage, the question is to see whether the uniqueness of n_ε^∞ is preserved at the limit even when there are several maximum points. If it is the case, the next goal is to identify the limit among the family of convex combination of Dirac masses.

2.3.1 In the case of symmetry

An appropriate setting for uniqueness is a case where the data has some symmetries, because principal eigenfunctions typically preserve such symmetries. We thus assume that Ω and r are symmetric with respect to some hyperplane, which without loss of generality we can assume to be $H := \{x_1 = 0\}$. For a point $x \in \Omega$, we write $x = (x_1, x')$, the assumption on Ω meaning $(x_1, x') \in \Omega \implies (-x_1, x') \in \Omega$. For a generic function v defined on $\overline{\Omega}$, we will always denote by $\hat{v}$ its symmetric, defined by $\hat{v}(x_1, x') = v(-x_1, x')$: by symmetric function we hence mean $\hat{v} = v$. When Ω and r are symmetric, the symmetry extends to the first eigenfunction of the operator $\varepsilon\Delta + r$.

Lemma 2.1. *The first eigenfunction ψ_ε is symmetric with respect to H.*

Proof. We simply write the equation defining ψ_ε: $\varepsilon\Delta\psi_\varepsilon(-x_1, x') + r(-x_1, x')\psi_\varepsilon(-x_1, x') = -\lambda_\varepsilon\psi_\varepsilon(-x_1, x')$. It rewrites $\varepsilon\Delta\hat{\psi}_\varepsilon(x_1, x') + r(x_1, x')\hat{\psi}_\varepsilon(x_1, x') = -\lambda_\varepsilon\hat{\psi}_\varepsilon(x_1, x')$, due to the symmetry of r. Thus, $\hat{\psi}_\varepsilon$ is also an eigenfunction of the operator $\varepsilon\Delta + r$, which by uniqueness must be equal to ψ_ε. □

Now, because the limit is expected to be a measure, we need to define what it means for a measure to be symmetric. It is as usual defined on the test-functions: a measure μ in $\mathcal{M}^1(\overline{\Omega})$ will be said to be symmetric if for all $v \in C(\overline{\Omega})$, it satisfies $\langle \mu, \hat{v} \rangle = \langle \mu, v \rangle$. It is easy to check that it extends the previous definition for L^1 functions: if $v \in L^1(\Omega)$ is symmetric, then the measure it defines is also symmetric. We finish these remarks with the following straightforward lemma.

Lemma 2.2. *Let (μ_n) be a sequence of symmetric Radon measures in $\mathcal{M}^1(\overline{\Omega})$ weakly converging to a measure μ. Then μ is also symmetric.*

From all the previous results, we are able to state the following proposition.

Proposition 2.3. *Assume that Ω and r are symmetric, with* $\arg\max(x) = \{x^1, x^2\}$*, both points lying inside Ω. Then*

$$n_\varepsilon^\infty \rightharpoonup \frac{1}{2}\rho^\infty \left(\delta_{x^1} + \delta_{x^2}\right),$$

weakly in $\mathcal{M}^1(\overline{\Omega})$, as ε goes to 0.

Proof. We first note that the symmetry of Ω and r imposes that the maxima must be symmetric: we can write $x^1 = (x_1^1, x')$, $x^2 = (-x_1^1, x')$ for some x_1^1, x'.

Since we already know that $-\lambda_\varepsilon$ converges to ρ^∞, we focus on (ψ_ε). Let μ be a (weak) limit point of (ψ_ε). From Proposition 2.2, we know that $\mu = (\alpha\delta_{x^1} + (1-\alpha)\delta_{x^2})$ for some $0 \leqslant \alpha \leqslant 1$. From Lemma 2.1, we know that the sequence is composed of symmetric functions, and by Lemma 2.2, the limit μ must also be symmetric, which entails $\alpha = \frac{1}{2}$. Since the sequence (ψ_ε) is bounded in $L^1(\Omega)$, it is weakly relatively compact in $\mathcal{M}^1(\overline{\Omega})$ from the Banach-Alaoglu theorem. As a consequence, the whole sequence must weakly converge to its unique limit point $\frac{1}{2}(\delta_{x^1} + \delta_{x^2})$. □

We illustrate these results still with an initial condition centred at 0.5, but now r has two maximum points at $x_1 := \frac{1}{3}$ and $x = \frac{2}{3}$ and is symmetric with respect to $x = \frac{1}{2}$, see Figure 2.3 below.

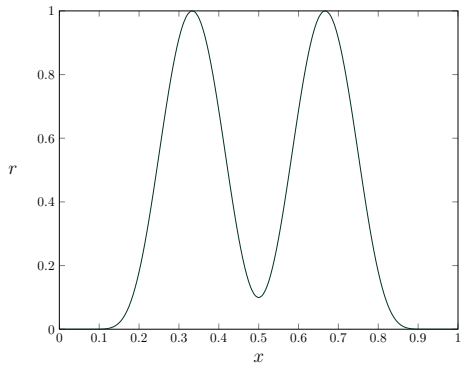


FIGURE 2.3: Plot of the symmetric function r with two maximum points at $\frac{1}{3}$ and $\frac{2}{3}$.

We also plot the time-evolution of n and n_ε in Figure 2.4. At time $T_1 = 500$, the integro-differential equation is already quite concentrated at x_1, and eventually all the mass will be at x_1. For n_ε, we have to take a large final time: the solution tends to concentrate at x_1 before the mass at x_1 is slowly "absorbed" at the other maximum point x_2, and at $T_2 = 20000$, the solution is almost symmetric, as expected theoretically.

2.3.2 In the absence of symmetry

What can be said in the absence of a very particular symmetry? The question can be addressed by slightly changing the viewpoint. Indeed, let us set $V = -r$ to emphasise that we are considering $-r$ as a potential, and we now see x as a space variable.

We are thus interested in the behaviour of the first eigenfunction ψ_ε of the operator $\varepsilon\Delta - V$, as ε tends to 0. From the previous results, we already know that it concentrates where V is minimal, namely on the set $\arg\max(r)$.

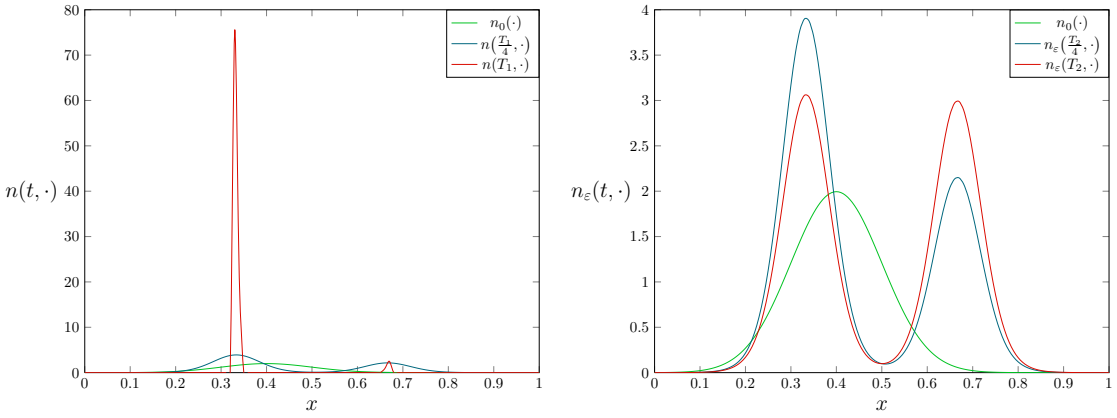


FIGURE 2.4: On the left, plot of functions $n(t,\cdot)$ at different times up to $T_1 = 500$, on the right plot of functions $n_\varepsilon(t,\cdot)$ at different times up to $T_2 = 20000$, with $\varepsilon = 5.10^{-4}$.

Further investigating in which points in $\arg\min(V) = \arg\max(r)$ there will actually be some mass is a fundamental question in semi-classical analysis. It is motivated by the analysis of the behaviour of a particle in a potential well, in the limit of small noise.

We assume that r is at least C^2, attains its maximum at a finite number of points $(x_1, \dots, x_N)$ and that they are non-degenerate maxima. For simplicity, we also assume that these points all lie inside Ω. We finally introduce some notations.

If x_i is a point of $\arg\max(r)$, we denote $(\lambda_i(x_i))_{1\leqslant i\leqslant d}$ the (positive) eigenvalues of the Hessian of $-r$ at the point x_i, and set

$$\zeta_i := \sum_{i=1}^{p} \sqrt{\lambda_i(x_i)}.$$

Translating the results obtained in semi-classical analysis, we obtain the following Theorem.

Theorem 2.1 ([79], Theorem 2)**.** *ψ_ε concentrates on the set $S := \arg\min_{1\leqslant i\leqslant N}(\zeta_i)$ as ε goes to 0.*

Returning to our problem, we have obtained the following result.

Corollary 2.1. *n_ε^∞ concentrates on the set $S = \arg\min_{1\leqslant i\leqslant N}(\zeta_i)$ as ε goes to 0.*

For example, in dimension 1, we obtain that n_ε^∞ will concentrate on the points x in $\arg\max(r)$ which minimise $-r''(x)$.

Remark 2.3. We stress that if the previous set S is not reduced to a singleton, the limit is still not completely identified. In [79], the analysis is continued to further reduce the possible points on which concentration occur. This reduced set has a very complicated expression, and we stick to the simpler one above which is enough for our purpose.

We illustrate these results (again with an initial condition centred at 0.5), and a non-symmetric r with two maximum points at $x_1 := \frac{1}{4}$ and $x_2 := \frac{5}{8}$, see Figure 2.5 below.

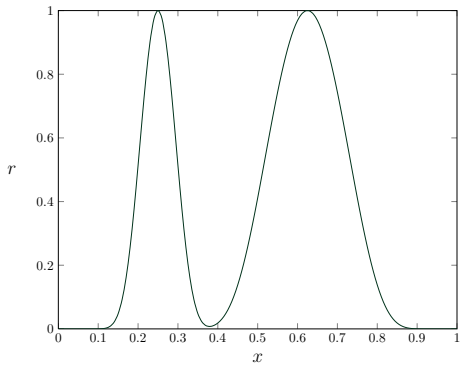


FIGURE 2.5: Plot of the non-symmetric function r with two maximum points at $\frac{1}{4}$ and $\frac{5}{8}$.

We also plot the time-evolution of n and n_ε in Figure 2.6. At time $T = 500$, the solution of the integro-differential equation is already quite concentrated at x_1, and eventually all the mass will be at x_1. For n_ε, the behaviour at the same time $T = 500$ is quite different, the solution is close to a smoothed Dirac at x_2, in perfect agreement with the theoretical results, because we have $-r''(x_2) \leqslant -r''(x_1)$.

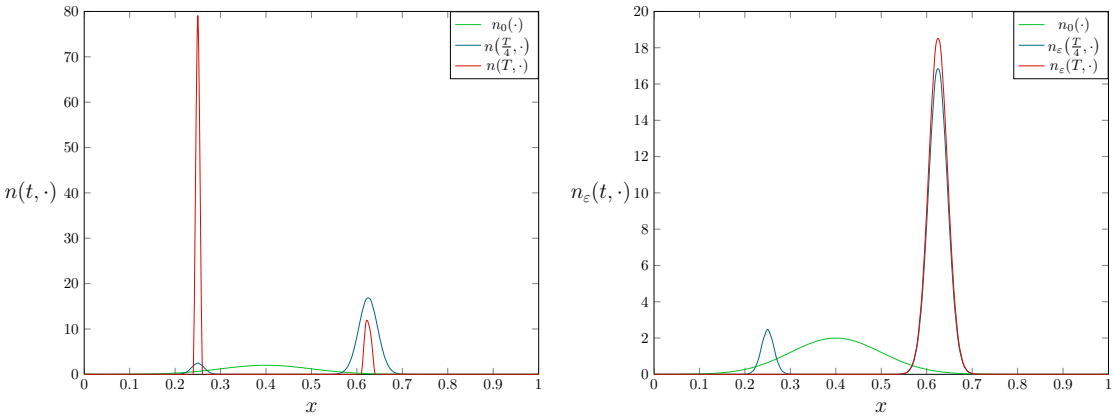


FIGURE 2.6: On the left, plot of functions $n(t, \cdot)$, on the right plot of functions $n_\varepsilon(t, \cdot)$ at different times, with $\varepsilon = 1.10^{-5}$. In both cases, $T = 500$.

2.3.3 On transient behaviours

We have seen that different behaviours can be expected at the same time T for equations (2.2) and (2.3), starting from the same initial condition. There are however cases for which reaching the unique stationary state is long, as in the symmetric case.

In the non-symmetric case, let us consider r with $x_1 = \frac{1}{4}$ and $x_2 = \frac{5}{8}$ as maximum points, in which case we know that in the end concentration will be on x_2. As evidenced by Figure 2.7, If the initial data is initially concentrated at $x = 0.1$, there will be a transient concentration on x_1 and a quick transition to concentration on x_2.

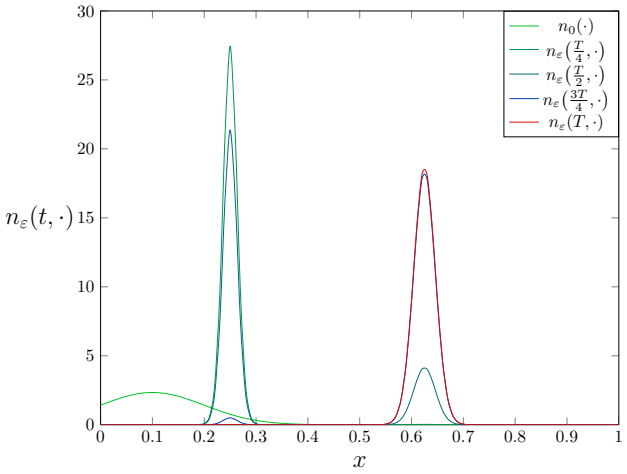


FIGURE 2.7: Plot of functions $n_\varepsilon(t,\cdot)$ at different times, with $\varepsilon = 1.10^{-5}$ and final time $T = 800$, for an initial condition centred at 0.1.

This is a relevant problem related to the actual biological time-scales at which such phenomena would occur. Capturing this mathematically requires to change time-scales, but it remains a challenging (and open) problem, as already explained in the introduction.

Part II

Population dynamics

Chapter 3

The non-local Fisher-KPP equation in a bounded domain

In this note accepted in the Comptes Rendus Mathematique, entitled *On the stability of the state* 1 *in the non-local Fisher-KPP equation in bounded domains* [138], we push the use of Lyapunov functionals based on the convex function $z \mapsto z - 1 - \ln(z)$ further to a case with diffusion. Under a strong competition assumption on the kernel, we prove that 1 attracts all trajectories. This condition is reminiscent of other ones found in the literature concerning the stability of 1.

3.1 Introduction

We consider the so-called non-local Fisher-KPP equation endowed with Neumann boundary conditons

$$\begin{aligned} &\frac{\partial u}{\partial t}(t,x) = \mu\left(1 - \int_\Omega K(x,y)u(t,y)\,dy\right)u(t,x) + \Delta u(t,x), \quad x \in \Omega,\ t > 0, \\ &\frac{\partial u}{\partial n}(t,x) = 0, \quad x \in \partial\Omega,\ t > 0, \\ &u(0,x) = u^0(x) \geqslant 0 \quad x \in \Omega, \end{aligned} \tag{3.1}$$

where Ω a regular bounded domain of $\mathbb{R}^d$ and $K > 0$ a kernel modelling an additional death rate due to non-local interactions.

We will sometimes write in short $K[u] = \int_\Omega K(x,y)u(y)\,dy$ for a generic function u.

Assuming

$$\forall x \in \Omega, \int_\Omega K(x,y)\,dy = 1, \tag{3.2}$$

and in the limit $K(x,y) \to \delta_{x-y}$, we recover the classical Fisher KPP-equation

$$\frac{\partial u}{\partial t} = \mu(1-u)u + \Delta u. \tag{3.3}$$

The assumption (3.2) ensures that 1 remains a homogeneous stationary solution of (3.1).

The classical Fisher-KPP equation (3.3) is often analysed on the whole space for the investigation of travelling waves, which are known to exist since the pioneering works of Fisher, Kolmogorov, Petrovsky and Piskunov [92] for any speed above $2\sqrt{\mu}$. Furthermore, any-non zero initial condition eventually converges locally uniformly to 1, which is therefore a globally asymptotically stable for non zero initial conditions.

When one adds a non-local term, it does not remain true that travelling waves exist and when they do, whether they link 0 to 1 or another non-homogeneous steady-state of the equation. 1 can indeed become unstable: Turing patterns appear [121, 130].

A natural question is thus to understand under which conditions the status of 1 is changed due to the non-local term. When $K(x,y)$ is given by a convolution $\phi(x-y)$, several results have already been obtained in the full space, in dimension $d = 1$. If the Fourier transform is everywhere positive of if μ is small enough, it is known that travelling waves necessarily connect 0 to 1 [12]. See also [3, 71]

In this note, we provide a general result on the global asymptotic stability on 1 on a bounded domain, based on a Lyapunov functional. The results holds provided that the following general assumption on the kernel K is satisfied:

$$\forall f \in L^2(\Omega), \int_{\Omega\times\Omega} K(x,y)f(x)f(y)\,dx\,dy \geqslant 0. \tag{3.4}$$

K is then referred to as being a positive kernel, and (3.4) can be thought of as a strong competition assumption. These types of Lyapunov functionals have been used successfully in selection equations in [82, 140, 139] and are inspired by Lyapunov functions for Lotka-Volterra ODEs [64].

It remains an open question to know whether this condition leads to the same conclusion on the whole space. As such, our Lyapunov function requires integrability for $u(t)-1-\ln(u(t))$ which is too much to ask in $\mathbb{R}^d$. We still believe that the condition (3.4) is highly relevant. Indeed, when $\Omega = \mathbb{R}^d$, and if K is a convolution $K(x,y) = \phi(x-y)$, then condition (3.4) becomes

$$\forall f \in L^2\big(\mathbb{R}^d\big), \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi(x-y)f(x)f(y)\,dx\,dy \geqslant 0. \tag{3.5}$$

It is easy to check that if ϕ has a non-negative Fourier transform, then condition (3.5) is satisfied, see [82]. The converse is almost true, as evidenced by Bochner's Theorem [143]:

if ϕ is bounded an continuous, then (3.4) holds if and only if it is the Fourier transform of a finite bounded measure on $\mathbb{R}^d$.

Consequently, condition (3.4) (or (3.5)) shows that the condition on the Fourier transform of ϕ used in dimension 1 in the literature can be appropriate in any dimension, and may not only be a sufficient but also a necessary condition when it comes to the stability of the state 1.

3.2 The Lyapunov function approach

We make the following regularity assumption on the kernel K:

$$K \in C^{0,1}\big(\overline{\Omega} \times \overline{\Omega}\big), \tag{3.6}$$

where $C^{0,1}\big(\overline{\Omega} \times \overline{\Omega}\big)$ denotes the set of Lipschitz continuous functions on $\overline{\Omega} \times \overline{\Omega}$.

Under the previous assumption (3.6), for $u^0 \in L^1(\Omega)$, we know from [45] that there exists a unique non-negative classical solution in $C([0,+\infty), L^1(\Omega)) \cap C^1((0,+\infty), C^{2,\alpha}(\Omega))$, which we denote $t \longmapsto S_t u^0$.

It will also be convenient to introduce the set $Z := \{u \in C^{2,\alpha}(\Omega),\ u \geqslant 0\}$. Finally, we define the non-negative function $H(w) := w - 1 - \ln(w)$ for $w > 0$, and for u in Z

$$V(u) := \int_\Omega (u(x) - 1 - \ln(u(x)))\ dx,$$

the last integral possibly being equal to $+\infty$.

Our result is then the following:

Theorem 3.1. *Assume (3.4), (3.6), (3.2). Then for any initial datum u^0 in $L^1(\Omega)$, $u^0 \geqslant 0$, $u^0 \neq 0$, the solution to (3.1) satisfies*

$$u(t,\cdot) \longrightarrow 1$$

uniformly in Ω.

Proof. First step: computation of the Lyapunov functional.

First, let us remark that by the parabolic strong maximum principle, $u(t,x) > 0$ for all $t > 0$, $x \in \Omega$. Now, let us check that this holds true also for $x \in \partial\Omega$, from which we will infer that $V(u(t))$ is well defined for all $t > 0$. By the parabolic strong maximum principle at the boundary, we have the following alternative for $x \in \partial\Omega$: either $u(t,x) > 0$ or $u(t,x) = 0$ and $\dfrac{\partial u}{\partial n}(t,x) < 0$. Only $u(t,x) > 0$ can hold due to the Neumann boundary conditions.

We now consider $g(t) := V(u(t))$ for $t > 0$, where $\{u(t)\}_{t\geqslant 0}$ is the trajectory emanating from by u_0. Let us prove that this is a Lyapunov functional, by computing for $t > 0$

$$\begin{aligned} g'(t) &= \int_\Omega \frac{\partial u}{\partial t}(t)\left(1-\frac{1}{u(t)}\right) \\ &= \int_\Omega \Delta u(t)\left(1-\frac{1}{u(t)}\right) - \mu\int_\Omega (1-K[u(t)])\,(1-u(t))) \\ &= -\int_\Omega \frac{|\nabla(u(t,x))|^2}{u^2(t,x)}\,dx - \mu\int_{\Omega^2} K(x,y)\,(1-u(t,x)))\,(1-u(t,y))\;dx\,dy. \end{aligned}$$

after integration by part for the first term. For the second one, we used $1-K[u] = K[1-u]$, owing to (3.2).

Thanks to (3.4), this yields $g'(t) \leqslant 0$ *i.e.*, that g is non-increasing over $\mathbb{R}_+$. Since $g \geqslant 0$, we infer the convergence of $g(t)$ as t tends to $+\infty$, and we denote l its limit.

Second step: compactness of trajectories.

Since $C^{2,\alpha}(\Omega)$ is compactly embedded into $C\big(\overline{\Omega}\big)$, the trajectory $\{S_t u^0\}_{t\geqslant\delta}$ (for some fixed $\delta > 0$) is relatively compact in $C(\overline{\Omega})$, meaning that one can find $\bar{u} \geqslant 0$ in $C(\overline{\Omega})$ and a sequence (t_k) tending to $+\infty$ in k, such that $u(t_k)$ converges to $\bar{u}$ as k goes to $+\infty$, in $C\big(\overline{\Omega}\big)$. Note that the limit cannot be identically 0 since otherwise $g(t)$ would go to $+\infty$, in contradiction with its convergence to l.

Our aim is to prove that $\bar{u} = 1$, which will mean that the whole trajectory converges to $\bar{u}$, hence the expected result.

Third step: identifying the limit.

Let us now consider the trajectory starting from the initial datum $\bar{u}$, namely $\{S_t\bar{u}\}_{t\geqslant 0}$, which we also denote by $\{\tilde{u}(t)\}_{t\geqslant 0}$. Because $\bar{u} \geqslant 0$, $\bar{u} \neq 0$, we again have $\tilde{u}(t,x) > 0$ for all $t > 0$, $x \in \overline{\Omega}$. Let us prove that V is constant along the trajectory $\{S_t\bar{u}\}_{t\geqslant 0}$ for $t > 0$.

For this, we write $V(\tilde{u}(t)) = V(S_t\bar{u}) = V\left(S_t \lim_{k\to+\infty} S_{t_k}u^0\right) = V\left(\lim_{k\to+\infty} S_{t+t_k}u^0\right)$. It is also easy to see that for any u in $C\big(\overline{\Omega}\big)$ which is furthermore positive on $\overline{\Omega}$, V (seen as acting on $C\big(\overline{\Omega}\big)$) is continuous at u, and this implies $V(\tilde{u}(t)) = \lim_{k\to+\infty} V\left(S_{t+t_k}u^0\right) = l$. As claimed the function $t \longmapsto V(S_t\bar{u})$ is constant (equal to l) for $t > 0$.

Hence its derivative must be zero for $t > 0$: from the computations made in the first step, it must hold that both $\int_\Omega \frac{|\nabla(\tilde{u}(t))|^2}{\tilde{u}^2(t)}$ and $\int_{\Omega^2} K(x,y)\,(\tilde{u}(t,x)-1))\,(\tilde{u}(t,y)-1)\;dx\,dy$ vanish identically for $t > 0$. Let us now fix $t > 0$, and from the first term, we know that $\tilde{u}(t)$ is a constant. From the second term and owing to $K > 0$, this constant must be equal to 1. By continuity of the trajectory, this also holds true at $t = 0$, *i.e.*, $\bar{u} = 1$, which ends the proof.

□

Chapter 4

Control of the 1D monostable and bistable reaction-diffusion equations

---

In this work with Emmanuel Trélat and Enrique Zuazua, we consider the problem of controlling the monostable and bistable equations on $(0, L)$ for a density of individuals $0 \leqslant y(t, x) \leqslant 1$, by means of Dirichlet controls taking their values in $[0, 1]$. We prove that the system can always be steered to invasion (steady state 1), while it is possible in the case of extinction (steady state 0) if and only if the length L of the interval domain is less than some threshold value $L^\star$, which can be computed as an integral. In the bistable case, controlling to the other homogeneous steady state $0 < \theta < 1$ relies on a staircase control strategy, and we prove that θ can be reached in finite time if and only if $L < L^\star$. The phase plane analysis of those equations is instrumental in the whole process, from reading obstacles to controllability and computing the threshold value for domain size, to the design of the path of steady states for the control strategy. This is the subject of a submitted article, entitled *Phase portrait control for 1D monostable and bistable reaction-diffusion equations.*

4.1 Introduction

For $L > 0$, $0 \leqslant T \leqslant +\infty$, we consider the following controlled reaction-diffusion equation on $(0, L) \times (0, T)$

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t, 0) = u(t),\ y(t, L) = v(t), \\ y(0) = y_0. \end{cases} \tag{4.1}$$

where f is a C^1 nonlinearity satisfying $f(0) = f(1) = 0$, with initial data $0 \leqslant y_0 \leqslant 1$ in $L^\infty(0, L)$. The Dirichlet controls u and v are measurable functions satisfying the constraints

$$0 \leqslant u(t) \leqslant 1, \quad 0 \leqslant v(t) \leqslant 1.$$

In such a setting, (4.1) admits a unique solution in

$$L^\infty((0,T) \times (0,L)) \cap C([0,T]; H^{-1}(0,L)),$$

see for instance [103].

We will consider two types of functions.

(H1) The *monostable* case: $f > 0$ on $(0,1)$. In such a case, we will also assume $f'(0) > 0$. The typical example is $f(y) = y(1-y)$.

(H2) The *bistable* case: $f < 0$ on $(0,\theta)$ and $f > 0$ on $(\theta,1)$ where $0 < \theta < 1$. In such a case, we will also assume $f'(0) < 0$ and $f'(1) < 0$. The typical example is $f(y) = y(1-y)(y-\theta)$.

We also set

$$F(y) := \int_0^y f(z)\,dz \text{ for } y \in [0,1].$$

In the case (H2), we will without loss of generality always assume $F(1) \geqslant 0$, which is equivalent to $\theta \leqslant \frac{1}{2}$ when $f(y) = y(1-y)(y-\theta)$. If $F(1) < 0$, one can set $z = 1-y$ to apply the results obtained when $F(1) \geqslant 0$.

By means of appropriately chosen Dirichlet controls $u(t)$ and $v(t)$ in $L^\infty(0,T;[0,1])$ at $x = 0$ and $x = L$ respectively, our goal is to control the equation towards either the steady states 0, 1, or in cases (H2), also towards the steady state θ.

Let us denote a a generic solution of $f(y) = 0$, namely $a = 0$, $a = 1$ or also $a = \theta$ in the case (H2). Our goal is to provide controls u, v steering the system to those homogeneous steady states. We will say that the controlled equation (4.1) is

- *controllable in finite time towards* a if for any initial condition $0 \leqslant y_0 \leqslant 1$ in $L^\infty(\Omega)$, there exist $0 \leqslant T < +\infty$, controls u, $v \in L^\infty(0,T;[0,1])$ such that

$$y(T,\cdot) = a.$$

- *controllable in infinite time towards* a if for any initial condition $0 \leqslant y_0 \leqslant 1$ in $L^\infty(\Omega)$, there exist controls u, $v \in L^\infty(0,+\infty;[0,1])$ such that

$$y(t,\cdot) \longrightarrow a$$

uniformly in $[0,L]$ as t tends to $+\infty$.

Motivations. These models are ubiquitous in population dynamics (see [5, 85, 130]) but they also appear in other contexts, e.g. in the theory of combustion. Let us use the point of view of population dynamics to introduce the main modelling aspects.

In case (H1), having in mind the example $f(y) = y(1-y)$, there is exponential increase of y whenever $y > 0$, but there is a saturation effect near $y = 1$ because the full capacity of the system has been reached. In case (H2), f takes negative values close to 0 to model the fact that a minimal density θ is required for reproduction and cooperation, under which the population will die out. The state θ is unstable in the absence of diffusion, since $f'(\theta) > 0$.

These models are also amenable to modelling invasion phenomena, because (when posed on the whole space) they typically have solutions called *travelling waves* in the form $y(x - ct)$ for certain speeds c, linking the states 0 and 1, see the pioneering work [92].

For such problems, it is thus a requirement for the solution to satisfy $y \geqslant 0$, a condition which is fulfilled with non-negative Dirichlet boundary conditions. We might consider using controls that are above 1, taking into account the possibility for releases at 0 or L to be above the capacity of the system.

However, there are contexts in which y is the proportion of individuals of type A over the total number of individuals of types A and B. This can be obtained as the suitable limit of a system of two reaction-diffusion equations for each type [160].

Thus, we shall also impose that the controls are below 1 to cover these cases, which will not be a restriction for the results. Imposing $0 \leqslant u(t), v(t) \leqslant 1$ leads to $0 \leqslant y(t, x) \leqslant 1$ by the parabolic maximum principle [142, 93].

In applications, it is common to target extinction or invasion of a given population: the goal is to reach the steady state 0 or the steady state 1. Converging to an intermediate steady state such as θ can also be desirable if the goal is to maintain the population all over the domain, but below invasion levels.

If one thinks of y as a proportion of one species over the total number of individuals in two species, reaching θ is one way of ensuring coexistence on the whole domain $(0, L)$. Doing it is a priori a more challenging task than for 0 and 1 since θ is an unstable equilibrium for the dynamical system $y' = f(y)$.

On the control for model (4.1). The literature for the control of semilinear parabolic equations such as (4.1) is abundant, whether it is by means of Dirichlet controls or controls acting inside the domain [41, 176]. The typical results (for nonlinearities small enough to avoid blow-up) when such controls are unbounded is the possibility to control towards 0 in any time $T > 0$ [94, 55, 57], but of course at the expense of controls becoming larger and larger as T becomes smaller [58].

Much effort has been recently put into studying controllability problems also in the presence of constraints on the controls, because it is a quite common assumption for applications. Such additional control constraints, and in particular non-negativity constraints, may dramatically change the types of results one can obtain [132, 107]. Controllability to 0 is no

longer granted, and when it is, a minimal time for controllability might appear. Also note that even with unbounded controls, controls on the state itself can result in difficulties for controllability and appearances of minimal times for it to hold true [108].

A simple static strategy. The simplest approach to steering the system to a homogeneous steady state a is by choosing constant controls $u(t) = v(t) = a$, a strategy we shall call *static* in what follows. A crucial result due to Matano is that any trajectory must converge to some stationary state.

Theorem 4.1 ([115]-Theorem B)**.** *Consider the equation*

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t,0) = \bar{u}, y(t,L) = \bar{v}, \\ y(0) = y_0. \end{cases} \tag{4.2}$$

with some constant controls $0 \leqslant \bar{u}, \bar{v} \leqslant 1$*. Then* $y(t,\cdot)$ *solving* (4.2) *converges uniformly to some stationary state as* $t \to +\infty$*, i.e., a solution* $\bar{y}$ *of*

$$\begin{cases} -\bar{y}_{xx} = f(\bar{y}), \\ \bar{y}(0) = \bar{u},\ \bar{y}(L) = \bar{v}. \end{cases} \tag{4.3}$$

This classical but nontrivial result is established thanks to the strong maximum principle, in the spirit of work that followed studying the number of oscillation points or of sign changes of solutions (lap-number, see [116]) as time evolves.

Note that the limit stationary state is not necessarily known: the above result only asserts its existence. Moreover, it is not necessarily unique. As a consequence, choosing $u(t) = v(t) = a$ will work asymptotically (independently of the initial condition) if the homogeneous solution a is the only solution to the above stationary problem (4.3) for $\bar{u} = \bar{v} = a$.

Threshold for domain size and obstacles. Intuitively, one can expect that if L is small, $y = a$ will be the only solution to the previous stationary problem, while if L is large, there might be others. It is indeed well-known that there exists a threshold for L under which a is the only solution, and above which there is at least one other [105]. Let us denote L_a this threshold.

When $\bar{u} = \bar{v} = a$, other stationary solutions $\bar{y}$ to (4.3) than a are obstacles for the static control strategy to work, since if $y_0 \geqslant \bar{y}$ (resp. $y_0 \leqslant \bar{y}$), then $y(t,\cdot) \geqslant \bar{y}$ (resp. $y(t,\cdot) \leqslant \bar{y}$) for the solution of the controlled model with constant controls $u(t) = v(t) = a$. This a consequence of the parabolic comparison principle.

Note that these obstacles also come up naturally for the construction of so-called *bubbles*, *i.e.*, initial conditions in the case (H2) on the whole space, which are big enough to induce invasion [161, 17].

Consequently, combining Matano's Theorem with this threshold phenomenon already yields that the static strategy (leaving aside the case of $L = L_a$ for the moment) is such that:

- for $L < L_a$, any initial condition converges asymptotically to a,
- for $L > L_a$, there exist some initial conditions for which the solution will not converge to a.

Application to invasion and extinction. Another application of the comparison principle shows that this actually settles the case of $a = 0$ and $a = 1$. We take $a = 0$ to illustrate the idea. The solution y of the controlled equation of (4.1) is such that $y(t, x) \geqslant z(t, x)$ where z solves the same equation but with $u(t) = v(t) = 0$. Thus a given control strategy will work if and only if the static strategy does.

Also note that the strong parabolic maximum principle entails that when $y_0 \neq 0$, then $y(t, x) > 0$ inside $(0, L)$: it is possible to reach the state 0 only asymptotically, and the same holds for 1. At this stage, for $a = 0$ or $a = 1$, we can state that system is not controllable to a in finite time, and that it is controllable in infinite time towards a depending on the position of L with respect to L_a.

Designing strategies for θ. The previous reasoning shows that the steady state θ will asymptotically attract all trajectories if L is small enough, more precisely if $L < L_\theta$, just by the static strategy of putting $u(t) = v(t) = \theta$ on both sides. One can then hope to reach θ in finite time, by waiting for the system to be close enough to θ in order to use a local controllability result.

Contrarily to the case of 0 and 1, the static strategy might be improved for the control towards θ since controls can take values both above and below θ. If either 0 or 1 attracts all trajectories, our idea is to try and use a path of steady states linking θ to 0 (or 1), in order to use the *staircase* method inspired by [42] and its development in [132]. It allows to steer (in finite time) any steady state to another one, as long as they are linked by a path of steady states.

Main results. In this chapter, we provide a complete understanding of controllability properties towards constant steady states for the equation (4.1), and the essential tool is phase plane analysis for the ODE $-y'' = f(y)$. First, it will allow us to prove that $L_1 = +\infty$ (due to $F(1) \geqslant 0$ which implies that 0 and 1 do not play symmetric roles), and that L_0, which we denote $L^\star$ from now on, is positive and can be computed explicitly as a transcendental integral. More precisely, we will show that

- (4.1) is controllable in infinite time towards 0 if and only if $L \leqslant L^\star$ in the case (H1) (resp. $L < L^\star$ in the case (H2)),
- (4.1) is controllable in infinite time towards 1 independently of L in both the cases (H1) and (H2).

Recall that, by parabolic comparison, controllability to 0 or 1 is never possible within finite time. Furthermore, $L^\star = \pi/\sqrt{f'(0)}$ under quite generic conditions in the case (H1). In the case (H2), let us stress that our integral formula for $L^\star$ was established for cubic nonlinearities, already with phase plane analysis in [158], but for other purposes.

Second, phase plane analysis will also be critical in understanding the controllability properties of θ. We already know from the reasonings above that θ can be reached asymptotically by the simple static strategy, which works for $L < L_\theta$. The main contribution of this chapters is the design of a control strategy which works not only for $L < L_\theta$, but more generally for $L < L^\star$. More precisely, we shall prove in the case (H2) that

$$\text{(4.1) is controllable in finite time towards } \theta \text{ if and only if } L < L^\star.$$

The proof of this equivalence as well as the design of an appropriate control strategy are instrumentally based on the phase plane analysis of the dynamical system $-y'' = f(y)$, in the region $0 \leqslant y \leqslant 1$, which involves the three steady states 0, θ and 1.

Such a strategy is far from obvious due to the instability of θ for the corresponding ODE. The main idea is to use the staircase method, together with a fine analysis of the phase plane showing that there is a path of steady states linking 0 and θ if and only if $L < L^\star$. Actually, because the controls must be non-negative, 0 is not an appropriate steady state and we shall need to find, again by phase plane analysis, another globally asymptotically stable steady state y_{init} close to 0 such that a path of steady states still links y_{init} to θ. Finally, we will also explain why there is a minimal time for controllability: one cannot hope to reach θ in arbitrarily small time.

Outline of the chapter. The chapter is organised as follows. In Section 4.2, we focus on the case of 0 and 1. Phase plane analysis allows us to recover the existence of a threshold and to find an explicit formula, together with some estimates. The problem of controllability towards θ is investigated in detail in Section 4.3, where we first recall the staircase method before using it with the help of phase plane analysis. Finally, Section 4.4 is devoted to confirming the theoretical results by numerical experiments, together with presenting some byproducts and perspectives which follow from our work.

4.2 Threshold length $L^\star$ for extinction and invasion

4.2.1 A general result for invasion

We recall that we assume $F(1) \geqslant 0$ (thus 0 and 1 do not play the same role in the bistable case).

Proposition 4.1. *Whether f satisfies (H1) or (H2), (4.1) is controllable in infinite time towards* 1.

Proof. As explained in the introduction, Matano's Theorem 4.1 and the parabolic comparison principle combined imply that (4.1) is controllable towards 1 in infinite time if and only if the only solution to

$$\begin{cases} -w_{xx} = f(w), \\ w(0) = 1,\ w(L) = 1, \end{cases} \tag{4.4}$$

with $0 \leqslant w \leqslant 1$ on $[0, L]$, is the constant 1. The equation $-w_{xx} = f(w)$ is a second-order ODE which can be rewritten as $w_x = z$, $z_x = -f(w)$, and there is a solution to the previous equation if and only if there are curves $(w(x), w'(x))$ in the phase portrait (w, w') starting and ending on the axis $w = 1$, which satisfy $0 \leqslant w \leqslant 1$. In both cases, (H1) and (H2), the only such a curve is the trivial one: $w \equiv 1$.

For completeness, we give an analytical proof. Assume there is a such a function $0 \leqslant w \leqslant 1$ which is not identically 1. Then there is $x_0 \in (0, L)$ such that w reaches its minimum, satisfying $w(x_0) < 1$. Since $w'(x_0) = 0$, the conservation of the energy $\frac{1}{2}w'^2 + F(w)$ yields $\frac{1}{2}(w'(0))^2 + F(1) = F(w(x_0))$ which implies $F(w(x_0)) \geqslant F(1)$. If f satisfies (H1) or (H2) with $F(1) > 0$, the last inequality imposes $w(x_0) = 1$, a contradiction. If f satisfies (H2) together with $F(1) = 0$, then $w'(0) = 0$. Then w would solve the second-order ODE $-w_{xx} = f(w)$ with $w(0) = 1$, $w'(0) = 0$, meaning that w would be identically 1 by Cauchy-Lipschitz uniqueness, a contradiction. □

Remark 4.1. In the case (H1), a Lyapunov functional exists and can be used to prove convergence to 1 [138]. Indeed, consider the solution to

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t, 0) = 1,\ y(t, L) = 1, \end{cases}$$

and, for $t > 0$, the functional $V(t) := \int_0^L \big(y(t, x) - 1 - \ln(y(t, x))\big)\, dx$. Then

$$\frac{dV}{dt} = -\int_0^L \left(\frac{y_x(t, x)}{y(t, x)}\right)^2 dx - \int_0^L f(y(t, x))\frac{1 - y(t, x)}{y(t, x)}\, dx \leqslant 0.$$

Up to our knowledge, however, no such Lyapunov functional has been exhibited in the case (H2).

4.2.2 A general result for extinction

Let us first note that in the case (H2) and if $F(1) = 0$, then the argument given for the state 1 in the previous section works similarly for 0, because the phase plane shows that 0 is the only solution to

$$\begin{cases} -w_{xx} = f(w), \\ w(0) = 0,\ w(L) = 0, \end{cases}$$

Thus, $F(1) = 0$ is a particular case for which (4.1) is controllable in infinite time towards 0 regardless of L. We now assume $F(1) > 0$ for the rest of this section.

Let us introduce some notations. In what follows, we will need to invert the function F.

In case (H1), F is increasing, and thus its inverse F^{-1} is well-defined, mapping $[0, F(1)]$ onto $[0, 1]$.

In case (H2) and if $F(1) > 0$, F decreases from 0 to $F(\theta)$, and then increases from $F(\theta)$ to $F(1) > 0$. There is thus a unique $\theta_1 \in (\theta, 1)$ such that $F(\theta_1) = 0$. In case (H2), we choose

to denote F^{-1} the inverse of F on $[\theta_1, 1]$ which maps $[0, F(1)]$ onto $[\theta_1, 1]$. If $F(1) = 0$, we set $\theta_1 = 1$.

Proposition 4.2. *In cases (H1) and (H2), there exists $L^\star$ such that*

- *if $L < L^\star$, (4.1) is controllable towards 0 in infinite time,*
- *if $L > L^\star$, (4.1) is not controllable towards 0 in infinite time.*

Furthermore,

$$L^\star = \inf_{\alpha \in (0, F(1))} \sqrt{2} \int_0^{F^{-1}(\alpha)} \frac{dy}{\sqrt{\alpha - F(y)}}.$$

At this stage, we do not know yet that $L^\star > 0$, nor what happens if $L = L^\star$, but this will be addressed in the next subsection.

Proof. We know that (4.1) is controllable towards 0 in infinite time if and only if the only solution to

$$\begin{cases} -w_{xx} = f(w), \\ w(0) = 0,\ w(L) = 0, \end{cases}$$

with $0 \leqslant w \leqslant 1$ on $[0, L]$, is the constant 0. There is a non-zero solution to the previous equation if and only if there are curves $(w(x), w'(x))$ in the phase portrait (w, w') starting and ending on the w'-axis having length L exactly, with the starting and ending points different from the origin.

Let us parametrise such curves by their starting point $\left(0, \sqrt{2\alpha}\right)$ where $\alpha \in (0, F(1)]$. If these curves end on the w'-axis, the end-point is $\left(0, -\sqrt{2\alpha}\right)$ and we denote $L(\alpha)$ the time required for them to reach this end-point. By symmetry, this is also twice the time for this trajectory to reach the w-axis, at a point which we denote $y^{max}(\alpha)$. To illustrate these curves, we refer to Figure 4.1 for a schematic view of the phase portrait, given in the case (H1).

Finally, we use the fact that y increases from 0 to $y^{max}(\alpha)$, which makes of y a C^1-diffeomorphism from $\left[0, \frac{1}{2}L(\alpha)\right]$ onto $[0, y^{max}(\alpha)]$, allowing us to compute

$$L(\alpha) = 2 \int_0^{\frac{L(\alpha)}{2}} dz = 2 \int_0^{y^{max}(\alpha)} \frac{dy}{y'}.$$

The energy $\frac{1}{2}(y')^2 + F(y)$ is conserved along trajectories, so that $F(y^{max}(\alpha)) = \alpha$, and inverting this yields $y^{max}(\alpha) = F^{-1}(\alpha)$. We also have $y' = \sqrt{2}\sqrt{\alpha - F(y)}$, and we arrive at

$$L(\alpha) = \sqrt{2} \int_0^{F^{-1}(\alpha)} \frac{dy}{\sqrt{\alpha - F(y)}}.$$

It is easy to check that this integral is finite, except, as we will see, for $\alpha = F(1)$. Thus, $L^\star$ is well-defined. From this formula, one clearly infers that if $L < L^\star$, there is no curve other

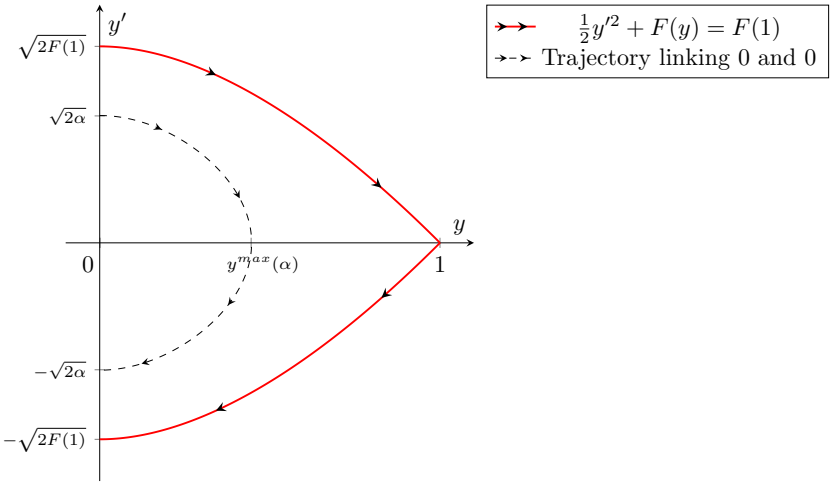


FIGURE 4.1: Phase portrait in the monostable case (here $f(y) = y(1-y)$), with the trajectory of energy $\frac{1}{2}y'^2 + F(y) = F(1)$ and an example of trajectory parametrised with α.

than 0 linking two points on the w'-axis such that the corresponding trajectory satisfies $0 \leqslant w \leqslant 1$. Thus, if $L < L^\star$, (4.1) is controllable towards 0 in infinite time.

To prove the second point, we compute $L(F(1)) = \sqrt{2}\int_0^1 \frac{dy}{\sqrt{F(1)-F(y)}} = +\infty$ because $F(1) - F(z) \underset{z\to 1}{\sim} \frac{F^{(k)}(1)}{k!}(1-z)^k$ where $k \geqslant 2$ since $F'(1) = f(1) = 0$. Consequently, $L(F(1)) = +\infty$ leading to $L(\alpha) \to +\infty$ as α tends to $F(1)$. Owing to the continuity of $\alpha \mapsto L(\alpha)$, this implies the existence of a non-zero stationary solution to (4.3) (with $\bar{u} = \bar{v} = 0$) and equivalently, the non-controllability of (4.1) towards 0 in infinite time, as soon as $L > L^\star$. □

Remark 4.2. Since $\alpha \mapsto y^{max}(\alpha)$ is increasing with α, we can instead parametrise by $\beta := y^{max}(\alpha)$ leading to the alternative formulae

$$L^\star = \inf_{\beta\in(0,1)} \sqrt{2}\int_0^\beta \frac{dy}{\sqrt{F(\beta)-F(y)}}$$

and

$$L^\star = \inf_{\beta\in(\theta_1,1)} \sqrt{2}\int_0^\beta \frac{dy}{\sqrt{F(\beta)-F(y)}}$$

in cases (H1) and (H2) respectively.

4.2.3 Estimating $L^\star$

Let us start by giving a global bound for $L^\star$, valid both in cases (H1) and (H2).

Proposition 4.3. *It holds that*

$$L^\star \geqslant \frac{\pi}{\sqrt{\max_{y\in[0,1]}\left(\frac{f(y)}{y}\right)}}.$$

Proof. Let $0 \leqslant y_0 \leqslant 1$ be given in $L^\infty(0,L)$ and consider the solution to (4.1) with null-boundary Dirichlet values. For $R := \max_{y\in[0,1]}\left(\frac{f(y)}{y}\right)$, we can bound

$$y_t - y_{xx} = f(y) = \left(\frac{f(y)}{y}\right) y \leqslant Ry \text{ on } (0,L).$$

Subsequently, y is a subsolution of the equation

$$\begin{cases} z_t - z_{xx} = Rz, \\ z(t,0) = 0,\ z(t,L) = 0, \\ z(0) = y_0. \end{cases}$$

From the comparison principle for parabolic equations, we deduce $y(t,x) \leqslant z(t,x)$. Now, using the Hilbertian basis of $L^2(0,L)$ formed by eigenvectors of the operator $A : z \mapsto -z_{xx} - Rz$, it is standard that

$$\|z(t,\cdot)\|_{L^2(0,L)} \leqslant \|y_0\|_{L^2(0,L)} e^{-\lambda_1 t}$$

where λ_1 is the first eigenvalue of A, given by $\lambda_1 := -R + \frac{\pi^2}{L^2}$. Thus it is clear that $z(t,\cdot) \to 0$ independently of y_0 as soon as $L < \frac{\pi}{\sqrt{R}}$. □

In case (H1), it is possible to obtain the actual value of $L^\star$ under an additional sufficient condition on the function f.

Proposition 4.4. *Let f satisfy (H1) be a C^2 function. Further assume*

$$f^2 \geqslant 2Ff' \text{ on } [0,1]. \tag{4.5}$$

Then

$$L^\star = \frac{\pi}{\sqrt{f'(0)}}.$$

The hypothesis clearly applies to concave functions, hence the following corollary.

Corollary 4.1. *Let f satisfy (H1). If f is strictly concave on $[0,1]$, then $L^\star = \frac{\pi}{\sqrt{f'(0)}}$. In particular, if $f(y) = y(1-y)$, then $L^\star = \pi$.*

Proof (of Proposition 4.4). Let us first prove that $\alpha \mapsto L(\alpha)$ is increasing on $(0, F(1))$. We first change variables by setting $u = \alpha F^{-1}(y)$, yielding $L(\alpha) = \sqrt{2\alpha}\int_0^1 \frac{(F^{-1})'(\alpha u)}{\sqrt{1-u}}\,du$. Now

we compute the derivative of the previous expression for $\alpha \in (0, F(1))$

$$\begin{aligned} L'(\alpha) &= \frac{1}{\sqrt{2\alpha}} \int_0^1 \frac{(F^{-1})'(\alpha u)}{\sqrt{1-u}}\, du + \sqrt{2\alpha} \int_0^1 \frac{u (F^{-1})^{(2)}(\alpha u)}{\sqrt{1-u}}\, dz \\ &= \frac{1}{\sqrt{2\alpha}} \int_0^1 \frac{(F^{-1})'(\alpha u) + 2(\alpha u)(F^{-1})^{(2)}(\alpha u)}{\sqrt{1-u}}\, du. \end{aligned}$$

If $F^{-1}(z) + 2z(F^{-1})^{(2)}(z) \geqslant 0$ for all $z \in (0, F(1))$ our claim is proved. Computing the derivatives, we find

$$(F^{-1})'(z) + 2z(F^{-1})^{(2)}(z) = \frac{1}{f(F^{-1}(z))} \left(1 - 2z \frac{f'(F^{-1}(z))}{(f(F^{-1}(z)))^2} \right).$$

Changing variables again through $z = F(y)$, the last quantity is non-negative on $(0, F(1))$ if and only if $1 - 2F(y)\frac{f'(y)}{f^2(z)} \geqslant 0$ on $(0,1)$, which is exactly the hypothesis (4.5).

At this stage, we can claim that

$$L^\star = \lim_{\alpha \to 0} L(\alpha),$$

and it remains to compute the limit. Recall that

$$L(\alpha) = \sqrt{2\alpha} \int_0^1 \frac{(F^{-1})'(\alpha u)}{\sqrt{1-u}}\, du = \sqrt{2\alpha} \int_0^1 \frac{1}{\sqrt{1-u}} \frac{1}{f(F^{-1}(\alpha u))}\, du.$$

Since $F(y) \underset{y \to 0}{\sim} \frac{F^{(2)}(0)}{2} y^2 = \frac{f'(0)}{2} y^2$, $F^{-1}(z) \underset{z \to 0}{\sim} \sqrt{\frac{2z}{f'(0)}}$. As a consequence, $f(F^{-1}(z)) \underset{z \to 0}{\sim} f'(0) F^{-1}(z) \underset{z \to 0}{\sim} \sqrt{2f'(0)z}$. Finally, we arrive at $\frac{1}{f(F^{-1}(\alpha u))} \underset{\alpha \to 0}{\sim} \frac{1}{\sqrt{2\alpha f'(0) u}}$ leading to

$$L(\alpha) \underset{\alpha \to 0}{\sim} \frac{1}{\sqrt{f'(0)}} \int_0^1 \frac{1}{\sqrt{u(1-u)}}\, du = \frac{\pi}{\sqrt{f'(0)}}$$

whence the result. □

From the previous result, we know what happens for $L = L^\star$: there is no obstacle to the convergence to 0.

Corollary 4.2. *Let f satisfy (H1) and (4.5). Then (4.1) is controllable towards 0 in infinite time if and only if $L \leqslant L^\star$.*

We now turn our attention towards the case (H2), for which there is no simple formula. When $F(1) = 0$, we set $L^\star = +\infty$ as a convention, *i.e.*, (4.1) is controllable in infinite time towards 0, whatever the value of L for this particular value of $F(1)$.

Proposition 4.5. *Let f satisfy (H2) and $F(1) > 0$. Then $\alpha \mapsto L(\alpha)$ reaches a minimum at some point of $(0, F(1))$.*

Proof. We define $g(\alpha) := \frac{1}{\sqrt{2}} L(\alpha) = \int_0^{F^{-1}(\alpha)} \frac{dy}{\sqrt{\alpha - F(y)}}$ to get rid of the constant, and split the integral in two on the intervals $[0, \theta_1]$ and $[\theta_1, F^{-1}(\alpha)]$, and change variables in the second integral as in the monostable case to uncover

$$g(\alpha) = \int_0^{\theta_1} \frac{dy}{\sqrt{\alpha - F(y)}} + \sqrt{\alpha} \int_0^1 \frac{(F^{-1})'(\alpha u)}{\sqrt{1-u}}\, du.$$

We first note that the second integral converges to 0 when α tends to 0, while the first one converges to $\int_0^{\theta_1} \frac{dy}{\sqrt{-F(y)}} = +\infty$ because $F(y) \underset{y \to 0}{\sim} \frac{f'(0)}{2} y^2$. Thus, the infimum of $L(\alpha) = \sqrt{2} g(\alpha)$ is a minimum, reached inside $(0, F(1))$. □

From the previous result, we know what happens for $L = L^\star$: there is an obstacle to the convergence to 0.

Corollary 4.3. *Let f satisfy (H2). Then (4.1) is controllable towards 0 in infinite time if and only if $L < L^\star$.*

4.3 Controlling towards θ in the bistable case

In this Section, we assume the function f to be of type (H2). We will see that if and only if $L < L^\star$, it is possible to build a control strategy steering any initial state in finite time towards the constant steady state θ.

Existence of a minimal time for controllability. Before doing so, a simple argument suffices to explain why it is not possible to steer the system to θ in arbitrarily small time. For simplicity, assume that $y_0 \in C([0, L])$, and first that y_0 is strictly above θ at least at one point in space inside Ω. As usual, whatever the controls, we can write $y(t, \cdot) \geqslant z(t, \cdot)$, where $z(t, \cdot)$ starts from y_0 but with zero Dirichlet boundary conditions.

Since the trajectory $z(t, \cdot)$ is smooth in time, it requires a positive time t_1 to be uniformly below θ, and so if there exists a time T and a control strategy such that $y(T, \cdot) = \theta$, there must hold that $T \geqslant t_1$. If y_0 is below θ somewhere inside Ω, we argue similarly by comparing to the trajectory associated with Dirichlet boundary controls equal to 1.

4.3.1 Control along a path of steady states

We will say that a steady state $\bar{y}$ associated with static controls $\bar{u}$, $\bar{v}$ is *admissible* if

$$0 < \bar{u}, \bar{v} < 1.$$

This property will be of great importance because we shall need to make small variations around the controls $\bar{u}$, $\bar{v}$ when making use of the staircase method.

Finally, we will say that there exists a *path of steady states* linking two steady states y_1 and y_2 if there is a set of steady states $\mathcal{S}$ and a continuous mapping

$$\gamma : [0,1] \longmapsto \mathcal{S}$$

such that $\gamma(0) = y_1$, $\gamma(1) = y_2$, where $\mathcal{S}$ is endowed with the $C([0,L])$-topology. The corresponding 1-parameter family of controls will be denoted by $(\bar{u}(s), \bar{v}(s))_{0\leqslant s\leqslant 1}$.

We start by giving a local exact controllability result, which holds uniformly given a family of steady states and rests on the local controllability for a single steady state, well known in the 1D case [146] and since then generalised [55, 94], see for example [132] for a full derivation. We stress that the controls provided by this result do not necessarily lie in $[0,1]$. We also emphasise that such a uniform result is possible because, by definition, steady states are taken to be between 0 and 1.

Lemma 4.1. *Assume that we have a set of steady states $\bar{y} \in \mathcal{S}$ associated with controls $\bar{u}, \bar{v}$. Let $T > 0$ be fixed. Then there exist constants $C(T) > 0$, $\delta(T) > 0$ such that for all $\bar{y} \in \mathcal{S}$, for all $0 \leqslant y_0 \leqslant 1$ in $L^\infty(0,L)$ with*

$$\|y_0 - \bar{y}\|_\infty \leqslant \delta(T),$$

there exist controls $u, v \in L^\infty(0,T;\mathbb{R})$ such that the solution of (4.1) starting at y_0 satisfies

$$y(T,\cdot) = \bar{y}.$$

Furthermore,

$$\max\left(|u(t) - \bar{u}|, |v(t) - \bar{v}|\right) \leqslant C(T)\,\|y_0 - \bar{y}\|_\infty \ \text{ on } (0,T).$$

With this lemma, we can now explain the staircase method. Applied with a path of admissible steady states, it ensures that one can steer any steady state to another one by controls with values in $[0,1]$.

Proposition 4.6. *Assume that there exists a path of admissible steady states $(\bar{y}_s)_{0\leqslant s\leqslant 1}$ associated with controls $(\bar{u}_s, \bar{v}_s)_{0\leqslant s\leqslant 1}$. Then there exists a time $T > 0$ and a control strategy $u, v \in L^\infty(0,T;[0,1])$ such that the solution of (4.1) starting at $\bar{y}_0$ satisfies*

$$y(T,\cdot) = \bar{y}_1.$$

The proof simply goes by applying a finite number of times the local controllability result along the (compact) path of steady states.

Proof. We take $T = 1$. By continuity, $\delta_0 := \min_{0\leqslant s\leqslant 1}(\bar{u}_s, \bar{v}_s, 1-\bar{u}_s, 1-\bar{v}_s) > 0$. We choose an integer N large enough such that for all $k = 1, \dots, N$,

$$\left\| \bar{y}_{\frac{k-1}{N}} - \bar{y}_{\frac{k}{N}} \right\|_\infty \leqslant \varepsilon,$$

where ε will be defined below.

For $k = 1, \dots, N$, let u_k, v_k be the controls in $L^\infty(0,1;\mathbb{R})$ such that the solution of (4.1) starting at $\bar{y}_{\frac{k-1}{N}}$ reaches exactly $\bar{y}_{\frac{k}{N}}$ at time 1. These controls are such that

$$\max\left(\left|u_k(t) - \bar{u}_{\frac{k}{N}}\right|, \left|v_k(t) - \bar{v}_{\frac{k}{N}}\right|\right) \leqslant C(1) \left\|\bar{y}_{\frac{k-1}{N}} - \bar{y}_{\frac{k}{N}}\right\|_\infty \leqslant C(1)\varepsilon$$

on $[0,1]$. In particular, $u_k(t) \geqslant \bar{u}_{\frac{k}{N}} - C(1)\varepsilon > 0$ for ε small enough. We prove similarly that u_k is bounded away from 1, and the reasoning for v_k is the same. At this stage, it suffices to define $(u, v) = (u_k(t-k), v_k(t-k))$ for $t \in (k, k+1)$ to obtain the desired control. □

4.3.2 Phase portrait in the case (H2)

To define a path of steady states linking an appropriate state (say y_{init}) to the state θ, phase plane analysis is again instrumental. We will indeed consider a path of steady states $(w_s)_{0\leqslant s\leqslant 1}$ (such that $w_0 = y_{init}$ and $w_1 = \theta$) by choosing a path of initial conditions $s \in [0,1] \mapsto (w_s(0), w_s'(0))$ in the phase plane. For some L fixed, the corresponding controls are $s \in [0,1] \mapsto (w_s(0), w_s(L))$, but there is no reason in general that $0 \leqslant w_s(L) \leqslant 1$.

However, if $0 \leqslant w_s(L) \leqslant 1$ for all $s \in [0,1]$, it yields a path of steady states, defined by the controls $s \in [0,1] \mapsto (w_s(0), w_s(L))$. The continuity of the mapping $s \mapsto w_s$ is ensured by continuity of solutions of ODEs with respect to initial conditions. To ensure that the chosen path of initial conditions does not violate $0 \leqslant w_s(L) \leqslant 1$, we must analyse further elementary properties of the phase portrait in the case (H2), an example of which we depict on Figure 4.2.

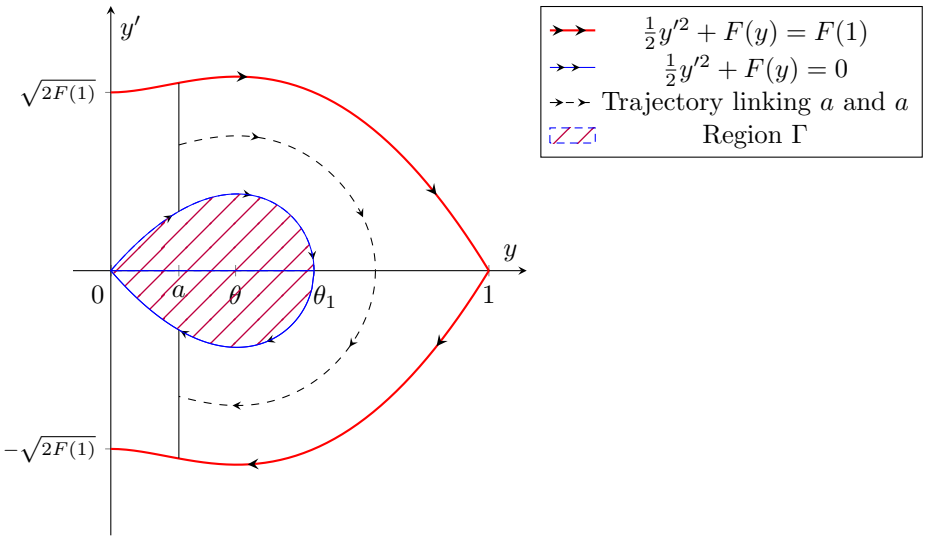


FIGURE 4.2: Phase portrait in the bistable case (H2). (Here, $f(y) = y(1-y)(y-\theta)$, $\theta = \frac{1}{3}$.) The hatched region is Γ, delimited by the trajectory of energy $\frac{1}{2}y'^2 + F(y) = 0$. Also depicted: the trajectory of energy $\frac{1}{2}y'^2 + F(y) = F(1)$ and an example of a trajectory starting and ending at a.

There are two curves of importance in the phase portrait for (H2). The first one is defined by the energy $\frac{1}{2}y'^2 + F(y) = F(1)$, while the second has energy $\frac{1}{2}y'^2 + F(y) = 0$. Note that if one starts with an initial condition along the first curve (resp. the second curve), it takes an infinite time (here, length), for the corresponding solution to the ODE $-w'' = f(w)$ to reach 1 (resp. 0).

We define Γ to be the region defined by the set of points (x, y) such that $|y| \leqslant \sqrt{-2F(x)}$, that is, those delimited by the second curve. The important result in what follows is that any initial condition $(w(0), w'(0))$ inside Γ is such that the corresponding trajectory $w(x)$ remains indefinitely between 0 and 1 (actually, between 0 and θ_1).

Finally, let us fix some $a \in [0, 1]$. We look at all the trajectories starting with $w(0) = a$ and outside the interior of Γ, namely with $\sqrt{-2F(a)} \leqslant w'(0) \leqslant \sqrt{2(F(1) - F(a))}$. In accordance with notations of the introduction, we define L_a to be the minimal time for such trajectories to reach a again. Note that with this definition, we clearly have $L_0 = L^\star$.

4.3.3 The control strategy induced by phase plane analysis

First recall the simple static strategy to try and reach θ, which consists in setting θ on the boundary. With the notations introduced in the previous Subsection 4.3.2, this is the case if and only if L is below the threshold $L_\theta < L^\star$. Consequently, this strategy is suitable but works only for smaller domains when compared with the one we are about to introduce.

Let us now define the control strategy for $L < L^\star$, based on the staircase method. The core idea is to find a path of steady states between 0 and θ, which, as we shall see, is possible if and only if $L < L^\star$. However 0 is not admissible so that we must instead resort to another close admissible steady state. We will build an admissible steady state y_{init} such that

- y_{init} can be reached asymptotically for any initial condition,
- there exists a path of admissible steady states linking y_{init} to θ.

The key lemma in order to obtain such a state is the following.

Lemma 4.2. *Let $L < L^\star$. Then for any $\varepsilon < \theta_1$ small enough, the solutions $0 \leqslant w \leqslant 1$ of*

$$\begin{cases} -w_{xx} = f(w), \\ w(0) = \varepsilon, \ w(L) = \varepsilon. \end{cases} \tag{4.6}$$

are in Γ, namely they must be such that $|w'(0)| \leqslant \sqrt{-2F(\varepsilon)}$.

Proof. With the notations of Subsection 4.3.2, L_ε tends to $L_0 = L^\star$ when ε tends to 0, and thus we can choose ε small enough such that $L < L(\varepsilon) < L^\star$. Consequently, by the very definition of $L(\varepsilon)$, there is no solution to (4.6) other than those in Γ. □

Theorem 4.2. (4.1) *is controllable towards θ in finite time (or infinite time) if and only if $L < L^\star$.*

Proof. We fix $L < L^\star$ and some initial data $0 \leqslant y_0 \leqslant 1$ in $L^\infty(0, L)$. Assume that $\varepsilon > 0$ is small enough so that the conclusions of Lemma 4.2 hold true. The idea is to first use static Dirichlet controls $u(t) = v(t) = \varepsilon$ for a long time, because Lemma 4.2 ensures that the trajectory will converge to some steady state y_{init} in Γ, independently of the initial condition. Such a steady state can then be reached exactly because of the local controllability result. Finally, the fact that y_{init} is in Γ allows us to find a path of steady states linking it to θ, so that it remains to use the staircase method.

First step. We start by approaching a steady state y_{init} in Γ. Consider the equation

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t,0) = \varepsilon,\ y(t,L) = \varepsilon, \\ y(0) = y_0. \end{cases}$$

Then, by Theorem 4.1, the solution must converge to a steady state with Dirichlet boundary conditions $(\varepsilon, \varepsilon)$. By Lemma 4.2, this is some state in Γ, which we denote y_{init}. In particular, for any $\eta > 0$, there exists $t_0 > 0$ such that for $t \geqslant t_0$, $\|y(t,\cdot) - y_{init}\|_\infty \leqslant \eta$. Thus, we start by taking $u(t) = \varepsilon$, $v(t) = \varepsilon$ on $(0, t_0)$ (η and the corresponding t_0 will be fixed appropriately in the next step).

Second step. We now make use of Lemma 4.1 with for example time 1 and choosing η (and corresponding t_0) such that $C(1)\eta$ is small enough for $\varepsilon - C(1)\eta > 0$ to hold. This provides controls $\tilde{u}, \tilde{v}$ in $L^\infty(0, 1; [0, 1])$ such that defining $u(t) = \tilde{u}(t - t_0)$, $v(t) = \tilde{v}(t - t_0)$ on $(t_0, t_0 + 1)$, we have $y(t_0 + 1, \cdot) = y_{init}$.

Third step. We build a path c of initial conditions linking the initial conditions associated with y_{init}, *i.e.* $(\varepsilon, y'_{init}(0))$, and θ, *i.e.*, $(\theta, 0)$. The simplest choice is the straight line, illustrated by Figure 4.3 below.

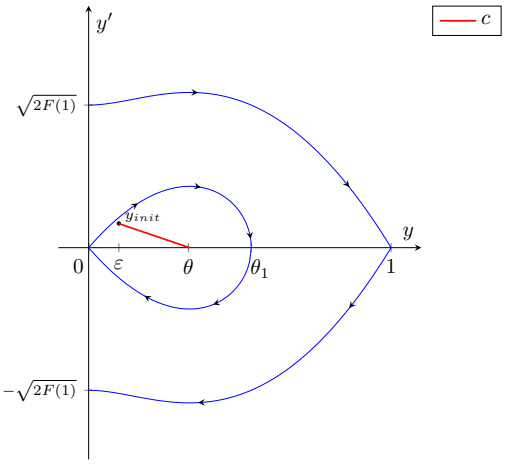


FIGURE 4.3: The path c linking the initial conditions for y_{init} and those for θ.

We denote γ the path of admissible steady states associated with c, and it now just remains to follow this path: by Theorem 4.6, there exist a time T_0 and controls u_0, v_0 in $L^\infty(0, T_0; [0,1])$ bringing y_{init} to θ. We set $u(t) = u_0(t - (t_0 + 1))$, $v(t) = v_0(t - (t_0 + 1))$ on $(t_0 + 1, t_0 + 1 + T_0)$ and $T = t_0 + 1 + T_0$. The controls u and v are indeed such that $y(t, \cdot)$ reaches exactly θ at time T.

We now prove the converse and assume $L \geqslant L^\star$. We know that there exists a non trivial solution $0 \leqslant w \leqslant 1$ to

$$\begin{cases} -w_{xx} = f(w), \\ w(0) = 0,\ w(L) = 0, \end{cases}$$

As already pointed out when it came to controlling towards 0, for any control strategy $u(t)$, $v(t)$, the solution of (4.1) with $y_0 \geqslant w$ satisfies $y(t, \cdot) \geqslant w$. If we had found a control strategy bringing us in finite (or infinite time) towards θ, we would have $w \leqslant \theta$. Let us prove that w must take values higher than θ. Let us denote $x_0 \in (0, L)$ such that w reaches its maximum. Then $w'(x_0) = 0$ and $w''(x_0) \leqslant 0$, from which we infer $f(w(x_0)) \geqslant 0$. Thus, $w(x_0)$ must either be 0, or in $[\theta_1, 1]$. It cannot be 0, θ (nor 1) because together with $w'(x_0) = 0$, Cauchy-Lipschitz uniqueness would yield $w = 0$ or $w = \theta$. Thus $w(x_0) > \theta$, and our claim is proved. □

4.4 Numerical simulations, comments and perspectives

4.4.1 A numerical optimal control approach

We consider the case (H2), and look for numerical control strategies to reach the state θ with the goal of both

- illustrating the theoretical results,
- investigating alternative strategies to the staircase one.

To this end, we consider the following optimal control problem for some final time $T > 0$:

$$\text{minimise } C_T(u, v) = \|y(T, \cdot) - \theta\|^2_{L^2(0,L)}$$

over controls $u, v \in L^\infty(0, T; [0,1])$, and where y solves (4.1).

We are interested in seeing whether, for a given $L > 0$, we can find some $T > 0$ such that this optimal control problem leads to a very small cost: this will correspond to a strategy such that $y(T, \cdot)$ is very close to θ. We do not need to reach θ exactly because we know that, once very close to it, there is a control strategy to reach it exactly, given by Lemma 4.1. In some instances, we will also force the controls to be equal to θ to illustrate when this control strategy suffices to reach θ.

To study this optimal control problem from a numerical point of view, we use direct methods. In a few words, the idea is to discretise the whole problem both in time and

space, through discretisation parameters N_t and N_x, and to solve the resulting high but finite-dimensional optimisation problem. This last step is done through the combination of automatic differentiation softwares (with the modelling language AMPL, see [60]) and expert optimisation routines (with the open-source package IpOpt, see [173]).

All the numerical experiments will be led with

$$f(y) = y(1-y)(y-\theta), \quad \theta = \frac{1}{3}, \quad y_0 = 0.1\left(\frac{x}{L}\right) + 0.8\left(1 - \frac{x}{L}\right)$$
$$N_x = 60, \quad N_t = 400.$$

For this function f and this value of θ, using the formula for $L^\star$, we find numerically $L^\star \approx 10.43$. As for the threshold L_θ (above which setting θ on the boundary makes θ globally asymptotically stable), we find $L_\theta \approx 6.29$.

We start by taking $L = 5 < L_\theta$ and impose θ on the boundary. For $T = 20$, we indeed find that this is enough to approach θ, see Figure 4.4.

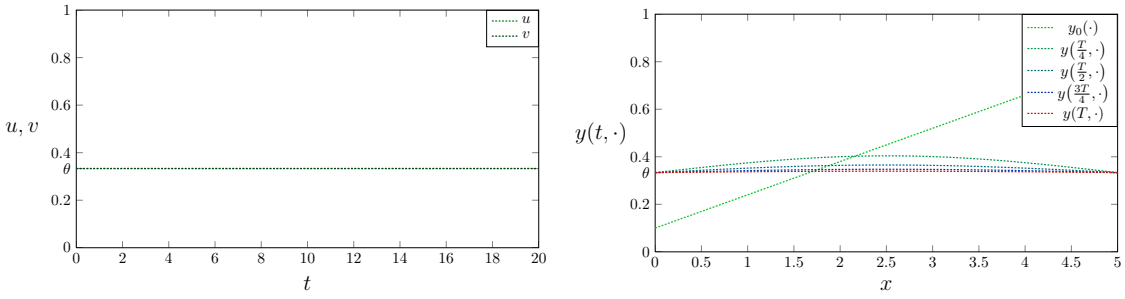


FIGURE 4.4: Static strategy $u = v = \theta$ and resulting state at times 0, $\frac{T}{4}$, $\frac{T}{2}$, $\frac{3T}{4}$ and T for $L = 5 < L_\theta$, $T = 20$.

For $L_\theta < L = 8 < L^\star$ and $T = 20$ (or larger final times), the static strategy is not enough as already known theoretically and evidenced by the upper graphs of Figure 4.5. The lower graphs show the optimal control, as obtained numerically, to reach θ: the interesting feature is that it oscillates very quickly around θ near the final time T. This is a common feature when controlling a heat equation to zero [104]. Also worth mentioning is the fact that controls take small values for a long time, which is reminiscent of the first long phase of our staircase strategy with $u(t) = v(t) = \varepsilon$ for a small ε.

For $L = 12 > L^\star$ and even for a large final time $T = 100$, the control strategy minimising the cost does not bring the final state close to θ, see Figure 4.6. One can see that the control is close to 0 for a long time, trying to bring the solution down but it remains blocked by a non-zero solution to the stationary problem with zero Dirichlet boundary conditions.

About taking same Dirichlet controls $u = v$. One important feature reflected by these simulations is that the optimal controls u and v are actually very close to one

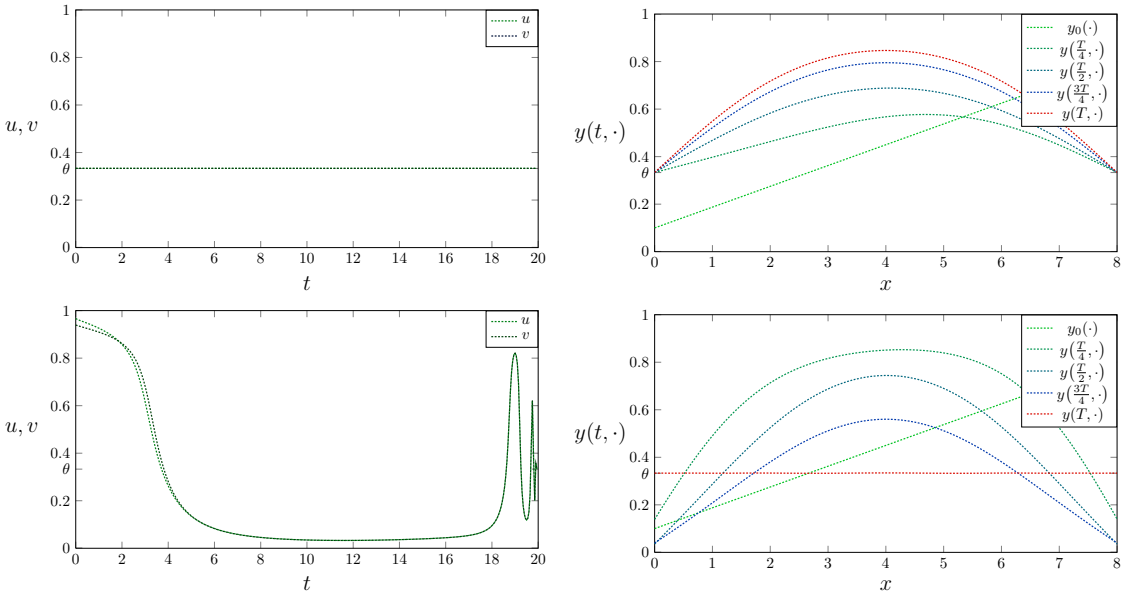


FIGURE 4.5: Constant controls $u = v = \theta$ (upper) or optimal controls u, v (lower) and resulting state at times 0, $\frac{T}{4}$, $\frac{T}{2}$, $\frac{3T}{4}$ and T for $L_\theta < L = 8 < L^\star$, $T = 20$.

another, almost equal after some time. Further simulations (not shown here) performed with $u = v$ indeed indicate that it is possible to design a control strategy with $u = v$ to reach θ, whenever $L < L^\star$. It remains an open problem to prove it, because we stress again that the strategy developed in Section 4.3 is such that, in general, $u \neq v$.

4.4.2 Comments and perspectives

Other boundary conditions and steady states. As a byproduct of our analysis, we also have proved results for

$$\begin{cases} y_t - y_{xx} = f(y), \\ y(t,0) = u(t),\ y_x(t,L) = 0, \\ y(0) = y_0, \end{cases}$$

namely the system where there is only one control at $x = 0$, while a Neumann boundary condition is enforced at the other end of the domain. Indeed, the same phase plane analysis shows that it is controllable towards 0 in infinite time if and only if (putting $u(t) = 0$ at the left end) $L \leqslant \frac{L^\star}{2}$ in the monostable case ($L < \frac{L^\star}{2}$ in the bistable case, respectively).

Simulations not shown here suggest that this system can be controlled to θ if $L < \frac{L^\star}{2}$ in the bistable case, and it is an open problem to prove it (as our control strategy requires to act on both ends).

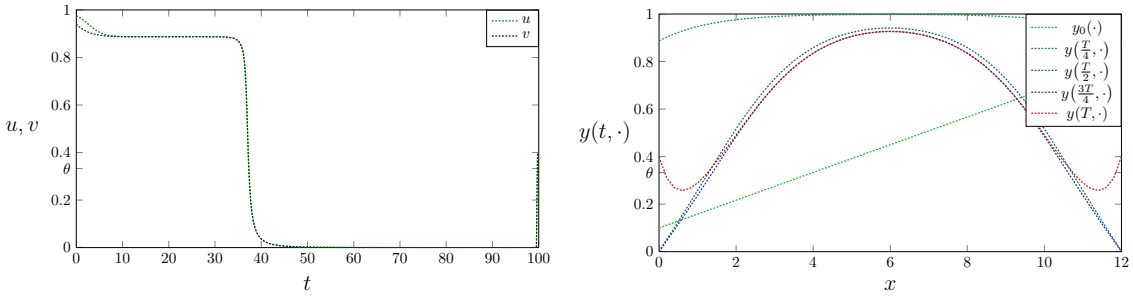


FIGURE 4.6: Optimal controls u, v and resulting state at times 0, $\frac{T}{4}$, $\frac{T}{2}$, $\frac{3T}{4}$ and T for $L = 12 > L^\star$, $T = 100$.

Also note that our approach would also work it we had Neumann controls instead of Dirichlet controls. It can also be used to reach other stationary states, while the strategy as well as possible obstacles and corresponding threshold values are all readable on the phase plane.

The multi-D case. Understanding what happens in the previous case would be critical in view of tackling the problem in higher dimension. It is indeed natural to think of situations where the control acts only on a part of the boundary, while the rest of the boundary is endowed with Neumann conditions.

If the control acts on the whole boundary, the problem of controllability towards 0 again leads to analysing whether only the trivial solution solves the stationary problem, because the result of Matano has been generalised [156]. Then, the threshold phenomenon is already known [105]. In this work, it is stated for

$$\begin{cases} -\Delta y = \lambda f(y) \text{ in } \Omega, \\ y = 0 \text{ on } \partial\Omega \end{cases}$$

where the parameter related to the domain size is λ. However, there are up to our knowledge no explicit formulae for the threshold value, although bounding like in Subsection 4.2.3 still works.

For the control towards 1 and in the monostable case (H1), the Lyapunov functional introduced in Remark 4.1 works in arbitrary dimension [138].

Feedback control. The control strategy designed for θ is not constructive when it comes to the staircase part, but it would be possible to design it in feedback form, by adapting the results of [42]. The whole control procedure would then be completely constructive, except for the time required for the solution to be close enough to y_{init} so that the local exact controllability result can be applied with controls between 0 and 1.

Part III

Optimal control for chemotherapy

Chapter 5

Theoretical and numerical study of the optimal control problem (OCP_1)

In this Chapter, we investigate the optimal control ($\mathbf{OCP_1}$) in the integro-differential case (namely $\beta_H = \beta_C = 0$), both theoretically and numerically. We manage to carry out the theoretical analysis in a smaller class, using the asymptotic analysis result from Chapter 1 which entails that with constant doses, we will reach Dirac masses. Then, the system becomes arbitrarily close to an ODE one for which the PMP with state constraints can be applied. The structure of the optimal controls is obtained and confirmed by the numerical results, conveying the idea that it is necessary to use low doses for a long time in order to let the tumour be sensitive enough for the maximum tolerated doses to work efficiently. All these results are presented in the article *Asymptotic analysis and optimal control of an integro-differential system modelling healthy and cancer cells exposed to chemotherapy*, jointy written with Jean Clairambault, Alexander Lorz and Emmanuel Trélat, and published in the Journal de Mathématiques Pures et Appliquées [139].

5.1 Introduction

One of the primary causes of death worldwide is cancer [154]. Cancer treatment encounters two main pitfalls: the emergence of drug resistance in cancer cells and toxic side effects to healthy cells. Given these causes of treatment failure, designing optimised therapeutic strategies is a major objective for oncologists. In this chapter, we develop a mathematical framework for modelling these phenomena and optimally combining therapies.

5.1.1 Overview and motivation

The most frequently used class of anti-cancer drugs are chemotherapeutic (cytotoxic) drugs, which are toxic to cells, leading to cell death. For example, platinum-based agents kill dividing cells by causing DNA damage and disrupting DNA replication [86]. Another class of drugs are cytostatic drugs, which slow down cell proliferation without killing cells. For example, trastuzumab is a cytostatic drug used in breast cancer treatment that targets growth factor receptors present on the surface of cells, and inhibits their proliferation [81]. Despite this obvious functional difference between the two classes, cytostatic drugs, such as tyrosine kinase inhibitors, can also be cytotoxic at high doses [145].

It is a well documented fact that cytotoxic agents can fail to control cancer growth and relapse [72, 127, 149]. First, eradication of the tumour cell population is compromised by the emergence of drug resistance, due to intrinsic or acquired genotypic and phenotypic heterogeneity in the cancer cell population [9, 24, 66, 123], because a subpopulation of resistant cells survives and proliferates, even in the presence of further treatment with identical [152], or higher doses [134]. Second, chemotherapeutic treatments have unwanted side effects on healthy cells, which precludes unconstrained treatment use for fear of unwanted toxicities to major organs. It is therefore a challenge for oncologists to optimally and safely treat patients with chemotherapy.

The medical objective of killing cancer cells together with preserving healthy cells from excessive toxicity is routinely translated in mathematical terms as finding the best therapeutic strategies (*i.e.*, below some maximum tolerated dose, referred to as MTD) in order to minimise an appropriately chosen cost function. There are many works in mathematical oncology focusing on the optimal modulation of chemotherapeutic doses and schedules designed to control cancer growth, *e.g.* [2, 43, 44, 91, 98, 96, 97, 95, 162, 163, 164].

Since using ordinary differential equations (ODEs) is a common technique for modelling the temporal dynamics of cell populations, the mathematical field of optimal control applied to ODEs has emerged as an important tool to tackle such questions (see for instance [151] for a complete presentation). In these ODE models, toxicity can either be incorporated in the cost functional as in [44], or by adding the dynamics of the healthy cells [14]. One simple, but rather coarse, paradigm used to represent drug resistance in such ODE models is by distinguishing between sensitive and resistant cancer cell subpopulations [44, 98]. Herein, the main tools available to obtain rigorous results are the Pontryagin maximum principle (PMP) and geometric optimal control techniques [1, 136, 150, 166].

Adaptive dynamics is a natural theoretical framework for the representation of phenotypic evolution in proliferating cell populations exposed to anti-cancer drugs and tumor microenvironmental factors. Non-Darwinian evolutionary principles have also been proposed to take into account drug resistance phenomena [133]. Adaptive dynamics is amenable to modelling these principles as well. To this end, stochastic or game-theoretic points of view (see [34, 78]) are standard in adaptive dynamics. Apart from ODEs, partial differential equations (PDEs) and integro-differential equations (IDEs) represent other deterministic approaches. The latter ones represent our focus. For an introduction to PDE and IDE

models in adaptive dynamics, we refer the interested reader to [129, 112].

We show that our model is consistent with clinical observations on the effect of constant infusion of high doses [20, 72], and we address the optimal control problem of such IDE models. Our study has a potential impact for oncologists and mathematical biologists, since it provides an accurate and robust understanding of possible optimal strategies.

5.1.2 Modelling and overview of the main results

We recall that we here consider the system presented in the General Introduction, in its integro-differential form, namely

$$\begin{aligned} \frac{\partial n_H}{\partial t}(t,x) &= R_H\left(x, \rho_H(t), \rho_C(t), u_1(t), u_2(t)\right) n_H(t,x), \\ \frac{\partial n_C}{\partial t}(t,x) &= R_C\left(x, \rho_C(t), \rho_H(t), u_1(t), u_2(t)\right) n_C(t,x), \end{aligned} \tag{5.1}$$

with the intrinsic growth rates defined as

$$\begin{aligned} R_H(x, \rho_H, \rho_C, u_1, u_2) &:= \frac{r_H(x)}{1+\alpha_H u_2} - d_H(x) I_H - u_1 \mu_H(x), \\ R_C(x, \rho_C, \rho_H, u_1, u_2) &:= \frac{r_C(x)}{1+\alpha_C u_2} - d_C(x) I_C - u_1 \mu_C(x), \end{aligned}$$

the non-local coupling as

$$I_H := a_{HH}\rho_H + a_{HC}\rho_C, \quad I_C := a_{CH}\rho_H + a_{CC}\rho_C,$$

with

$$\rho_H(t) = \int_0^1 n_H(t,x)\,dx, \quad \rho_C(t) = \int_0^1 n_C(t,x)\,dx.$$

The system starts from the initial conditions

$$n_H(0,x) = n_H^0(x) \geqslant 0, \qquad n_C(0,x) = n_C^0(x) \geqslant 0.$$

Also recall that we assume

$$\alpha_H < \alpha_C. \tag{5.2}$$

and

$$0 < a_{HC} < a_{HH}, \qquad 0 < a_{CH} < a_{CC}. \tag{5.3}$$

Asymptotic behaviour for controls in $BV([0,+\infty))$. Our first aim is to show that our model reproduces the following clinical observations: when high drug doses are administered, the tumour first reduces in size before regrowing, insensitive to further treatment.

The following statement is our first main result: using Lyapunov functionals as in Chapter 1 we achieve a complete description of the asymptotic behaviour of system (5.1), with a class of asymptotically constant controls.

Theorem 5.1. *Let u_1, u_2 be any functions in BV $([0,+\infty))$, and let $\bar{u}_1$, $\bar{u}_2$ be their limits at $+\infty$. Then, for any positive initial population of healthy and of tumour cells, $(\rho_H(t), \rho_C(t))$ converges to some equilibrium point $(\rho_H^\infty, \rho_C^\infty)$, which can be explicitly computed. Furthermore, n_H and n_C concentrate on a set of points which can also be explicitly computed.*

The explicit values can be found in Section 5.2, where this result is proved. If $\bar{u}_1 = 0$, the sets of points on which n_H and n_C concentrate are independent of $\bar{u}_2$. This is due to the fact that the phenotypic variable x models resistance to cytotoxic drugs. If μ_C vanishes identically on some interval $[1-\varepsilon, 1]$ (meaning that full resistance is possible), this theorem explains why, in the long run, high doses are not optimal. This means that our mathematical conclusions are in agreement with the idea that the standard method used in the clinic, namely administering maximum tolerated doses, should be reconsidered. Alternatives are currently extensively being investigated by oncologists, e.g., metronomic scheduling, which relies on frequent and continuous low doses of chemotherapy [10, 29, 127].

Theorem 5.1 thus motivates the optimal control problem ($\mathbf{OCP_1}$) of searching for the best possible functions u_1 and u_2 to minimise the number of cancer cells within a given horizon of time, with

$$0 \leqslant u_1(t) \leqslant u_1^{\max}, \qquad 0 \leqslant u_2(t) \leqslant u_2^{\max}, \tag{5.4}$$

and we recall the two state constraints

$$\frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)} \geqslant \theta_{HC}, \tag{5.5}$$

together with

$$\rho_H(t) \geq \theta_H \rho_H(0). \tag{5.6}$$

It might seem more natural to study the problem in free final time, but as explained later on, the mapping $T \longmapsto \rho_C(T)$ (where $\rho_C(T)$ is the optimal value obtained by solving ($\mathbf{OCP_1}$) on $[0,T]$) is decreasing in T. This implies that the optimal control problem in free final time T is ill-posed and does not admit any solution. The other implication is that when solving the optimal control problem in free final time t_f under the constraint $t_f \leqslant T$ (where T is a horizon), then the optimal solution will be such that $t_f = T$. This is why we focus on an optimal control problem in fixed final time.

In this chapter, we perform a thorough study of ($\mathbf{OCP_1}$), both theoretically and numerically. Recall that the theoretical analysis is made is made in

$$\mathcal{B}_T := \left\{ (u_1,u_2) \in \mathcal{A}_T,\ (u_1(t),u_2(t)) = (\bar{u}_1,\bar{u}_2) \text{ on } (0,T_1),\ T - T_1 \leqslant T_2^M \right\}$$

where T_2^M is given.

The reason for this restriction to this class of controls comes from the answer to the following question: given a specific tumour size (*i.e.*, a given number of cancer cells), what would be the optimal phenotypic cellular distribution in order to minimise the tumor burden at the end of the time interval? Proposition 5.1 shows that, for a very short time,

it is always better that the cancer cell population be concentrated on some appropriate phenotype, *i.e.*, that the initial population be a Dirac mass at some appropriate point.

From Theorem 5.1, we know that it is possible to asymptotically reach Dirac masses with constant controls. The combination of these two results justifies the analysis in $\mathcal{B}_T$. In this class of controls, our second main result characterises a quasi-optimal strategy in large time. The precise result and hypotheses are given by Theorem 5.2 in Section 5.3.

In order for Theorem 5.2 to hold, an important assumption we make is that when cancer cells are concentrated on a sensitive phenotype, the maximum tolerated doses will kill more cancer cells than healthy ones. Without this assumption, it is not clear whether one can expect the same strategy to be optimal, nor whether the patient can efficiently be treated.

We also emphasise that, for these IDEs, a PMP can be established but would not lead to tractable equations. The key property to still be able to identify the optimal strategy in $\mathcal{B}_T$ is that the long first phase allows us to use Theorem 5.1: both populations concentrate and their dynamics on the last phase are (approximately) governed by ODEs, as proved in Lemma 5.4. The second phase can thus be analysed with ODE techniques, here the Pontryagin maximum principle (see [1, 136, 167]). This is done in Proposition 5.2.

More concretely, Theorem 5.2 says that:

> the quasi-optimal strategy consists of:
>
> - first, administering constant doses to the patient, over a long time. The role of the first long-time arc is to allow the cancer cell population to concentrate on a sensitive phenotype. From a mathematical point of view, this means that the healthy and tumour cell populations have (almost) converged to a Dirac mass.
> - second, during a short-time phase, following a strategy composed of at most three arcs. If the first phase is such that the constraint (5.5) is saturated, then there can be a first arc along this constraint. The maximal amount of drugs is administered until the constraint (5.6) saturates. The last arc is along this constraint, with an appropriately chosen cytotoxic drug infusion which leads to a further decrease of the number of cancer cells.

Numerically, we solve the problem ($\mathbf{OCP_1}$) in $\mathcal{A}_T$. The simulations confirm the theoretical results and show that, with the chosen set of parameters, the strategy indeed approximately consists of these two phases for T large. We also compare the optimal strategy with a periodic one, and verify that the former performs better than the latter.

Furthermore, the numerical results suggest that for generic parameters, the optimal choice of constant controls on the first phase is such that the constraint (5.5) is saturated. Thus, the second phase possibly starts on this constraint.

Another important property highlighted by the numerical simulations is that, given the choice of parameters made, ρ_C can decrease arbitrarily close to 0 once the cancer cell

population has concentrated on a sensitive enough phenotype. We thus find a strategy for which $T \longmapsto \rho_C(T)$ is decreasing to 0; hence, there would be no solution to ($\mathbf{OCP_1}$) if the final time T were let free.

This is the first time that a mathematical model based on integro-differential equations demonstrates that, within our modelling framework, immediate administration of maximal tolerated drug doses, or a periodic treatment schedule, is not an optimal solution for eradicating cancer. Here, we prove that it is better to allow the phenotypes to concentrate, before administering maximal doses.

The chapter is organised as follows. Section 5.2 is devoted to the proof of Theorem 5.1 and to numerical simulations showing how the model can reproduce the regrowth of a cancer cell population. Using these results, we have theoretical and numerical grounds for our claim that constant doses are sub-optimal and we then turn our attention to ($\mathbf{OCP_1}$). In Section 5.3, several arguments are given to justify the restriction to the class $\mathcal{B}_T$, with a long first phase. The rest of the section is then devoted to proving Theorem 5.2. The numerical solutions of ($\mathbf{OCP_1}$) in $\mathcal{A}_T$ are provided in Section 5.4. They are compared to periodic strategies. In Section 5.5, we conclude with several comments and open questions.

5.2 Constant infusion strategies

This section is devoted to the asymptotic analysis of the IDE model (5.1), in order to specifically understand the effect of giving constant doses on the long run.

5.2.1 Asymptotics for the complete model: proof of Theorem 5.1

Now, let us take into account the complete coupling between healthy and tumour cells. For the remaining part of this chapter, we assume for simplicity that both n_H^0 and n_C^0 are continuous and positive on $[0,1]$. A further technical assumption is needed to prove that convergence and concentration hold, namely that the functions are Lipschitz continuous:

$$r_H, r_C, d_H, d_C, \mu_H, \mu_C \in C^{0,1}(0,1). \tag{5.7}$$

We use the Lyapunov functional technique introduced in Chapter 1, which needs to be slightly adapted to the fact that controls are asymptotically constant, not only constant.

Recall that the limits of ρ_H, ρ_C and the sets on which n_H, n_C concentrate are defined as follows. We invert the system

$$\begin{aligned} a_{HH}\rho_H^\infty + a_{HC}\rho_C^\infty &= I_H^\infty, \\ a_{CH}\rho_H^\infty + a_{CC}\rho_C^\infty &= I_C^\infty, \end{aligned}$$

where $I_H^\infty \geqslant 0$ is the smallest nonnegative real number such that

$$\frac{r_H(x)}{1+\alpha_H\bar{u}_2} - \bar{u}_1\mu_H(x) \leqslant d_H(x) I_H^\infty,$$

and $I_C^\infty \geqslant 0$ is the smallest nonnegative real number such that

$$\frac{r_C(x)}{1+\alpha_C \bar{u}_2} - \bar{u}_1 \mu_C(x) \leqslant d_C(x) I_C^\infty.$$

Furthermore, if this convergence holds true, then n_H (resp. n_C) concentrate on A_H (resp. A_C) defined as

$$A_H = \left\{ x \in [0,1],\ \frac{r_H(x)}{1+\alpha_H \bar{u}_2} - \bar{u}_1 \mu_H(x) - d_H(x) I_H^\infty = 0 \right\},$$
$$A_C = \left\{ x \in [0,1],\ \frac{r_C(x)}{1+\alpha_C \bar{u}_2} - \bar{u}_1 \mu_C(x) - d_C(x) I_C^\infty = 0 \right\}.$$

Proof of Theorem 1. We adapt the proof of Chapter 1 for these BV controls: we choose any couple of measures (n_H^∞, n_C^∞) in $\mathcal{M}^1(0,1)$ satisfying $\int_0^1 n_i^\infty(x)\,dx = \rho_i^\infty$, and also

$$\mathrm{supp}(n_H^\infty) \subset A_H,\ \mathrm{supp}(n_C^\infty) \subset A_C. \tag{5.8}$$

For $i = H, C$, and $m_i := \frac{1}{d_i}$, we define the Lyapunov functional as

$$V(t) := \lambda_H V_H(t) + \lambda_C V_C(t),$$

where

$$V_i(t) = \int_0^1 m_i(x) \left[n_i^\infty(x) \ln\left(\frac{1}{n_i(t,x)} \right) + (n_i(t,x) - n_i^\infty(x)) \right] dx,$$

with positive constants λ_H and λ_C to be adequately chosen later.

In what follows, we skip dependence in t in the functions R_H and R_C to increase readability. This time, we have

$$\frac{dV_H}{dt} = \int_0^1 m_H(x)\ \left(R_H\left(x, \rho_H, \rho_C, u_1, u_2\right) - R_H\left(x, \rho_H^\infty, \rho_C^\infty, u_1, u_2\right)\right) \left[n_H(t,x) - n_H^\infty(x)\right]\ dx$$
$$+ \int_0^1 m_H(x)\, R_H\left(x, \rho_H^\infty, \rho_C^\infty, u_1, u_2\right) \left[n_H(t,x) - n_H^\infty(x)\right]\ dx$$

The first term is simply

$$\int_0^1 m_H(x)\ \left(R_H\left(x, \rho_H, \rho_C, u_1, u_2\right) - R_H\left(x, \rho_H^\infty, \rho_C^\infty, u_1, u_2\right)\right) \left[n_H(t,x) - n_H^\infty(x)\right]$$
$$= \int_0^1 m_H(x) d_H(x)\ \left[a_{HH}(\rho_H^\infty - \rho_H) + a_{HC}(\rho_C^\infty - \rho_C)\right] \left[n_H(t,x) - n_H^\infty(x)\right]\ dx$$
$$= -a_{HH}(\rho_H^\infty - \rho_H)^2 - a_{HC}(\rho_C^\infty - \rho_C)(\rho_H^\infty - \rho_H)$$

The second term can also be written as

$$\begin{aligned}
B_H(t) &:= \int_0^1 m_H(x)\, R_H\left(x, \rho_H^\infty, \rho_C^\infty, u_1, u_2\right) \left[n_H(t,x) - n_H^\infty(x)\right]\, dx \\
&= \int_0^1 m_H(x)\, R_H\left(x, \rho_H^\infty, \rho_C^\infty, \bar{u}_1, \bar{u}_2\right) \left[n_H(t,x) - n_H^\infty(x)\right]\, dx \\
&+ \int_0^1 m_H(x)\, \left(R_H\left(x, \rho_H^\infty, \rho_C^\infty, u_1, u_2\right) - R_H\left(x, \rho_H^\infty, \rho_C^\infty, \bar{u}_1, \bar{u}_2\right)\right) \left[n_H(t,x) - n_H^\infty(x)\right]\, dx \\
&= \int_0^1 m_H(x) R_H\left(x, \rho_H^\infty, \rho_C^\infty, \bar{u}_1, \bar{u}_2\right) n_H(t,x)\, dx \\
&+ \int_0^1 m_H(x) \left[r_H(x) \left(\frac{1}{1+\alpha_H u_2} - \frac{1}{1+\alpha_H \bar{u}_2}\right) + \mu_H(x)(\bar{u}_1 - u_1)\right] \left[n_H(t,x) - n_H^\infty(x)\right]\, dx,
\end{aligned}$$

where we use (5.8) for the last equality. Note that the first term in the last expression is nonpositive by definition of $(\rho_H^\infty, \rho_C^\infty)$, and the second goes to 0 as t goes to $+\infty$. Consequently, the decomposition

$$B_i = \tilde{B}_i + E_i, \quad i = H, C, \tag{5.9}$$

holds, with $\tilde{B}_H$, $\tilde{B}_C$ nonpositive, and E_H, E_C which asymptotically vanish. This decomposition will be important in the last step.

Eventually, we have:

$$\frac{dV}{dt} = -\frac{1}{2} X^T M X + \lambda_H B_H + \lambda_C B_C$$

with $M = A^T D + DA$, $X = \begin{pmatrix} \rho_H^\infty - \rho_H \\ \rho_C^\infty - \rho_C \end{pmatrix}$, $D = \begin{pmatrix} \lambda_H & 0 \\ 0 & \lambda_C \end{pmatrix}$ and $A = \begin{pmatrix} a_{HH} & a_{HC} \\ a_{CH} & a_{CC} \end{pmatrix}$.
We choose $\lambda_H := \frac{1}{a_{HC}}$ and $\lambda_C := \frac{1}{a_{CH}}$ as in Proposition 1.2 of Chapter 1, so that $\det(M) > 0$ by assumption (5.3).
Our aim is to prove that $-\frac{1}{2} X^T M X$ converges to 0 as t goes to $+\infty$, which will yield the convergence of (ρ_H, ρ_C).

We have as in Chapter 1 the lower estimate $V(t) \geqslant -C\left(\ln(t) + 1\right)$, which uses assumption (5.7).

We now set

$$G := -\frac{1}{2} X^T M X + 2\left(\lambda_H B_H + \lambda_C B_C\right).$$

Differentiating G, we no longer find that G is non-decreasing as in Chapter 1, as new terms come up from the derivatives of u_1, u_2:

$$\frac{dG}{dt} \geqslant -\left(a(t)\frac{du_2}{dt} + b(t)\frac{du_1}{dt}\right) \tag{5.10}$$

where a and b are bounded functions defined by

$$a(t) := 2\lambda_H \frac{1}{(1+\alpha_H u_2)^2} \int_0^1 m_H(x) r_H(x)(n_H(t,x) - n_H^\infty(x))\,dx + 2\lambda_C \frac{1}{(1+\alpha_C u_2)^2} \int_0^1 m_C(x) r_C(x)(n_C(t,x) - n_C^\infty(x))\,dx,$$

$$b(t) := 2\lambda_H \int_0^1 m_H(x)\mu_H(x)(n_H(t,x) - n_H^\infty(x))\,dx + 2\lambda_C \int_0^1 m_C(x)\mu_C(x)(n_C(t,x) - n_C^\infty(x))\,dx.$$

We conclude by noting that, by $\frac{dV}{dt} \leqslant \frac{1}{2}G$, it follows that $V(t) - V(0) \leqslant \frac{1}{2}\int_0^t G(s)\,ds$. From $G(s) = G(t) - \int_s^t \frac{dG}{dt}(z)\,dz$, using (5.10) and by integrating the previous inequality, we have

$$\frac{V(t)-V(0)}{t} \leqslant \frac{1}{2}G(t) + \frac{1}{2t}\int_0^t \int_s^t \left(a(z)\frac{du_2}{dt}(z) + b(z)\frac{du_1}{dt}(z)\right) dz\,ds.$$

Now, using the decomposition (5.9) introduced previously, we obtain:

$$2\frac{V(t)-V(0)}{t} - \frac{1}{t}\int_0^t \int_s^t \left(a(z)\frac{du_2}{dt}(z) + b(z)\frac{du_1}{dt}(z)\right) dz\,ds - 2\left(\lambda_H E_H + \lambda_C E_C\right) \leqslant -\frac{1}{2}X^T M X + 2\left(\lambda_H \tilde{B}_H + \lambda_C \tilde{B}_C\right).$$

In other words, since the right-hand side of this inequality consists of nonpositive terms, the claim on the convergence of ρ_H and ρ_C is proved if we establish that the left-hand side tends to 0.

As a consequence of the lower estimate V, $2\frac{V(t)-V(0)}{t}$ converges to 0. This is also true for $2\left(\lambda_H E_H + \lambda_C E_C\right)$. It thus remains to analyse the asymptotic behaviour of the function $\frac{1}{t}\int_0^t \int_s^t \left(a(z)\frac{du_2}{dt}(z) + b(z)\frac{du_1}{dt}(z)\right) dz\,ds$. The analysis relies on the following lemma.

Lemma 5.1. *Let ϕ in $L^\infty(0,+\infty)$, and u in $BV([0,+\infty))$. Then*

$$\lim_{t\to+\infty} \frac{1}{t}\int_0^t \int_s^t \phi(z)u'(z)\,dz\,ds = 0.$$

Proof. Let us start by writing

$$\begin{aligned}\frac{1}{t}\int_0^t \int_s^t \phi(z)u'(z)\,dz\,ds &= \frac{1}{t}\int_0^t \int_0^t \phi(z)u'(z)\,dz\,ds - \frac{1}{t}\int_0^t \int_0^s \phi(z)u'(z)\,dz\,ds \\ &= \int_0^t \phi(z)u'(z)\,dz - \frac{1}{t}\int_0^t \int_0^s \phi(z)u'(z)\,dz\,ds \\ &= \Gamma(t) - \frac{1}{t}\int_0^t \Gamma(s)\,ds\end{aligned}$$

where $\Gamma(t) := \int_0^t \phi(z)u'(z)\,dz$. The expression above can thus be decomposed as the function Γ minus its Cesàro average. To conclude, it suffices that Γ has a limit at $+\infty$, which in turn is true as soon as $\phi\,u'$ is integrable on the half-line. This fact is a direct consequence of the boundedness of ϕ and the integrability of the derivative of a BV function on $[0,+\infty)$. □

This ends the proof of Theorem 5.1.

Remark 5.1. This theorem means that under general conditions, both populations concentrate and the total number of healthy (resp., cancer) cells converge. In the case of constant controls and when there is selection of a unique phenotype in both populations, it provides a complete understanding of the mapping

$$(\bar{u}_1, \bar{u}_2) \longmapsto (x_H^\infty, x_C^\infty, \rho_H^\infty, \rho_C^\infty),$$

where $\rho_H^\infty \delta_{x_H^\infty}$ and $\rho_C^\infty \delta_{x_C^\infty}$ are the respective limits of $n_H(t,\cdot)$ and $n_C(t,\cdot)$ in $\mathcal{M}^1(0,1)$, as t goes to $+\infty$. In particular, if we restrict ourselves to constant controls and a large time T, the problem of minimising $\rho_C(T)$ is equivalent to minimising ρ_C^∞ as a function of $(\bar{u}_1, \bar{u}_2)$.

5.2.2 Mathematical simulations of the effect of constant drug doses

Throughout the study, we will consider the following numerical data, taken from [111]:

$$r_H(x) = \frac{1.5}{1+x^2}, \quad r_C(x) = \frac{3}{1+x^2},$$

$$d_H(x) = \frac{1}{2}(1-0.1x), \quad d_C(x) = \frac{1}{2}(1-0.3x),$$

$$a_{HH} = 1, \quad a_{CC} = 1, \quad a_{HC} = 0.07, \quad a_{CH} = 0.01,$$

$$\alpha_H = 0.01, \quad \alpha_C = 1,$$

$$u_1^{\max} = 3.5, \quad u_2^{\max} = 7,$$

and the initial data

$$n_H(0,x) = K_{H,0}\exp(-(x-0.5)^2/\varepsilon), \quad n_C(0,x) = K_{C,0}\exp(-(x-0.5)^2/\varepsilon),$$

with $\varepsilon > 0$ small (typically, we will take either $\varepsilon = 0.1$ or $\varepsilon = 0.01$), and where $K_{H,0} > 0$ and $K_{C,0} > 0$ are such that

$$\rho_H(0) = 2.7, \quad \rho_C(0) = 0.5.$$

The value $\rho_H(0)$ is not the same as in [111]: it is chosen to be slightly below the equilibrium value of the system with $n_C \equiv 0$, $u_1 \equiv 0$, $u_2 \equiv 0$, in accordance with the fact that there is *homeostasis* in a healthy tissue. Indeed, we start with a non-negligible tumour which must have (due to competition) slightly lowered the number of healthy cells with comparison to a normal situation.

We also define $\rho_{CS}(t) := \int_0^1 (1-x) n_C(t,x)\,dx$, which may be seen as the total number at time t of tumour cells that are sensitive, and $\rho_{CR}(t) := \int_0^1 x n_C(t,x)\,dx$, which may be seen as the total number at time t of tumour cells that are resistant.

Of course, sensitivity/resistance being by construction a non-binary variable, the weights x and $1-x$ are an example of a partition between a relatively sensitive class and a relatively resistant class in the cancer cell population; other choices might be made for these weights, e.g., x^2 and $1-x^2$.

Discussion of the choice for μ_H and μ_C. These functions measure the efficiency of the drugs treatment. The choice done in [111] is

$$\mu_H(x) = \frac{0.2}{0.7^2 + x^2}, \quad \mu_C(x) = \frac{0.4}{0.7^2 + x^2}.$$

However, with this choice of functions, if we take constant controls u_1 and u_2, with

$$\bar{u}_1 = u_1^{\max} = 3.5, \qquad \bar{u}_2 = 2,$$

then we can kill all tumour cells (at least, they decrease exponentially to 0), and no optimisation is necessary. The results of a simulation can be seen on Figure 5.1. The

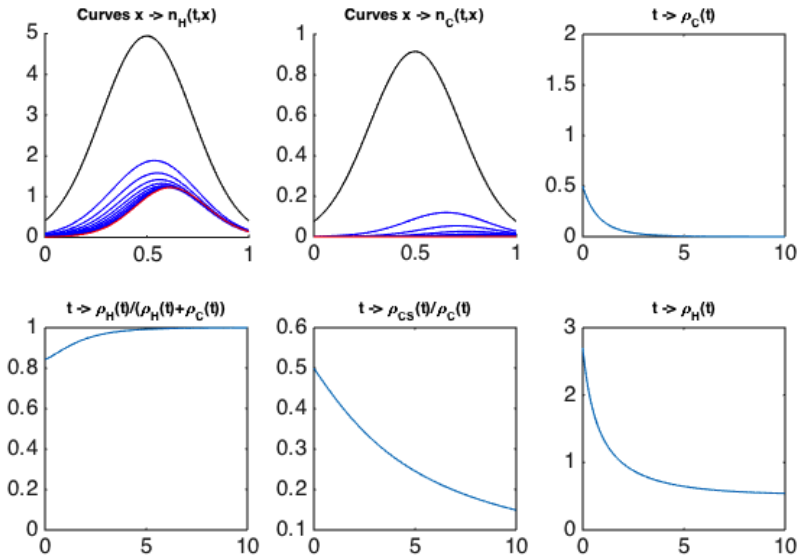


FIGURE 5.1: Simulation with $\bar{u}_1 = 3.5$ and $\bar{u}_2 = 2$, in time $T = 10$. At the top, left and middle: evolution in time of the curves $x \mapsto n_H(t,x)$ and $x \mapsto n_C(t,x)$, with the initial conditions in black, and the final ones in red. At the right, top and bottom: graphs of $t \mapsto \rho_C(t)$ and of $t \mapsto \rho_H(t)$. At the bottom, left and middle: graphs of $t \mapsto \frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)}$ and of $t \mapsto \frac{\rho_{CS}(t)}{\rho_C(t)}$.

population of tumour cells is Gaussian-shaped, decreases exponentially to 0 while its center is being shifted to the right: it means that tumour cells become more and more resistant as

time goes by. This is in agreement with the fact that cells acquire resistance to treatment when drugs are given constantly. However, although the proportion of sensitive cells $t \mapsto \frac{\rho_{CS}(t)}{\rho_C(t)}$ is quickly decreasing, the drugs are still efficient at killing the cells. This is not realistic, as it does not match the clinically observed saturation phenomenon. Most cancer cells have acquired resistance and any immediate further treatment should have no effect.

In the simulation above, there is no saturation because the function μ_C is continuous and positive over the whole interval $[0, 1]$ and is not small enough close to 1. In order to model this saturation phenomenon, we choose to modify the model used in [111], by modifying slightly the function μ_C. The new function μ_C that will throughout be considered is defined by

$$\mu_C(x) = \max\left(\frac{0.9}{0.7^2 + 0.6x^2} - 1, 0\right).$$

On Figure 5.2, the former function μ_C is in blue, and the new one is in red. This new

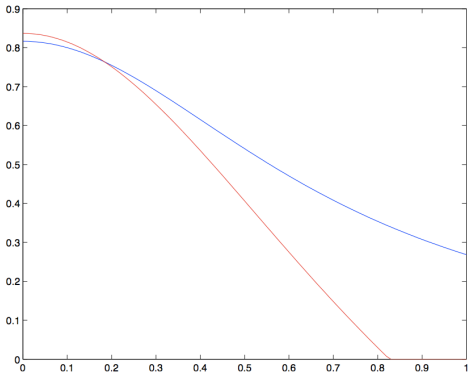


FIGURE 5.2: Former function μ_C in blue, and new function μ_C in red.

function μ_C is nonnegative and decreasing on $[0, 1]$, and vanishes identically on a subinterval containing $x = 1$.

With this new function, the simulation of Figure 5.1, with $\bar{u}_1 = 3.5$ and $\bar{u}_2 = 2$, is completely modified, as can be seen on Figure 5.3. Indeed, this time, the strategy consisting of taking constant controls $\bar{u}_1 = 3.5$ and $\bar{u}_2 = 2$ is not efficient anymore and does not allow for (almost) total eradication of the tumour. In sharp contrast, we observe on Figure 5.1 that the tumour cells are growing again, moreover concentrating around some resistant phenotype.

Conclusion on constant controls. The simulations show that choosing constant doses too high leads to the selection of resistant cells, and then, to regrowth of the cancer cell population if these cells can become insensitive to the treatment. With the notations of Theorem 5.1, it is because among constant controls, (u_1^{max}, u_2^{max}) does not minimise ρ_C^∞. However, it is quite clear that choosing the optimal constant dose $(\bar{u}_1, \bar{u}_2)$ to minimise

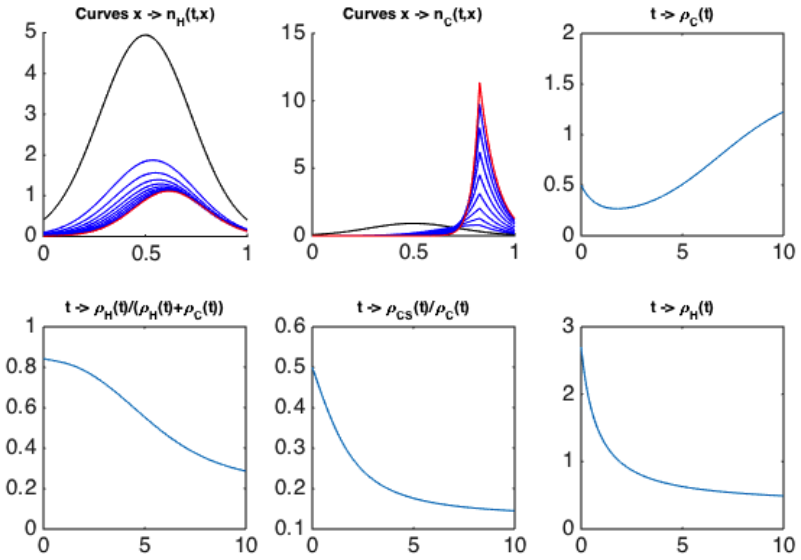


FIGURE 5.3: Simulation with $\bar{u}_1 = 3.5$ and $\bar{u}_2 = 2$, in time $T = 10$, with the new function μ_C.

ρ_C^∞ leaves room for improvement, as the choice (u_1^{max}, u_2^{max}) is still the optimal one for sensitive cells. Therefore, it makes sense to allow u_1 and u_2 to be any functions satisfying (5.1) as in ($\mathbf{OCP_1}$), which we will study both from the theoretical and numerical points of view in the next two sections.

5.3 Theoretical analysis of ($\mathbf{OCP_1}$)

Before analysing ($\mathbf{OCP_1}$), let us first consider a much simpler ODE model for which we can find the solution explicitly in order to develop some intuition.

5.3.1 Simplified optimal control problems

We consider the ODE

$$\begin{aligned} \frac{d\rho}{dt} &= (r - d\rho(t) - \mu u(t))\rho(t), \\ \rho(0) &= \rho_0 > 0. \end{aligned} \tag{5.11}$$

(C1) Optimal control problem: minimise $\rho(T)$ over all possible solutions of (5.11) with a L^1-constraint on u, *i.e.*,

$$\int_0^T u(t)\,dt \leq u^{1,max}. \tag{5.12}$$

Lemma 5.2. *The optimal solution for problem (C1) is*

$$u_{opt} = u^{1,max}\delta_{t=T}.$$

Remark 5.2. The statement can be misleading: we actually prove that there is no optimal solution, but rather that the problem leads to an infimum. Still, writing $u_{opt} = u^{1,max}\delta_{t=T}$ makes sense as a way to obtain the infimum is to take a family $(u_\varepsilon)_{\varepsilon>0}$ in L^1 which converges to u_{opt}, for example $u_\varepsilon := \frac{1}{\varepsilon}u^{1,max}\mathbb{1}_{[T-\varepsilon,T]}$.

Adding another the constraint, we have a second optimal control problem

(C2) minimise $\rho(T)$ over all possible solutions of (5.11) with the L^1-constraint (5.12) and a L^∞-constraint

$$u \leq u^{\infty,max}.$$

We assume $u^{\infty,max}T > u^{1,max}$, since otherwise it is clear that the optimal strategy is $u^{\infty,max}\mathbb{1}_{\{[0,T]\}}$.

Lemma 5.3. *We define* $T_1(T) := T - \frac{u^{1,max}}{u^{\infty,max}}$. *The optimal solution for problem (C2) is*

$$u_{op} = u^{\infty,max}\mathbb{1}_{\{[T_1,T]\}}.$$

The proofs of these two results can be found in Appendix 5.6.

Remark 5.3. The previous lemmas on simplified equations give some insight on two important features:

• for large times, constant controls lead to concentration, as evidenced by Theorem 5.1. As explained more rigorously further below in Lemma 5.4, when the populations are concentrated on some single phenotypes, the integro-differential equations boil down to ODEs, for which the last results and standard techniques from optimal control theory apply.

• for ODE models, it is optimal to use the maximal amount of drug at the end of the time-window if there is a L^1 constraint on the control. Avoiding the emergence of resistance will indirectly act as some L^1 constraint, which is why this result also provides some interesting intuition on the optimal control problem ($\mathbf{OCP_1}$).

5.3.2 Assumptions and further remarks

Let us start by mentioning a possible alternative state constraint for ($\mathbf{OCP_1}$).

Remark 5.4. Alternatively to (5.5), we might want to directly control the number of cancer cells and replace (5.5) by

$$\rho_C(t) \leqslant C^{max} \tag{5.13}$$

for some $C^{max} > 0$. The set of constraints (5.5)-(5.6) on the one hand, and (5.13)-(5.6) on the other hand, are similar. Although we focus on the first one in the sequel, our analysis applies to the other set of constraints.

We now make several important additional assumptions which will be used throughout this section, all relying on the notations of Theorem 5.1. Our first assumption is that

for any $0 \leqslant \bar{u}_1 \leqslant u_1^{max}$, $0 \leqslant \bar{u}_2 \leqslant u_2^{max}$, A_H *and* A_C *are reduced to singletons.* (5.14)

In this case, recall that Theorem 5.1 provides a mapping $(\bar{u}_1, \bar{u}_2) \mapsto (x_H^\infty, x_C^\infty, \rho_H^\infty, \rho_C^\infty)$, and with a slight abuse of notation, we will omit the dependence in $(\bar{u}_1, \bar{u}_2)$ in the following final assumptions:

whenever $\bar{u}_1$, $\bar{u}_2$ *are admissible (i.e., such that neither the constraint (5.5) nor the constraint (5.6) is violated), we require that the solution of the ODE system*

$$\begin{aligned}\frac{d\rho_H}{dt} &= R_H\left(x_H^\infty, \rho_H, \rho_C, u_1^{max}, u_2^{max}\right)\rho_H,\\ \frac{d\rho_C}{dt} &= R_C\left(x_C^\infty, \rho_C, \rho_H, u_1^{max}, u_2^{max}\right)\rho_C,\end{aligned}$$

with initial data $(\rho_C^\infty, \rho_H^\infty)$, *has the following property:*

$$\frac{d}{dt}\rho_C < 0,\ \frac{d}{dt}\rho_H < 0 \ \text{ and } \ \frac{d}{dt}\frac{\rho_C}{\rho_H} < 0. \tag{5.15}$$

The assumption (5.15) means that both populations of cells decrease but that the treatment is more efficient on cancer cells. In some sense, this is a curability assumption and it will be crucial in the sequel.

We now motivate the choice of restricting our attention to the class $\mathcal{B}_T$ by giving two results.

5.3.3 Optimality of a concentrated initial population for a small time

Here, we assume that for any $0 \leqslant \rho_C \leqslant \rho_C^{max}$, $0 \leqslant \rho_H \leqslant \rho_H^{max}$, $0 \leqslant u_1 \leqslant u_1^{max}$ and $0 \leqslant u_2 \leqslant u_2^{max}$,

$$x \mapsto R_C(x, \rho_C, \rho_H, u_1, u_2) \text{ has a unique minimum.} \tag{5.16}$$

For a given initial amount of cancer cells $\rho_C^0 > 0$, we define:

$$A_{\rho_C^0} := \left\{ n_C^0 \in \mathcal{M}^1(0,1) \text{ such that } \int_0^1 n_C^0(x)\,dx = \rho_C^0 \right\}.$$

For $n_C^0 \in A_{\rho_C^0}$, and given n_H^0 in $\mathcal{M}^1(0,1)$, final time $t_f > 0$, and controls u_1, u_2 in $BV(0,t_f)$ satisfying (5.4), we consider the associated trajectory $(n_H(\cdot,x), n_C(\cdot,x))$ on $[0,t_f]$ solution to the system (5.1) starting from (n_H^0, n_C^0).

We consider the following minimisation problem

$$\inf_{\substack{0\leqslant u_1(t)\leqslant u_1^{max}\\ 0\leqslant u_2(t)\leqslant u_2^{max}}} \ \inf_{n_C^0 \in A_{\rho_{C^0}}} \rho_C(t_f).$$

In other words, for a fixed initial tumour size, we aim at tackling the following question:

what is the cancer cells' best possible repartition in phenotype?

A simpler (and instantaneous) version of the previous optimisation problem for t_f small is

$$\inf_{\substack{0\leqslant u_1\leqslant u_1^{max}\\ 0\leqslant u_2\leqslant u_2^{max}}} \inf_{n_C^0\in A_{\rho_{C^0}}} \frac{d\rho_C}{dt}(0), \tag{5.17}$$

for which the solution is easily obtained, and given in the following proposition.

Proposition 5.1. *Let $g := R_C\left(\cdot, \rho_C^0, \rho_H^0, u_1^{max}, u_2^{max}\right)$. We define x_C by $\{x_C\} := \arg\min g$ and $\tilde{n}_C^0 := \rho_C^0 \delta_{x_C}$. The optimal solution for the optimisation problem (5.17) is given by*

$$\left(u_1^{max}, u_2^{max}, \rho_C^0 \delta_{x_C}\right). \tag{5.18}$$

Proof. For any $0 \leqslant u_1 \leqslant u_1^{max}$, $0 \leqslant u_2 \leqslant u_2^{max}$, $n_C^0 \in A_{\rho_{C^0}}$,

$$\frac{d\rho_C}{dt}(0) = \int_0^1 R_C\left(x, \rho_C^0, \rho_H^0, u_1, u_2\right) n_C^0(x)\, dx \geqslant \int_0^1 g(x)\, n_C^0(x)\, dx$$

with equality if and only if $u_1 = u_1^{max}$, $u_2 = u_2^{max}$.

We also have $\int_0^1 g(x)\, n_C^0(x)\, dx \geqslant \int_0^1 g(x_C)\, n_C^0(x)\, dx = g(x_C)\rho_C^0$ and it remains to prove that there is equality if and only if $n_C^0 = \rho_C^0 \delta_{x_C}$. If $n_C^0 \neq \rho_C^0 \delta_{x_C}$there exists $a \in \text{supp}\left(n_C^0\right)$, $a \neq x_C$: it is therefore possible to find $\varepsilon > 0$ such that both $x_C \notin [a-\varepsilon, a+\varepsilon]$ and $\int_{[a-\varepsilon,a+\varepsilon]} n_C^0(x)\, dx > 0$.

This implies

$$\int_0^1 (g(x) - g(x_C))\, n_C^0(x)\, dx \geqslant \int_{[a-\varepsilon,a+\varepsilon]} (g(x) - g(x_C))\, n_C^0(x)\, dx > 0,$$

which concludes the proof. □

For ($\mathbf{OCP_1}$), the previous Proposition means that, very close to T, the best shape of the cancer cell density $n_C(t,\cdot)$ is a Dirac mass. As it was proved in Theorem 5.1, it is possible (in arbitrarily large time) to reach Dirac masses with constant controls. The combination of these two results is our motivation for the restriction to the set $\mathcal{B}_T$.

5.3.4 Reduction of IDEs to ODEs at the end of the long first phase

Because of the previous result, it makes sense to steer the cancer dell density as close as possible to a Dirac mass. As it was proved in Theorem 5.1, it is possible (in large time limit) to reach Dirac masses with constant controls. Our aim is now to prove that if we give constant controls $(\bar{u}_1, \bar{u}_2)$ for a long time, the dynamics of the total number of cells (ρ_H, ρ_C) are arbitrarily close to being driven by a system of ODEs, a result which comes from the concentration of the IDE on (x_H^∞, x_C^∞). The rigorous statement is given hereafter:

Lemma 5.4. *We fix $T_2 > 0$, $0 \leqslant \bar{u}_1 \leqslant u_1^{max}$ and $0 \leqslant \bar{u}_2 \leqslant u_2^{max}$. We consider any controls (u_1, u_2) defined on $[0, T_1 + T_2]$ as follows: they are constant equal to $(\bar{u}_1, \bar{u}_2)$ on $[0, T_1]$, and any BV functions on $[T_1, T_1 + T_2]$ which satisfy (5.4).*

Let (n_H, n_C) be the solution of (5.1) on $[0, T_1 + T_2]$, with corresponding (ρ_H, ρ_C). Then

$$\lim_{T_1 \to +\infty} \sup_{[T_1, T_1+T_2]} \max\left(|\rho_H - \tilde{\rho}_H|, |\rho_C - \tilde{\rho}_C|\right) = 0,$$

where $(\tilde{\rho}_H, \tilde{\rho}_C)$ solves the controlled ODE system

$$\begin{aligned}
\frac{d\tilde{\rho}_H}{dt} &= R_H\left(x_H^\infty, \tilde{\rho}_H, \tilde{\rho}_C, u_1, u_2\right)\tilde{\rho}_H, \\
\frac{d\tilde{\rho}_C}{dt} &= R_C\left(x_C^\infty, \tilde{\rho}_C, \tilde{\rho}_H, u_1, u_2\right)\tilde{\rho}_C,
\end{aligned}$$

defined on $[T_1, T_1 + T_2]$, starting at T_1 from $(\rho_H(T_1), \rho_C(T_1))$.

Proof. Let $\varepsilon > 0$. We focus on the equation on n_H which we integrate in x for any $t \in [T_1, T_1 + T_2]$:

$$\begin{aligned}
\frac{d\rho_H}{dt} &= \int_0^1 R_H(x, \rho_H, \rho_C, u_1, u_2) n_H(t, x)\, dx \\
&= R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\rho_H \\
&\qquad + \int_0^1 \left(R_H(x, \rho_H, \rho_C, u_1, u_2) - R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\right) n_H(t, x)\, dx
\end{aligned}$$

For the first term, we write

$$\begin{aligned}
R_H(x_H^\infty, \rho_H, \rho_C, u_1, u_2)\rho_H &= R_H(x_H^\infty, \rho_H, \rho_C, u_1, u_2)\tilde{\rho}_H + R_H(x_H^\infty, \rho_H, \rho_C, u_1, u_2)(\rho_H - \tilde{\rho}_H) \\
&= \frac{d\tilde{\rho}_H}{dt} + \tilde{\rho}_H d_H(x_H^\infty)\left(-a_{HH}(\rho_H - \tilde{\rho}_H) - a_{HC}(\rho_C - \tilde{\rho}_C)\right) \\
&\qquad + R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\left(\rho_H - \tilde{\rho}_H\right)
\end{aligned}$$

This means we end up with

$$\begin{aligned}
\frac{d}{dt}(\rho_H - \tilde{\rho}_H) &= \tilde{\rho}_H d_H(x_H^\infty)\left(-a_{HH}(\rho_H - \tilde{\rho}_H) - a_{HC}(\rho_C - \tilde{\rho}_C)\right) \\
&\quad + R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\left(\rho_H - \tilde{\rho}_H\right) \\
&\qquad + \int_0^1 \left(R_H(x, \rho_H, \rho_C, u_1, u_2) - R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\right) n_H(t, x)\, dx.
\end{aligned}$$

We look at the last term separately: the first two ones are linked to the discrepancy between ρ and $\tilde{\rho}$, while the last one will be small because n_H is concentrated if T_1 is large enough. Setting $w := \max\left(|\rho_H - \tilde{\rho}_H|, |\rho_C - \tilde{\rho}_C|\right)$, we have the differential inequality

$$\frac{d}{dt}|\rho_H - \tilde{\rho}_H| \leqslant Cw + \int_0^1 \left(R_H(x, \rho_H, \rho_C, u_1, u_2) - R_H\left(x_H^\infty, \rho_H, \rho_C, u_1, u_2\right)\right) n_H(t, x)\, dx \tag{5.19}$$

for some constant $C > 0$. The last term can be decomposed as

$$\frac{1}{1+\alpha_H u_2}\int_0^1 (r_H(x) - r_H(x_H^\infty))\, n_H(t,x)\, dx \\ - u_1 \int_0^1 (\mu_H(x) - \mu_H(x_H^\infty))\, n_H(t,x)\, dx - I_H \int_0^1 (d_H(x) - d_H(x_H^\infty))\, n_H(t,x)\, dx.$$

Note that u_1, $\frac{1}{1+\alpha_H u_2}$ and I_H are all bounded on $[T_1, T_1+T_2]$. Thus, if for any generic function ϕ, $\int_0^1 (\phi_H(x) - \phi_H(x_H^\infty))\, n_H(t,x)\, dx$ is arbitrarily small, so is the last quantity. To that end, we write the solution of the IDE in exponential form

$$n_H(t,x) = n_H(T_1,x) \exp\left(\int_{T_1}^t R_H\left(x, \rho_H(s), \rho_C(s), u_1(s), u_2(s)\right)\, ds\right),$$

where the exponential is uniformly bounded on $[0,1]\times[T_1, T_1+T_2]$, which means that $\left|\int_0^1 (\phi_H(x) - \phi_H(x_H^\infty))\, n_H(t,x)\, dx\right| \leqslant C \int_0^1 |\phi_H(x) - \phi_H(x_H^\infty)|\, n_H(T_1,x)\, dx$. Since $n_H(T_1,\cdot)$ converges to $\rho_H^\infty \delta_{x_H^\infty}$ in $\mathcal{M}^1(0,1)$ as T_1 goes to $+\infty$, this quantity is arbitrarily small. Plugging this estimate into (5.19) and writing a similar inequality for the equations on the cancer cells, we obtain for T_1 large enough $\frac{dw}{dt} \leqslant Cw + \varepsilon$. We conclude by applying the Gronwall lemma, together with the fact that $w(T_1) = 0$. □

5.3.5 Analysis of the second phase

According to the previous results, for large T and admissible constant controls $(\bar{u}_1, \bar{u}_2)$, we arrive at concentrated populations whose dynamics are driven by a system of ODEs. This naturally leads to considering the following optimal control problem, on the resulting ODE concentrated in (x_H^∞, x_C^∞), starting from $(\rho_H^\infty, \rho_C^\infty)$ at $t = 0$. For readability, we write g_H for $g_H(x_H^\infty)$ (resp., g_C for $g(x_C^\infty)$) for any function g_H (resp., g_C), and we stress that all assumptions made in this subsection are made for all possible admissible constant controls $(\bar{u}_1, \bar{u}_2)$.

The ODE system of equations now reads

$$\frac{d\rho_H}{dt} = \Big(\underbrace{\frac{r_H}{1+\alpha_H u_2} - d_H I_H - u_1 \mu_H}_{R_H}\Big)\rho_H, \quad \frac{d\rho_C}{dt} = \Big(\underbrace{\frac{r_C}{1+\alpha_C u_2} - d_C I_C - u_1 \mu_C}_{R_C}\Big)\rho_C. \tag{5.20}$$

For a given $T_2^M > 0$, we investigate the optimal problem of minimising $\rho_C(t_f)$ for $t_f \leqslant T_2^M$ and controls (u_1, u_2) which satisfy (5.4), as well as the constraints (5.6) and (5.5). The constraint (5.5) rewrites $\frac{\rho_C}{\rho_H} \leqslant \gamma$ with

$$\gamma := \frac{1-\theta_{HC}}{\theta_{HC}}.$$

Assume that there exists an optimal solution which is the concatenation of free and constrained arcs (either on the constraint (5.6) or (5.5)), with associated times $(t_i)_{1\leqslant i\leqslant M}$. In particular, we thus assume without loss of generality that the parameters are such that

both constraints do not saturate simultaneously on an optimal arc. (5.21)

Then, by the Pontryagin maximum principle for an optimal control problem with state constraints (see [170]), there exists a bounded variation adjoint vector $p = (p_H, p_C)$ defined on $[0, t_f]$, a scalar $p^0 \leqslant 0$, non-negative functions η_1 and η_2 and non-negative scalars ν_i, $i = 1, \dots, M$ such that if we define the Hamiltonian function by

$$\begin{aligned} &H(\rho_H, \rho_C, p_H, p_C, u_1, u_2) \\ &\quad := p_H R_H \rho_H + p_C R_C \rho_C + \eta_1(\theta_H \rho_H^0 - \rho_H) + \eta_2(\rho_C - \gamma\rho_H) \\ &\quad = -p_H d_H I_H \rho_H - p_C d_C I_C \rho_C + \left(\frac{r_H p_H \rho_H}{1+\alpha_H u_2} + \frac{r_C p_C \rho_C}{1+\alpha_C u_2} \right) \\ &\qquad\qquad - (\mu_H p_H \rho_H + \mu_C p_C \rho_C)\, u_1 + \eta_1(\theta_H \rho_H^0 - \rho_H) + \eta_2(\rho_C - \gamma\rho_H), \end{aligned}$$

we have

1. p, p^0, η_1, η_2 and the $(\nu_i)_{i=1,\dots,M}$ are not all zero.

2. The adjoint vector satisfies

$$\begin{aligned} \frac{dp_H}{dt} &= -\frac{\partial H}{\partial \rho_H} = -p_H\left(-a_{HH} d_H \rho_H + R_H\right) + a_{CH} d_C p_C \rho_C + \eta_1 + \gamma\eta_2, \\ \frac{dp_C}{dt} &= -\frac{\partial H}{\partial \rho_C} = -p_C\left(-a_{CC} d_C \rho_C + R_C\right) + a_{HC} d_H p_H \rho_H - \eta_2, \end{aligned}$$

with $p_H(t_f) = 0$, $p_C(t_f) = p^0$.

3. $t \longmapsto \eta_1(t)$ (resp. $t \longmapsto \eta_2(t)$) is continuous along (5.6) (resp. (5.5)), and is such that $\eta_1(\theta_H \rho_H^0 - \rho_H) = 0$ (resp. $\eta_2(\rho_C - \gamma\rho_H) = 0$) on $[0, t_f]$.

4. For any $i = 1, \dots, M$, the Hamiltonian is continuous at t_i. If t_i is a junction or contact[2] point with the boundary (5.6) (resp. with the boundary (5.5)), $p_H(t_i^+) = p_H(t_i^-) + \nu_i$, $p_C(t_i^+) = p_C(t_i^-)$ (resp. $p_H(t_i^+) = p_H(t_i^-) + \gamma\nu_i$, $p_C(t_i^+) = p_C(t_i^-) - \nu_i$).

5. The controls u_1, u_2 maximise the Hamiltonian almost everywhere.

We now make several technical assumptions (for all admissible constant controls $(\bar{u}_1, \bar{u}_2)$) by requiring

$$\gamma < \frac{\mu_H}{\mu_C} \frac{\mu_C a_{HH} d_H - \mu_H a_{CH} d_C}{\mu_H a_{CC} d_C - \mu_C a_{HC} d_H} \tag{5.22}$$

(assuming first $\mu_C a_{HH} d_H > \mu_H a_{CH} d_C$, $a_{CC}\mu_H d_C > a_{HC}\mu_C d_H$),

$$\mu_H,\ \mu_C > 0, \tag{5.23}$$

[2]The starting and ending points of a boundary arc are called junction points if they are distinct, and contact points if they coincide (*i.e.*, if the arc is reduced to a singleton).

$$\alpha_H \mu_C r_H < \alpha_C \mu_H r_C,\ \alpha_H \mu_H r_C < \alpha_C \mu_C r_H \tag{5.24}$$

$$(\alpha_C r_H \mu_C - \alpha_H r_C \mu_H)\,(u_2^{max})^2 + 2(r_H \mu_C - r_C \mu_H) u_2^{max} + \frac{\alpha_H r_H \mu_C - \alpha_C r_C \mu_H}{\alpha_H \alpha_C} < 0. \tag{5.25}$$

Note that the two last assumptions are satisfied as soon as $\frac{\alpha_H}{\alpha_C}$ is very small, at least compared to $\frac{\mu_H}{\mu_C}$. This amounts to saying that cytostatic drugs specifically target the cancer cells better than cytotoxic drugs do.

This last necessary condition motivates the definitions

$$\phi_1 := \mu_H p_H \rho_H + \mu_C p_C \rho_C,$$

and (abusively, since this quantity also depends on t)

$$\psi(u_2) := \frac{r_H p_H \rho_H}{1 + \alpha_H u_2} + \frac{r_C p_C \rho_C}{1 + \alpha_C u_2}.$$

Let us first analyse a constrained arc on (5.6), whenever it is not reduced to a singleton.

Arc on the constraint (5.6).

First note that $\rho_C = \frac{\rho_C}{\rho_H}\rho_H = \frac{\rho_C}{\rho_H}\theta_H \rho_H^0$ is bounded from above by $\gamma\,\theta_H \rho_H^0$. If we differentiate the constraint, we find that u_1 and u_2 are determined by

$$\frac{r_H}{1 + \alpha_H u_2} - d_H(a_{HH}\theta_H \rho_H^0 + a_{HC}\rho_C) - u_1 \mu_H = 0,$$

together with the fact that

$$\begin{aligned}(u_1, u_2) &\in \arg\max \left(\frac{r_H p_H \rho_H}{1 + \alpha_H u_2} + \frac{r_C p_C \rho_C}{1 + \alpha_C u_2} - (\mu_H p_H \rho_H + \mu_C p_C \rho_C)\, u_1 \right) \\ &= \arg\max \frac{p_C \rho_C}{\mu_H} \left(\frac{r_C \mu_H}{1 + \alpha_C u_2} - \frac{r_H \mu_C}{1 + \alpha_H u_2} \right)\end{aligned}$$

One can check that (5.24) and (5.25) are sufficient conditions to have decrease of the function $u_2 \mapsto \frac{r_C \mu_H}{1+\alpha_C u_2} - \frac{r_H \mu_C}{1+\alpha_H u_2}$. In particular, $u_2 = u_2^{max}$ if $p_C < 0$, $u_2 = 0$ if $p_C > 0$. Thus, the maximisation condition is equivalent to maximising $-\phi_1 u_1$ if p_C does not vanish on the arc. Hence, $\phi_1 = 0$ when this condition on p_C is fulfilled. We also obtain u_1 in feedback form along the arc, and when p_C does not vanish it is given by:

$$u_1^{b,v} := \frac{1}{\mu_H} \left(\frac{r_H}{1 + \alpha_H v} - d_H(a_{HH}\theta_H \rho_H^0 + a_{HC}\rho_C) \right)$$

where $v = 0$ or $v = u_2^{max}$ depending on the sign of p_C. We assume that this is an admissible control, *i.e.*, that it satisfies

$$0 < u_1^{b,v} < u_1^{max} \tag{5.26}$$

for $v = 0$ and $v = u_2^{max}$, and any $0 \leqslant \rho_C \leqslant \gamma\theta_H \rho_H^0$. If $p_C > 0$ and $u_2 = 0$, the dynamics of ρ_C on the arc (5.6) are given by

$$\frac{d\rho_C}{dt} = \frac{1}{\mu_H}\,(r_b - d_b \rho_C)\,\rho_C$$

with

$$r_d := (r_C\mu_H - r_H\mu_C) + (a_{HH}d_H\mu_C - \mu_H a_{CH}d_C)\,\theta_H\rho_H^0,\ d_b := (a_{CC}\mu_H d_C - a_{HC}\mu_C d_H)\,,$$

which we assume to be positive. This autonomous ODE leads to a monotonic behaviour of ρ_C. In order to ensure that the boundary control $u_1 = u_1^{b,0}$ is not enough to prevent the increase of ρ_C we assume the following

$$\gamma\,\theta_H\rho_H^0 < \frac{r_d}{b_d}. \tag{5.27}$$

The previous hypothesis implies that ρ_C will increase on an arc on (5.6) when $p_C > 0$.

Arc on the constraint (5.5).

If we differentiate the constraint, we find that $R_H = R_C$, *i.e.*, u_1 and u_2 are related to one another by

$$\frac{r_H}{1+\alpha_H u_2} - d_H\rho_H(a_{HH}+\gamma a_{HC}) - u_1\mu_H = \frac{r_C}{1+\alpha_C u_2} - d_C\rho_H(\gamma a_{CC}+a_{CH}) - u_1\mu_C.$$

We are now set to prove the result:

Proposition 5.2. *Assume (5.2), (5.14), (5.15), (5.21), (5.22), (5.23), (5.24), (5.25), (5.26), (5.27) and that there exists an optimal solution which is the concatenation of free and constrained arcs (either on the constraint (5.6) or (5.5)), with associated times $(t_i)_{1\leqslant i\leqslant M}$. Then, the last three possible arcs are:*

- *a boundary arc along the constraint (5.5).*
- *a free arc with controls $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$,*
- *a boundary arc along the constraint (5.6) with $u_2 = u_2^{max}$.*

The proof is technical and can be found in Appendix 5.7.

5.3.6 Solution of ($\mathbf{OCP_1}$) in $\mathcal{B}_T$ for large T: proof of Theorem 5.2

Recall that we want to solve (**OCP**) for controls $(u_1,u_2)\in\mathcal{B}_T$ for large T and small T_2^M, a choice motivated by the previous results. For a given T, we denote $\left(\bar u_1^{(T)},\bar u_2^{(T)}\right)$ a choice of optimal values for the constant controls during the first phase.

Theorem 5.2. *Assume the hypotheses of Proposition 5.2. Then asymptotically in T and for T_2^M small, there exists at least one solution to ($\boldsymbol{OCP_1}$) in $\mathcal{B}_T$. More precisely, there exists $\left(\bar u_1^{opt},\bar u_2^{opt},T_2^{ode}\right)$, $\left(u_1^{ode},u_2^{ode}\right)\in BV\left(0,T_2^{ode}\right)$ such that if we define the control (u_1,u_2) by*

$$(u_1,u_2)(t) = \begin{cases} \left(\bar u_1^{opt},\bar u_2^{opt}\right) & on\quad \left(0,T-T_2^{ode}\right)\\ \left(u_1^{ode}(t-T+T_2^{ode}),u_2^{ode}(t-T+T_2^{ode})\right) & on\quad \left(T-T_2^{ode},T\right), \end{cases}$$

then up to a subsequence we have

$$\lim_{T\to+\infty}\left(C_T(u_1,u_2)-\inf_{(u_1,u_2)\in\mathcal{B}_T}C_T(u_1,u_2)\right)=0,$$

meaning that (u_1,u_2) *is quasi-optimal if* T *is large enough. Furthermore, on* $\left(T-T_2^{ode},T\right)$ *the optimal trajectory trajectory obtained with* (u_1,u_2) *is the concatenation of at most three arcs:*

- *a quasi-boundary arc along the constraint (5.5),*
- *a free arc with controls* $u_1=u_1^{max}$ *and* $u_2=u_2^{max}$,
- *a quasi-boundary arc along the constraint (5.6), with* $u_2=u_2^{max}$.

Remark 5.5. By *quasi-boundary arc*, we mean that the quasi-optimal control is such that (ρ_H,ρ_C) almost saturates the constraints, *i.e.*, up to an error vanishing as T goes to $+\infty$.

Proof.
Up to a subsequence, still denoted T, we can find $\left(\bar{u}_1^{opt},\bar{u}_2^{opt}\right)$ such that $\left(\bar{u}_1^{(T)},\bar{u}_2^{(T)}\right)$ converges to $\left(\bar{u}_1^{opt},\bar{u}_2^{opt}\right)$ as $T\to+\infty$. These values for the constant controls yield asymptotic phenotypes $\left(x_H^{opt},x_C^{opt}\right)$ thanks to Theorem 5.1. Then, for any choice of time $T_2\leqslant T_2^M$ and BV controls (u_1,u_2) on $(T-T_2,T)$,

$$\lim_{T\to+\infty}\sup_{[T-T_2,T]}\max\left(|\rho_H-\tilde{\rho}_H|,|\rho_C-\tilde{\rho}_C|\right)=0, \tag{5.28}$$

with the notations of Lemma 5.4: ρ is obtained from the IDE system, while $\tilde{\rho}$ is obtained from the ODE concentrated on $\left(x_H^{opt},x_C^{opt}\right)$. This is a consequence of a slight refinement of Lemma 5.4. Indeed, for T large, the IDE is almost concentrated on some $\left(x_H^{(T)},x_C^{(T)}\right)$ associated to $\left(\bar{u}_1^{(T)},\bar{u}_2^{(T)}\right)$. The formulae for these quantities given by Theorem 5.1 show that $\left(x_H^{(T)},x_C^{(T)}\right)$ converges to $\left(x_H^{opt},x_C^{opt}\right)$, hence the concentration of the IDE on $\left(x_H^{opt},x_C^{opt}\right)$ and the result (5.28).

As a consequence, the optimal strategy for the ODE, obtained by Proposition 5.2 is also optimal for the IDE, up to an error vanishing as T goes to infinity. We denote $T_2^{ode}\leqslant T_2^M$, $\left(u_1^{ode},u_2^{ode}\right)\in BV\left(0,T_2^{ode}\right)$ the solutions of this optimal control problem. The last statements of the theorem are then a direct consequence of Proposition 5.2 and the assumption that T_2^M is small, since the IDE and ODE trajectories are arbitrarily close. □

5.4 Numerical simulations

In this section, we solve ($\mathbf{OCP_1}$) numerically in the full class $\mathcal{A}_T$. We will compare the results with the previous section, and check that alternative strategies to the one given in Theorem 5.2 are indeed sub-optimal when T is large.

5.4.1 Numerical simulations of the solution to $(\mathbf{OCP_1})$

For a survey on numerical methods in optimal control of ODEs, we refer to [167].

Here, we use *direct* methods which consist in discretising the whole problem and reducing it to a "standard" constrained optimisation problem, and we refer to Chapter 6 for a more detailed presentation of the approach. For the simulations, we take $\theta_{HC} = 0.4$, $\theta_H = 0.6$, $\varepsilon = 0.1$. We let T take the value $T = 60$ to complement the graph already given in the Introduction for $T = 30$. The results are reported on Figure 5.4. These simulations

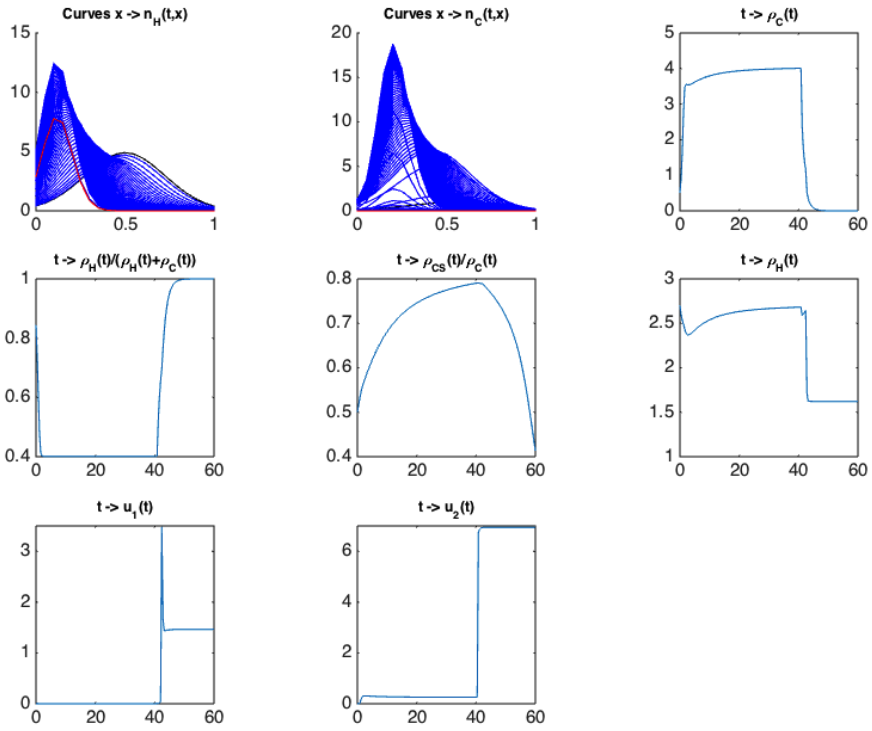


FIGURE 5.4: Simulation of $(\mathbf{OCP_1})$ for $T = 60$.

clearly indicate that for the chosen numerical data, if T is large enough, then the optimal controls are such that:

- the optimal control u_1 is first equal to 0 on a long arc. Then, on a short-time arc, $u_1 = u_1^{\max}$ and then to a value such that the constraint (5.6) saturates;
- the optimal control u_2 has a three-part structure, with a long-time starting arc which is a *boundary arc*, that is, an arc along which the state constraint (5.5) is (very quickly) saturated. It corresponds to an almost constant value for the control u_2. The last short-time arc coincides with that of u_1, and along this arc $u_2 = u_2^{max}$.

We denote by $t_s(T)$ the switching time, defined by largest time such that $u_1(t) = 0$ for all

$t < t_s(T)$.

According to the numerical simulations, as T tends to $+\infty$, both $x \mapsto n_C(t_s(T), x)$, $x \mapsto n_H(t_s(T), x)$ converge to (weighted) Dirac masses. Since the controls u_1 and u_2 are almost constant on $(0, t_s(T))$, this is in accordance with Theorem 5.1. The cancer cell population is then concentrated on a phenotype on which the drugs are very efficient.

More precisely, as T tends to $+\infty$, the optimal strategy seems to tend to a two-piece trajectory, consisting of:

- a first long-time arc, along the boundary $\frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)} = \theta_{HC}$, with $u_1(t) = 0$ and with a constant control u_2, at the end of which the populations of healthy and of cancer cells have concentrated on some given sensitive phenotype;
- a second short-time arc along which the populations of healthy and cancer cells are very quickly decreasing with $u_1 = u_1^{max}$, $u_2 = u_2^{max}$, before the constraint on ρ_H saturates and u_1 switches to a boundary value allowing ρ_C to keep decreasing.

We also find that the mapping $T \longmapsto \rho_C(T)$ (where $\rho_C(T)$ is the value obtained by solving ($\mathbf{OCP_1}$) on $[0, T]$) is decreasing. This is because our parameters are such that, once concentrated on a sensitive phenotype, the cancer cell population satisfies a controlled ODE for which there exists a strategy letting ρ_C converge to 0. Because our model is exponential, we cannot reach 0 exactly but for very small values of ρ_C, one can consider that the tumour has been eradicated.

Remark 5.6. In order to avoid additional lengthy hypotheses, we did not give conditions under which the strategy established in Theorem 5.2 can further be identified. However, the numerical solutions show that, for generic parameters, it can be expected that:

• the constant controls on the first phase are such that at the end of the first phase, we have saturation of (5.5),

• the second phase is of time duration T_2^M and starts with a constrained arc along (5.5).

5.4.2 Comparison with clinical settings

As explained before, our results advocate for a first long phase which must be all the more long for an initially heterogeneous tumour (with respect to resistance). They also apply to 'born to be bad' tumours [159], with high initial heterogeneity with respect to genes or phenotypes in general. Indeed, the heterogeneity or homogeneity we address here is related to one phenotype defined by resistance towards one category of cytotoxic drug. In this sense, our use of the term heterogeneity is unambiguous, functionally defined, and cannot be superimposed on other more classical uses, defined by the accumulation of mutations, such as in [52, 62, 159].

This being said, we are ultimately concerned with the application of our optimal control methods to the improvement of classical therapeutic regimens in which repeated courses of chemotherapy are delivered to patients with cancer. To this end, we keep the previous

parameters, that are in particular relevant to represent an initially heterogeneous tumour, and we propose for possible implementation in the clinic a quasi-periodic strategy such as in the example defined below:

- As long as $\frac{\rho_H}{\rho_H+\rho_C} \geqslant \theta_{HC}$, we follow the drug-holiday strategy by choosing $u_1 = \bar{u}_1 = 0$, $u_2 = \bar{u}_2 = 0.5$ obtained in the previous numerical simulations.
- Then, as long as $\rho_H > \theta_H \rho_H(0)$, we use the maximal amount of drugs. As soon as $\rho_H = \theta_H \rho_H(0)$, go back to the drug-holiday strategy.

The implementation is straightforward, Figure 5.5 shows an example for $T = 60$. This strategy allows to maintain the tumour size below some upper value and to prevent resistant cells from taking over the whole population. However, the tumour is not eradicated and this strategy is far from being optimal: $\rho_C(T)$ is slightly below 1, to be compared to the value obtained with $T = 60$ (see Figure 5.4) with the optimal strategy, which is around 1.10^{-5}. It is another proof of the importance of a long first phase. It also shows that, at least with our parameters, the last arc on the constraint (5.6) obtained in the previous simulations is instrumental in view of significantly decreasing the tumour size. To assess

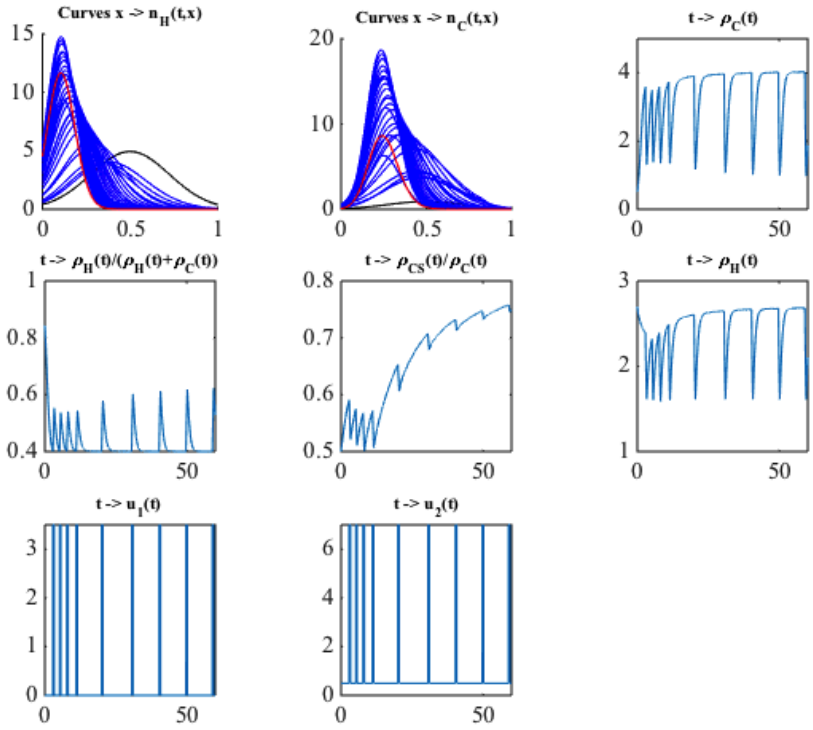


FIGURE 5.5: Quasi-periodic strategy, for $T = 60$.

the importance of the saturation of the constraint $\rho_H = \theta_H \rho_H(0)$, we complement the previous strategy with an arc on this constraint, with $u_2 = u_2^{max}$, and adequately chosen

feedback control u_1 obtained from the equality $\frac{d\rho_H}{dt} = 0$. We go back to the drug-holiday strategy as soon as ρ_C starts increasing again, since it is a sign that the tumour has become too resistant. We choose $T = 100$ to have enough cycles; the corresponding results are reported on Figure 5.6 below. They tend to show that ρ_C can be brought arbitrarily close to 0 after enough cycles, meaning that there is a chance for total eradication of the tumour.

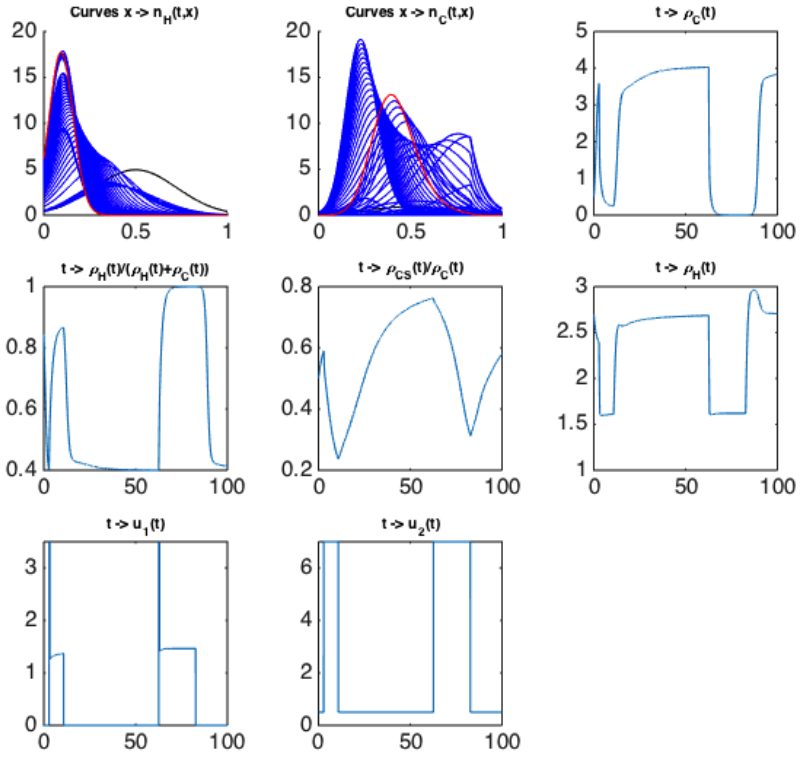


FIGURE 5.6: Second quasi-periodic strategy, for $T = 100$.

5.5 Conclusion

5.5.1 Summary of the results

By analysing a controlled integro-differential system of cancer and healthy cells structured by a resistance phenotype, we have mathematically investigated the effect of combined chemotherapeutic (cytotoxic and cytostatic) drugs on a tumour. Since we chose a biologically grounded modelling for the resistance phenomenon and took the healthy tissue into account, our approach is tailored for understanding and circumventing the two main pitfalls in cancer therapy: resistance to drugs and toxicity to healthy tissue. The goal of our

analysis was indeed twofold: check that our model can reproduce the possible deleterious effect of chemotherapy when MTDs are used (the standard clinical strategy), and propose alternative (optimised) infusion protocols.

Since MTD can first strongly reduce the size of the tumour which then starts growing again, we addressed the first question through an asymptotic analysis of the model. This was performed in Theorem 5.1, which showed that both populations converge, while the cells concentrate on some phenotypes. The proof of convergence and concentration, presented in Section 5.2, relies instead on a suitably defined Lyapunov function. Interestingly, the approach could incorporate controls which are not only constant, but also asymptotically constant.

The rest of the chapter was then devoted to addressing the second question, by considering the optimal control problem ($\mathbf{OCP_1}$) of minimising the number of cancer cells on a given time interval $[0, T]$, keeping the tumour size in check and limiting damage to the healthy tissue. In Section 5.3, we gave several rigorous mathematical arguments to explain why, when T is large, a good strategy is to first steer the cancer cell population on an appropriate phenotype by first giving constant doses for a long time. These arguments justified a restriction to a smaller class of controls for which we managed to identify an asymptotically optimal strategy in large time, presented in Theorem 5.2.

In Section 5.4, we showed through numerical simulations that, when T increases, the optimal solution is indeed increasingly close to a two-phase trajectory. The first very long phase consists in giving low doses of drugs in order to let the cancer cell population concentrate on a given sensitive phenotype. The doses are chosen as low as the constraint on the relative tumour size allows it. Our results advocate for a first long phase which must be all the more long for an initially heterogeneous tumour (with respect to resistance). During the second phase, we numerically recover the expected trajectory, given by Theorem 5.2: high doses are given (MTD as long as the constraint on the healthy tissue does not saturate) and the cancer cell population quickly decreases.

5.5.2 Possible generalisations

Other extensions than those already introduced in the general presentation of the thesis (advection term, mutation term which will be considered in the next Chapter) are worth mentioning. A possibility is a mixed deterministic/stochastic framework, namely using a piecewise deterministic Markov process (PDMP [47], see [144] for the optimal control of this class of equations). In these models, mutations are stochastic jumps between deterministic (and phenotypically reversible) models, each jump becoming less and less rare in the course of phenotypic evolution in the deterministic processes. Furthermore, the probability of jump would depend exclusively on (and as an increasing function of) the phenotype structure variable, that would thus bear a quantitative meaning of malignancy, or phenotype plasticity entraining genetic instability (this last point is discussed with references in [36]).

A final extension would stem from the fact that tumours are also very heterogeneous in space (for example, because cells at the outer rim and cells at the centre of a tumour spheroid encompass very different metabolic conditions; more genenerally, high heterogeneity depending on space has been experimentally shown in solid tumours [62, 159], which should lead to also structure the populations of cells according to an added space variable. Another modelling advantage of such representation is that the interaction between the tumour and the healthy tissue is itself spatial, since part of it essentially happens at the boundary of the tumour, through direct contact. For possible cancer models taking both phenotype and space into account, we refer to [83, 110, 118].

5.6 Appendix A: proofs for the simplified optimal control problems

Proof of Lemma 5.2

Proof. Using the family u_ε defined in Remark 5.2, we obtain the corresponding $\rho_\varepsilon(T)$, which can be computed exactly, as well as its limit. It is given by

$$\rho_{opt}(T) := \rho_{opt}(T^-)\exp(-\mu u^{1,max})$$

where ρ_{opt} is the function obtained through $\frac{d}{dt}\rho_{opt}(t) = (r - d\rho_{opt}(t))\rho_{opt}(t)$ for $t < T$, and $\rho_{opt}(0) = \rho_0$.
Now, let any u satisfy (5.12). The solution of (5.11) with u is thus a subsolution of that satisfied by ρ_{opt}, leading to $\rho \leqslant \rho_{opt}$ on $[0, T)$. Using $u \geqslant 0$, we also have

$$\rho(T) \geqslant \rho_0 \exp\left(\int_0^T (r - d\rho(s))\,ds\right)\exp(-\mu u^{1,max}).$$

Since $\rho_{opt}(T^-) = \rho_0 \exp\left(\int_0^T (r - d\rho_{opt}(s))\,ds\right)$ and $\rho \leqslant \rho_{opt}$, this implies $\rho(T) \geqslant \rho_{opt}(T)$.
Let us now investigate the possible case of equality to prove that the infimum is not attained: the foregoing equality implies that there is equality if and only if $\int_0^T u\,ds = u^{1,max}$ (the contraint is saturated) and

$$\exp\left(\int_0^T (r - d\rho(s))\,ds\right) = \exp\left(\int_0^T (r - d\rho_{opt}(s))\,ds\right),$$

whence $\rho \equiv \rho_{opt}$ on $[0, T)$. As ρ is continuous, $\rho(T)$ would be given by taking $u \equiv 0$, which is not optimal. □

Proof of Lemma 5.3

Proof. To account for the L^1 constraint (C1), we augment the system by defining another state variable y, whose dynamics are given by $\frac{dy}{dt} = u$, leading to:

$$\begin{aligned} \frac{d\rho}{dt} &= (r - d\rho - \mu u)\rho, & \frac{dy}{dt} &= u, \\ \rho(0) &= \rho_0, & y(0) &= 0. \end{aligned}$$

The constraint (C1) thus rewrites $y(T) \leqslant u^{1,max}$.

According to the Pontryagin maximum principle (see [136]), there exist absolutely continuous adjoint variables p_ρ and p_y on $[0,T]$, and $p^0 \leqslant 0$, such that:

$$\frac{dp_\rho}{dt} = -\frac{\partial H}{\partial \rho} = -(r - 2d\rho - \mu u)\, p_\rho, \quad \frac{dp_y}{dt} = -\frac{\partial H}{\partial y} = 0$$

where the Hamiltonian is

$$H(\rho, y, p_\rho, p_y, u) := p_\rho \left(r - d\rho - \mu u\right) \rho + p_y u = \left(r - d\rho\right) p_\rho + u \left(p_y - \mu p_\rho \rho\right).$$

Thus, p_y is some constant, and p_ρ does not change sign on $[0,T]$.
The maximisation of the Hamiltonian leads to defining the switching function $\phi := p_y - \mu p_\rho \rho$. u is thus equal to $u^{\infty,max}$ whenever $\phi > 0$, equal to 0 whenever $\phi < 0$.

The transversality condition is that the vector $\begin{pmatrix} p_\rho \\ p_y \end{pmatrix}(T) - p_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ must be orthogonal to the tangent space of $\{(p,y) \in \mathbb{R}^2, y \leqslant u_{1,max}\}$ at the point $(\rho(T), y(T))$.

First case.

If $y(T) < u^{1,max}$, then the transversality conditions imply $p_\rho(T) = p_0$ and $p_y \equiv 0$. $p^0 \neq 0$ since otherwise we would have $(p_\rho, p_y, p^0) \equiv 0$. Thus, in this case, $p_\rho(T) < 0$ and p_ρ is negative on the interval $[0,T]$. The switching function ϕ is therefore positive on the whole $[0,T]$, which would imply $u \equiv u^{\infty,max}$. This is a contradiction since a consequence is $\int_0^T u(s)\, ds = u^{\infty,max} T > u^{1,max}$.

Second case.

If $y(T) = u^{1,max}$, we still have $p_\rho(T) = p_0$. As in the first case, we cannot have $p_y = 0$.
Let us first remark that ϕ cannot be positive nor negative on the whole interval, since otherwise $u \equiv u^{\infty,max}$, a contradiction, or $u \equiv 0$, which is clearly not optimal. If $p_0 = 0$, $p_\rho \equiv 0$, so that ϕ has the sign of $p_y \neq 0$, a contradiction. Therefore, $p_\rho < 0$ on $[0,T]$ as before, and this implies $p_y < 0$ to ensure that ϕ changes sign.

The derivative of ϕ is given by $\frac{d\phi}{dt} = -\mu d p_\rho \rho^2 > 0$. Thus, ϕ is increasing and u is bang-bang with one switching only. The fact that $y(T) = \int_0^T u(s)\, ds = u^{1,max}$ imposes that this switching happens at $T_1(T)$ as announced, which ends the proof. □

5.7 Appendix B: proof of Proposition 5.2

Proof. If the constraint (5.5) does not saturate on the whole $[0, t_f]$, we distinguish on whether the last arc is a free arc or a boundary arc on (5.6).

First case: the last arc is a boundary arc on (5.6), not reduced to a singleton.

In this case, $t_M = t_f$ and there can be a jump on the adjoint vector at t_f.

Let us start by proving the following:

Lemma 5.5. $p^0 < 0$.

Proof. We argue by contradiction and assume $p^0 = 0$. We first look at the interval $[t_{M-1}, t_f]$, and assume, also by contradiction, that $\nu_M > 0$. Then $p_H(t_f^-) = -\nu_M < 0$, hence $p_C'(t_f^-) < 0$, leading to $p_C > 0$ in a right neighbourhood of t_f. From assumption (5.27), this means that ρ_C decreases locally around t_f, a contradiction since t_f is free (a better strategy would be to stop before ρ_C starts increasing): $\nu_M = 0$.

Now, let us prove that p_H, p_C and η_1 vanish identically on $[t_{M-1}, t_f]$. If we have $p_C(t_0) > 0$ (resp., $p_C(t_0) < 0$) for some $t_0 \in [t_{M-1}, t_f)$, we define the maximal interval $[t_0, t^\star)$ on which $p_C > 0$ (resp., $p_C < 0$), with $p_C(t^\star) = 0$. In this case, we know that the switching function ϕ_1 vanishes on $[t_0, t^\star]$, hence p_H factorises with p_C. Coming back to the equation on p_C, we have $p_C' = \beta_C p_C$ on $(t_0, t^\star)$, for some function β_C. Since $p_C(t^\star) = 0$, this imposes $p_C \equiv 0$ on the interval, a contradiction. Thus p_C is identically 0 on the whole (t_{M-1}, t_f), and so are p_H (from the equation on p_C) and η_1 (from the equation on p_H).

We now analyse the arc $[t_{M-2}, t_{M-1}]$. From the previous step, we know that $\phi_1(t_{M-1}) = 0$. If $\nu_{M-1} > 0$, then $\phi_1(t_{M-1}^-) < \phi_1(t_{M-1}) = 0$, thus $u_1 = u_1^{max}$ locally on the left of t_{M-1}. Similarly, maximising $\psi(u_2)$ imposes $u_2 = u_2^{max}$. Also, $H(t_{M-1}) = 0$, and $H\left(t_{M-1}^-\right) = -\nu_{M-1} R_H\left(t_{M-1}^-\right) \rho_H(t_{M-1})$. By continuity of the Hamiltonian, we get $R_H\left(t_{M-1}^-\right) = 0$. At the left of t_{M-1}, u_1 and u_2 saturate at their maximal values. At the right of t_{M-1}, $R_H = 0$ but this imposes $u_1 < u_1^{max}$ or $u_2 < u_2^{max}$ since, owing to (5.15), ρ_H decreases for the maximal values. Thus, $0 = R_H\left(t_{M-1}^-\right) < R_H(t_{M-1}) = 0$, a contradiction. Finally, we have proved $\nu_{M-1} = 0$.

Standard Cauchy-Lispchitz arguments, together with the result $p_H(t_{M-1}^-) = p_C(t_{M-1}^-) = 0$ yield that p_H and p_C are also identically null on the interval $[t_{M-2}, t_{M-1}]$. Repeating these arguments on the whole $[0, t_f]$, we find that p, p^0, η_1, η_2 and the $(\nu_i)_{i=1,\dots,M}$ are all zero, in contradiction with condition 1 given by the PMP (see Section 5.3). □

Thus $p^0 < 0$ and we set $p^0 = -1$. This normalisation is allowed because the final adjoint vector $(p(t_f), p^0)$ is defined up to scaling. Again, we start by analysing the PMP on $[t_{M-1}, t_f]$. From $p_C(t_f) < 0$, we know that $u_2 = u_2^{max}$ and $\phi_1 = 0$ locally around t_f. This implies $p_H > 0$ also locally around t_f. In particular, $\nu_M = 0$. Using the same reasoning as before with $p_C' = \beta_C p_C$, we get this time that p_C and p_H have constant sign on (t_{M-1}, t_f): $p_C < 0$ and $p_H > 0$.

Let us now first assume $\nu_{M-1} > 0$. Then $\phi_1(t_{M-1}^-) < 0$, leading to $u_1 = u_1^{max}$ close to t_{M-1}. If ν_{M-1} is such that $p_H\left(t_{M-1}^-\right) \leqslant 0$, then clearly the maximisation of $\psi(u_2)$ leads to $u_2 = u_2^{max}$. At t_{M-1}, we would thus have continuity of u_2 and not u_1 since $u_1 < u_1^{max}$ on $[t_{M-1}, t_f]$ from assumption (5.26). In such a case, it holds true that there can be no jump on the adjoint vector (see for instance [18]), which is contradictory unless ν_{M-1} is such that $p_H\left(t_{M-1}^-\right) > 0$, which we assume from now on.

Let us now analyse the interval $[t_{M-2}, t_{M-1}]$, on which we will prove that $u_1 = u_1^{max}$, $u_2 = u_2^{max}$. Because η_1 and η_2 vanish on such an interval, it is easy to prove from standard Cauchy-Lipschitz uniqueness arguments that $p_C < 0$ and $p_H > 0$ on $[t_{M-2}, t_{M-1}]$. Also, because of (5.22) the inequality

$$\frac{\rho_C}{\rho_H} < \frac{\mu_H}{\mu_C}\frac{\mu_C a_{HH} d_H - \mu_H a_{CH} d_C}{\mu_H a_{CC} d_C - \mu_C a_{HC} d_H} \tag{5.29}$$

is satisfied on $[0, t_f]$. Let us prove that this implies $\phi_1 < 0$ on (t_{M-2}, t_{M-1}). For that purpose, we will prove that whenever $\phi_1(t_0) = 0$, its derivative satisfies $\phi_1'(t_0) > 0$. Note that we already know that $\phi_1(t_{M-1}^-) \leqslant \phi_1(t_{M-1}) = 0$. For such a time t_0 we indeed obtain

$$\phi_1'(t_0) = -\frac{(p_C\rho_C)(t_0)}{\mu_H}\Big[\mu_H\big(\mu_C a_{HH} d_H - \mu_H a_{CH} d_C\big)\rho_H(t_0) \\ - \mu_C\big(\mu_H a_{CC} d_C - \mu_C a_{HC} d_H\big)\rho_C(t_0)\Big].$$

Combined with (5.29), this yields $\phi_1'(t_0) > 0$, as announced. Thus $u_1 = u_1^{max}$ on the whole $[t_{M-2}, t_{M-1}]$.

For u_2, the proof is a bit more involved because the dependence is not linear. In what follows, we generically denote $\phi_{(\lambda_H,\lambda_C)} = \lambda_H p_H \rho_H + \lambda_C p_C \rho_C$ for positive constants λ_H, λ_C. With this notation the previous established result writes $\phi_{(\mu_H,\mu_C)} < 0$ on (t_{M-2}, t_{M-1}).

We need to maximise $\psi(u_2) = \frac{r_H p_H \rho_H}{1+\alpha_H u_2} + \frac{r_C p_C \rho_C}{1+\alpha_C u_2}$ as a function of u_2, whose derivative has the opposite sign of $P(u_2)$, where

$$P(u) := \alpha_H\alpha_C\phi_{(\alpha_C r_H, \alpha_H r_C)}\, u^2 + 2(\alpha_H\alpha_C)\phi_{(r_H, r_C)}\, u + \phi_{(\alpha_H r_H, \alpha_C r_C)},$$

which has discriminant $\Delta = -\alpha_H\alpha_C r_H p_H \rho_H r_C p_C \rho_C(\alpha_C - \alpha_H)^2 > 0$ on $(0, t_f)$. We consider two cases, depending on the sign of $\phi_{(\alpha_C r_H, \alpha_H r_C)}$. Note that (5.2) implies the order $\phi_{(\alpha_H r_H, \alpha_C r_C)} < \phi_{(r_H, r_C)} < \phi_{(\alpha_C r_H, \alpha_H r_C)}$. From (5.24) and $\phi_1 < 0$, $P(0) = \phi_{(\alpha_H r_H, \alpha_C r_C)} < 0$.

Let us first assume $\phi_{(\alpha_C r_H, \alpha_H r_C)} < 0$, in which case all the coefficients of the polynomial are negative. Let us denote u_+ the greater root of this polynomial. Since the coefficient in front of u^2 is negative, the function ψ is increasing with u_2 on $(u_+, +\infty)$. We cannot have $u_+ \geqslant 0$ because of the signs of the coefficients: u_2^{max} maximises the function of interest. If $\phi_{(\alpha_C r_H, \alpha_H r_C)} = 0$, it is easy to see that the same result holds.

Now, let us assume that $\phi_{(\alpha_C r_H, \alpha_H r_C)} > 0$. Because $P(0) < 0$, $P(u_2^{max}) < 0$ is a sufficient condition for u_2^{max} to maximise $\psi(u_2)$. For any $\lambda_H > 0$, $\lambda_C > 0$, $\phi_1 < 0$ leads to $\phi_{(\lambda_H,\lambda_C)} <$

$(\lambda_H\mu_C - \lambda_C\mu_H)\frac{p_H\rho_H}{\mu_C}$. Applying this to $P(u_2^{max})$, we find

$$P(u_2^{max}) < \frac{p_H\rho_H}{\mu_C}\big(\alpha_H\alpha_C(\alpha_C r_H\mu_C - \alpha_H r_C\mu_H)\,(u_2^{max})^2 \\ +2(\alpha_H\alpha_C)(r_H\mu_C - r_C\mu_H)\,u_2^{max} + (\alpha_H r_H\mu_C - \alpha_C r_C\mu_H)\big).$$

We conclude that $P(u_2^{max}) < 0$ thanks to (5.25).

Thus, we have proved that, on (t_{M-2}, t_{M-1}), $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$. Note that the result actually implies $\nu_{M-1} = 0$. However the same reasoning with $\nu_{M-1} = 0$ works and we obtain $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$. From assumption (5.15), $\frac{\rho_C}{\rho_H}$ increases backwards. If this ratio reaches the value γ, *i.e.*, if the system saturates the constraint (5.5) (if not, $t_{M-2} = 0$), then we have a potential boundary arc on (5.5) on (t_{M-3}, t_{M-2}).

Second case: the last arc is a boundary arc on (5.6), reduced to a singleton.

Note that, again, $t_M = t_f$. This case is handled as the previous one: p^0 cannot be 0 and $\phi_1(t_f^-) \leqslant 0$. Because of this result, the whole reasoning made above in the previous case applies: there is an unconstrained arc with $u_1 = u_1^{max}$ and u_2^{max}. If there is a previous arc, it is a constrained arc on (5.5).

Third case: the last arc is a free arc.

Again, the same kind of arguments are enough to prove that $p^0 < 0$, and $u_1 = u_1^{max}$ and u_2^{max} on this arc. If there is a previous arc, it is a constrained arc on (5.5). □

Chapter 6

A homotopy strategy in numerical optimal control, application to (OCP_1).

The goal of this Chapter is to provide a suitable approach for the numerical solution for the full optimal control problem (**OCP_1**). With A. Olivier, we develop a general method in numerical optimal control, relying on direct methods but consisting in working first at the continuous level. The idea is to simplify the problem by setting appropriate parameters to 0, in order for a PMP to yield precise information. At the discrete level, this provides us with a suitable starting point for a homotopy. When applied to the problem (**OCP_1**), this technique allows to discretise the equations much more finely, and to check that our results in the integro-differential case are robust to the addition of modelling genetic instability. This work has led to the submitted article *Combination of direct methods and homotopy in numerical optimal control: application to the optimisation of chemotherapy in cancer* [124].

6.1 Introduction

The motivation for this work is the article [139], namely the previous Chapter 5. In this work, the numerical resolution of the optimal control problem was made through a direct method, thanks to a discretisation both in time and in the phenotypic variable. It led to a complex nonlinear constrained optimisation problem, for which even efficient algorithms will fail for large discretisation parameters because they require a good initial guess. To

overcome this, the idea was to perform (with AMPL and IpOpt, see below) a continuation on the discretisation parameters, starting from low values (*i.e.*, a coarse discretisation) for which the optimisation algorithm converges regardless of the starting point.

A clear optimal strategy emerged from these numerical simulations when the final time was increased. It roughly consists of first using as few drugs as possible during a long first phase to avoid the emergence of resistance. Cancer cells would hence concentrate on a sensitive phenotype, allowing for an efficient short second phase with the maximum tolerated doses.

The model of the previous chapter did not include any genetic instability, having in mind that epimutations which are believed to be very frequent in the life-time of a tumor. Our aim here is to numerically address the optimal control problem with the mutations modeled through diffusion operators, *i.e.*, the full problem ($\mathbf{OCP_1}$) with $\beta_H, \beta_C > 0$.

However, the previous numerical technique already failed (even without Laplacians) to get fine discretisations when the final time is very large: the optimisation stops converging when the discretisation parameters are large. The values reached for the discretisation in time were enough to observe the optimal structure, in particular all the arcs that were expected for theoretical reasons.

The addition of Laplacians significantly increases the run-time and again fails to work once the discretisation parameters are too large when the final time itself is large, and some arcs become difficult to observe. We thus have to find an alternative method to see whether the optimal strategy found in Chapter 5 is robust with respect to adding the effect of mutations.

This chapter is devoted to the presentation of a method which, up to our knowledge, is new. In our case, it provides a significant improvement in run-time and precision, and shows that the optimal strategy keeps an analogous structure when mutations are considered. The method relies on the two following steps:

- first, simplify the optimal control problem up to a point where we can show that, thanks to a Pontryagin Maximum Principle (PMP) in infinite dimension, the optimal controls are bang-bang and thus can be reduced to their switching times, which are very easy to estimate numerically. This is equivalent to setting several coefficients to 0 in the model.
- second, perform a continuation on these parameters on the optimisation problems obtained with a direct method, starting from the simplified problem all the way back to the full optimal control problem.

It allows us to start the homotopy method on this simplified optimisation problem with an already fine discretisation, actually much finer than the maximal values which could be obtained with the previous homotopy method.

Numerical optimal control and novelty of the approach. Discretising the time variable, control and state variables to approximate a control problem for an ODE (which

is an optimisation problem in infinite dimension) by a finite-dimensional optimisation problem has now become the most standard way of proceeding. These so-called direct methods thus lead to using efficient optimisation algorithms, for example through the combination of automatic differentiation softwares (such as the modelling language AMPL, see [60]) and expert optimisation routines (such as the open-source package IpOpt, see [173]).

Another approach is to use indirect methods, where the whole process relies on a PMP, leading to a shooting problem on the adjoint vector. Numerically, one thus needs to find the zeros of an appropriate function, which is usually done through a Newton-like algorithm. For a comparison of the advantages and drawbacks of direct and indirect methods, we refer to the survey [167].

For both direct and indirect methods, the numerical problem shares at least the difficulty of finding an initial guess leading to convergence of the optimisation algorithm or the Newton algorithm, respectively (it is well known that Newton algorithms can have a very small domain of convergence). To tackle this issue in the case of indirect methods, it is very standard to use homotopy techniques, for instance to simplify the problem so that one can have a good idea for a starting point as in [33, 39], or to change the cost in order to benefit from convexity properties, as in [61, 26]. Besides, when studying optimal control problem for ODE systems, a common approach is to use of so-called hybrid methods, in order to take advantage from the better convergence properties of the direct method and the high accuracy provided by the indirect method. We refer to [167, 23, 131, 172] for further developments on this subject.

We have found the combination of direct methods and continuation (such as the one done in Chapter 5) to be much less common in the literature, see however [23]. For a mathematical investigation of why continuation methods are mathematically valid, see [167].

It is however believed that direct methods typically lead to optimisation problems with several local minima [167], as it could happen for the starting problem (with low discretisation), which has yet no biological meaning. This implies one important drawback of a continuation on discretisation parameters with direct methods: although the algorithm will quickly converge in such cases, one cannot *a priori* exclude that one will get trapped in local minima that are meaningless, with the possibility for such trapping to propagate through the homotopy procedure.

Our approach of simplifying the optimal control problem so that it can be analyzed with theoretical tools such as a PMP is a way to address the previous problem and to decrease the computation time. The simplified optimal control problem, once approximated by a direct method, will indeed efficiently be solved even with a very refined discretisation. Therefore, another original aspect of our work, due to the complex PDE structure of the model, is the use of the PMP in view of building an initial guess for the direct method, in contrast with the hybrid approach we described for ODE systems, where direct methods serve to initialise shooting problems.

More generally, we advocate for the strategy of trying to simplify the problem, testing whether a PMP can provide a good characterisation of the optimal controls. Then con-

tinuation with direct methods are performed to get back to the original and more difficult one. We believe that this can always be tried as a possible strategy to solve any optimal control problem (ODE or PDE) numerically.

Recalling the optimal control problem. The system of equations describing the time-evolution of the density of healthy cells $n_H(t,x)$ and cancer cells $n_C(t,x)$ is given by

$$\frac{\partial n_H}{\partial t}(t,x) = \left[\frac{r_H(x)}{1+\alpha_H u_2(t)} - d_H(x)I_H(t) - u_1(t)\mu_H(x)\right] n_H(t,x) + \beta_H \Delta n_H(t,x),$$
$$\frac{\partial n_C}{\partial t}(t,x) = \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)I_C(t) - u_1(t)\mu_C(x)\right] n_C(t,x) + \beta_C \Delta n_C(t,x),$$

starting from an initial condition (n_H^0, n_C^0) in $C([0,1])^2$, with Neumann boundary conditions in $x=0$ and $x=1$.

Finally, for a fixed final time T we consider the optimal control problem (**OCP$_1$**) of minimising the number of cancer cells at the end of the time-frame

$$\inf \rho_C(T)$$

as a function of the L^∞ controls u_1, u_2 subject to L^∞ constraints for the controls and two state constraints on (ρ_H, ρ_C), for all $0 \leqslant t \leqslant T$:

- The maximum tolerated doses cannot be exceeded:

$$0 \leqslant u_1(t) \leqslant u_1^{max}, \;\; 0 \leqslant u_2(t) \leqslant u_2^{max}.$$

- The tumor cannot be too big compared to the healthy tissue:

$$\frac{\rho_H(t)}{\rho_H(t)+\rho_C(t)} \geqslant \theta_{HC}, \tag{6.1}$$

  with $0 < \theta_{HC} < 1$.

- Toxic side-effects must remain controlled:

$$\rho_H(t) \geqslant \theta_H \rho_H(0), \tag{6.2}$$

  with $0 < \theta_H < 1$.

Remark 6.1. In practice, other objective functions can be deemed pertinent. For example, if there is no hope of actually getting rid of the tumour, the goal might be to try and control its size on the whole interval. Thus, we will also consider objective functions of the form of convex combinations

$$\lambda_0 \int_0^T \rho_C(s)\, ds + (1-\lambda_0)\rho_C(T),$$

where $0 \leqslant \lambda_0 \leqslant 1$. For $\lambda_0 = 0$, we recover the previous objective function, while for $\lambda_1 = 1$ only the L^1 norm of ρ_C is considered.

Outline of the chapter. The chapter is organised as follows. Section 6.2 presents the simplified optimal control problem together with the application of a Pontryagin Maximum Principle in infinite dimension which almost completely determines the optimal controls. In Section 6.3, we thoroughly explain how direct methods for the optimal control of PDEs and continuations can be combined to solve a given PDE optimal control problem. These techniques and the result of Section 6.2 are then used to build an algorithm solving the complete optimal control problem. In Section 6.4 the numerical simulations obtained thanks to the algorithm are presented.

6.2 Resolution of a Simplified Model

6.2.1 Simplified Model for one Population with no State Constraints

We here introduce the simpler optimal control problem. Its precise link with the initial optimal control ($\mathbf{OCP_1}$) will be explained in Section 6.3. It is based on the equation

$$\frac{\partial n_C}{\partial t}(t,x) = \left[\frac{r_C(x)}{1+\alpha_C u_2(t)} - d_C(x)\rho_C(t) - \mu_C(x)u_1(t)\right] n_C(t,x), \tag{6.3}$$

starting from n_C^0, where $\rho_C(t) = \int_0^1 n_C(t,x)\,dx$. We denote by ($\mathbf{OCP_0}$) the optimal control problem

$$\inf_{(u_1,u_2)\in\mathcal{U}} \rho_C(T)$$

where $\mathcal{U}$ is the space of admissible controls

$$\mathcal{U} := \{(u_1,u_2) \in L^\infty([0,T],\mathbb{R}) \text{ such that } 0 \leqslant u_1 \leqslant u_1^{max},\ 0 \leqslant u_2 \leqslant u_2^{max},\ \text{a.e. on } [0,T]\}\,.$$

6.2.2 A Maximum Principle in Infinite Dimension

General statement. Let T be a fixed final time, X be a Banach space and $n_0 \in X$, U be a separable metric space. We also consider two mappings $f : [0,T]\times X\times U \to X$ and $f^0 : [0,T]\times X\times U \to \mathbb{R}$.

We consider the optimal control problem of minimising an integral cost, with a free final state $n(T)$:

$$\inf_{u\in\mathcal{U}} J(u(\cdot)) := \int_0^T f^0(t,n(t),u(t))\,dt,$$

where $y(\cdot)$ is the solution[3] of

$$\dot{n}(t) = f(t,n(t),u(t)), \quad n(0) = n_0.$$

[3]Note that the evolution equation has to be understood in the mild sense

$$n(t) = n_0 + \int_0^t f(s,n(s),u(s))\,ds.$$

In [102, Chapter 4], necessary conditions for optimality are presented, for such problems (they are actually presented in [102] in a more general setting, but for the sake of simplicity, we restrict ourselves to the material required to solve (**OCP$_0$**)). The set of these conditions is referred to as a Pontryagin Maximum Principle (PMP).

Under appropriate regularity assumptions on f and f^0, it states that any optimal pair $(\overline{n}(\cdot), \overline{u}(\cdot))$ must be such that there exists a nontrivial pair $(p^0, p(\cdot)) \in \mathbb{R} \times C([0,T], X)$ satisfying

$$p^0 \leqslant 0, \tag{6.4}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial n}(t, \overline{n}(t), \overline{u}(t), p^0, p(t)), \tag{6.5}$$

$$H(t, \overline{n}(t), \overline{u}(t), p^0, p(t)) = \max_{v \in U} H(t, \overline{n}(t), v, p^0, p(t)), \tag{6.6}$$

where the hamiltonian H is defined as $H(t,n,u,p,p^0) := p^0 f^0(t,n,u) + \langle p, f(t,n,u)\rangle$.

Remark 6.2. If the final state is free, (6.4) can be improved to $p_0 < 0$[4] and we have the additional transversality condition:

$$p(T) = 0. \tag{6.7}$$

Besides, if the final state were fixed, there would be additional assumptions to check in order to apply the PMP, assumptions that are automatically fulfilled whenever $n(T)$ is free. We refer to [102, Chapter 4 - Section 5] for more details on this issue.

Application to the problem (*OCP$_0$*). By applying the PMP, we derive the following theorem on the optimal control structure.

Theorem 6.1. *Let $(n_C(\cdot), u(\cdot))$ be an optimal solution for (**OCP$_0$**). There exists $t_1 \in [0,T[$ and $t_2 \in [0,T[$ such that*

$$u_1(t) = u_1^{max} \mathbb{1}_{[t_1,T]}, \qquad u_2(t) = u_2^{max} \mathbb{1}_{[t_2,T]}.$$

Proof. Let us define $U := \{u = (u_1, u_2) \text{ such that } 0 \leqslant u_1 \leqslant u_1^{max},\ 0 \leqslant u_2 \leqslant u_2^{max}\}$. Given a function $u \in L^\infty([0,T], U)$, the associated solution of the equation (6.3) belongs to $C([0,T], C(0,1))$, which can be seen as a subset of $C([0,T], L^2(0,1))$. We define $X = L^2(0,1)$.

First, we notice that minimising the cost $\rho_C(T)$ is equivalent to minimising the cost $\rho_C(T) - \rho_C(0)$ (as the initial number of cells is prescribed), and it can be written under the integral form:

$$\begin{aligned}
\rho_C(T) - \rho_C(0) &= \int_0^T \rho_C'(t)\, dt \\
&= \int_0^T \int_0^1 \partial_t n_C(t,x)\, dx\, dt \\
&= \int_0^T \int_0^1 \left[\frac{r_C(x)}{1 + \alpha_C u_2(t)} - d_C(x)\rho_C(t) - \mu_C(x) u_1(t) \right] n_C(t,x)\, dx\, dt
\end{aligned}$$

[4] An extremal in the PMP is said to be normal (resp. abnormal) whenever $p^0 \neq 0$ (resp. $p^0 = 0$). Here, it means that there is no abnormal extremal.

Thus, in view of applying the PMP, we define the function $f^0 : X \times U \to \mathbb{R}$ by

$$f^0(n, u_1, u_2) := \int_0^1 \left[\frac{r_C(x)}{1 + \alpha_C u_2} - d_C(x)\rho - \mu_C(x)u_1 \right] n(x)\, dx,$$

where $\rho := \int_0^1 n$, and the hamiltonian is then defined by

$$H(n, u_1, u_2, p, p^0) := p^0 f^0(n, u_1, u_2) + \int_0^1 p(x) \left[\frac{r_C(x)}{1 + \alpha_C u_2} - d_C(x)\rho - \mu_C(x)u_1 \right] n(x)\, dx.$$

Since $(n_C(\cdot), u(\cdot))$ is optimal, there exists a non trivial pair $(p^0, p(\cdot)) \in \mathbb{R} \times C([0,T], X)$, such that the adjoint equation (6.5) writes:

$$\frac{\partial p}{\partial t}(t,x) = -\left[\frac{r_C(x)}{1 + \alpha_C u_2(t)} - d_C(x)\rho - \mu_C(x)u_1(t) \right] \left(p(t,x) + p^0\right) + \int_0^1 d(x)n(t,x) \left(p(t,x) + p^0\right)\, dx.$$

Owing to Remark 6.2, we know that $p^0 < 0$.

Let us set $\tilde{p} := p + p^0$, which satisfies

$$\frac{\partial \tilde{p}}{\partial t}(t,x) = -\left[\frac{r_C(x)}{1 + \alpha_C u_2(t)} - d_C(x)\rho - \mu_C(x)u_1(t) \right] \tilde{p}(t,x) + \int_0^1 d(x)n(t,x)\tilde{p}(t,x)\, dx.$$

The transversality equation (6.7) yields $p(T, \cdot) = 0$, i.e., $\tilde{p}(T) = p^0$.

Then, in order to exploit the maximisation condition (6.6), we split the hamiltonian as

$$H(t, n_C(t), u_1(t), u_2(t), p(t), p^0) = -\Big(\int_0^1 p(t,x) d_C(x) n_C(t,x)\, dx \Big)\rho(t) - u_1(t)\phi_1(t) + \frac{\phi_2(t)}{1 + \alpha_C u_2(t)},$$

where the two switching functions are defined by

$$\phi_1(t) := \int_0^1 \mu_C(x) n_C(t,x) \tilde{p}(t,x)\, dx, \quad \phi_2(t) := \int_0^1 r_C(x) n_C(t,x) \tilde{p}(t,x)\, dx.$$

Thus, we derive the following rule to compute the controls:

- If $\phi_1(t) > 0$ (resp. $\phi_2(t) > 0$), then $u_1(t) = 0$ (resp. $u_2(t) = 0$).
- If $\phi_1(t) < 0$ (resp. $\phi_2(t) < 0$), then $u_1(t) = u_1^{max}$ (resp. $u_2(t) = u_2^{max}$).

We compute the derivative of the switching function:

$$\begin{aligned}\phi_1'(t) &= \int_0^1 \mu_C(x)\left(\partial_t n_C(t,x)\tilde{p}(t,x) + n_C(t,x)\partial_t\tilde{p}(t,x)\right)\,dx \\ &= \left(\int_0^1 \mu_C(x) n_C(t,x)\,dx\right)\left(\int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx\right).\end{aligned}$$

We know that $\int_0^1 \mu_C(x) n_C(t,x)\,dx > 0$, so that the sign of $\phi_1'(t)$ is given by the sign of

$$\int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx.$$

Let us set $\psi_1(t) := \int_0^1 d_C(x) n_C(t,x)\tilde{p}(t,x)\,dx$. The same computation as before yields

$$\psi_1'(t) = \left(\int_0^1 d_C(x) n_C(t,x)\,dx\right)\psi_1(t).$$

Therefore, the sign of $\psi_1(t)$ is constant, given by the sign of

$$\psi_1(T) = \int_0^1 d_C(x) n_C(T,x)\tilde{p}(T,x)\,dx = \int_0^1 d_C(x) n_C(T,x) p^0\,dx < 0$$

since $p^0 < 0$. This implies that the function ϕ_1 is decreasing on $[0,T]$. Since at the final time, $\phi_1(T) < 0$, we deduce the existence of a time $t_1 \in [0,T)$ such that $\phi_1(t) \geqslant 0$ on $[0,t_1]$, and $\phi_1(t) < 0$ on $[t_1,T]$. The same computation yields the same result for ϕ_2, for some time $t_2 \in [0,T]$. □

6.3 The Continuation Procedure

6.3.1 General Principle

We here recall the principle of direct methods and of continuations for optimisation problems. Together with Theorem 6.1, we then derive an algorithm to solve the problem (**OCP$_1$**).

On direct methods for PDEs. Let us give an informal presentation of the principle of a direct method for the resolution of the optimal control of a PDE. Assume that we have some evolution equation written in a general form on $[0,T]\times[0,1]$ as

$$\frac{\partial n}{\partial t}(t,x) = f(t,n(t),u(t)) + An(t,x),\ n(0) = n^0,$$

where T is a fixed time, A is some operator on the state space, f some function which might depend non-locally on n, u a scalar control, $t \in [0,T]$, and $x \in [0,1]$ is the space or phenotype variable. The possible boundary conditions are contained in the operator A, which in our case will be the Neumann Laplacian.

Consider the optimal control problem

$$\inf_{u\in\mathcal{U}} g(n(T)),$$

where T is fixed, as a function of $u \in \mathcal{U} := \{u \in L^\infty([0,T],\mathbb{R}),\ 0 \leqslant u(t) \leqslant u^{max} \text{ on } [0,T]\}$.

Further assume that we have discretised this PDE both in time and space through uniform meshes $0 < t_0 < t_1 < \ldots < t_{N_t} := T$, $0 =: x_0 < x_1 < \ldots < x_{N_x} := 1$, and that we are given some discretisations of the operator A (resp. the function f, g) denoted by A_h (resp. f_h, g_h), where $h := \frac{1}{N_x}$. With a Euler scheme in time, if one writes formally $n(t_i, x_j) \approx n_{i,j}$, $u(t_i) \approx u_i$ and $n_i := (n_{i,j})_{0\leqslant j\leqslant N_x}$, we are faced with the optimisation problem

$$\inf_{u_i,\, 0\leqslant i\leqslant N_t} g_h\left(n_{N_t}\right),$$

subject to the constraints

$$n_{i+1,j} = n_{i,j} + h f_{h,j}(t_i, n_{i,j}, u_i) + h A_h(n_i),\ n_{i,0} = n^0(x_i),\ 0 \leqslant u_i \leqslant u^{max}$$

for all $0 \leqslant i \leqslant N_t,\ 0 \leqslant j \leqslant N_x$. Note that $f_{h,j}(t_i, n_{i,j}, u_i)$ stands for the function $f_h(t_i, n_{i,j}, u_i)$ evaluated at x_j.

On continuation methods for optimisation problems. The optimal control problem of a PDE becomes a finite-dimensional optimisation problem once approximated through a direct method, such as the one presented above. Let us denote $\mathcal{P}_1$ this problem. As already mentioned in the introduction, the numerical resolution of such a problem requires a good initial guess for the optimal solution. The idea of a continuation is to deform the problem to an easier problem $\mathcal{P}_0$ for which we either have a very good a priori knowledge of the optimal solution, or expect the problem to be solved efficiently.

One then progressively transforms the problem back to the original one thanks to a continuation parameter λ, thus passing through a series of optimisation problems $(\mathcal{P}_\lambda)$. At each step of the procedure, the optimisation problem $\mathcal{P}_{\lambda+d\lambda}$ is solved by taking the solution to $\mathcal{P}_\lambda$ as an initial guess.

6.3.2 From ($\mathbf{OCP_1}$) to ($\mathbf{OCP_0}$)

Let us consider ($\mathbf{OCP_1}$) and formally set the following coefficients to 0:

$$\beta_H,\ \beta_C,\ a_{CH},\ \theta_H,\ \theta_{HC}.$$

Note that by setting β_H and β_C to 0, we also imply that the Neumann boundary conditions are no longer enforced.

When doing so, the equations on n_C and n_H are no longer coupled since the constraints do not play any role and the interaction itself (through a_{CH}) is switched off. Consequently, the optimal control problem with all these coefficients set to 0 is precisely ($\mathbf{OCP_0}$).

We now define a family of optimal controls (**OCP**$_\lambda$) where $\lambda \in \mathbb{R}^4$ has each of its components between 0 and 1. It is a vector because several consecutive continuations will be performed (in an order to be chosen) on the different parameters. For $\lambda = (\lambda_i)_{1\leqslant i\leqslant 4}$, we use the subscript λ for the parameters associated to the optimal control problem (**OCP**$_\lambda$), and they are defined by:

$$\beta_H^{(\lambda)} := \lambda_1\beta_H, \quad \beta_C^{(\lambda)} := \lambda_1\beta_C, \quad a_{CH}^{(\lambda)} := \lambda_2 a_{CH}, \quad \theta_{CH}^{(\lambda)} := \lambda_3\theta_{CH}, \quad \theta_H^{(\lambda)} := \lambda_4\theta_H.$$

In other words, λ_1, λ_2, λ_3 and λ_4 stand for the continuations on the mutations rates, the interaction coefficient a_{CH}, the constraint (6.1) and the constraint (6.2), respectively.

6.3.3 General Algorithm

Let us now explain the general approach based on the previous considerations.

Final objective and discretisation. Our final aim is to solve (**OCP$_1$**) numerically, with T large, and a very fine discretisation in time (N_t is taken to be large): T, N_t and N_x are thus fixed to certain given values. To do so, we will solve successively several problems (**OCP**$_\lambda$) with the same discretisation paremeters. Following the general method introduced about direct methods for PDEs, numerically solving an intermediate optimal control problem (**OCP**$_\lambda$) for a given λ will mean solving the resulting optimisation problem. To be more specific, we briefly explain below how the different terms are discretised. Recall that our discretisation is uniform both in time t and in phenotype x, with respectively N_t and N_x points.

- The non-local terms ρ_H, ρ_C are discretised with the rectangle method:

$$\rho(t_i) = \int_0^1 n(t_i, x)\, dx \approx \frac{1}{N_x}\sum_{j=0}^{N_x-1} n_{i,j}.$$

- The Neumann Laplacian is discretised by its classical discrete explicit counterpart:

$$\Delta n(t_i, x_j) \approx \frac{n_{i,j+1} - 2n_{i,j} + n_{i,j-1}}{(\Delta x)^2}.$$

  We manage to take N_t large enough to make sure that the CFL

$$\beta_C T\frac{(N_x)^2}{N_t} < \frac{1}{2},$$

  is verified. Using an implicit discretisation could allow us to get rid of the CFL condition but an implicit scheme happens to be more time-consuming. Therefore, we preferred using an explicit discretisation, as our procedure enables us to discretise the equations finely enough to satisfy the CFL.

- The selection term (whose sign can be both positive or negative) is discretised through an implicit-explicit scheme to ensure unconditional stability.

Sketch of the algorithm.

Step 1. We start the continuation by solving ($\mathbf{OCP_0}$). Thanks to the result 6.1, finding the minimiser of the end-point mapping $(u_1, u_2) \longmapsto \rho_C(T)$ is equivalent to finding the minimiser of the application $(t_1, t_2) \longmapsto \rho_C(T)$ where t_1 (resp. t_2) are the switching times of u_1 (resp. u_2) from 0 to u_1^{max} (resp. u_2^{max}), as introduced in Theorem 6.1.

Numerically, we can use an arbitrarily refined discretisation of ($\mathbf{OCP_0}$), since the resulting optimisation problem has to be made on a $\mathbb{R}^2$-valued function, which leads to a quick and efficient resolution.

Step 2. Once ($\mathbf{OCP_0}$) has been solved numerically, we get an excellent initial guess to start performing the continuation on the parameter λ. Its different components will successively be brought from 0 to 1, either directly or, when needed, through a proper discretisation of the interval $[0, 1]$. The order in which the successive coefficients are brought to their actual values is chosen so as to reduce the run-time of the algorithm. The precise order and way in which the continuation has been carried out are detailed together with the numerical results in Section 6.3.

Let us make a few remarks on possible further continuations:

- Since the goal is to take large values for T, one might think of performing a continuation on the final time. We again emphasise that the interest and coherence of the method requires to start with a fine discretisation at Step 1, but we note that it is also possible to further refine the discretisation after Step 2.
- Finally, it is also possible to consider the cost as introduced in Remark 6.1, which can be done through a continuation on the parameter λ_0.

6.4 Numerical Results

Let us now apply the algorithm with AMPL and IpOpT.

For our numerical experiments, we will use the following values, taken from [109]:

$$r_C(x) = \frac{3}{1+x^2}, \quad r_H(x) = \frac{1.5}{1+x^2},$$

$$d_C(x) = \frac{1}{2}(1 - 0.3x), \quad d_H(x) = \frac{1}{2}(1 - 0.1x),$$

$$a_{HH} = 1, \quad a_{CC} = 1, \quad a_{HC} = 0.07, \quad a_{cH} = 0.01$$

$$\alpha_H = 0.01, \quad \alpha_C = 1,$$

$$\mu_H = \frac{0.2}{0.7^2 + x^2}, \quad \mu_C = \max\left(\frac{0.9}{0.7^2 + 0.6x^2} - 1, 0\right),$$

$$u_1^{max} = 2, \quad u_2^{max} = 5.$$

Also, we consider the initial data:

$$n_H(0,x) = K_{H,0}\exp\left(-\frac{(x-0.5)^2}{\varepsilon}\right), \quad n_C(0,x) = K_{C,0}\exp\left(-\frac{(x-0.5)^2}{\varepsilon}\right),$$

with $\varepsilon = 0.1$ and $K_{H,0}$ and $K_{C,0}$ are chosen such that:

$$\rho_H(0) = 2.7, \quad \rho_C(0) = 0.5.$$

The rest of the parameters (namely β_H, β_C, θ_H and θ_{HC}) will depend on the case we consider, and we will specify them in what follows.

Remark 6.3. Note that we have taken u_1^{max} and u_2^{max} to be slightly below their values chosen in Chapter 5 (which makes the problem harder from the applicative point of view). This is because we are here able to let T take larger values, for which the final cost obtained with the optimal strategy $\rho_C(T)$ becomes too small, see below for the related numerical difficulties.

As for the mutations rates, we have proceeded as follows: we have simulated the effect of constant doses and observed the long-time behaviour. In the case $\beta_H = \beta_C = 0$, we know from the previous chapter that both cell densities must converge to Dirac masses. With mutations, we expect some Gaussian-like approximation of these Diracs, the variance of which was our criterion to select a suitable mutation rate in terms of modelling. It must be large enough to observe a real variability due to the mutations, but small enough to avoid seeing no selection effects (diffusion dominates and the steady state looks almost constant).

Test case 1: $T = 60$**.** We set the parameters for the diffusion to $\beta_H = 0.001$ and $\beta_C = 0.0001$. The coefficients for the constraints are $\theta_{HC} = 0.4$ and $\theta_H = 0.6$. For such numerical values, the optimal cost satisfies $\rho_C(T) << 1$, which can be source of numerical difficulties. To overcome this, we introduce the following trick: let us define $u_1^{max,0} = 1$ and $u_2^{max,0} = 4$. We apply the procedure described in Section 6.2 with the values $u_1^{max,0}$ and $u_2^{max,0}$. We then add another continuation step by raising them to the original desired values $u_1^{max} = 2$ and $u_2^{max} = 5$. In the formalism previously introduced, it amounts to adding two continuation parameters λ_5 and λ_6 to the vector $\lambda = (\lambda_i)_{1\leqslant i\leqslant 4}$. The parameters associated to the optimal control problem ($\mathbf{OCP}_\lambda$) are then defined as:

$$u_1^{max,(\lambda)} := (1-\lambda_5)u_1^{max,0} + \lambda_5 u_1^{max}, \quad u_2^{max,(\lambda)} := (1-\lambda_6)u_2^{max,0} + \lambda_6 u_2^{max}.$$

More precisely, we perform the continuation in the following way, summarised in Figure 6.1:

- First, we solve ($\mathbf{OCP_0}$), with $u_1^{max,0} = 1$ and $u_2^{max,0} = 4$.
- Second, we add the interaction between the two populations, the diffusion parameters, and the constraint on the number of healthy cells. That is, the parameters a_{CH}, β_H, β_C and θ_H are set to their values.
- Then, we add the constraint measuring the ratio between the number of healthy cells and the total number of cells, that is θ_{HC}.

- Lastly, we raise the maximum values for the controls from $u_i^{max,0}$ to u_i^{max} ($i \in \{1,2\}$), and we solve $(\mathbf{OCP_1})$ for $T = 60$.

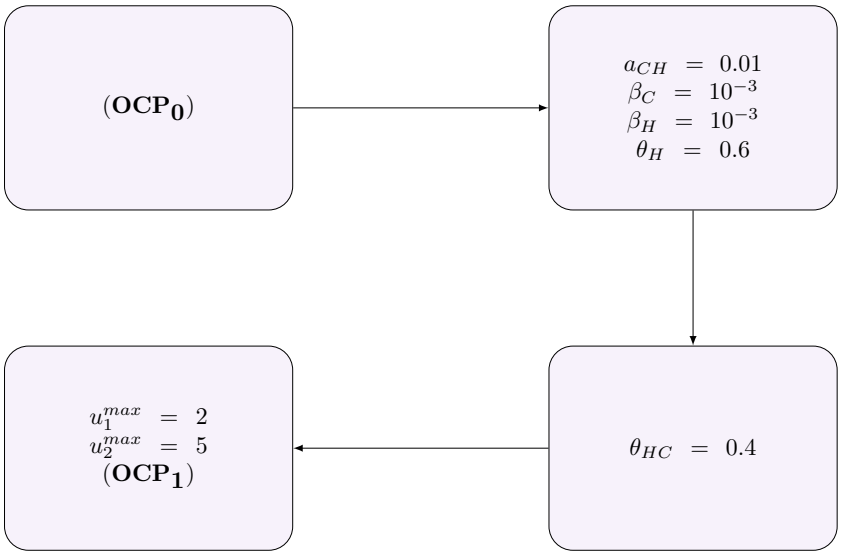


FIGURE 6.1: Continuation procedure to solve $(\mathbf{OCP_1})$ for $T = 60$.

Actually, for this set of parameters, only four consecutive resolutions are required to solve $(\mathbf{OCP_1})$ starting from $(\mathbf{OCP_0})$. That is, the components of the continuation vector $\lambda = (\lambda_i)_{1 \leqslant i \leqslant 6}$ are brought directly from 0 to 1, taking no intermediate value, in the order schematised on Figure 6.1. We will study further in the chapter a case for a larger final time, for which having a more refined discretisation is mandatory.

On Figure 6.2, we plot the optimal controls u_1 and u_2 at the four steps of the continuation procedure. We also display the evolution of the constraint on the size of the tumor compared to the healthy tissue (6.1). We can clearly identify the emergence of the expected structure for the controls, namely a long phase along which the constraint (6.1) saturates, followed by a bang arc with $u_1 = u_1^{max}$ and $u_2 = u_2^{max}$, and a last boundary arc along which the constraint (6.2) saturates. Throughout this section, we will use a red solid line in our figures for $(\mathbf{OCP_1})$, a green solid line for $(\mathbf{OCP_0})$ and a dotted style for anything refering to $(\mathbf{OCP}_\lambda)$.

Remark 6.4. We would like to emphasise here that our procedure enables us to use a much more refined discretisation of the problem than what was done in Chapter 5. More precisely, we discretise with $N_t = 500$ and $N_x = 20$ points in our direct method. For such a discretisation, directly tackling $(\mathbf{OCP_1})$ with the direct method fails.

Remark 6.5. Note that the constraint $\rho_H/\rho_H(0) > 0.6$ does not saturate until the last step of the continuation, when raising the maximal value of the controls. Therefore, when we add it at the beginning of the procedure, it is not actually active.

Test case 2: $T = 80$. Whereas one could believe that raising the final time from $T = 60$

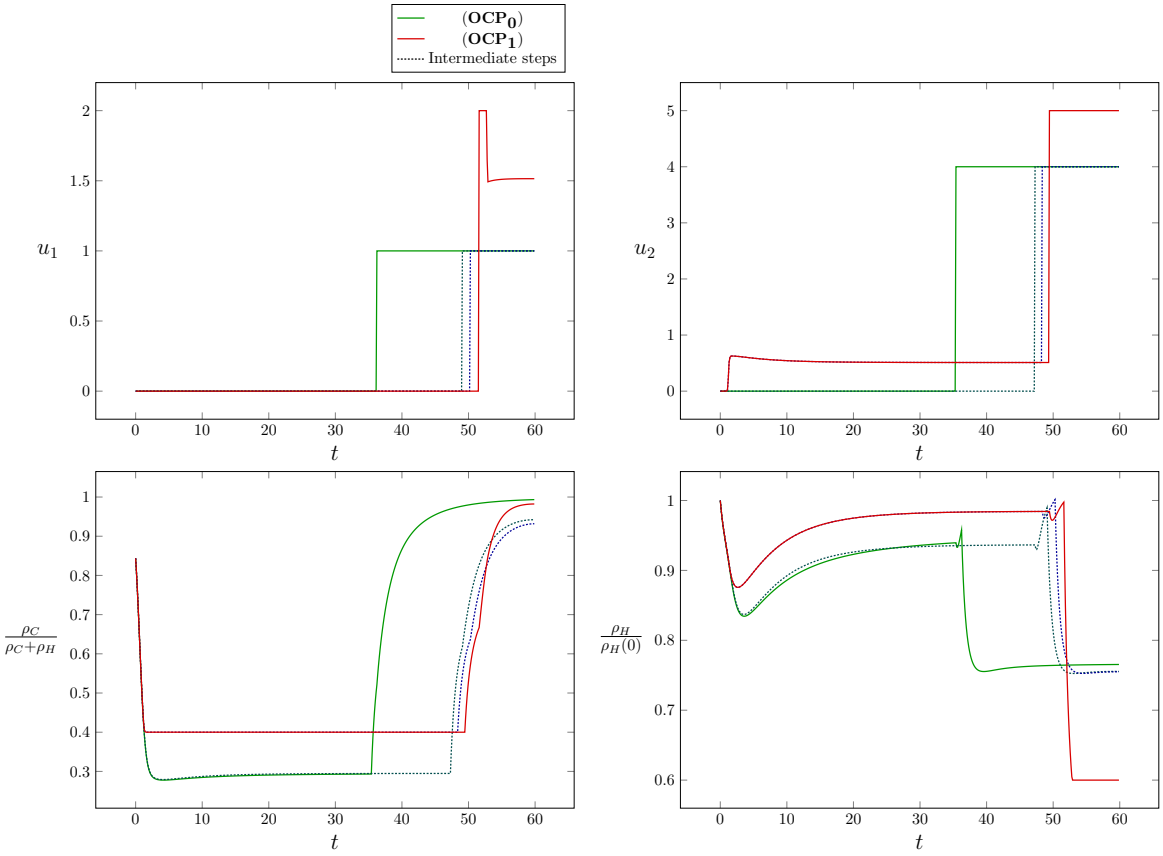


FIGURE 6.2: Intermediate steps of the continuation procedure for the test case 1.

to $T = 80$ does not much increase the difficulty of the problem, we noticed that several numerical obstacles appeared. In the following, we consider a discretisation with $N_t = 250$ and $N_x = 12$ points, in order to keep the optimisation run-time reasonable. Besides, in order to test the robustness of our procedure, we consider a more restrictive constraint on the number of healthy cells: we choose $\theta_H = 0.75$ (0.6 in the first example).

First, we use the same numerical trick as explained in our first example, reducing the maximal value for the controls to $u_1^{max,0} = 0.7$ and $u_2^{max,0} = 3.5$. For given values of u_1^{max} and u_2^{max}, the optimal cost $\rho_C(T)$ decreases when T increases. This is why we now use smaller values of $u_1^{max,0}$ and $u_2^{max,0}$, compared to the first example where we set them to respectively 1 and 4.

We performed the continuation in the following way, summarised in Figure 6.3:

- First, we solve (**OCP$_0$**), with $u_1^{max,0} = 0.7$ and $u_2^{max,0} = 3.5$.
- Second, we add the interaction between the two populations (via the parameter a_{CH}), and the constraint measuring the ratio between the number of healthy cells and the total number of cells (6.1) is introduced at the intermediate value $\theta_{HC}^{(\lambda)} = 0.3$.

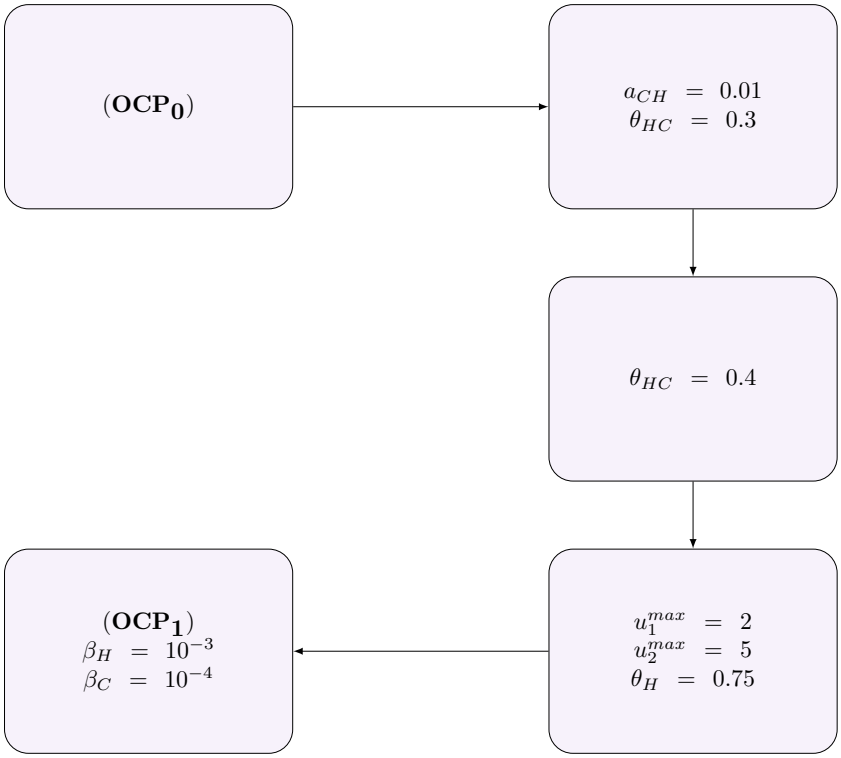


FIGURE 6.3: Continuation procedure to solve $(\mathbf{OCP_1})$ for $T = 80$.

- We then raise it to its final value of $\theta_{HC} = 0.4$.
- As a fourth step, we simultaneously add the constraint (6.2) on the healthy cells and raise the maximal values for the controls from $u_i^{max,0}$ to u_i^{max} $(i \in \{1, 2\})$.
- Lastly, we add diffusion to the model, via the parameters β_H and β_C, and we solve $(\mathbf{OCP_1})$ for $T = 80$.

At this point, we need to make two important remarks concerning this continuation procedure.

Remark 6.6. The order in which we make the components of the continuation vector $\lambda = (\lambda_i)_{1 \leqslant i \leqslant 6}$ vary from 0 to 1 is different from the order we presented for $T = 60$. For instance, we noticed that the diffusion makes the problem significantly harder to solve, although the Laplacians where discretised using the simplest explicit finite-difference approximation. Therefore, we only added it at the last step of the continuation.

Remark 6.7. Whereas for $T = 60$, raising the $(\lambda_i)_{1 \leqslant i \leqslant 6}$ directly from 0 to 1 was enough to solve $(\mathbf{OCP_1})$, it became necessary to use a more refined discretisation for $T = 80$. This fact justifies the principle of our continuation procedure, as each step is necessary to solve the next one, and thus $(\mathbf{OCP_1})$ in the end. For instance, on Figure 6.4, we display

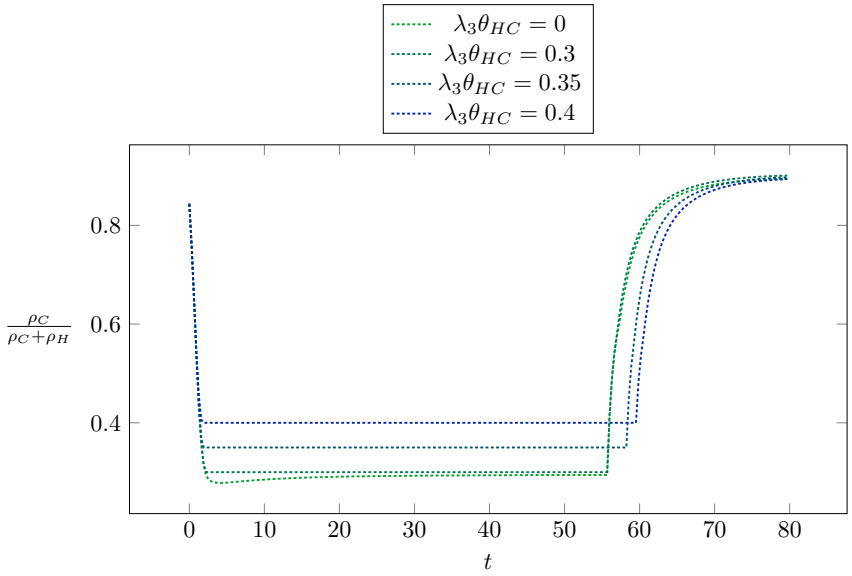


FIGURE 6.4: Evolution of the constraint (6.1) during continuation.

the evolution of the constraint (6.1):

$$\frac{\rho_H(t)}{\rho_C(t)+\rho_H(t)} \geqslant \lambda_3\theta_{HC}$$

when raising the continuation parameter λ_3 from 0 to 1. On Figure 6.5, we display the evolution of the controls u_1 and u_2 when raising their maximal allowed values from $(u_1^{max,0}, u_2^{max,0})$ to (u_1^{max}, u_2^{max}). For the sake of readability, we do not show all the steps of the continuation, but only some of them. It clearly shows how the structure of the optimal solution evolves from the simple one of ($\mathbf{OCP_0}$) to the much more complex one of ($\mathbf{OCP_1}$).

Finally, we display on Figure 6.6 the evolution of n_C, when applying the optimal strategy we found solving ($\mathbf{OCP_1}$). In black we represent the initial condition $n_C(0,\cdot)$, and with lighter shades of red, the evolution of $n_C(t,x)$ as time increases.

One clearly sees that the optimal strategy has remained the same: the cancer cell population concentrates on a sensitive phenotype, which is the key idea to then use the maximal tolerated doses. In other words, the strategy identified in Chapter 5 is robust with respect to addition of mutations. An important remark is that the cost obtained with the optimal strategy is higher with the mutations than without them: this is because we cannot have convergence to a Dirac located at a sensitive phenotype, but to a smoothed (Gaussian-like) version of that Dirac. There will always be residual resistant cells which will make the second phase less successful.

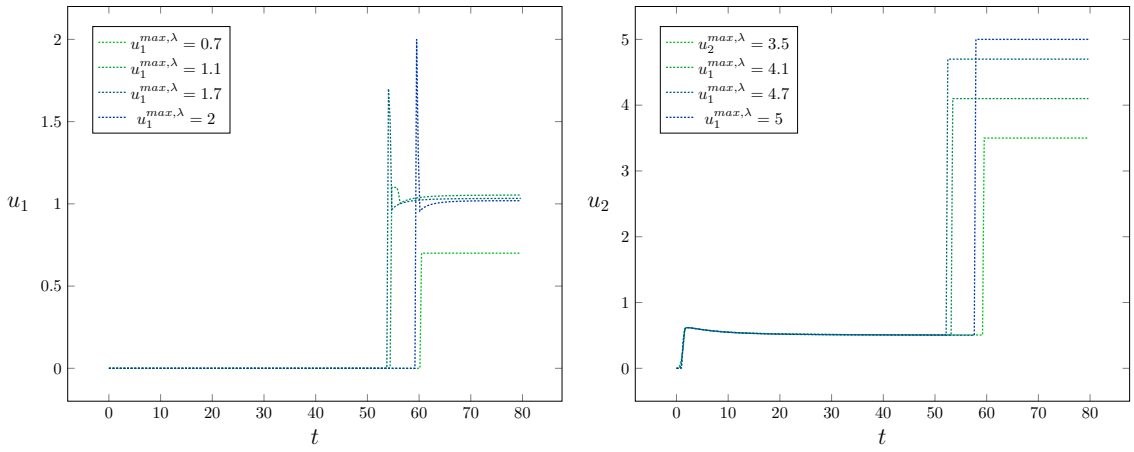


FIGURE 6.5: Raising the maximal values u_1^{max}, u_2^{max} for the controls.

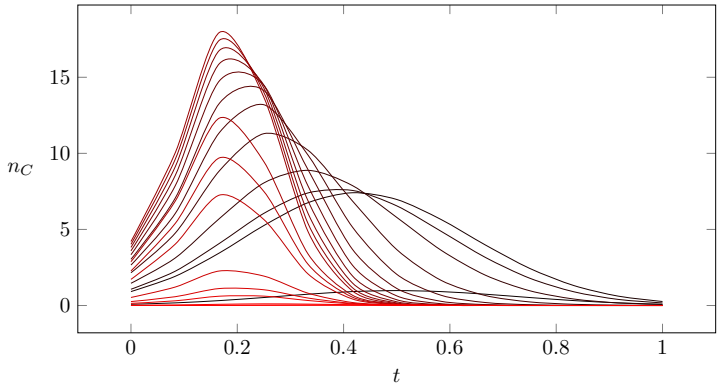


FIGURE 6.6: Evolution of n_C for the optimal solution of ($\mathbf{OCP_1}$).

Test case 3: $T = 60$**, more general objective function.** The optimal strategy obtained with the previous objective function $\rho_C(T)$ might seem surprising, in particular because it advocates for very limited action at the beginning: giving no cytotoxic drugs and low loses of cytostatic drugs. To further investigate the robustness of this strategy, let us also consider the objective function $\lambda_0 \int_0^T \rho_C(s)\, ds + (1-\lambda_0)\rho_C(T)$ as introduced in Remark 6.1, for different values of λ_0. To ease numerical computations, we take $\beta_H = \beta_C = 0$, $u_1^{max} = 2$, $u_2^{max} = 5$, and finally $N_x = 20$, $N_t = 100$. The results are reported on Figure 6.7. For $\lambda_0 = 0.5$, the L^1 term is dominant in the optimisation and the variations of ρ_C are smaller over the interval $]0, T[$. However, although there is a significant change in the control u_2 which is always equal to u_2^{max}, u_1 has kept the same structure: an arc with no drugs, a short arc with maximal doses and a final arc with intermediate doses. The only (though important) difference is that the first arc is not a long one as before.

We infer from these numerical simulations that the optimal structure is inherent in the equations: there is no choice but to let the cancer cell density concentrate on a sensitive

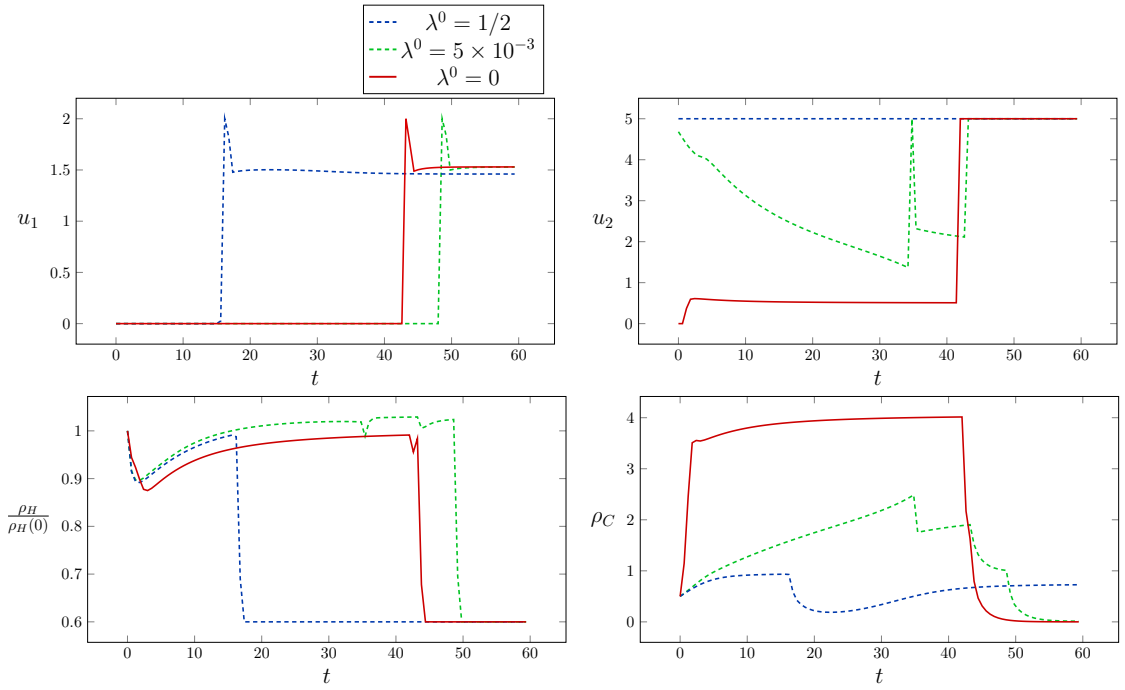


FIGURE 6.7: Plot of the optimal controls and optimal trajectories for different values of λ_0, for $T = 60$.

phenotype if the goal is to eradicate the tumour at time T. Since at $\lambda_0 = 0.5$, the integral term dominates, we also consider other convex combinations with smaller values of λ_0 up to 0, for which u_2 takes intermediate values before being equal to u_2^{max}, while $u_1 = 0$ on a longer arc.

Among these families of objectives (depending on λ_0) and their outcomes, it is up to the oncologist to decide which one is best, depending also on what is not modeled here, for example metastases. We see that a balance between $\rho_C(T)$ and $\int_0^T \rho_C(s)\,ds$ provide interesting strategies curbing tumour growth, but not killing it.

Further comments on the continuation principle. A continuation procedure is an option for many related problems. Let us illustrate our point with an example: we have presented a procedure to solve (**OCP$_1$**), for some initial conditions n_H^0 and n_C^0. Suppose that we wish to solve (**OCP$_1$**) for some different initial conditions $\tilde{n}_H^0$ and $\tilde{n}_C^0$. Biologically, this could correspond to finding a control strategy for a different tumor. A natural idea is then to use a continuation procedure to deform the problem from the initial conditions (n_H^0, n_C^0) to $(\tilde{n}_H^0, \tilde{n}_C^0)$, rather than applying again the whole procedure to solve (**OCP$_1$**) with $\tilde{n}_H^0$ and $\tilde{n}_C^0$. We successfully performed some numerical tests to validate this idea: if we dispose of a set of initial conditions for which we want to solve (**OCP$_1$**), it is indeed faster to solve (**OCP$_1$**) for one of them and then perform a continuation on the initial data, rather than solving (**OCP$_1$**) for each of the initial conditions. More generally, any parameter in the model could lend itself to a continuation.

Part IV

Spheroid formation and Keller-Segel equations

Chapter 7

Turing instabilities in chemotaxis for spheroid formation

The purpose of this chapter is to explain patterns observed in cultures of cells in a 3D structure mimicking the extracellular matrix. We show that chemotaxis is a good candidate by analysing a Keller-Segel system with nonlinear sensitivity, which exhibits Turing patterns. These patterns indeed prove to be a good qualitative match with the experimental ones, as evidenced by 2D simulations. This chapter is a very preliminary version of an article in preparation, written jointly with Federica Bubba, Nathalie Ferrand, Luis Neves de Almeida, Benoît Perthame and Michèle Sabbah, a tentative title being *A chemotaxis-based explanation of spheroid formation in 3D structures mimicking the extracellular matrix.*

7.1 Introduction and biological data

The 3D culture of cells is progressively becoming preferred over 2D cultures for in vitro experiments, because it is closer to the environment cells encounter in vivo, namely that of the extra-cellular matrix (ECM). Regardless of the chemical engineering process for the creation of these 3D scaffolds, the typical behaviour of cells inside them is aggregation [99]. The fact that different types of patterns can emerge depending on the type of cells is now well documented [90, 157].

The aggregation process and type of aggregates is already interesting from the point of view of theoretical biology, since it can shed light on how normal cells move and organise spatially in the ECM. It has also great importance in cancer biology, in order to understand

tumour progression in a more realistic experimental scenario. The focus is in particular on metastases, since it is crucial to understand how cells can organise to escape a given organ, having in mind that cells creating metastases move in clusters and not alone [80]. In this direction, a natural question is the following: how do normal healthy cells, epithelial cancer cells or invasive cancer cells compare regarding the type of pattern exhibited?

Experimental setting. The experiments performed at the Laboratoire de Biologie et Thérapeutique des Cancers by Michèle Sabbah and Nathalie Ferrand were made with 3D hydrogels. They are cylinders of radius $a = 5\,mm$, of small height $h = 2\,mm$.

The cell lines used are breast cancer cell lines, the MCF7 breast cancer cell line (epithelial cells) and the MCF7-sh-WISP2 cell line of invasive breast cancer cells (mesenchymal cells). Note that the latter cell line is obtained by the former after knock-down of the protein WISP2, which regulates cell adhesion, migration, proliferation and differentiation. We refer to [59] for a detailed study of invasive properties of the MCF7-sh-WISP2 cell line.

The MCF7 or MCF7-sh-WISP2 cells are put at the top of the hydrogel so that they can enter the 3D structure by gravity and spread inside it. The different initial numbers of cells considered are 10000, 25000, 50000, 75000 and 100000. The hydrogel is put in a growth medium renewed every 2 or 3 days. At days 4, 8, and 14 (denoted respectively by J4, J8 and J14), a 2D image is obtained thanks to an EVOS imaging station. The number and size of patterns are estimated by a dedicated software.

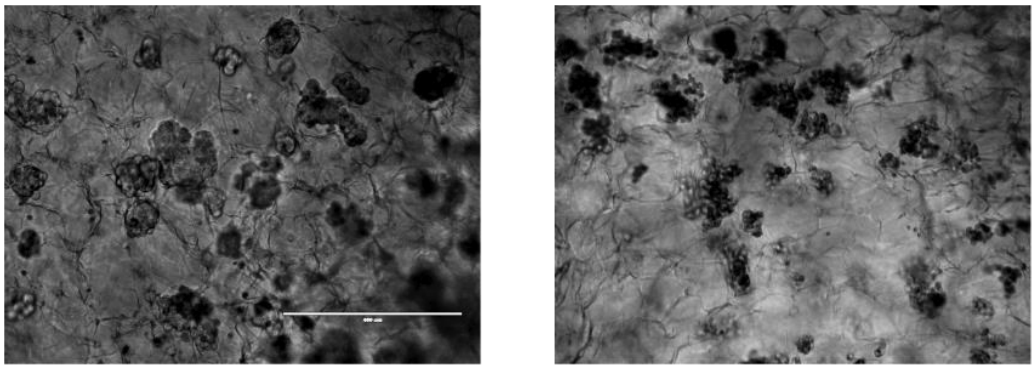
An overview of the experimental results. First, cells tend to spread uniformly in the whole structure. It is only after this very quick first phase that, in most cases, cells have formed patterns at day J4, in accordance with the results observed in the literature. These patterns are close to being round in the MCF7 case, while for the other cell line, the structure is less compact, more elongated, see a comparison in Figure 7.1 below. This is why, as in previous works, we shall slightly abusively refer to these patterns as *spheroids*. A typical spheroid has a radius of around $100\,\mu m$, and contains an estimated number of 5000 cells.

The common dynamical features exhibited by all the experiments (apart from the special case of 100000 cells which we shall discuss below):

- From J4 to J8, the number of spheres is almost constant, while the average size of spheres slightly increases.
- At J14, the number of spheroids has almost doubled, while their average size has remained constant.

In the meantime, cells grow unboundedly as long as they have enough space since the medium is renewed. Note, however, that this growth is less important from J4 to J8 than it is after J8. It has already been reported that growth is typically inhibited by the formation of spheroids, except at their periphery [175].

The case of 100000 seeded cells is a bit specific: no patterns are observed and all cells tend



(A) (B)

FIGURE 7.1: 2D image of spheroids in the hydrogel, formed by MCF7 cells after 8 days of culture (left panel) and by cells from MCF7-sh-wisp2 after 4 days of culture (right panel), courtesy of N. Ferrand and M. Sabbah.

aggregate to outside the hydrogel. Such aggregates are reported to escape the hydrogel also for lower seeding numbers, but only after enough time. Finally, let us mention that although the number and size of spheroids increases with the initial number of cells for 10000, 25000 to 50000 cells, it remains constant from 50000 to 75000: some plateau has been reached.

Mathematical modelling for patterns in the ECM. The aim of this work is to propose a minimally-parametrised continuous model amenable to reproducing these patterns. Mathematical models for cells in the extracellular matrix are numerous, and belong to two main distinct categories, see for instance [141] for a review. The first category is that of discrete models, or agent-based models (lattice-based or not), where each cell is individually modelled and its behaviour is defined by probabilistic laws. The second one is that of continuum models, typically based on ODEs or PDEs which can be either phenomenological or derived from physical or chemical laws [125].

To the best of our knowledge, however, works on the formation of patterns by cells in the ECM have focused on finely describing the cell-matrix adhesion, while chemotaxis is not necessarily a feature of these models. In the PDE models of [125], for example, there is no chemotaxis and patterns are obtained, but at the expense of assuming that ECM fibers have a specific orientation. No patterns are reported when the ECM is isotropic. In the lattice-based models of [175] developed precisely for patterns of hepatocarcinoma cells in 3D scaffolds, cell diffusion, cell division, cell-cell adhesion and cell-scaffold adhesion are taken into account, but not chemotaxis.

Here, we choose a different approach by modelling chemotaxis, reducing the cell-hydrogel modelling to a diffusion term (thus assuming that the 3D structure is isotropic). In our case, well-documented chemoattractants are the chemokines CXCL12 and CXCL8 [119], the first one (resp. the second) being believed to be mostly expressed by MCF7 cells (resp. MCF7-sh-wisp2 cells).

The model has very few parameters and, as we shall see, the condition for pattern formation has a simple expression in terms of the parameters.

7.2 Chemotaxis model and spheroid formation

7.2.1 The Keller-Segel model

We now introduce the model, namely the Keller-Segel system for the density of cells n and chemoattractant density c:

$$\begin{cases} \frac{\partial n}{\partial t} - D_1 \Delta n + \chi \nabla \cdot (\varphi(n) \nabla c) = rn, \\ \frac{\partial c}{\partial t} - D_2 \Delta c = \alpha n - \beta c, \end{cases} \tag{7.1}$$

in Ω endowed with Neumann boundary conditions, starting from (n^0, c^0), with initial mass $\int_\Omega n^0 = M$. Here $\Omega \subset \mathbb{R}^3$ is the hydrogel, a cylinder with small height (thus taking $\Omega \subset \mathbb{R}^2$ as a disk is a good approximation).

The first main ingredient is diffusion: cells can move randomly at rate D_1, thanks to the scaffold. The second ingredient is chemotaxis: cells move in the direction of the gradient of chemoattractant, of density c, the movement strength being measure by χ, and depends on a possibly nonlinear function $\varphi(n)$, which we shall specify later on. The chemoattractant is produced by the cells themselves at rate α, is naturally depleted at rate β, and diffuses at rate D_2. At variance with other works, we again emphasise that we do not explicitly model cell-scaffold adhesion and stick to a simpler diffusion term.

Finally, cells grow at rate r, and we will keep a simple exponential growth term rather than a logistic one because nutrients are constantly being brought to them, and for the time-scale of interest, there is enough space for cells to grow. There is also a hidden parameter in the model, namely the total mass of the initial condition $M := \frac{1}{|\Omega|} \int_\Omega n^0(x)\, dx$. We will assume that the initial condition is close to being homogeneous, namely $n^0 \approx M$, as observed quickly after cells have been seeded.

Upon changes of time and space variables $\tilde{t} = \beta \frac{D_1}{D_2} t$, $\tilde{x} = \sqrt{\frac{\beta}{D_2}} x$ and appropriate scalings for n and c, namely

$$n(t,x) = \tilde{n}\left(\beta \frac{D_1}{D_2} t, \sqrt{\frac{\beta}{D_2}} x\right), \; c(t,x) = \frac{\alpha}{\beta} \tilde{c}\left(\beta \frac{D_1}{D_2} t, \sqrt{\frac{\beta}{D_2}} x\right),$$

and writing again n for $\tilde{n}$, c for $\tilde{c}$ we find a minimally parametrised version:

$$\begin{cases} \frac{\partial n}{\partial t} - \Delta n + A \nabla \cdot (\varphi(n) \nabla c) = r_0 n, \\ \varepsilon \frac{\partial c}{\partial t} - \Delta c = n - c, \end{cases} \tag{7.2}$$

where only three parameters A, ε and r_0 remain, given by

$$A = \frac{\alpha\chi}{\beta D_1},\ \varepsilon = \frac{D_1}{D_2},\ r_0 = \frac{r}{\beta\varepsilon}.$$

ε is thus typically small because the chemoattractant diffuses much faster than cells, while A depends on the ratio $\frac{\chi}{D_1}$, which measures the relative importance of diffusion and attraction.

Although we will consider the growth term in simulations, we will neglect it for the investigation of Turing instabilities. Then, when it starts near the homogeneous initial condition $n^0 \approx M$, $c^0 \approx M$, this system has a homogenous (in space) solution, given by

$$\bar{n} = M,\ \bar{c} = M,$$

the (linear) stability of which we now investigate in detail.

7.2.2 Condition for Turing instabilities

Around the homogeneous steady-state $(\bar{n}, \bar{c})$, the linearised system without growth reads

$$\begin{cases} \frac{\partial n}{\partial t} - \Delta n + A\varphi(M)\Delta c = 0, \\ \varepsilon\frac{\partial c}{\partial t} - \Delta c = n - c, \end{cases} \tag{7.3}$$

We denote $(\psi_k)_{k\geqslant 1}$ the orthonormal basis of $L^2(\Omega)$ made of the eigenfunctions of the Neumann Laplace operator associated with eigenvalues $(\lambda_k)_{k\geqslant 1}$, namely

$$\begin{cases} -\Delta\psi_k = \lambda_k\psi_k, \\ \frac{\partial\psi_k}{\partial\nu} = 0. \end{cases}$$

Projecting the linearised equation (7.3) on the orthonormal basis $(\psi_k)_{k\geqslant 1}$ through

$$n(t,\cdot) = \sum a_k(t)\psi_k,\ c(t,\cdot) = \sum b_k(t)\psi_k,$$

we find

$$\begin{cases} a_k'(t) = -\lambda_k a_k(t) + A\varphi(M)\lambda_k b_k(t), \\ \varepsilon b_k'(t) = -\lambda_k b_k(t) + a_k(t) - b_k(t). \end{cases}$$

Looking for exponentially increasing (in time) solutions, we must test whether there are solutions in the form $(a_k(t), b_k(t)) = e^{\lambda t}(a_k^0, b_k^0)$ with real part $\Re(\lambda) > 0$. This is equivalent to finding a solution λ with $\Re(\lambda) > 0$ for $(A_k - \lambda Id)X = 0$ with

$$X = \begin{pmatrix} a_k^0 \\ b_k^0 \end{pmatrix},\ A_k = \begin{pmatrix} -\lambda_k & A\varphi(M)\lambda_k \\ \frac{1}{\varepsilon} & -\frac{1}{\varepsilon}(\lambda_k + 1). \end{pmatrix}.$$

Since $\mathrm{Tr}(A_k) < 0$, it easy to check that A_k has such an eigenvalue λ if and only if $\det(A_k) < 0$, which is equivalent to

$$\lambda_k + 1 < A\varphi(M) = \varphi(M)\frac{\alpha\chi}{\beta D_1},$$

a condition which does not depend on D_2.

Thus, we have the following result (owing to $\lambda_1 = 0$):

Proposition 7.1. *The state (M, M) is Turing unstable if and only if*

$$\varphi(M) > \frac{\beta D_1}{\alpha\chi}. \tag{7.4}$$

We denote $\mu = A\varphi(M) - 1$, so that Turing instability is equivalent to $\mu > 0$.

Remark 7.1. If we were to try and analyse Turing instabilities for the model with the growth term, the system should be linearised around the homogeneous (in space) solution given by

$$\bar{n}(t) := Me^{r_0 t}, \quad \bar{c}(t) := \frac{M}{1+\varepsilon r_0}\left(\varepsilon r_0 e^{-\frac{t}{\varepsilon}} + e^{r_0 t}\right).$$

Turing instabilities around time-dependent homogeneous steady states has attracted attention in the case of growing domains, at the expense of more involved computations and concepts that are beyond the scope of this chapter [114].

7.2.3 Turing unstable modes

The observed mode. In practice, when there are k_0 unstable modes, those that will be observed are the ones associated with the highest eigenvalue λ. For each $0 \leqslant k \leqslant k_0$, we denote λ_k^+ the highest eigenvalue of A_k, and we investigate how it depends on λ_k. For convenience, we set $x := \lambda_k$, and we are led to analyse the variations of the function $x \mapsto \lambda^+(x) := \frac{1}{2}(t(x) + \sqrt{t^2(x) - 4d(x)})$ where $t(x) = \mathrm{Tr}(A_k) = -\left(\frac{1}{\varepsilon} + 1\right)x - \frac{1}{\varepsilon}$, $d(x) = \det(A_k) = \frac{x(x-\mu)}{\varepsilon}$.

At the zeroth order in ε, we find

$$\lambda^+(x) \simeq x\frac{\mu - x}{x+1},$$

the derivative of which has the sign of $-x^2 - 2x + \mu$. Recall that $0 < x < \mu$ (since we are assuming that the kth mode is unstable, namely $\lambda_k < \mu$), and we find that the maximum of $\lambda^+(x)$ on $(0, \mu)$ is reached at $-1 + \sqrt{1+\mu}$. As a consequence, we find that when $\lambda_k < \mu$ for $1 \leqslant k \leqslant k_0$ and at the zeroth order in ε, either the mode $k^\star - 1$ or $k^\star$ will be observed, where $k^\star$ is defined by $\lambda_{k^\star - 1} \leqslant -1 + \sqrt{1+\mu} < \lambda_{k^\star}$.[5]

Computing the modes explicitly. Since Ω has a particular shape, eigenvalues and eigenfunctions can actually be explicitly computed. We first consider the case of the 2D simulations, namely when Ω is a disk of radius a. It is then standard that, after separation of variables in polar coordinates $\psi(x, y) = f(r)g(\theta)$, the equation $-\Delta\psi = \lambda\psi$ with Neumann boundary conditions is equivalent to $g(\theta) = A\cos(m\theta) + B\sin(m\theta)$ for some $m \in \mathbb{Z}$ and $\rho \mapsto f(\frac{\rho}{\sqrt{\lambda}})$ must solve the Bessel equation $\rho^2 y''(\rho) + \rho y'(\rho) + (\rho^2 - m^2)y(\rho) = 0$

[5] The actual mode observed is $k^\star - 1$ (resp. $k^\star$) if $\lambda^+(\lambda_{k^\star - 1}) > \lambda^+(\lambda_{k^\star})$ (resp. $\lambda^+(\lambda_{k^\star - 1}) \leqslant \lambda^+(\lambda_{k^\star})$).

with $y'(0) = y'(\sqrt{\lambda}a) = 0$. This yields, up to a constant, to the result $f(r) = J_m(\sqrt{\lambda}r)$ where J_m is the first kind Bessel function of order m. The boundary conditions impose $m \neq \pm 1$ (because $J'_m(0) = 0$ for all m except 1 and -1), while, denoting $\gamma_{m,p}$ the pth zero of the derivative of J_m, we find $\lambda = \left(\frac{\beta_{m,p}}{a}\right)^2$.

Summing up, we obtain

$$\lambda_{m,p} = \left(\frac{\beta_{m,p}}{a}\right)^2,$$
$$\psi_{m,p}(r,\theta) = J_m\left(\frac{\beta_{m,p}}{a}r\right)(A\cos(m\theta) + B\sin(m\theta)),$$

a family indexed by $m \in \mathbb{Z} \setminus \{\pm 1\}$, $p \in \mathbb{N}^\star$. Unless $m = 0$ for which the eigenfunction is unique after normalisation, the eigenspace associated to $\lambda_{m,p}$ is of dimension 2.

Similar computations for the case of a cylinder of height h and radius a lead to the result

$$\lambda_{m,p,l} = \left(\frac{\beta_{m,p}}{a}\right)^2 + \left(\frac{l\pi}{h}\right)^2,$$
$$\psi_{m,p,l}(r,\theta,z) = J_m\left(\frac{\beta_{m,p}}{a}r\right)(A\cos(m\theta) + B\sin(m\theta))\cos\left(\frac{l\pi z}{h}\right),$$

a family indexed by $m \in \mathbb{Z} \setminus \{\pm 1\}$, $p \in \mathbb{N}^\star$, $l \in \mathbb{N}$. The multiplicity of eigenfunctions is the same as in the previous case (1 if $m = 0$ and 2 if not).

If h is small, we will typically not see the eigenvalues with high frequency in the z variable, and thus only the modes with $l = 0$ will be observed. An interesting and relevant consequence is that from the point of view of Turing instabilities, it is a good approximation to neglect the z variable and focus on $\Omega \subset \mathbb{R}^2$ as a disk for simulations.

Choosing the nonlinearity. From the previous computations, we uncover that $\mu = A\varphi(M) - 1$ has the same monotony as φ in terms of the initial mass: hence, we expect that there should be more spheroids for higher values of $\varphi(M)$, all other parameters being fixed. Experiments have shown that the number of spheroids increases with M, and reaches a plateau at 50000, 75000 seeded cells. For 100000 cells, there are essentially no spheroids in the hydrogel since most of them have escaped it. It is not clear whether it proves that no patterns are exhibited in this situation.

However, we can already assert that the classical linear sensitivity $\varphi(n) = n$ is not an appropriate modelling choice. This is why we will instead consider either the *logistic* function $\varphi(n) = n(1 - \frac{n}{K})$ or the *exponential* one $\varphi(n) = ne^{-\frac{n}{K}}$. The first one has been proposed long ago to prevent overcrowding [75, 137], since the density K plays the role of a maximal packing above which movement by chemotaxis is shut down. The second is less classical but has the advantage of flexibility because it does not impose an a priori maximal density.

7.3 Comparison of 2D simulations with experiments

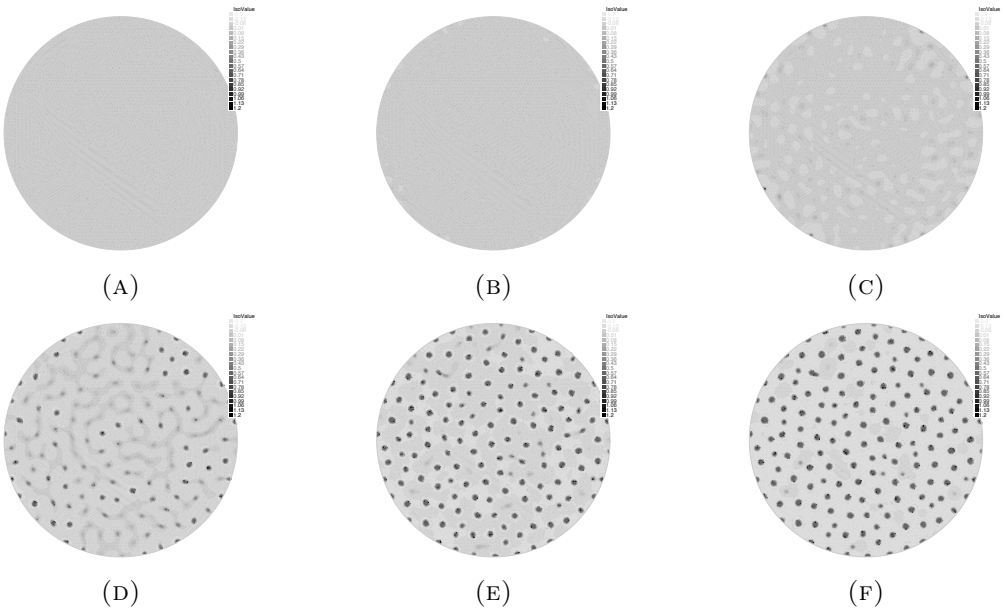
In this section, we perform several numerical simulations, and for simplicity, all simulations will be done with the adimensionalised system, with parameters, K, M, ε and A, as well as the radius a of the disk in these 2D simulations. Simulations for the model with growth have r_0 as an additional parameter. We shall always fix in the forthcoming simulations the following values:

$$M = 0.1, \ \varepsilon = 1.10^{-2}, \ K = 1, \ a = 40.$$

The initial condition is always taken to be a small perturbation of the homogeneous steady state (M, M). The value of A will vary depending on the simulation. All simulations have been done using the software Freefem++ [73], with finite elements $P1$ basis functions, $dt = 0.01$ and a mesh made of around 40000 triangles.

7.3.1 Pattern formation and agreement with the theoretical results

Let us start by providing a typical simulation exhibiting Turing instability with spheroidal patterns, in the logistic case $\varphi(n) = n(1 - \frac{n}{K})$. We take $A = 50$, for which we expect patterns since $\mu = A\varphi(M) - 1 = 3.5 > 0$. We report the results from times $t = 5$ to $t = 8$ on Figure 7.2. From time $t = 0$ to $t = 5$, only the constant state is visible. From $t = 6$ onwards, patterns very quickly become visible, first at the periphery and then, centripetally, everywhere in the domain. Most of them have formed at $t = 8$.



(A) (B) (C)

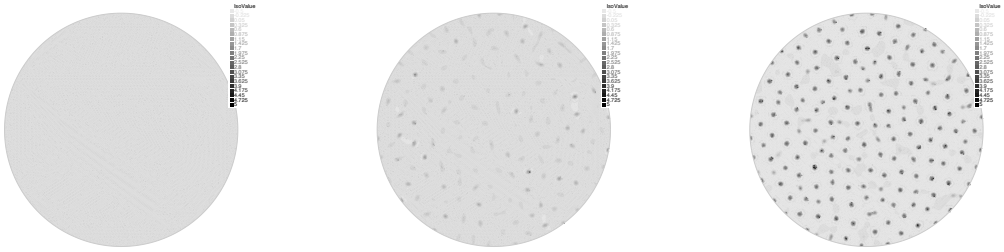
(D) (E) (F)

FIGURE 7.2: Simulation of the cancer cell density (satisfying $0 \leqslant n \leqslant 1$ up to numerical errors) for the Keller-Segel model (7.2) without growth, with $A = 50$ and a logistic sensitivity function, from times $t = 5$ to $t = 8$.

For the exponential function, the results are close and it is not clear whether one of the two is better. Simulations not shown here also exhibit round patterns are also obtained, but there are more of them and they are smaller. This is also in agreement with the Turing instability condition, since for the same K, we always have $ne^{-\frac{n}{K}} \geqslant n(1-\frac{n}{K})$. Thus, all other parameters being equal, μ is larger in the exponential case than in the logistic case.

7.3.2 Pattern formation with growth

Once patterns have formed, a long transient phase starts, during which spheroids slowly merge, in accordance with some theoretical results obtained in 1D [137]. From the experimental point of view, this peculiar phenomenon is not relevant since it requires long time scales during which growth will take over. We now provide some simulations with the growth term $r_0 n$, in Figure 7.3. The exponential sensitivity function is preferred over the logistic one, since growth can make densities exceed the value 1 (and then the advection term would make cells go in the opposite direction of ∇c in zones of densities higher than 1), which is not an appropriate modelling.

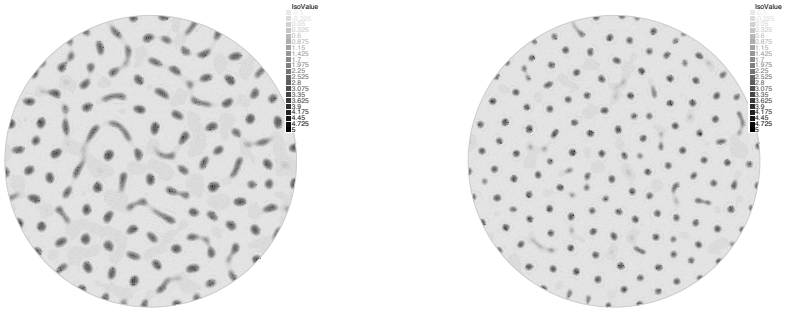


(A) (B) (C)

FIGURE 7.3: Simulation of the cancer cell density (which approximately ranges from 0 to 5 up to numerical errors)for the Keller-Segel model (7.2) with growth, with $A = 40$, $r_0 = 0.1$ and exponential sensitivity function, at times $t = 5$, $t = 6$ and $t = 7$.

We finally provide a comparison with a simulation with a lower value for A, in order to investigate the effect of a change of parameters modelling differences between the more adhesive MCF7 cells with the more invasive MCF7-sh-WISP2. The rationale is that, assuming that the sensitivity strength measured by χ is roughly the same for both cell lines, the diffusion constant D_1 is expected to be higher for the MCF7-sh-WISP2. In the absence of growth, the Turing pattern constant μ will thus be larger for MCF7 cells than it is for MCF7-sh-WISP2: we expect less spheroids, and of bigger size for the latter.

Figure 7.4 compares spheroids for values $A = 20$ and $A = 30$. It indeed shows that, even with the growth term, the spheroids are bigger and less numerous for a smaller A. It can be interpreted as them being less compact, a feature which is coherent with experiments. Another interesting feature is that the geometry has also changed, as structures look less round.



(A) (B)

FIGURE 7.4: Simulation of the cancer cell density (which approximately ranges from 0 to 5 up to numerical errors) for the Keller-Segel model (7.2) with growth, with $A = 20$ (A) and $A = 30$ (B) at times $t = 15$ and $t = 10$ respectively.

7.3.3 Conclusions and perspectives

Our 2D simulations show a good ability at qualitatively reproducing the patterns obtained experimentally. Contrarily to other modelling approaches, the salient feature is not cell-scaffold adhesion, which we reduce to a diffusion term, but instead chemotaxis. We thus hypothesise that it is a key phenomenon responsible for these aggregates.

Once a nonlinear sensitivity function has been chosen, the monotony of the Turing instability constant μ, as a function of the diffusion D_1, is qualitatively consistent with the differences observed between MCF7 cells and MCF7-sh-WISP2 cells. The number, size, and to a lesser extent, the shape of spheroids match well, at least for days J4 and J8. Going further in this direction would require to analyse Turing instabilities in the presence of the growth term, since it is not the same for both cell lines, building up on works carried out for growing domains [114].

The model has however difficulties in showing more flexibility in the size of patterns. To quantify that, the numerical results have also been analysed with the software counting spheroids, and the distribution (in size) of spheroids does not match very well: while the standard deviation was of the order of the average size for experiments, we find that standard deviation is about one third of the mean size in simulations. Numerically, we also did not manage to account for the important increase in the number of spheres at day J14.

Finally, we remark that the numerical schemes (based on finite element methods) do not preserve desirable properties, such as positivity of the solution, nor that it should be below 1 in the absence of growth and with a logistic sensitivity. The difficulties at preserving these bounds are mostly seen at the sharp interfaces between zones of high densities (the spheroids), and zones of low densities (the rest). We address this problem in the next chapter.

Chapter 8

Numerical analysis of schemes for the 1D Keller-Segel equation

---

In this work with Federica Bubba, Luis Neves de Almeida and Benoît Perthame, we provide two finite-volume schemes for the 1D Fokker-Planck equation and parabolic-elliptic Keller-Segel equation with a nonlinear sensitivity. They are obtained either by following the Gradient-Flow structure or by a rewriting inspired by the Scharfetter-Gummel discretisation, so that the schemes preserve energy, steady states and bounds at the semi-discrete level. An implicit-explicit (in time) scheme is then proposed by tuning the discretisation of each term in order to obtain appropriate monotony ensuring that the scheme is well-posed and still preserves the important properties at the discrete level. This has been accepted under the name *Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations* [22].

8.1 Introduction

Taxis-diffusion and aggregation equations are widely studied in the context of biological populations (see [120, 77, 74, 32] for instance). They describe cell communities which react to external stimuli and form aggregates of organisms (pattern formation), such as bacterial colonies, slime mold or cancer cells. The Patlak-Keller-Segel model [87] is the most famous

system and we are interested in the following generalisation

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[\frac{\partial u}{\partial x} - \varphi(u)\frac{\partial v}{\partial x}\right] = 0, & x \in (0,1),\ t > 0, \\ \frac{\partial u}{\partial x} - \varphi(u)\frac{\partial v}{\partial x} = 0, & \text{for } x = 0 \text{ or } 1, \\ u(x,0) = u^0(x) \geqslant 0, & x \in [0,1]. \end{cases} \tag{8.1}$$

Here, $u(x,t) \geqslant 0$ represents the density of a given quantity (e.g., cells or bacteria population) and the initial data $u^0(x)$ is a given nonnegative smooth function. As for the function v, which models a molecular concentration, we choose either the case of the Fokker-Planck (FP in short) equation, where $v(x)$ is known

$$v := v(x) \geqslant 0, \qquad \frac{\partial v}{\partial x} \in L^\infty(0,1),$$

or the case of the generalised Keller-Segel (GKS in short) equation, where

$$v(x,t) = \int K(x,y)u(y,t)dy, \qquad K(x,y) \text{ a smooth symmetric kernel.} \tag{8.2}$$

Depending on the modelling choice for $\varphi(u)$, solutions to (8.1) can blow-up in finite time depending upon a critical mass (see [122, 16]) or reach stationary profiles in the form of peaks or plateaus [137] (pattern formation by Turing instability). The high nonlinearities due to the advection term make problem (8.1) mainly intractable through analytical methods. Thus, it is important to dispose of reliable numerical methods avoiding non-physical oscillations and numerical instabilities even when dealing with non-smooth solutions. The main properties that one wishes to preserve in a numerical method are

(P1) *positivity property*, since we are dealing with densities or concentrations,

$$u(x,t) \geqslant 0, \tag{8.3}$$

(P2) *mass conservation*, because no-flux boundary conditions are imposed,

$$\int_0^1 u(x,t) = \int_0^1 u^0(x)dx, \tag{8.4}$$

(P3) *preservation of discretised steady states* of the form

$$g(u) = \mu + v, \qquad g'(u) = \frac{1}{\varphi(u)}, \tag{8.5}$$

where μ is a parameter related to the mass of u, and

(P4) *energy dissipation*

$$\frac{d}{dt}\mathcal{E}(t) \leqslant 0, \qquad \mathcal{E}(t) = \int_0^1 \left[G(u) - \kappa uv\right] dx, \tag{8.6}$$

where $G(u)$ is a primitive of $g(u)$ and the value of κ differs for the two cases we study here, namely

$$\kappa = 1 \quad \text{(FP case)}, \qquad \kappa = \frac{1}{2} \quad \text{(GKS case)}. \tag{8.7}$$

The aims of our work are first to recall two points of view for the derivation of the above energy inequality, second to use them for the construction of conservative, finite volume numerical schemes preserving energy dissipation to solve equation (8.1), third to make numerical comparisons in the case of complex patterns in order to distinguish physical instabilities from numerical artifacts. The two different derivations of the energy dissipation use two symmetrisation strategies: the gradient flow or the Scharfetter-Gummel approaches. It turns out that they lead to two strategies for discretisation of problem (8.1). We prove that the proposed schemes satisfy properties (8.3)–(8.6) and because we build implicit schemes, there is no limitation on the time step in the fully discrete case.

There exist other works which propose schemes for the resolution of problems in the form (8.1). For instance, finite elements methods are used, see [147] and references therein. Optimal transportation schemes for Keller-Segel systems are introduced in [15]. The papers [31] and [30] propose a finite-volume method able to preserve the above properties, including energy dissipation, at the semi-discrete level or with an explicit in time discretisation, using the gradient flow approach. The symmetrisation using the Scharfetter-Gummel approach is used in [106] where properties similar to ours are proved. However, the results do not include sensitivity saturation. To the best of our knowledge, our work is the first to propose implicit in time methods, without time step limitation (CFL condition), for which we are able to prove that, under generic conditions, the energy decreases at both semi-discrete and discrete level. Moreover, we build an alternative to the gradient flow approach applying the Scharfetter-Gummel strategy [148] for the discretisation of drift-diffusion equations (8.1) with a general saturation function φ.

The chapter is organised as follows. In Section 8.2, we present in more details our assumptions for the equation (8.1). We also explain some modelling choices in particular for the nonlinearity $\varphi(u)$ and on the choice of the kernel K. 8.3 is devoted to the introduction of the two approaches, gradient flow or Scharfetter-Gummel, and to how we use the continuous version of energy dissipation to derive the schemes. In Sections 8.4 and 8.5, we show how the aforementioned two approaches lead to two different numerical methods, developed from the semi-discrete (only in space discretisation) level to the fully discretised scheme. In particular, using a general result recalled in 8.8 about monotone schemes, we prove that the proposed schemes are well-posed and satisfy the fundamental properties (8.3)–(8.6). These theoretical results are illustrated in Section 8.6 by numerical simulations: we compare the gradient flow and the Scharfetter-Gummel schemes with the upwind approach, typically used to solve this kind of models.

8.2 Assumptions and notations

The standard biological interpretation of (8.1) ([76, 120, 130]) provides us with some further properties of the nonlinearities which we describe now.

Chemotactic sensitivity. The function $\varphi(u)$ is called chemotactic sensitivity. It determines how the random movement of particles of density u is biased in the direction of the gradient of v. In order to include the different choices of φ, as $\varphi(u) = u$ as in the Keller-Segel or drift-diffusion model, or the logistic case $\varphi(u) = u(1-u)$, or the generalised case $\varphi(u) = ue^{-u}$, we use the formalism

$$\varphi(u) = u\psi(u), \qquad \text{with} \qquad \psi(u) \geqslant 0, \qquad \psi'(u) \leqslant 0.$$

More precisely, we consider two cases for the smooth function ψ,

$$\psi(u) > 0, \qquad \forall u > 0, \tag{8.8}$$

or

$$\psi(u) > 0 \quad \text{for} \quad 0 < u < M, \qquad \psi(M) = 0. \tag{8.9}$$

In the case (8.9) we only consider solutions which satisfy $u \in [0, M]$.

It is convenient to introduce the notations

$$g(u) = \int_a^u \frac{1}{\varphi(v)} dv, \qquad G(u) = \int_0^u g, \tag{8.10}$$

where a is a constant that enables to get rid of the integration constant and depends on the choice of φ. For instance, for the standard case $\varphi(u) = u$, $a = 1$ and one obtains $g(u) = \ln(u)$ and $G(u) = u\ln(u) - u$. For functions φ satisfying (8.8), a natural hypothesis which is related to blow-up is the following

$$\frac{1}{\varphi} \notin L^1(1, +\infty), \qquad g(u) \underset{u\to\infty}{\longrightarrow} +\infty, \tag{8.11}$$

an assumption which, as we see it later, appears naturally when it comes to the well-posedness of numerical schemes.

Note that under assumption (8.8) and if ψ is bounded, solutions exist globally and are uniformly bounded [40]. Under assumption (8.9), if $0 \leqslant u^0 \leqslant M$, the solution is also globally defined and satisfies $0 \leqslant u(t, \cdot) \leqslant M$ for all times [74].

Expression of the drift v. The convolution expression for v as a function of u has been widely used in recent studies [4, 8, 13]. It also comes from the Keller-Segel model [77, 87, 130], where the equation for the cells density in (8.1) is complemented with a parabolic equation for the chemoattractant concentration v. Since the chemoattractant is supposed to diffuse much quicker than the cells density, we can consider a simplified form of the Keller-Segel system and couple (8.1) with the elliptic equation for v

$$\begin{cases} -\frac{\partial^2 v}{\partial x^2} = u - v, & x \in (0, 1), \\ \frac{\partial v}{\partial x} = 0, & x = 0 \text{ or } 1. \end{cases}$$

This equation leads to (8.2) using the Green function given by the positive and symmetric kernel $K(x,y)$ defined as

$$K(x,y) = \lambda\left(e^{x}+e^{-x}\right)\left(e^{y}+e^{2-y}\right), \quad x \leqslant y, \qquad \lambda = \frac{1}{2\left(e^{2}-1\right)}.$$

Notations for numerical schemes. We give here our notations for discretisation. We consider a (small) space mesh size $\Delta x = \frac{1}{I}$, $I \in \mathbb{N}$. The mesh is centered at $x_i = i\Delta x$, with endpoints $x_{i+1/2} = (i+1/2)\Delta x$ for $i = 1, ..., I-1$. Therefore, our computational domain is always shifted and takes the form $\left(\frac{\Delta x}{2}, (I+\frac{1}{2})\Delta x\right)$. Finally, the mesh is formed by the intervals

$$I_i = \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right), \qquad x_{i+\frac{1}{2}} = \left(i+\frac{1}{2}\right)\Delta x.$$

As for time discretisation, we consider (small) time steps $\Delta t > 0$, and set $t^n = n\Delta t$. The approximation of $u(x,t)$, interpreted in the finite volume sense ([19, 101]), is denoted by

$$u_i^n \approx \frac{1}{\Delta x}\int_{I_i} u(x,t^n)dx.$$

Integration on the interval I_i yields fluxes $F_{i+\frac{1}{2}} \approx \left(\frac{\partial u}{\partial x} - \varphi(u)\frac{\partial v}{\partial x}\right)\Big|_{x_{i+\frac{1}{2}}}$ for $i = 0, \ldots, I-1$ through the interval interfaces.

8.3 Energy dissipation

Energy dissipation is the most difficult property to preserve in a discretisation and methods might require corrections [84]. Therefore, it is useful to recall how it can be derived simply for the continuous equation. We focus on two different strategies, that lead to two different discretisation approaches, the gradient flow approach and the Scharfetter-Gummel approach.

The gradient flow approach to energy. Using the notations (8.10), the equation for u can be rewritten as

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[\varphi(u)\frac{\partial\left(g(u)-v\right)}{\partial x}\right] = 0, \tag{8.12}$$

so that

$$\begin{aligned}\left(g(u)-v\right)\frac{\partial u}{\partial t} &= \left(g(u)-v\right)\frac{\partial}{\partial x}\left[\varphi(u)\frac{\partial\left(g(u)-v\right)}{\partial x}\right] \\ &= \frac{1}{2}\frac{\partial}{\partial x}\left[\varphi(u)\frac{\partial(g(u)-v)^2}{\partial x}\right] - \varphi(u)\left[\frac{\partial(g(u)-v)}{\partial x}\right]^2.\end{aligned}$$

Consequently, we find, in the Fokker-Planck case

$$\frac{d}{dt}\int_0^1 [G(u) - uv(x)]dx = -\int_0^1 \varphi(u)\left[\frac{\partial(g(u)-v)}{\partial x}\right]^2 dx \leqslant 0,$$

and in the generalised Keller-Segel case

$$\frac{d}{dt}\int_0^1 [G(u) - \frac{1}{2}uv(x,t)]dx = -\int_0^1 \varphi(u)\left[\frac{\partial(g(u)-v)}{\partial x}\right]^2 dx \leqslant 0,$$

because, thanks to the symmetry assumption on K and by using the definition (8.2) of v, we have

$$\int\int K(x,y)u(y,t)\frac{\partial u(x,t)}{\partial t} = \int\int K(x,y)\frac{\partial u(y,t)}{\partial t}u(x,t) = \frac{1}{2}\frac{d}{dt}\int\int K(x,y)u(y,t)u(x,t).$$

The Scharfetter-Gummel approach to energy. Inspired from the case of electric forces in semi-conductors, the equation for u can be rewritten as

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[e^{v-g(u)}\varphi(u)\frac{\partial e^{g(u)-v}}{\partial x}\right] = 0, \tag{8.13}$$

so that

$$(g(u)-v)\frac{\partial u}{\partial t} = (g(u)-v)\frac{\partial}{\partial x}\left[e^{v-g(u)}\varphi(u)\frac{\partial e^{g(u)-v}}{\partial x}\right]$$

$$= \frac{\partial}{\partial x}\left[(g(u)-v)e^{v-g(u)}\varphi(u)\frac{\partial e^{g(u)-v}}{\partial x}\right] - e^{v-g(u)}\varphi(u)\frac{\partial e^{g(u)-v}}{\partial x}\frac{\partial\,(g(u)-v)}{\partial x}.$$

It is immediate to see that the last term has the negative sign while the time derivative term is exactly the same as in the gradient flow approach.

At the continuous level, these two calculations are very close to each other. However, they lead to the construction of different discretisations. The gradient flow point of view is used for numerical schemes by [32], the Scharfetter-Gummel approach is used in [106].

8.4 Semi-discretisation

The mass conservative form of (8.1) leads to a finite volume type semi-discrete scheme

$$\begin{cases} \frac{du_i(t)}{dt} + \frac{1}{\Delta x}[F_{i+1/2}(t) - F_{i-1/2}(t)] = 0, \qquad i = 0,\dots,I, \;\; t > 0, \\ \\ F_{1/2}(t) = F_{I+1/2}(t) = 0. \end{cases} \tag{8.14}$$

We use the definition (8.7) for κ and set for $i = 0, \dots, I$

$$E_i(t) = G(u_i) - \kappa u_i v_i.$$

The semi-discrete energy is then

$$\mathcal{E}_{\mathrm{sd}}(t) = \Delta x \sum_{i=0}^{I} E_i(t).$$

8.4.1 The gradient flow approach

Using the form (8.12) of equation (8.1), we define the semi-discrete flux as

$$F_{i+1/2}(t) = -\frac{\varphi_{i+1/2}}{\Delta x}\big[g(u_{i+1}) - v_{i+1} - \big(g(u_i) - v_i\big)\big],\ i = 1, \dots, I-2. \tag{8.15}$$

The precise expression of $\varphi_{i+1/2}$ is not relevant for our present purpose which is to preserve the energy dissipation property. However, for stability considerations it is useful to upwind, an issue which we shall tackle when we consider the full discretisation.
Then, the semi-discrete energy form is obtained after multiplication by $(g(u_i) - v_i)$ and yields

$$\begin{aligned}\frac{d}{dt}\Delta x \sum_{i=0}^{I} E_i(t) &= -\sum_{i=1}^{I-1}(g(u_i) - v_i)[F_{i+1/2} - F_{i-1/2}] \\ &= \sum_{i=1}^{I-1} F_{i+1/2}\big[\big(g(u_{i+1}) - v_{i+1}\big) - \big(g(u_i) - v_i\big)\big].\end{aligned}$$

Therefore, we find the semi-discrete form of energy dissipation

$$\frac{d\mathcal{E}_{\mathrm{sd}}}{dt} = -\Delta x \sum_{i=1}^{I-1} \varphi_{i+1/2}\left[\frac{\big(g(u_{i+1}) - v_{i+1}\big) - \big(g(u_i) - v_i\big)}{\Delta x}\right]^2 \leqslant 0.$$

8.4.2 The Scharfetter-Gummel approach

We choose to discretise the form (8.13), defining the semi-discrete flux as

$$F_{i+1/2}(t) = -\frac{\big(e^{v-g(u)}\varphi(u)\big)_{i+1/2}}{\Delta x}\big[e^{g(u_{i+1})-v_{i+1}} - e^{g(u_i)-v_i}\big],\ i = 1, \dots, I-2, \tag{8.16}$$

where, again, the specific form of the interpolant $\big(e^{v-g(u)}\varphi(u)\big)_{i+1/2}$ is not relevant here.

As above, the semi-discrete energy form follows upon multiplication by $g(u_i)-v_i$ and reads

$$\begin{aligned}\frac{d}{dt}\Delta x\sum_{i=0}^{I}E_i(t) &= -\sum_{i=1}^{I-1}\big(g(u_i)-v_i\big)\left[F_{i+1/2}-F_{i-1/2}\right]\\ &= \sum_{i=1}^{I-1}F_{i+1/2}\left[\big(g(u_{i+1})-v_{i+1}\big)-\big(g(u_i)-v_i\big)\right].\end{aligned}$$

Summing up, the semi-discrete form of energy dissipation here writes

$$\frac{d\mathcal{E}_{\rm sd}}{dt} = -\Delta x\sum_{i=1}^{I-1}\big(e^{v-g(u)}\varphi(u)\big)_{i+1/2}\frac{e^{g(u_{i+1})-v_{i+1}}-e^{g(u_i)-v_i}}{\Delta x}\,\frac{\big(g(u_{i+1})-v_{i+1}\big)-\big(g(u_i)-v_i\big)}{\Delta x},$$

and thus we also have $\frac{d\mathcal{E}_{\rm sd}}{dt}\leqslant 0$.

8.4.3 Discrete steady states

Steady states make the energy dissipation vanish which imposes both in the gradient flow and the Scharfetter-Gummel approaches that $\big(g(u_{i+1})-v_{i+1}\big)=\big(g(u_i)-v_i\big)$. In other words they are given, up to a constant μ, as the discrete version of (8.5),

$$g(u_i)=v_i+\mu,\quad \forall i=0,\dots,I\,. \tag{8.17}$$

We recall from [137] that in the GKS case, there are several steady states and the constant ones can be unstable.

8.5 Fully discrete schemes

To achieve the time discretisation, and restricting our analysis to the Euler scheme, we write the time difference $\frac{du_i(t)}{dt}$ as $\frac{u_i^{n+1}-u_i^n}{\Delta t}$. Therefore, the full discretisation of (8.14) writes

$$\begin{cases} u_i^{n+1}-u_i^n+\frac{\Delta t}{\Delta x}\left[F_{i+1/2}^{n+1}-F_{i-1/2}^{n+1}\right]=0, \qquad i=0,\dots,I,\\ F_{1/2}^{n+1}=F_{I-1/2}^{n+1}=0.\end{cases} \tag{8.18}$$

The issue here is to decide which terms (in u and v) should be discretised with implicit or explicit schemes based on fully discrete energy dissipation. We claim that, apart from the interpolant, we need to make the terms in u_i implicit and, for the GKS case, the terms in v_i explicit, a fact on which we now elaborate.

The computation made in the semi-discrete case, $\frac{dE_i(t)}{dt}=\frac{du_i(t)}{dt}(g(u_i(t))-v_i(t))$, extends to the fully discrete setting and leads to the following constraint on the energy

$$\sum_{i=0}^{I}\big(E_i^{n+1}-E_i^n\big)\leqslant\sum_{i=0}^{I}(u_i^{n+1}-u_i^n)(g(u_i^{\alpha_n})-v_i^{\beta_n}).$$

Here, $u_i^{\alpha_n} := \alpha u_i^n + (1-\alpha)u_i^{n+1}$ $v_i^{\beta_n} := \beta v_i^n + (1-\beta)u_i^{n+1}$. The convexity of $G(\cdot)$ imposes the choice of an implicit discretisation for u, namely $\alpha = 0$, because

$$G(u_i^{n+1}) - G(u_i^n) \leqslant g(u_i^{n+1})(u_i^{n+1} - u_i^n).$$

Regarding the term in uv, only the case of GKS needs to be fixed and we thus require

$$-\sum_{i=0}^{I} \left[(uv)_i^{n+1} - (uv)_i^n\right] \leqslant -2\sum_{i=0}^{I} v_i^{\beta_n}(u_i^{n+1} - u_i^n).$$

It is natural to try and balance the terms by choosing a semi-explicit discretisation with $\beta = \frac{1}{2}$, which yields

$$\begin{aligned} 2\sum_{i=0}^{I}\left[v_i^{\beta_n}(u_i^{n+1} - u_i^n) - \left(u_i^{n+1}v_i^{n+1} - u_i^n v_i^n\right)\right] &= \sum_{i=0}^{I}\left(u_i^{n+1}v_i^n - u_i^n v_i^{n+1}\right) \\ &= \sum_{i,j} K_{ij}\left(u_i^{n+1}u_j^n - u_i^n u_j^{n+1}\right) \end{aligned}$$

with the last term vanishing due to the symmetry of K.

However, implicit and explicit time discretisations for v can also be considered at the expense of adding hypotheses on the kernel K. Indeed, for a given $0 \leqslant \beta \leqslant 1$, we find

$$2\sum_{i=0}^{I}\left[v_i^{\beta_n}(u_i^{n+1} - u_i^n) - \left(u_i^{n+1}v_i^{n+1} - u_i^n v_i^n\right)\right] = (2\beta - 1)\sum_{i,j} K_{ij}(u_i^{n+1} - u_i^n)(u_j^{n+1} - u_j^n).$$

As a consequence, an explicit (resp. implicit) scheme is suitable for the time discretisation of v provided that K is a non-negative (resp. non-positive) symmetric kernel. Since K is a non-negative symmetric kernel for the Keller-Segel equation (8.2), for simplicity we choose an explicit discretisation for v.

Finally, we note that the interpolant does not play any role for energy discretisation and we can use the simplest explicit or implicit discretisation (both in u and v), so as to make the analysis of the scheme as simple as possible.

8.5.1 The gradient flow approach

We consider the full discretisation of (8.15) and define the fully discrete flux in (8.18) as

$$F_{i+1/2}^{n+1} = -\frac{\varphi(u)_{i+1/2}^{n+1}}{\Delta x}\left[\left(g(u_{i+1}^{n+1}) - v_{i+1}^n\right) - \left(g(u_i^{n+1}) - v_i^n\right)\right], \qquad i = 1, ..., I-2. \tag{8.19}$$

At this level, we need to define the form of the interpolant $\varphi(u)_{i+1/2}^{n+1}$. From the theorem in 8.8, we use an upwind technique in order to ensure well-posedness and monotonicity properties of the scheme. Thus, for $i = 1, \ldots, I-2$, we define

$$\varphi(u)_{i+1/2}^{n+1} := \begin{cases} u_i^{n+1}\psi(u_{i+1}^{n+1}) & \text{when} \quad g(u_i^{n+1}) - g(u_{i+1}^{n+1}) + v_{i+1}^n - v_i^n \geqslant 0, \\ u_{i+1}^{n+1}\psi(u_i^{n+1}) & \text{when} \quad g(u_i^{n+1}) - g(u_{i+1}^{n+1}) + v_{i+1}^n - v_i^n < 0. \end{cases} \tag{8.20}$$

Proposition 8.1 (Fully discrete gradient flow scheme)**.** *We assume either* (8.8) *and* (8.11)*, or* (8.9) *and give the* $u_i^0 \geqslant 0$*. Then, the scheme* (8.19)–(8.20) *has the following properties:*

(i) *the solution* u_i^n *exists and is unique, for all* $i = 0, \dots, I$*, and* $n \geqslant 1$*;*

(ii) *it satisfies* $u_i^n \geqslant 0$*, and* $u_i^n \leqslant M$ *for the case* (8.9)*, if it is initially true;*

(iii) *the steady states* $g(u_i) - v_i = \mu$ *are preserved;*

(iv) *the energy dissipation inequality is satisfied*

$$\mathcal{E}^{n+1} - \mathcal{E}^n \leqslant -\frac{\Delta t}{\Delta x} \sum_{i=1}^{I-1} \varphi(u)_{i+1/2}^n \left[\left(g(u_{i+1}^{n+1}) - v_{i+1}^n \right) - \left(g(u_i^{n+1}) - v_i^n \right) \right]^2 .$$

Notice that this theorem does not state a uniform bound in the case (8.8) and (8.11).

Proof. **(i)** We prove that the scheme satisfies the hypotheses of the theorem in 8.8. We set

$$A_{i+1/2}(u_i^{n+1}, u_{i+1}^{n+1}) = \frac{\Delta t}{(\Delta x)^2} F_{i+1/2}^{n+1}.$$

Then, the simplest case is when φ satisfies (8.9), since clearly $u_i^{n+1} \equiv 0$ and $u_i^{n+1} \equiv M$ are respectively a sub- and supersolution. When φ satisfies (8.8) and (8.11), $u_i^{n+1} \equiv 0$ is again a subsolution, while for the supersolution we choose $\bar{U}_i^{n+1} = g^{-1}(C + v_i^n)$. Such a choice indeed makes the flux terms vanish:

$$F_{i+1/2}^{n+1} = -\frac{\varphi(u)_{i+1/2}^{n+1}}{\Delta x} \left[\left(g(\bar{U}_{i+1}^{n+1}) - v_{i+1}^n \right) - \left(g(\bar{U}_i^{n+1}) - v_i^n \right) \right] = -\frac{\varphi(u)_{i+1/2}^{n+1}}{\Delta x} \left[C - C \right] = 0.$$

Thus $\bar{U}_i^{n+1}$ is a supersolution as soon as $g^{-1}(C + v_i^n) \geqslant u_i^n$, which holds when C is taken to be large enough because we recall that assumption (8.11) ensures that $g(u)$ tends to $+\infty$ as u tends to $+\infty$.

Moreover, the scheme is monotone since

$$\begin{aligned} \partial_1 A_{i+\frac{1}{2}}(u_i^{n+1}, u_{i+1}^{n+1}) = &- \frac{\Delta t}{(\Delta x)^2} u_{i+1}^{n+1} \psi'(u_i^{n+1}) \left[g(u_{i+1}^{n+1}) - v_{i+1}^n - \left(g(u_i^{n+1}) - v_i^n \right) \right]_+ \\ &- \frac{\Delta t}{(\Delta x)^2} \psi(u_{i+1}^{n+1}) \left[g(u_{i+1}^{n+1}) - v_{i+1}^n - \left(g(u_i^{n+1}) - v_i^n \right) \right]_- \\ &- \frac{\Delta t}{(\Delta x)^2} \varphi(u)_{i+\frac{1}{2}}^{n+1} \left[-g'(u_i^{n+1}) \right] \geqslant 0, \end{aligned}$$

where

$$0 \leqslant [x]_+ = \begin{cases} x & \text{for} \quad x \geqslant 0 \\ 0 & \text{for} \quad x < 0 \end{cases} \qquad \text{and} \qquad 0 \geqslant [x]_- = \begin{cases} 0 & \text{for} \quad x \geqslant 0 \\ x & \text{for} \quad x < 0 \end{cases} .$$

(ii) Positivity of discrete solutions and the upper bound in the logistic case follow from the subsolution and supersolution build in step (i).

(iii) Preservation of steady states at the discrete level follows immediately from the form we have chosen for the fully discrete fluxes.
(iv) For the energy inequality, we remark that the contribution regarding time discretisation is treated in the introduction of the present section. The space term is exactly treated as in the corresponding subsection of Section 8.4. □

8.5.2 The Scharfetter-Gummel approach

In (8.18), the fully discrete Scharfetter-Gummel flux reads as

$$F_{i+1/2}^{n+1} = \left(e^{v^n - g(u^n)}\varphi\left(u^{n+1}\right)\right)_{i+1/2}\left[e^{g(u_{i+1}^{n+1})-v_{i+1}^n} - e^{g(u_i^{n+1})-v_i^n}\right], \qquad i = 1, ..., I-2. \tag{8.21}$$

As for the gradient flow approach, we need the upwind technique to get a scheme which satisfies the hypotheses in 8.8. Thus, we set for $i = 1, \dots, I-2$

$$\left(e^{v^n - g(u^n)}\varphi\left(u^{n+1}\right)\right)_{i+1/2} = u_{i+1}^{n+1}\psi(u_i^{n+1})e^{v_{i+1}^n - g(u_{i+1}^n)} \text{ if } e^{\left(g(u_{i+1}^{n+1})-v_{i+1}^n\right)} - e^{\left(g(u_i^{n+1})-v_i^n\right)} \geqslant 0, \tag{8.22}$$

and

$$\left(e^{v^n - g(u^n)}\varphi\left(u^{n+1}\right)\right)_{i+1/2} = u_i^{n+1}\psi(u_{i+1}^{n+1})e^{v_i^n - g(u_i^n)} \text{ if } e^{\left(g(u_{i+1}^{n+1})-v_{i+1}^n\right)} - e^{\left(g(u_i^{n+1})-v_i^n\right)} < 0. \tag{8.23}$$

Proposition 8.2 (Fully Scharfetter-Gummel scheme)**.** *We assume either* (8.8) *and* (8.11)*, or* (8.9) *and give the* $u_i^0 \geqslant 0$*. Then, the scheme* (8.21)–(8.23) *has the following properties:*
(i) *the solution* u_i^n *exists and is unique, for all* $i = 1, \dots, I$*, and* $n \geqslant 1$*;*
(ii) *it satisfies* $u_i^n \geqslant 0$*, and* $u_i^n \leqslant M$ *for the case* (8.9)*, if it is initially true;*
(iii) *the steady states* $g(u_i) - v_i = \mu$ *are preserved;*
(iii) *the energy dissipation inequality is satisfied*

$$\mathcal{E}^{n+1} - \mathcal{E}^n \leqslant -\frac{\Delta t}{\Delta x}\sum_{i=1}^{I-1}\left(e^{v^n - g(u^n)}\varphi\left(u^n\right)\right)_{i+1/2} \left[e^{g(u_{i+1}^{n+1})-v_{i+1}^n} - e^{g(u_i^{n+1})-v_i^n}\right]\left[\left(g\left(u_{i+1}^{n+1}\right) - v_{i+1}^n\right) - \left(g\left(u_i^{n+1}\right) - v_i^n\right)\right] \leqslant 0.$$

Proof. We argue exactly as for the gradient flow approach. □

8.5.3 The upwinding approach

The upwind scheme is driven by simplicity and, in (8.18), the fluxes are defined by

$$F_{i+1/2}^{n+1} = -\frac{1}{\Delta x}\left[u_{i+1}^{n+1} - u_i^{n+1} - \varphi(u)_{i+1/2}^n\left(v_{i+1}^n - v_i^n\right)\right], \qquad i = 1, ..., I-2,$$

with

$$\varphi(u)_{i+1/2}^{n+1} := \begin{cases} u_i^{n+1}\psi(u_{i+1}^{n+1}) & \text{when} \quad v_{i+1}^n - v_i^n \geqslant 0, \\ u_{i+1}^{n+1}\psi(u_i^{n+1}) & \text{when} \quad v_{i+1}^n - v_i^n < 0, \end{cases} \tag{8.24}$$

as in (8.20), but this time depending on the sign of $v_{i+1}^n - v_i^n$.

Proposition 8.3 (Fully discrete upwind scheme)**.** *We assume either* (8.8) *and* (8.11)*, or* (8.9) *and give the* $u_i^0 \geqslant 0$*. Then, the scheme* (8.19)–(8.20) *has the following properties:*

(i) *the solution* u_i^n *exists and is unique, for all* $i = 0, \dots, I$*, and* $n \geqslant 1$*;*

(ii) *it satisfies* $u_i^n \geqslant 0$*, and* $u_i^n \leqslant M$ *for the case* (8.9)*, if it is initially true.*

Proof. As for the gradient flow approach, the above choice makes the scheme monotone, because

$$\partial_1 F_{i+\frac{1}{2}}(u_i^{n+1}, u_{i+1}^{n+1}) = -\left(-1 - u_{i+1}^{n+1}\psi'(u_i^{n+1})\left[v_{i+1}^n - v_i^n\right]_- - \psi(u_{i+1}^{n+1})\left[v_{i+1}^n - v_i^n\right]_+\right) \geqslant 0.$$

Thus, arguing as for the gradient flow approach and relying on the results in Appendix 8.8, existence and uniqueness of the discrete solution as well as preservation of the initial bounds follow immediately. □

Thus, choice (8.24) enables to prove that the scheme is well-defined, satisfies $u_i^n \geqslant 0$ and preserves the bound $u_i^n \leqslant M$ for the case (8.9), but the energy dissipation inequality is lost. Also the steady states, in this case, are defined by the nonlinear relation $u_{i+1} - u_i = \varphi(u)_{i+1/2}(v_{i+1} - v_i)$ which are usually not in the form (8.17).

8.6 Numerical simulations

8.6.1 The Fokker-Planck equation, $\varphi(u) = u$

We first present the numerical implementation of the Fokker-Planck equation with $\varphi(u) = u$. We do not consider the gradient flow approach which requires to solve a nonlinear fixed point unlike the other methods. The Scharfetter-Gummel and upwind approaches are linear and can be compared, in fact they only require the solution of a tri-diagonal system at each time step.

We consider a first case with $\chi/D = 24$, with $I = 100$ and an initial density $u^0 = 1$. We take the velocity field as

$$v = x(1-x)|x - 0.3|.$$

In 8.1, we compare the approximate stationary solutions obtained with the upwind scheme (blue, dashed line) and the Scharfetter-Gummel scheme (red line) with the exact stationary solution (black line), which in this case has the form $u(x) = Ce^{\chi v(x)/D}$, with $C = \left(\int_0^1 e^{\chi v(x)/D} dx\right)^{-1}$ so that the mass of the solution is constant in time. In this first case, the two schemes have no significant differences; this is a major difference with the Keller-Segel equation, as we show it in the next subsection.

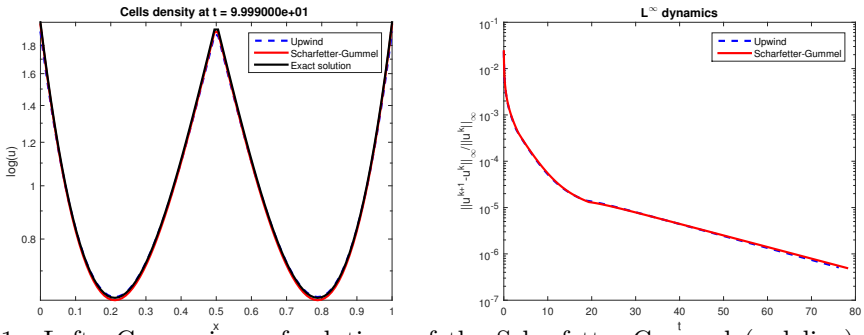


FIGURE 8.1: Left: Comparison of solutions of the Scharfetter-Gummel (red line) and upwind (blue, dashed line) schemes at time $t = 100$ with the exact stationary solution (black line) for the linear Fokker-Planck equation with $\varphi(u) = u$. Right: Dynamics of discrete solutions.

8.6.2 The nonlinear Keller-Segel equation

We turn to the equation (8.1) coupled with (8.2) for two nonlinear forms of the chemotactic sensitivity function: the logistic form $\varphi(u) = u(1-u)$ and the exponential form $\varphi(u) = ue^{-u}$. The goal is to compare the discrete solutions obtained with the three numerical approaches presented above when patterns arise, namely when Turing instabilities drive the formation of spatially inhomogeneous solutions (we refer to [120] for an introduction to this topic). To this end, we slightly modify the original equation (8.1) to

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left[D\frac{\partial u}{\partial x} - \chi\varphi(u)\frac{\partial v}{\partial x}\right] = 0, & x \in (0,1),\ t > 0, \\ D\frac{\partial u}{\partial x} - \chi\varphi(u)\frac{\partial v}{\partial x} = 0, & \text{for } x = 0 \text{ or } 1, \\ u(x,0) = u^0(x) \geqslant 0, & x \in [0,1], \end{cases} \tag{8.25}$$

in order to emphasise the coefficients driving the instabilities: $D > 0$, the constant diffusion coefficient and $\chi > 0$, the chemosensitivity. The concentration of the chemoattractant v remains driven by (8.2).

We first consider the logistic case with $\chi/D = 40$. We take as initial condition a random spatial perturbation of the constant steady state $u^0 = 0.5$ and solve the equation with 100 equidistant points in $[0,1]$.

Figure 8.2 shows the evolution in time of the density u_i^n obtained with the Scharfetter-Gummel (red line) and the gradient flow approach (black, dashed line). After a rather short-time period, the initial spatial perturbation evolves, as expected, in spatially inhomogeneous patterns: a series of "steps" arise in the regions where the concentration of the chemoattractant is greater. After some time, a structure with a smaller number of steps forms when the two central plateaus merge. It is worth noticing that, even if the transitions from one structure to another happen very quickly, the time period during which these structures remain unchanged grows with the number of transitions that occurred. In [137],

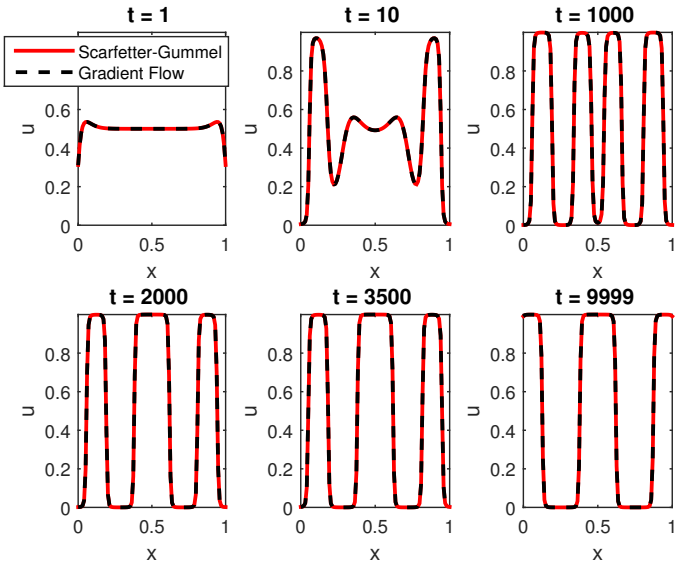


FIGURE 8.2: EVOLUTION IN TIME OF SOLUTIONS to (8.25) in the logistic case $\varphi(u) = u(1-u)$ with $\chi/D = 40$. We solved the equation with the Scharfetter-Gummel (red line) and the gradient flow scheme (black dashed line) with $I = 100$ and $\Delta t = 1$. There is no major difference between the solutions given by the two approaches.

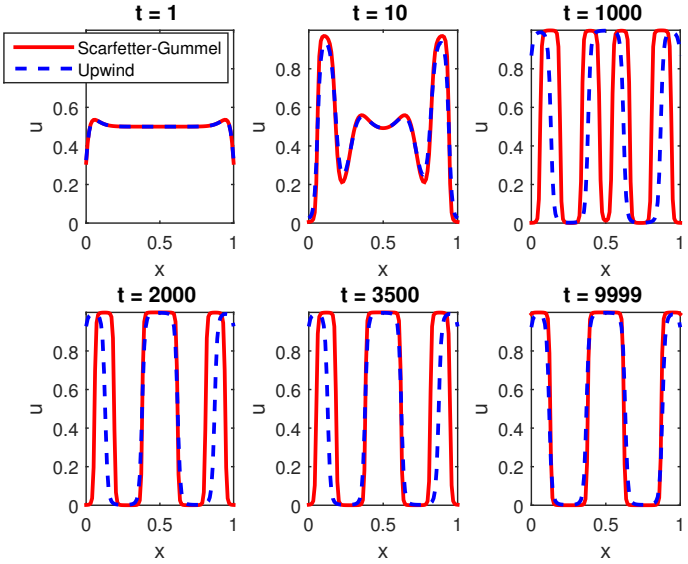
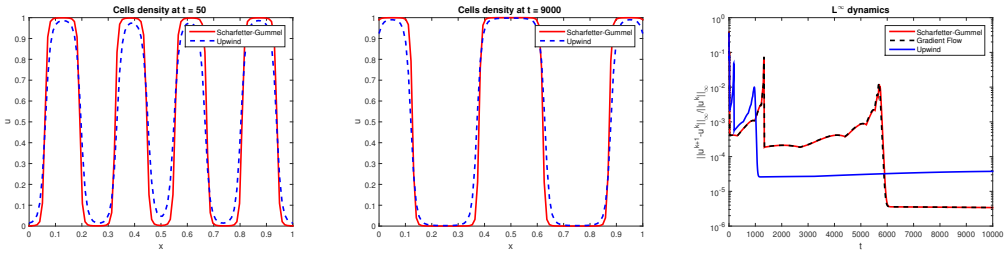


FIGURE 8.3: EVOLUTION IN TIME OF SOLUTIONS to (8.25) in the logistic case $\varphi(u) = u(1-u)$ with $\chi/D = 40$. We solved the equation with the Scharfetter-Gummel (red line) and the upwind scheme (blue, dashed line) with $I = 100$ and $\Delta t = 1$. The upwind solution transitions faster from one metastable state to the following, while the Scharfetter-Gummel scheme preserves discrete stationary profiles.

where the inter-transitions patterns are called *metastable*, this peculiar phenomenon is explained in details. As for the schemes, Figure 8.2 shows that the Scharfetter-Gummel



(A) (B) (C)

FIGURE 8.4: STATIONARY PROFILES AND DYNAMICS. (A), (B) Comparison of the stationary profiles of solutions to the Scharfetter-Gummel (red line) and the upwind (blue, dashed line) schemes at $t = 50$ and $t = 9000$. The Scharfetter-Gummel scheme approximates the expected 0-1 plateaus better, while the solution to the upwind scheme has smoother edges. (C) The dynamics of solutions represented by the quantity $\frac{\|u^n - u^{n-1}\|_\infty}{\|u^{n-1}\|_\infty}$. Peaks correspond to transitions. The Scharfetter-Gummel and the gradient flow solutions have smaller errors when they reach a metastable profile.

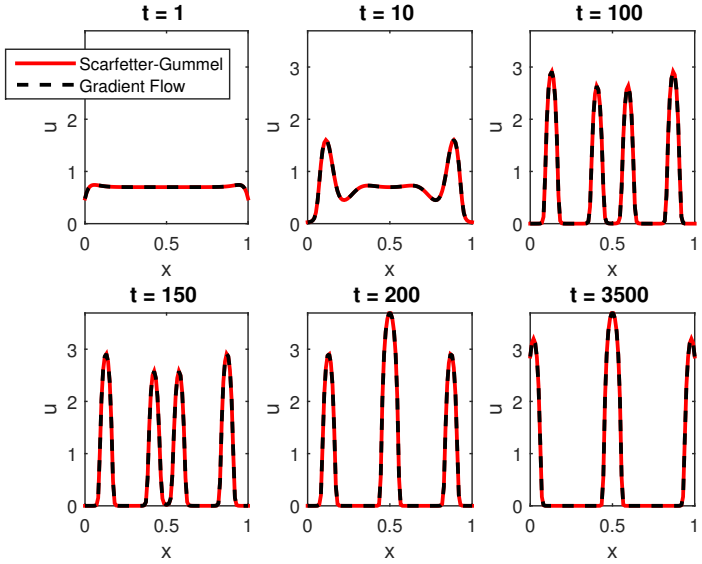


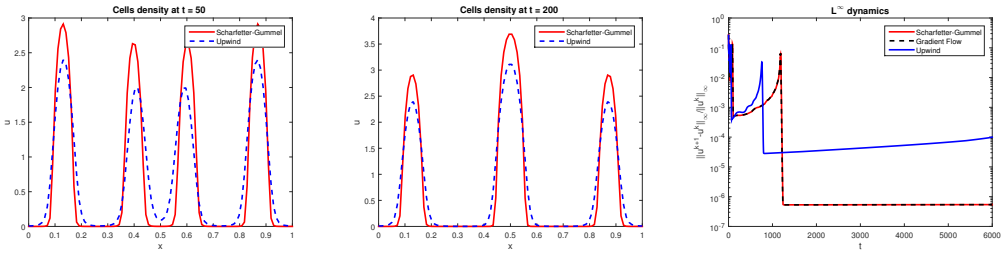
FIGURE 8.5: EVOLUTION IN TIME OF SOLUTIONS to (8.25) in the exponential case $\varphi(u) = ue^{-u}$ with $\chi/D = 24$. We solved the equation with the Scharfetter-Gummel (red line) and the gradient flow schemes (black, dashed line) with $I = 100$ and $\Delta t = 1$. As for the logistic model, the two schemes give the same solution.

and the gradient flow approaches give the same solution; no difference can be spotted. This is not true for the upwind approach. In Figure 8.3, we compare the solutions to

the Scharfetter-Gummel (red line) and the upwind (blue, dashed line) schemes. The upwind solution transitions faster from one metastable structure to the following than the Scharfetter-Gummel one. In fact, as proved above, the latter preserves discrete stationary profiles which, using the no-flux boundary conditions, solve the equation

$$\frac{\partial u}{\partial x} = \frac{\chi}{D}\varphi(u)\frac{\partial v}{\partial x}. \tag{8.26}$$

From (8.26), it is clear that, in the logistic case, the expected stationary solutions are 0-1 plateaus (or "steps") connected by a sigmoid curve which is increasing or decreasing when v is. We refer again to [137] and also to [53] for a detailed study of the stationary solutions and their properties for the logistic Keller-Segel system. In Figures 8.4a and 8.4b, we compare two stationary solutions to the Scharfetter-Gummel and upwind schemes, at time $t = 50$ and $t = 9000$ respectively. The Scharfetter-Gummel scheme approximates the 0-1 plateaus better than the upwind scheme, whose solutions have smoother edges. Moreover, in 8.4c we compare the L^∞ dynamics of the three schemes, computing the quantity $\left\|u^n - u^{n-1}\right\|_\infty / \left\|u^{n-1}\right\|_\infty$ for each n. The peaks shown by this figure correspond to the transitions from one profile to another. Observe that, for both solutions of the Scharfetter-Gummel and the gradient flow scheme, the two peaks are further away in time than the ones from the upwind scheme: this confirms that the upwind solution is in advance when it comes to transitioning. Nevertheless, from $t \approx 6000$, the relative errors of the upwind solution are consistently greater that the ones from the two other approaches, thus confirming that only the Scharfetter-Gummel and the gradient flow schemes preserve the exact discrete stationary profiles. Notice however that none of the schemes produce overshoot, due to our upwinding of the term in $\psi(u)$. Next, we consider an exponentially



(A) (B) (C)

FIGURE 8.6: STATIONARY PROFILES AND DYNAMICS. (A), (B) Comparison of the stationary profiles obtained with the Scharfetter-Gummel (red line) and the upwind scheme (blue, dashed line) at $t = 50$ (left) and $t = 200$. (C) Dynamics of solutions to the three schemes: observe that the upwind approach, even though it is faster, has greater errors when stationary profiles are reached.

decreasing form of the chemotactic sensitivity function with $\chi/D = 24$. Again, we take as initial condition a random spatial perturbation of the constant steady state $u^0 = 0.7$ and solve the equation on 100 equidistant points. The evolution in time of discrete solutions obtained with the three numerical approaches are compared in Figures 8.5 and 8.7. In

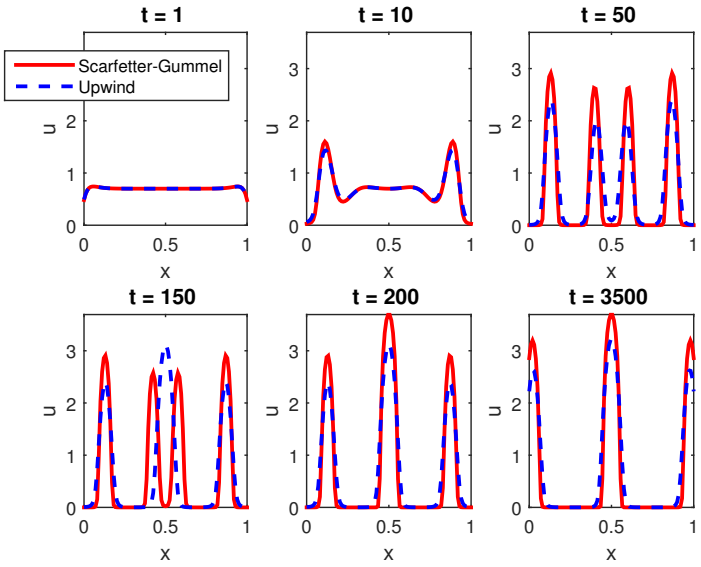


FIGURE 8.7: EVOLUTION IN TIME OF SOLUTIONS to (8.25) in the exponential case $\varphi(u) = ue^{-u}$ with $\chi/D = 24$. We compare the solutions of the Scharfetter-Gummel (red line) and the upwind schemes (blue, dashed line) obtained with $I = 100$ and $\Delta t = 1$ for different times.

this model, no initial upper bound for the solution is imposed, so that the cells aggregate "naturally" where the chemoattractant has the greatest concentration, resulting in profiles without the plateaus observed in the logistic model. However, solutions face the same kind of transitions observed before, evolving from one stationary profile to another. As before, the Scharfetter-Gummel and the gradient flow approaches give the same solutions (8.5), while the solution of the upwind scheme evolves faster since we cannot expect conservation of the stationary profiles for this approach (8.7). In Figures 8.6a and 8.6b, we compare stationary profiles obtained with the different approaches while in 8.6c we compare dynamics of the solutions. This last figure shows that, as for the logistic model, smaller errors can be expected for the Scharfetter-Gummel and gradient flow approaches when steady states are reached.

8.7 Conclusion

In the context of the Generalised Keller-Segel system, we have presented constructions of numerical schemes which extend previous works [106, 32], built on two different views of energy dissipation. Our construction unifies these two views, the gradient flow and Scharfetter-Gummel symmetrisations. Our schemes preserve desirable continuous properties: positivity, mass conservation, exact energy dissipation, discrete steady states. Being correctly tuned between implicit and explicit discretisation, they can handle large time steps without CFL condition.

The present work is motivated by experiments of breast cancer cells put in a 3D structure

mimicking the conditions they meet in vivo, namely in the extracellular matrix. After a few days, confocal Images/ of 2D sections show that cells have organised as spheroids, a phenomenon believed to be driven by chemotaxis. The spheroids can then be interpreted as Turing patterns for Keller-Segel type models and it is crucial to use appropriate schemes for them to be distinguishable from actual steady states or numerical artifacts. Comparing 2D simulations of such models with these experimental images will be the subject of future work.

In fact, it is important to remark that the schemes we presented here in 1D could easily be extended to rectangular domains, without loss of properties (8.3)–(8.6). However, it remains a perspective to treat more general geometries in a multi-dimensional setting with our approach.

8.8 Appendix C: well-posedness for monotone schemes

We recall sufficient conditions for which an implicit Euler discretisation in time can be solved, independently of the step-size. This is the case for a monotone scheme. The proof relies on the existence of sub and supersolutions, and thus also yields the preservation of positivity and other pertinent bounds as we have used in Section 8.5.

We consider the problem of finding a unique solution $\left(u_i^{n+1}\right)$ to the nonlinear equation arising in Section 8.5 which reads

$$\frac{u_i^{n+1}-u_i^n}{\Delta t}+\frac{1}{\Delta x^2}\Big[\underbrace{F(u_i^n,u_{i+1}^n,v_i^n,v_{i+1}^n,u_i^{n+1},u_{i+1}^{n+1})}_{F_{i+\frac12}^{n+1}}-F_{i-\frac12}^{n+1}\Big]=0,\ i=1,\dots,I.$$

With simplified notations, we thus consider the problem of finding a solution (u_i) (which stands for u_i^{n+1}) to

$$u_i+A_{i+\frac12}(u_i,u_{i+1})-A_{i-\frac12}(u_{i-1},u_i)=f_i,\qquad i=1,\dots,I, \tag{8.27}$$

where we assume that the f_i are given (it stands for u_i^n) and that the A_i are Lipschitz continuous and, a.e.,

$$\partial_1 A_{i+\frac12}(\cdot,\cdot)\geqslant 0,\qquad \partial_2 A_{i+\frac12}(\cdot,\cdot)\leqslant 0, \tag{8.28}$$

and there are a supersolution $(\bar U_i)_{i=1\ \dots,I}$ and a subsolution $(\underline{U}_i)_{i=1\ \dots,I}$

$$\bar U_i+A_{i+\frac12}(\bar U_i,\bar U_{i+1})-A_{i-\frac12}(\bar U_{i-1},\bar U_i)\geqslant f_i,\qquad \underline{U}_i+A_{i+\frac12}(\underline{U}_i,\underline{U}_{i+1})-A_{i-\frac12}(\underline{U}_{i-1},\underline{U}_i)\leqslant f_i.$$

We build a solution of (8.27) using an evolution equation

$$\frac{du_i(t)}{dt}+u_i(t)+A_{i+\frac12}(u_i(t),u_{i+1}(t))-A_{i-\frac12}(u_{i-1}(t),u_i(t))=f_i. \tag{8.29}$$

Theorem 8.1. *Assume (8.28) and the existence of a subsolution and of a supersolution. Then,*

(i) *For a supersolution (resp. subsolution) initial data, the dynamics (8.29) satisfies* $\frac{d\bar{u}_i(t)}{dt} \leqslant 0$ *(resp.* $\frac{d\underline{u}_i(t)}{dt} \geqslant 0$*) for all times* $t \geqslant 0$*, and thus* $\bar{u}_i(t)$ *is a supersolution (resp. subsolution) for all times.*

(ii) *A subsolution is smaller than a supersolution.*

(iii) $\bar{u}_i(t)$ *and* $\underline{u}_i(t)$ *converge to the same solution of (8.27).*

Proof. (i) We prove the statement with the supersolution. We set

$$z_i(t) = \frac{d\bar{u}_i(t)}{dt}, \qquad z_i(0) \leqslant 0.$$

Differentiating the equation, we obtain

$$\frac{dz_i(t)}{dt} + z_i(t) + [\partial_1 A_{i+\frac{1}{2}} - \partial_2 A_{i-\frac{1}{2}}]\, z_i(t) = -\partial_2 A_{i+\frac{1}{2}}\, z_{i+1}(t) + \partial_1 A_{i-\frac{1}{2}}\, z_{i-1}(t).$$

The solution cannot change sign and thus $z_i(t) \leqslant 0$ for all times.
(ii) Consider u, (v) sub (super) solutions. Set $w = u - v$ and we want to prove that $w \leqslant 0$. We write

$$\begin{aligned} w_i &+ [A_{i+\frac{1}{2}}(u_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, u_{i+1})] + [A_{i+\frac{1}{2}}(v_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, v_{i+1})] \\ &- [A_{i-\frac{1}{2}}(u_{i-1}, u_i) - A_{i-\frac{1}{2}}(v_{i-1}, u_i)] - [A_{i-\frac{1}{2}}(v_{i-1}, u_i) - A_{i-\frac{1}{2}}(v_{i-1}, v_i)] \leqslant 0. \end{aligned}$$

Multiply by $\mathrm{sgn}_+(w_i)$ and add the relations to conclude that

$$\sum_i^I (w_i)_+ + \sum_i^{I-1} J_{i+\frac{1}{2}} + \sum_i^{I-1} K_{i+\frac{1}{2}} = 0,$$

with

$$\begin{aligned} J_{i+\frac{1}{2}} &= [A_{i+\frac{1}{2}}(u_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, u_{i+1})]\, [\mathrm{sgn}_+(w_i) - \mathrm{sgn}_+(w_{i+1})], \\ K_{i+\frac{1}{2}} &= [A_{i+\frac{1}{2}}(v_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, v_{i+1})]\, [\mathrm{sgn}_+(w_i) - \mathrm{sgn}_+(w_{i+1})]. \end{aligned}$$

For each of the these terms, we show that $J_{i+\frac{1}{2}} \geqslant 0$, $K_{i+\frac{1}{2}} \geqslant 0$, as follows. Only the case when the $+$ signs in the right brackets are different has to be considered. Assume for instance that

$$u_i \geqslant v_i, \qquad \text{and} \quad u_{i+1} \leqslant v_{i+1}.$$

Then, we have by assumption (8.28),

$$\begin{aligned} [A_{i+\frac{1}{2}}(u_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, u_{i+1})] \geqslant 0 \;&\Rightarrow J_{i+\frac{1}{2}} \geqslant 0, \\ [A_{i+\frac{1}{2}}(v_i, u_{i+1}) - A_{i+\frac{1}{2}}(v_i, v_{i+1})] \geqslant 0 \;&\Rightarrow K_{i+\frac{1}{2}} \geqslant 0. \end{aligned}$$

Therefore $\sum_i (w_i)_+ \leqslant 0$ and this implies $w_i \leqslant 0$ for all i.
(iii) This is clear since the limits are solutions. □

Bibliography

[1] Agrachev, A. A., and Sachkov, Y. L. Control Theory from the Geometric Viewpoint, vol. 87 of Encyclopaedia of Mathematical Sciences. Control Theory and Optimization, II, 2004.

[2] Agur, Z., Hassin, R., and Levy, S. Optimizing chemotherapy scheduling using local search heuristics. *Operations Research 54*, 5 (Oct 2006), 829–846.

[3] Alfaro, M., and Coville, J. Rapid traveling waves in the nonlocal Fisher equation connect two unstable states. *Applied Mathematics Letters 25*, 12 (2012), 2095–2099.

[4] Armstrong, N. J., Painter, K. J., and Sherratt, J. A. A continuum approach to modelling cell–cell adhesion. *Journal of Theoretical Biology 243*, 1 (2006), 98–113.

[5] Aronson, D. G., and Weinberger, H. F. Multidimensional nonlinear diffusion arising in population genetics. *Advances in Mathematics 30*, 1 (1978), 33–76.

[6] Baigent, S. Lotka-Volterra Dynamics: an introduction. Preprint, 2010.

[7] Barles, G., Mirrahimi, S., Perthame, B., et al. Concentration in Lotka-Volterra parabolic or integral equations: a general convergence result. *Methods and Applications of Analysis 16*, 3 (2009), 321–340.

[8] Barré, J., Carrillo, J., Degond, P., Peurichard, D., and Zatorska, E. Particle interactions mediated by dynamical networks: assessment of macroscopic descriptions. *Journal of Nonlinear Science 28*, 1 (2018), 235–268.

[9] Bedard, P. L., Hansen, A. R., Ratain, M. J., and Siu, L. L. Tumour heterogeneity in the clinic. *Nature 501*, 7467 (Sep 2013), 355–364.

[10] Benzekry, S., and Hahnfeldt, P. Maximum tolerated dose versus metronomic scheduling in the treatment of metastatic cancers. *Journal of theoretical biology 335* (2013), 235–244.

[11] Berestycki, H., and Hamel, F. *Reaction-diffusion equations and propagation phenomena*. Springer, 2007.

[12] Berestycki, H., Nadin, G., Perthame, B., and Ryzhik, L. The non-local Fisher-KPP equation: travelling waves and steady states. *Nonlinearity 22*, 12 (2009), 2813.

[13] Bertozzi, A. L., Carrillo, J. A., and Laurent, T. Blow-up in multidimensional aggregation equations with mildly singular interaction kernels. *Nonlinearity 22*, 3 (2009), 683.

[14] Billy, F., Clairambault, J., and Fercoq, O. Optimisation of cancer drug treatments using cell population dynamics. In *Mathematical Models and Methods in Biomedicine*, A. Friedman, E. Kashdan, U. Ledzewicz, and H. Schättler, Eds., Lecture Notes on Mathematical Modelling in the Life Sciences. Springer, 2013, pp. 265–309.

[15] Blanchet, A., Calvez, V., and Carrillo, J. A. Convergence of the mass-transport steepest descent scheme for the subcritical patlak–keller–segel model. *SIAM Journal on Numerical Analysis 46*, 2 (2008), 691–721.

[16] Blanchet, A., Dolbeault, J., and Perthame, B. Two-dimensional keller-segel model: Optimal critical mass and qualitative properties of the solutions. *Electronic Journal of Differential Equations (EJDE)[electronic only] 2006* (2006), Paper–No.

[17] Bliman, P.-A., and Vauchelet, N. Establishing traveling wave in bistable reaction-diffusion system by feedback. *IEEE control systems letters 1*, 1 (2017), 62–67.

[18] Bonnard, B., Faubourg, L., Launay, G., and Trélat, E. Optimal control with state constraints and the space shuttle re-entry problem. *Journal of Dynamical and Control Systems 9*, 2 (Apr. 2003), 155.

[19] Bouchut, F. *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*. Birkhauser, 2004.

[20] Browder, T., Butterfield, C. E., Kräling, B. M., Shi, B., Marshall, B., O'Reilly, M. S., and Folkman, J. Antiangiogenic scheduling of chemotherapy improves efficacy against experimental drug-resistant cancer. *Cancer research 60*, 7 (2000), 1878–1886.

[21] Brutovsky, B., and Horvath, D. Structure of intratumor heterogeneity: Is cancer hedging its bets? arxiv, page 1307.0607, 2013.

[22] Bubba, F., Neves de Almeida, L., Perthame, B., and Pouchol, C. Energy and implicit discretization of the fokker-planck and keller-segel type equations. *arXiv preprint arXiv:1803.10629* (2018).

[23] Bulirsch, D. R., Nerz, D. M. E., Pesch, P.-D. D. H. J., and von Stryk, D. M. O. Combining direct and indirect methods in optimal control: Range maximization of a hang glider. In *Optimal control*. Springer, 1993, pp. 273–288.

[24] Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature 501*, 7467 (Sep 2013), 338–345.

[25] Busse, J.-E., Gwiazda, P., and Marciniak-Czochra, A. Mass concentration in a nonlocal model of clonal selection. *Journal of mathematical biology* (2016), 1–33.

[26] Caillau, J.-B., Daoud, B., and Gergaud, J. Minimum fuel control of the planar circular restricted three-body problem. *Celestial Mechanics and Dynamical Astronomy 114*, 1 (Oct 2012), 137–150.

[27] Calvez, V. *Modeles et analyses mathématiques pour les mouvements collectifs de cellules*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2007.

[28] Cañizo, J. A., Carrillo, J. A., and Cuadrado, S. Measure solutions for some models in population dynamics. *Acta Applicandae Mathematicae, Vol. 123, No. 1, pp. 141-156* (2013).

[29] Carrère, C. Optimization of an in vitro chemotherapy to avoid resistant tumours. *Journal of Theoretical Biology 413* (Jan 2017), 24–33.

[30] Carrillo, J. A., Chertock, A., and Huang, Y. A finite-volume method for nonlinear nonlocal equations with a gradient flow structure. *Communications in Computational Physics 17*, 1 (2015), 233–258.

[31] Carrillo, J. A., Kolbe, N., and Lukacova-Medvidova, M. A hybrid mass-transport finite element method for keller-segel type systems. *arXiv:1709.07394* (2017).

[32] Carrillo de la Plata, J., Huang, Y., and Schmidtchen, M. Zoology of a non-local cross-diffusion model for two species. *SIAM Journal on Applied Mathematics* (2018).

[33] Cerf, M., Haberkorn, T., and Trélat, E. Continuation from a flat to a round earth model in the coplanar orbit transfer problem. *Optimal Control Applications and Methods 33*, 6 (2012), 654–675.

[34] Champagnat, N., Ferrière, R., and Méléard, S. From individual stochastic processes to macroscopic models in adaptive evolution. *Stochastic Models 24*, S1 (2008), 2–44.

[35] Champagnat, N., Jabin, P.-E., and Raoul, G. Convergence to equilibrium in competitive Lotka-Volterra and chemostat systems. *Comptes Rendus Mathematique 348*, 23 (2010), 1267–1272.

[36] Chisholm, R. H., Lorenzi, T., and Clairambault, J. Cell population heterogeneity and evolution towards drug resistance in cancer: Biological and mathematical assessment, theoretical treatment optimisation. *Biochimica et Biophysica Acta (BBA) - General Subjects 1860*, 11 (Nov 2016), 2627–2645.

[37] Chisholm, R. H., Lorenzi, T., and Lorz, A. Effects of an advection term in nonlocal lotka–volterra equations. *Communications in Mathematical Sciences 14*, 4 (2016), 1181–1188.

[38] Chisholm, R. H., Lorenzi, T., Lorz, A., Larsen, A. K., de Almeida, L. N., Escargueil, A., and Clairambault, J. Emergence of Drug Tolerance in Cancer Cell Populations: An Evolutionary Outcome of Selection, Nongenetic Instability, and Stress-Induced Adaptation. *Cancer research 75*, 6 (2015), 930–939.

[39] Chupin, M., Haberkorn, T., and Trélat, E. Low-Thrust Lyapunov to Lyapunov and Halo to Halo with L^2-Minimization . *ESAIM: Mathematical Modelling and Numerical Analysis 51*, 3 (2017), 965–996.

[40] Cieślak, T., Morale-Rodrigo, C., et al. Quasilinear non-uniformly parabolic-elliptic system modelling chemotaxis with volume filling effect. existence and uniqueness of global-in-time solutions. *Topological Methods in Nonlinear Analysis 29*, 2 (2007), 361–381.

[41] Coron, J.-M. *Control and nonlinearity*. No. 136 in Mathematical surveys and monographs. American Mathematical Soc., 2007.

[42] Coron, J.-M., and Trélat, E. Global steady-state controllability of one-dimensional semilinear heat equations. *SIAM journal on control and optimization 43*, 2 (2004), 549–569.

[43] Costa, M., Boldrini, J., and Bassanezi, R. Optimal chemical control of populations developing drug resistance. *Mathematical Medicine and Biology 9*, 3 (1992), 215–226.

[44] Costa, M., Boldrini, J., and Bassanezi, R. Optimal chemotherapy: a case study with drug resistance, saturation effect, and toxicity. *Mathematical Medicine and Biology 11*, 1 (1994), 45–59.

[45] Coville, J. Convergence to equilibrium for positive solutions of some mutation-selection model. *Preprint arXiv:1308.6471* (2013).

[46] Coville, J., and Fabre, F. Convergence to the equilibrium in a Lotka-Volterra ODE competition system with mutations. *Preprint arXiv:1301.6237* (2013).

[47] Davis, M. Piecewise-deterministic Markov processes: a general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B 46*, 3 (1984), 353–388.

[48] Desvillettes, L., Jabin, P. E., Mischler, S., Raoul, G., et al. On selection dynamics for continuous structured populations. *Communications in Mathematical Sciences 6*, 3 (2008), 729–747.

[49] Diekmann, O., et al. A beginner's guide to adaptive dynamics. *Banach Center Publications 63* (2004), 47–86.

[50] Diekmann, O., Gyllenberg, M., and Metz, J. Steady-state analysis of structured population models. *Theoretical population biology 63*, 4 (2003), 309–338.

[51] Diekmann, O., Jabin, P.-E., Mischler, S., and Perthame, B. The dynamics of adaptation: an illuminating example and a Hamilton-Jacobi approach. *Theoretical Population Biology 67*, 4 (2005), 257–271.

[52] Ding, L., Ley, T., Larson, D., Miller, C., Koboldt, D., Welch, J., Ritchey, J., Young, M., Lamprecht, T., McLellan, M., et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole genome sequencing. *Nature 481* (Jan 2012), 506–510.

[53] Dolbeault, J., Jankowiak, G., and Markowich, P. Stationary solutions of keller-segel type crowd motion and herding models: multiplicity and dynamical stability. *Mathematics and Mechanics* (2015), 211.

[54] Dolbeault, J., and Perthame, B. Optimal critical mass in the two dimensional keller–segel model in r2. *Comptes Rendus Mathematique 339*, 9 (2004), 611–616.

[55] Emanuilov, O. Y. Controllability of parabolic equations. *Sbornik: Mathematics 186*, 6 (1995), 879–900.

[56] Evans, L. C., and Gariepy, R. F. *Measure theory and fine properties of functions*. CRC press, 2015.

[57] Fernández-Cara, E., and Zuazua, E. Null and approximate controllability for weakly blowing up semilinear heat equations. In *Annales de l'Institut Henri Poincare (C) Non Linear Analysis* (2000), vol. 17, Elsevier, pp. 583–616.

[58] Fernández-Cara, E., Zuazua, E., et al. The cost of approximate controllability for heat equations: the linear case. *Advances in Differential equations 5*, 4-6 (2000), 465–514.

[59] Ferrand, N., Gnanapragasam, A., Dorothee, G., Redeuilh, G., Larsen, A. K., and Sabbah, M. Loss of wisp2/ccn5 in estrogen-dependent mcf7 human breast cancer cells promotes a stem-like cell phenotype. *PloS one 9*, 2 (2014), e87878.

[60] Fourer, R., Gay, D. M., and Kernighan, B. W. A modeling language for mathematical programming. *Duxbury Press 36*, 5 (2002), 519–554.

[61] Gergaud, Joseph, and Haberkorn, Thomas. Homotopy method for minimum consumption orbit transfer problem. *ESAIM: COCV 12*, 2 (2006), 294–310.

[62] Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med 366*, 10 (Mar 2012), 883–892.

[63] Goh, B. Stability in models of mutualism. *American Naturalist* (1979), 261–275.

[64] Goh, B. S. Global stability in many-species systems. *American Naturalist* (1977), 135–143.

[65] Greaves, M. Cancer stem cells: back to Darwin? *Semin Cancer Biol 20*, 2 (Apr 2010), 65–70.

[66] Greaves, M. Evolutionary determinants of cancer. *Cancer Discov 5*, 8 (Aug 2015), 806–820.

[67] Greaves, M., and Maley, C. C. Clonal evolution in cancer. *Nature 481*, 7381 (Jan 2012), 306–313.

[68] Greene, J., Lavi, O., Gottesman, M. M., and Levy, D. The impact of cell density and mutations in a model of multidrug resistance in solid tumors. *Bulletin of mathematical biology 76*, 3 (2014), 627–653.

[69] Greene, J. M., Levy, D., Fung, K. L., Souza, P. S., Gottesman, M. M., and Lavi, O. Modeling intrinsic heterogeneity and growth of cancer cells. *Journal of theoretical biology 367* (2015), 262–277.

[70] Gyllenberg, M., and Meszéna, G. On the impossibility of coexistence of infinitely many strategies. *Journal of mathematical biology 50*, 2 (2005), 133–160.

[71] Hamel, F., and Ryzhik, L. On the nonlocal Fisher-KPP equation: steady states, spreading speed and global bounds. *Nonlinearity 27*, 11 (2014), 2735.

[72] Hanahan, D., Bergers, G., and E, B. Less is more, regularly: metronomic dosing of cytotoxic drugs can target tumor angiogenesis in mice. *J Clin Invest 105* (Apr 2000), 1045–7.

[73] Hecht, F. New development in freefem++. *Journal of numerical mathematics 20*, 3-4 (2012), 251–266.

[74] Hillen, T., Painter, K., and Schmeiser, C. Global existence for chemotaxis with finite sampling radius. *Discrete and Continuous Dynamical Systems Series B 7*, 1 (2007), 125.

[75] Hillen, T., and Painter, K. J. Global existence for a parabolic chemotaxis model with prevention of overcrowding. *Advances in Applied Mathematics*, 26 (2001), 280–301.

[76] Hillen, T., and Painter, K. J. Volume-filling and quorum-sensing in models for chemosensitive movement. *Canadian Applied Mathematics Quarterly*, 4 (2002), 501–543.

[77] Hillen, T., and Painter, K. J. A user's guide to pde models for chemotaxis. *Journal of mathematical biology 58*, 1-2 (2009), 183.

[78] Hofbauer, J., and Sigmund, K. Adaptive dynamics and evolutionary stability. *Applied Mathematics Letters 3*, 4 (1990), 75–79.

[79] Holcman, D., and Kupka, I. Singular perturbation for the first eigenfunction and blow-up analysis. *Forum Mathematicum 18*, 3 (2006), 445–518.

[80] Hong, Y., Fang, F., and Zhang, Q. Circulating tumor cell clusters: What we know and what we expect. *International journal of oncology 49*, 6 (2016), 2206–2216.

[81] Hortobagyi, G. N. Trastuzumab in the treatment of breast cancer. *N Engl J Med 353* (Oct 2005), 1734–1736.

[82] Jabin, P.-E., and Raoul, G. On selection dynamics for competitive interactions. *Journal of Mathematical Biology 63*, 3 (2011), 493–517.

[83] Jabin, P.-E., and Schram, R. S. Selection-Mutation dynamics with spatial dependence. *arXiv preprint arXiv:1601.04553* (2016).

[84] Jin, S., and Yan, B. A class of asymptotic-preserving schemes for the fokker–planck–landau equation. *Journal of Computational Physics 230*, 17 (2011), 6420–6437.

[85] Kanarek, A. R., and Webb, C. T. Allee effects, adaptive evolution, and invasion success. *Evolutionary Applications 3*, 2 (2010), 122–135.

[86] Kelland, L. The resurgence of platinum-based cancer chemotherapy. *Nature Reviews Cancer 7* (Aug 2007), 573–584.

[87] Keller, E. F., and Segel, L. A. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26 (1970), 399–415.

[88] Keller, E. F., and Segel, L. A. Model for chemotaxis. *Journal of theoretical biology 30*, 2 (1971), 225–234.

[89] Keller, E. F., and Segel, L. A. Traveling bands of chemotactic bacteria: a theoretical analysis. *Journal of theoretical biology 30*, 2 (1971), 235–248.

[90] Kenny, P. A., Lee, G. Y., Myers, C. A., Neve, R. M., Semeiks, J. R., Spellman, P. T., Lorenz, K., Lee, E. H., Barcellos-Hoff, M. H., Petersen, O. W., et al. The morphologies of breast cancer cell lines in three-dimensional assays correlate with their profiles of gene expression. *Molecular oncology 1*, 1 (2007), 84–96.

[91] Kimmel, M., and Świerniak, A. Control theory approach to cancer chemotherapy: Benefiting from phase dependence and overcoming drug resistance. In *Tutorials in Mathematical Biosciences III*, A. Friedman, Ed., vol. 1872 of *Lecture Notes in Mathematics*. Springer Berlin / Heidelberg, 2006, pp. 185–221.

[92] Kolmogorov, A. N., Petrovsky, I. G., and Piskounov, N. Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Bull Univ État Moscou Sér Int A 1* (1937), 1–26.

[93] Ladyzhenskaya, O. A., Solonnikov, V. A., and Uraltseva, N. N. *Linear and quasi-linear equations of parabolic type*. American Mathematical Society, 1968.

[94] Lebeau, G., and Robbiano, L. Contrôle exact de léquation de la chaleur. *Communications in Partial Differential Equations 20*, 1-2 (1995), 335–356.

[95] Ledzewicz, U., Maurer, H., and Schaettler, H. Optimal and suboptimal protocols for a mathematical model for tumor anti-angiogenesis in combination with chemotherapy. *Mathematical Biosciences and Engineering 8* (2011), 307–323.

[96] Ledzewicz, U., and Schättler, H. Optimal controls for a model with pharmacokinetics maximizing bone marrow in cancer chemotherapy. *Math Biosci 206* (2007), 320–342.

[97] Ledzewicz, U., and Schättler, H. Optimal and suboptimal protocols for a class of mathematical models of tumor anti-angiogenesis. *Journal of Theoretical Biology 252*, 2 (2008), 295–312.

[98] Ledzewicz, U., and Schättler, H. M. *Analysis of models for evolving drug resistance in cancer chemotherapy*. Watam Press, 2006.

[99] Lee, J., Cuddihy, M. J., and Kotov, N. A. Three-dimensional cell culture matrices: state of the art. *Tissue Engineering Part B: Reviews 14*, 1 (2008), 61–86.

[100] Leman, H., Méléard, S., and Mirrahimi, S. Influence of a spatial structure on the long time behavior of a competitive Lotka-Volterra type system. *Discrete and Continuous Dynamical Systems. Series B. A Journal Bridging Mathematics and Sciences 20*, 2 (2015), 469–493.

[101] LeVeque, and Randall, J. *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[102] Li, X., and Yong, J. *Optimal control theory for infinite dimensional systems*. Springer Science & Business Media, 2012.

[103] Lions, J. L. *Optimal control of systems governed by partial differential equations*. Springer, 1971.

[104] Lions, J.-L. Exact controllability, stabilization and perturbations for distributed systems. *SIAM review 30*, 1 (1988), 1–68.

[105] Lions, P.-L. On the existence of positive solutions of semilinear elliptic equations. *SIAM review 24*, 4 (1982), 441–467.

[106] Liu, J.-G., Wang, L., and Zhou, Z. Positivity-preserving and asymptotic preserving method for 2d keller-segal equations. *Mathematics of Computation 87*, 311 (2018), 1165–1189.

[107] Lohéac, J., Trélat, E., and Zuazua, E. Minimal controllability time for the heat equation under unilateral state or control constraints. *Mathematical Models and Methods in Applied Sciences 27*, 09 (2017), 1587–1644.

[108] Lohéac, J., Trélat, E., and Zuazua, E. Minimal controllability time for finite-dimensional control systems under state constraints. *preprint hal-01710759 27*, 09 (2018), 1587–1644.

[109] Lorz, A., Lorenzi, T., Clairambault, J., Escargueil, A., and Perthame, B. Effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors. *arXiv preprint arXiv:1312.6237* (2013).

[110] Lorz, A., Lorenzi, T., Clairambault, J., Escargueil, A., and Perthame, B. Modeling the effects of space structure and combination therapies on phenotypic heterogeneity and drug resistance in solid tumors. *Bulletin of mathematical biology 77*, 1 (2015), 1–22.

[111] Lorz, A., Lorenzi, T., Hochberg, M. E., Clairambault, J., and Perthame, B. Populational adaptive evolution, chemotherapeutic resistance and multiple anti-cancer therapies. *ESAIM: Mathematical Modelling and Numerical Analysis 47*, 02 (2013), 377–399.

[112] Lorz, A., Mirrahimi, S., and Perthame, B. Dirac mass dynamics in multidimensional nonlocal parabolic equations. *Communications in Partial Differential Equations 36*, 6 (2011), 1071–1098.

[113] Lotka, A. J. Elements of physical biology, reprinted 1956 as elements of mathematical biology, 1924.

[114] Madzvamuse, A., Gaffney, E. A., and Maini, P. K. Stability analysis of non-autonomous reaction-diffusion systems: the effects of growing domains. *Journal of mathematical biology 61*, 1 (2010), 133–164.

[115] Matano, H. Convergence of solutions of one-dimensional semilinear parabolic equations. *Journal of Mathematics of Kyoto University 18*, 2 (1978), 221–227.

[116] Matano, H. Nonincrease of the lap-number of a solution for a one-dimensional semilinear parabolic equation. *Journal of the Faculty of Science. University of Tokyo. Section IA. Mathematics 29*, 2 (1982), 401–441.

[117] Metz, J. A., and Diekmann, O. *The dynamics of physiologically structured populations*, vol. 68. Springer, 2014.

[118] Mirrahimi, S., and Perthame, B. Asymptotic analysis of a selection model with space. *Journal de Mathématiques Pures et Appliquées 104*, 6 (2015), 1108–1118.

[119] Müller, A., Homey, B., Soto, H., Ge, N., Catron, D., Buchanan, M. E., McClanahan, T., Murphy, E., Yuan, W., Wagner, S. N., et al. Involvement of chemokine receptors in breast cancer metastasis. *nature 410*, 6824 (2001), 50.

[120] Murray, J. D. Mathematical Biology I: An Introduction, vol. 17 of Interdisciplinary Applied Mathematics, 2002.

[121] Nadin, G., Perthame, B., and Tang, M. Can a traveling wave connect two unstable states? The case of the nonlocal Fisher equation. *Comptes Rendus Mathematique 349*, 9-10 (2011), 553–557.

[122] Nagai, T. Blow-up of radially symmetric solutions to a chemotaxis system. *Advances in Mathematical Sciences and Applications*, 5 (1995), 581–601.

[123] Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Månér, S., Zetterberg, A., Hicks, J., and Wigler, M. Inferring tumor progression from genomic heterogeneity. *Genome Res 20*, 1 (Jan 2010), 68–80.

[124] Olivier, A., and Pouchol, C. Combination of direct methods and homotopy in numerical optimal control: application to the optimization of chemotherapy in cancer. *arXiv preprint arXiv:1707.08038* (2017).

[125] Painter, K. Modelling cell migration strategies in the extracellular matrix. *Journal of mathematical biology 58*, 4-5 (2009), 511.

[126] Painter, K. J., and Hillen, T. Volume-filling and quorum-sensing in models for chemosensitive movement. *Can. Appl. Math. Quart 10*, 4 (2002), 501–543.

[127] Pasquier, E., Kavallaris, M., and André, N. Metronomic chemotherapy: new rationale for new directions. *Nature reviews Clinical oncology 7*, 8 (2010), 455–465.

[128] Patlak, C. S. Random walk with persistence and external bias. *The bulletin of mathematical biophysics 15*, 3 (1953), 311–338.

[129] Perthame, B. *Transport equations in biology*. Springer Science & Business Media, 2006.

[130] Perthame, B. Parabolic equations in biology. In *Parabolic Equations in Biology*. Springer, 2015, pp. 1–21.

[131] Pesch, H. J. A practical guide to the solution of real-life optimal control problems. *Control and cybernetics 23*, 1 (1994), 2.

[132] Pighin, D., and Zuazua, E. Controllability under positivity constraints of semilinear heat equations. *arXiv preprint arXiv:1711.07678* (2017).

[133] Pisco, A. O., Brock, A., Zhou, J., Moor, A., Mojtahedi, M., Jackson, D., and Huang, S. Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nat Commun 4* (2013), 2467.

[134] Pisco, A. O., and Huang, S. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br J Cancer 112*, 11 (May 2015), 1725–1732.

[135] Plemmons, R. J. M-matrix characterizations. I-nonsingular M-matrices. *Linear Algebra and its Applications 18*, 2 (1977), 175–188.

[136] Pontryagin, L., Boltyanskii, V., Gamkrelidze, R., and Mishchenko, E. *Mathematical theory of optimal processes*. Translated by D. E. Brown. A Pergamon Press Book. The Macmillan Co., New York, 1964.

[137] Potapov, A. B., and Hillen, T. Metastability in chemotaxis models. *Journal of Dynamics and Differential Equations*, 2 (2005), 293–330.

[138] Pouchol, C. On the stability of the state 1 in the non-local Fisher-KPP equation in bounded domains. *arXiv preprint arXiv:1801.05653* (2018).

[139] Pouchol, C., Clairambault, J., Lorz, A., and Trélat, E. Asymptotic analysis and optimal control of an integro-differential system modelling healthy and cancer cells exposed to chemotherapy. *Journal de Mathématical Pures et Appliquées* (2017).

[140] Pouchol, C., and Trélat, E. Global stability with selection in integro-differential Lotka-Volterra systems modelling trait-structured populations. *arXiv preprint arXiv:1702.06187* (2017).

[141] Preziosi, L. A review of mathematical models for the formation of vascular networks. *Journal of Theoretial Biology* (2012).

[142] Protter, M., and Weinberger, H. Maximum principles in differential equations, printce-hall. *INC., NJ* (1967).

[143] Reed, M., and Simon, B. Methods of Modern Mathematical Physics, vol II. *Academic Press* (1975).

[144] Renault, V., Thieullen, M., and Trélat, E. Optimal control of infinite-dimensional piecewise deterministic markov processes and application to the control of neuronal dynamics via optogenetics. *arXiv preprint arXiv:1607.05574* (2016).

[145] Rixe, O., and Fojo, T. Is cell death a critical end point for anticancer therapies or is cytostasis sufficient? *Clinical Cancer Research 13* (Dec 2007), 7280–7288.

[146] Russell, D. L. Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. *Siam Review 20*, 4 (1978), 639–739.

[147] Saito, N., and Suzuki, T. Notes on finite difference schemes to a parabolic-elliptic system modelling chemotaxis. *Applied mathematics and computation 171*, 1 (2005), 72–90.

[148] Scharfetter, D. L., and Gummel, H. K. Large signal analysis of a silicon read diode. *IEEE Transactions on Electron Devices*, 16 (1969), 64–77.

[149] Scharovsky, O., Mainetti, L., and Rozados, V. Metronomic chemotherapy: changing the paradigm that more is better. *Curr Oncol. 16* (Mar 2009), 7–15.

[150] Schättler, H., and Ledzewicz, U. *Geometric optimal control: theory, methods and examples*, vol. 38. Springer Science & Business Media, 2012.

[151] Schättler, H., and Ledzewicz, U. *Optimal Control for Mathematical Models of Cancer Therapies*. Springer New York, 2015.

[152] Sharma, S. V., Lee, D. Y., Li, B., Quinlan, M. P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M. A., et al. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell 141*, 1 (Apr 2010), 69–80.

[153] Shibue, T., and Weinberg, R. A. Emt, cscs, and drug resistance: the mechanistic link and clinical implications. *Nature reviews Clinical oncology 14*, 10 (2017), 611.

[154] Siegel, R., Miller, K., and Jemal, A. Cancer Statistics, 2016. *CA: A Cancer Journal for Clinicians 66*, 1 (Jan-Feb 2016), 7–30.

[155] Simon, B. Semiclassical analysis of low lying eigenvalues. i. nondegenerate minima: asymptotic expansions. *Ann. Inst. H. Poincaré Sect. A (NS) 38*, 3 (1983), 295–308.

[156] Simon, L. Asymptotics for a class of non-linear evolution equations, with applications to geometric problems. *Annals of Mathematics* (1983), 525–571.

[157] Singh, M., Mukundan, S., Jaramillo, M., Oesterreich, S., and Sant, S. Three-dimensional breast cancer models mimic hallmarks of size-induced tumor progression. *Cancer research 76*, 13 (2016), 3732–3743.

[158] Smoller, J. A., and Wasserman, A. G. Global bifurcation of steady-state solutions. *Journal of Differential Equations* (1981).

[159] Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., and Curtis, C. A big bang model of human colorectal tumor growth. *Nature Genetics 47*, 3 (Feb 2015), 209–216.

[160] Strugarek, M., and Vauchelet, N. Reduction to a single closed equation for 2-by-2 reaction-diffusion systems of lotka–volterra type. *SIAM Journal on Applied Mathematics 76*, 5 (2016), 2060–2080.

[161] Strugarek, M., Vauchelet, N., and Zubelli, J. Quantifying the survival uncertainty of wolbachia-infected mosquitoes in a spatial model. *arXiv preprint arXiv:1608.06792* (2016).

[162] Swan, G., and Vincent, T. Optimal control analysis in the chemotherapy of igg multiple myeloma. *Bulletin of Mathematical Biology 39* (1977), 317–337.

[163] Swan, G. W. *Applications of optimal control theory in biomedicine*. Marcel Dekker, 1984.

[164] Swan, G. W. Role of optimal control theory in cancer chemotherapy. *Marthematical Biosciences 101* (1990), 237–284.

[165] Tao, Y., and Winkler, M. Boundedness in a quasilinear parabolic–parabolic keller–segel system with subcritical sensitivity. *Journal of Differential Equations 252*, 1 (2012), 692–715.

[166] Trélat, E. *Contrôle optimal: théorie & applications*. Vuibert, 2008.

[167] Trélat, E. Optimal control and applications to aerospace: some results and challenges. *Journal of Optimization Theory and Applications 154*, 3 (2012), 713–758.

[168] Turing, A. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society*, 237 (1952), 37–72.

[169] Vance, R. Predation and Resource Partitioning in One Predator – Two Prey Model Communities. *American Naturalist* (1978), 797–813.

[170] Vinter, R. B. *Optimal control*. Systems and control : foundations & applications. Birkhäuser, Boston [u.a.], 2000.

[171] Volterra, V., and Brelot, M. *Leçons sur la théorie mathématique de la lutte pour la vie*, vol. 1. Gauthier-Villars Paris, 1931.

[172] von Stryk, O., and Bulirsch, R. Direct and indirect methods for trajectory optimization. *Annals of Operations Research 37*, 1 (Dec 1992), 357–373.

[173] Wächter, A., and Biegler, L. T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming 106*, 1 (2006), 25–57.

[174] Wang, G., and Wang, L. State-constrained optimal control in hilbert space. *Numerical Functional Analysis and Optimization 22*, 1-2 (2001), 255–276.

[175] Wang, Y., Kim, M. H., Tabaei, S. R., Park, J. H., Na, K., Chung, S., Zhdanov, V. P., and Cho, N.-J. Spheroid formation of hepatocarcinoma cells in microwells: experiments and monte carlo simulations. *PloS one 11*, 8 (2016), e0161915.

[176] Zuazua, E. Controllability and observability of partial differential equations: some results and open problems. In *Handbook of differential equations: evolutionary equations*, vol. 3. Elsevier, 2007, pp. 527–621.