ISBN 978-82-326-5008-8 (printed ver.)
ISBN 978-82-326-5009-5 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

Doctoral thesis

Sondre Tesdal Galtung

Doctoral theses at NTNU, 2020:331

NTNU
Norwegian University of
Science and Technology
Thesis for the degree of
Philosophiae Doctor
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU

Doctoral theses at NTNU, 2020:331

Sondre Tesdal Galtung

Discretizations of Wave Equations and Applications of Variational Principles

NTNU
Norwegian University of
Science and Technology

Sondre Tesdal Galtung

Discretizations of Wave Equations and Applications of Variational Principles

Thesis for the degree of Philosophiae Doctor

Trondheim, October 2020

Norwegian University of Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-5008-8 (printed ver.)
ISBN 978-82-326-5009-5 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2020:331

PRINTED IN
NORWAY
NO - 1598

NORDIC SWAN ECOLABEL
Printed matter
2041 0731

Printed by Skipnes Kommunikasjon AS

Preface

I have submitted this thesis as part of the fulfillment for the degree of Philosophiae Doctor (PhD) at the Norwegian University of Science and Technology – NTNU. My position has been financed by the Department of Mathematical Sciences, and I gratefully acknowledge the support and excellent working conditions provided by the department. I have enjoyed these four years as a PhD student, and I am left with many fond memories.

First of all, I would like to thank my main advisor Helge Holden for his excellent guidance. Despite having a busy schedule, he has always found time to answer my questions. Moreover, he has encouraged me to, and supported me in, applying for and attending conferences to present my work. This has provided me with much valuable experience, and my participation in the *Heidelberg Laureate Forum* in 2017 is a memorable highlight.

I am also grateful to my first co-advisor Katrin Grunert for her expert supervision. She has always followed up my questions thoroughly, and her meticulous readings of my texts have been very helpful. Then, I would also like to thank my second co-advisor Xavier Raynaud for excellent collaboration, his keen intuition and contagious enthusiasm for our work have been much appreciated.

Furthermore, I would like to thank Alberto Bressan for inviting me to the Pennsylvania State University for a research stay in the academic year 2018–2019. His careful supervision and intuitive explanations have been truly inspirational. I am also grateful for the hospitality shown by the faculty at the Department of Mathematics, in particular Alberto Bressan and Wen Shen, which made my stay very enjoyable.

On the same note, I would like to thank the U.S.-Norway Fulbright Foundation for their outstanding job in facilitating my time in the U.S., and always being responsive and helpful. In addition, I am thankful to the Fulbright program for the opportunity to participate in their well-organized and engaging events.

I also thank my parents Arild Sigve and Borghild, and my brothers

Anders, Kristoffer, and Ivar, for their support.

Finally, I would like to thank all my friends for keeping me company through these years. You have been essential in upholding a sound work–life balance, be it through coffee breaks, pub quizzes, board game nights, cabin trips, or other escapades.

Sondre Tesdal Galtung

Sondre Tesdal Galtung

Trondheim, 06.07.2020

Contents

Part I

Introduction

Chapter 0

Introduction

Mathematical modeling plays an important role in our attempt to understand the world around us. An ubiquitous and challenging objective for such models is to make them complicated enough to account for the properties we want to model, but simple enough that we are able to study them with the tools we have available. As such, one could say that this is yet another instance of the famous Occam's razor. However, the simplifications and assumptions do not necessarily end when one finally has arrived at some equation. On the contrary, most equations of practical interest cannot be explicitly solved, and one has to resort to further simplifications in order to obtain sufficiently good approximations of the solutions. These simplifications are then of a purely mathematical nature, rather than the physical considerations made in the derivation of the model.

Approximate solutions of the modeling equation are typically obtained through some form of discretization, and there are several possibilities here, but the ultimate goal is typically to end up with a finite-dimensional system which can be solved numerically by a computer. One alternative is to solve the equation exactly, but in a finite-dimensional subspace of the usually infinite-dimensional space of functions, and this is for instance the basic idea of the finite element methods. A different strategy is to derive a discrete version of the equation, which then has finite-dimensional solutions, as is the case for finite difference discretizations. Irrespective of the method employed, one usually employs some form of compactness argument to deduce that the approximate solutions yield a solution of the original equation as the refinement of the discretization is increased. Loosely speaking, this means that the set of discrete solutions is dense enough in the space of solutions, that for any solution of the original equation one can find a discrete solution arbitrarily close to it.

This thesis concerns mathematical models for two quite different phenomena arising in nature: The first part, Papers 1–3, concerns discretizatons of equations which have been derived as models for water waves. In particular, Paper 1 is a study of convergence rates for a finite element method applied to the Benjamin–Ono equation, which was derived in [1, 36] as a model for internal waves in stratified fluids. Papers 2 and 3 concern a finite difference-type discretization for a Camassa–Holm system. This equation has been derived as a model for shallow water waves, and is described in more detail in the next section. The second part of the thesis, Papers 4 and 5, are on simplified models for biological shape growth. Here we do not employ any discretizations to solve the highly nonlinear model equations, but we rather study existence and uniqueness properties of their solutions.

0.1 Selected background theory

In this section we shall present a selection of theory and results related to the five papers which constitute the thesis. This is by no means an attempt at providing a thorough theoretical background for the papers. Instead, the aim is to give a short introduction of properties and results which have been of importance in the writing of this thesis. For more detailed background theory, we refer to the introductions of the papers and the references therein.

Some properties of the Camassa–Holm equation

The Camassa–Holm (CH) equation,

$$u_t - u_{txx} + 2\kappa u_x + 3uu_x - 2u_x u_{xx} - uu_{xxx} = 0, \qquad (0.1.1)$$

for a time- and position-dependent velocity $u = u(t,x)$, is named after the authors of [9], who derived the equation as a shallow water limit of the Green–Naghdi equations from hydrodynamics. The CH equation is first known to have appeared in a work of Fuchssteiner and Fokas [20] as a somewhat anonymous particular case in a parameterized family of completely integrable evolution equations, and in a different form from (0.1.1). To be specific, it comes from combining equations (26e), (30a), and (30b) with parameters $\alpha = \beta = 0$, $\gamma = -1$, and $n = 1$ which yields the following equation for $u = u(t,x)$,

$$u_t = -\left(\partial_x u \partial_x^{-1} + u\right)\left(1 - \partial_x^2\right)^{-1} u_x. \qquad (0.1.2)$$

Here ∂_x is the usual partial derivative with respect to x, and ∂_x^{-1} is interpreted as the antiderivative giving functions vanishing asymptotically as $x \to -\infty$, thus

$$\partial_x^{-1} v(x) = \int_{-\infty}^{x} v(s)\, ds.$$

In particular, we will make use of the identity $\partial_x^{-1} u_x = u$. Introducing the change of variables $u = v - v_{xx} + \kappa$ in (0.1.2) and rearranging, we recover exactly (0.1.1) with v replacing u. For more details on how the techniques from [20] connect to (0.1.1), the interested reader is referred to [19].

The CH equation is completely integrable for any $\kappa \in \mathbb{R}$, see [9, 10], as it has a corresponding Lax pair for which the compatibility condition yields exactly (2.1.1). However, the case which has drawn most attention is the so-called dispersionless limit for which $\kappa = 0$. The main reason for this is most likely that the corresponding solitons then have a particularly simple form. Indeed, these are the famous peakons

$$u(t,x) = ce^{-|x-ct|}, \qquad c \in \mathbb{R}, \tag{0.1.3}$$

the name of which originates from the discontinuity in the spatial derivative for any fixed time t, yielding a peaked crest for the wave profile. From (0.1.3) we observe that the peakon travels at constant velocity c corresponding exactly to its elevation at the peak. It is clear from this discontinuity that peakons are solutions of (0.1.1) for $\kappa = 0$ only in the weak sense. In fact, traveling waves cannot be strong solutions of the CH equation, cf. [12, Ex. 5.2]. From here on we will consider only the case $\kappa = 0$ and the resulting equation

$$u_t - u_{txx} + 3uu_x - 2u_x u_{xx} - uu_{xxx} = 0, \tag{0.1.4}$$

with weak solutions u such that $u(t,\cdot)$ belongs to the Sobolev space $\mathbf{H}^1(\mathbb{R})$. As a consequence of being completely integrable, (0.1.4) has an infinite number of conserved quantities, see [34], one of which is the so-called energy

$$\frac{1}{2}\int_{\mathbb{R}} (u^2(t,x) + u_x^2(t,x))\, dx. \tag{0.1.5}$$

We also mention that the CH equation can be seen as a geodesic equation, cf. [33, 15].

Being a nonlinear, integrable evolution equation derived in the context of water waves, the CH equation is often grouped together with the Korteweg–de Vries (KdV) and, e.g., the Benjamin–Ono equations as a KdV-type equation. There is however a particular feature of the

CH equation which is not present in the prototypical KdV equation, which has sparked much research interest, namely the so-called wave breaking. That is, initially smooth solutions may develop singularities in finite time in the sense that the slope of the wave profile in a point becomes unbounded from below. For this phenomenon where the wave profile remains bounded while the slope turns vertical we say that the wave "breaks", or alternatively, alluding to u_x turning unbounded from below, we say the solution exhibits blow-up. Simultaneously, there is a concentration of energy, cf. (0.1.5), at the location where u_x becomes unbounded. This property was already pointed out in the original paper [9] of Camassa and Holm, and was verified in detail by Constantin and Escher [12, 13] where the former paper also contains a global existence result for (0.1.4). Since such wave breaking of is readily observed in nature, think for instance of the behavior of waves approaching a beach, one may argue that a faithful mathematical model for shallow water waves should incorporate such effects. We refer to [32] for a thorough discussion on the validity of the CH and KdV equations as models for water waves.

The singularity formation in finite time introduces an ambiguity in how to extend solutions beyond the time of blow-up, and this has resulted in considerable research interest. In the end, this has led to the dichotomy between conservative and dissipative weak solutions of (0.1.4), two solution concepts which differ in how the associated energy is treated as the wave breaks. Indeed, to illustrate the idea, let t_{c} be the first time when the solution blows up, and x_c be an associated position where u_x becomes unbounded. For $t < t_{\mathrm{c}}$, the two solution concepts remain equal, as $u(t, \cdot) \in \mathbf{H}^1(\mathbb{R})$ and the energy (0.1.5) is well defined. However, at $t = t_{\mathrm{c}}$ there is a concentration of energy which amounts to u_x^2 turning into a singular measure in $x = x_{\mathrm{c}}$. The conservative solution of the CH equation is then characterized by the energy being conserved for almost every time: that is, for $t > t_{\mathrm{c}}$, all energy, including the part concentrating at $x = x_{\mathrm{c}}$, is redistributed to $u(t, \cdot) \in \mathbf{H}^1$, and thus the value of (0.1.5) remains equal to what it was for $t < t_{\mathrm{c}}$. Conservative solutions of the CH equation have been studied in [3] and [29]. On the other hand, for dissipative solutions, some, if not all, of the energy concentrating at $x = x_c$ is dissipated, or removed, from the equation. Thus, for dissipative solutions the value of (0.1.5) for $t > t_{\mathrm{c}}$ is strictly smaller than for $t < t_{\mathrm{c}}$, and these solutions have been investigated in [4] and [31].

There have also been proposed several two-component extensions of the CH equation, one of which is the two-component Camassa–Holm

(2CH) system

$$
\begin{aligned}
u_t - u_{txx} + 3uu_x - 2u_x u_{xx} - uu_{xxx} + \rho\rho_x &= 0, \\
\rho_t + (\rho u)_x &= 0
\end{aligned}
\tag{0.1.6}
$$

derived by Olver and Rosenau [35, Eq. (43)]. This system has also been derived as a model for shallow water in [14]. One can think of (0.1.6) as (0.1.4) having been augmented with a term accounting for the effect of a fluid density variable $\rho = \rho(t,x)$, as well as a conservation law for this density. Assuming the density ρ has the asymptotic value $\rho_\infty \geq 0$ such that $\rho - \rho_\infty \in \mathbf{L}^2(\mathbb{R})$, the associated energy for (0.1.6) becomes

$$
\frac{1}{2}\int_{\mathbb{R}} \left[u^2(t,x) + u_x^2(t,x) + (\rho(t,x) - \rho_\infty)^2\right] dx. \tag{0.1.7}
$$

As for the CH equation, (0.1.6) can also be seen as a geodesic equation, see [17]. The 2CH system shares several properties with the CH equation, such as being completely integrable and allowing for wave breaking. We refer to [14, 23] for details on initial data for which one can or cannot have blow-up for (0.1.6). Since the 2CH system also exhibits wave breaking, it is then perhaps not surprising that it features conservative and dissipative solutions as well, and these have been studied in, e.g., [25, 26].

Concepts from the calculus of variations and control theory

Here we will briefly present some concepts from the calculus of variations and control theory which have been used in the papers of this thesis, but then typically in a more general form.

The direct method

The *direct method* is a procedure for proving the existence of an optimal solution for an optimization problem, see [7, Chap. 5]. To fix the ideas, let us formally consider the problem of finding x in a set of admissible solutions which minimizes the goal function $\phi(x)$, possibly under some additional constraints on x. Then the direct method can be summarized in four sequential steps:

1. Construct a minimizing sequence x_n, $n \in \mathbb{N}$.

2. Show that some subsequence converges to an x^*.

3. Prove that x^* is an admissible solution which satisfies the constraints.

4. Prove that x^* attains the minimum of $\phi(x)$.

If all these steps can be performed, one has proved the existence of an optimal solution x^*. In order to carry out the direct method, one must establish some continuity properties for the goal function ϕ. For instance, it would be desirable for $\phi(x)$ to be continuous in x, but this can be relaxed to lower semicontinuity for minimization problems, or upper semicontinuity for maximization problems.

The first variation in the calculus of variations

In the calculus of variations, the possibly simplest prototypical example is to minimize an expression of the form

$$J(x) = \int_{t_0}^{t_1} L(t, x(t), \dot{x}(t))\, dt, \tag{0.1.8}$$

where $x : [t_0, t_1] \to \mathbb{R}$ is a continuously differentiable curve, $\dot{x}$ denotes the derivative dx/dt, and L is a real-valued function of t, x, and $\dot{x}$. Let us also for simplicity impose fixed endpoints $x(t_0) = x_0$ and $x(t_1) = x_1$. To establish a first necessary condition for optimality, we assume x to be a minimizer and consider perturbations of the form $x_\varepsilon = x + \varepsilon y$, where $\varepsilon > 0$ and y is a continuously differentiable function for $t \in [t_0, t_1]$ which satisfies $y(t_0) = y(t_1) = 0$. These endpoint conditions for y are needed for the perturbation x_ε to satisfy the same endpoint conditions as x, i.e., for x_ε to be an admissible curve. The *first variation* of (0.1.8) is then given by

$$\delta J(x; y) = \frac{d\varepsilon}{d} J(x + \varepsilon y)\Big|_{\varepsilon=0}, \tag{0.1.9}$$

and we say that a first necessary condition for x to be optimal, is that $\delta J(x; y) = 0$ for any admissible y. Assuming L sufficiently smooth and denoting its partial derivatives with respect to x and $\dot{x}$ by respectively L_x and $L_{\dot{x}}$, we may integrate by parts to obtain the following expression for the first variation,

$$\begin{aligned}\delta J(x; y) &= \int_{t_0}^{t_1} \left[L_x(t, x(t), \dot{x}(t)) y(t) + L_{\dot{x}}(t, x(t), \dot{x}(t)) \dot{y}(t)\right] dt \\ &= \int_{t_0}^{t_1} \left[L_x(t, x(t), \dot{x}(t)) - \frac{dt}{d} L_{\dot{x}}(t, x(t), \dot{x}(t))\right] y(t)\, dt.\end{aligned}$$

Since the final expression is supposed to be zero for any admissible y, we claim that x must satisfy the identity

$$L_x(t, x(t), \dot{x}(t)) = \frac{dt}{d} L_{\dot{x}}(t, x(t), \dot{x}(t)) \tag{0.1.10}$$

for $t \in [t_0, t_1]$. Equation (0.1.10) is called the Euler–Lagrange equation, and an admissible x satisfying it is called an extremal.

The above presentation is based on [18, Chap. 1], and more details are found there. Note that the optimization problem considered here is rather simple, and that it could be made more intricate by imposing different endpoint conditions for x, or even letting the endpoints t_0 and t_1 be variable by including them as part of the solution. For more variants of such problems, see e.g., [37].

The Pontryagin maximum principle

Here we consider the optimization problem known as the Mayer problem with terminal constraints, as presented in [7, Chap. 6.3]. This can be stated as

$$\max_{u \in \mathcal{U}} \phi_0(x(T, u)) \tag{0.1.11}$$

subject to

$$\dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = \bar{x}, \quad u(t) \in \mathbf{U}, \quad t \in [0, T], \tag{0.1.12}$$

for the family of admissible controls

$$\mathcal{U} = \{u \colon [0, T] \to \mathbf{U}, \;\; u \text{ measurable}\}, \tag{0.1.13}$$

with $\mathbf{U} \subseteq \mathbb{R}^m$. In addition the terminal time T is fixed, and the terminal point $x(T)$ satisfies the constraints

$$x(T) \in S = \{x \in \mathbb{R}^n \;\; : \;\; \phi_i(x) = 0, \;\; i = 1, \dots, k\} \tag{0.1.14}$$

for some $k \in \mathbb{N}$. The Pontryagin maximum principle provides necessary conditions for an optimal solution of (0.1.11) given the control system (0.1.12) and the terminal constraints (0.1.14). We shall state its result as presented in [7, Thm. 6.3.1] below, under the following assumptions:

- The set $\Omega \subseteq \mathbb{R} \times \mathbb{R}^n$ is open.
- The function $f = f(t, x, u)$ is continuous on $\Omega \times \mathbf{U}$ and continuously differentiable w.r.t. x.

- The functions $\phi_i \colon \mathbb{R}^n \to \mathbb{R}$ for $i = 0, \dots, k$ are continuously differentiable.

Theorem 0.1.1 (The Pontryagin maximum principle with terminal constraints)**.** *Let u^* be a bounded admissible control, whose corresponding trajectory $x^*(\cdot)$ is optimal for the maximization problem* (0.1.11)–(0.1.14)*. Assume that the gradients $\nabla\phi_i$ for $i = 0, \dots, k$ are linearly independent at the terminal point $x^*(T)$. Then there exists a nontrivial, absolutely continuous vector function $p(\cdot)$ which satisfies the equations*

$$\dot{p}(t) = -p(t) \cdot D_x f(t, x^*(t), u^*(t)), \tag{0.1.15}$$

$$p(t) \cdot f(t, x^*(t), u^*(t)) = \max_{\omega \in \mathbf{U}} \{p(t) \cdot f(t, x^*(t), \omega)\} \tag{0.1.16}$$

at almost every time $t \in [0, T]$, together with the terminal conditions

$$p(T) = \sum_{i=0}^{k} \lambda_i \nabla\phi_i(x^*(T)) \tag{0.1.17}$$

for some constants $\lambda_0, \dots, \lambda_k$, with $\lambda_0 \geq 0$.

Note that in Theorem 0.1.1 x, f and u are column vectors, p is a row vector, and $D_x f$ denotes the Jacobian of f w.r.t. x. Moreover, we mention that the Pontryagin principle for a more general form of the Mayer problem is presented in [18, Chap. 2] in the setting of a minimization problem.

0.2 A variational discretization of the Camassa–Holm equation

Considering the numerous works on the CH equation, it is no surprise that several discretizations and numerical methods have been proposed for (0.1.4), and we refer to the introduction of Paper 3 for an outline of such numerical methods. An interesting discretization of the CH equation was already proposed in [9] and studied in more detail in [10], namely the multipeakon solution

$$u(t, x) = \sum_{i=1}^{n} p_i(t) e^{-|x - q_i(t)|}. \tag{0.2.1}$$

Here q_i and p_i satisfy the canonical Hamiltonian equations

$$\begin{aligned} \dot{q}_i &= \sum_{j=1}^{n} p_j e^{-|q_i - q_j|}, \\ \dot{p}_i &= p_i \sum_{j=1}^{n} \operatorname{sgn}(q_i - q_j) p_j e^{-|q_i - q_j|} \end{aligned} \tag{0.2.2}$$

for $i \in \{1, \dots, n\}$ with Hamiltonian

$$\frac{1}{2} \sum_{i,j=1}^{n} p_i p_j e^{-|q_i - q_j|}. \tag{0.2.3}$$

This is in fact a generalization of the single peakon solution (0.1.3), with an associated energy given by (0.2.3). As pointed out in [9, 10], (0.2.2) can be seen as a geodesic equation for a particle labeled i with position q_i and momentum p_i. We also mention that for conservative solutions of (0.2.2), i.e., those which preserve (0.2.3) for almost every time, this system has been proved to be completely integrable in [16].

As suggested by the existence of conservative solutions, wave breaking can also happen for the multipeakon solution (0.2.1). In particular, this happens when two particles with momenta of opposite signs collide. In these cases, the momenta diverge to plus and minus infinity, as studied in [38]. An alternative method for characterizing the conservative multipeakon solutions was established in [28]. This is based on the observation that (0.2.1) satisfies the boundary value problem $u - u_{xx} = 0$ between $q_i(t)$ and $q_{i+1}(t)$ for $i \in \{1, \dots, n-1\}$, with boundary values $u(t, q_i(t)) =: u_i(t)$ and $u(t, q_{i+1}(t)) =: u_{i+1}(t)$. Replacing q_i with y_i, the authors introduce an ODE system for y_i, u_i, and H_i, where the latter variable tracks the cumulative energy at position y_i. As opposed to the momentum variable in (0.2.2), these variables remain bounded even during wave breaking. By approximating initial data by multipeakons, one can obtain a numerical method for the CH equation, and for the conservative multipeakons this is done in [30]. Similar numerical methods based on (0.2.2) are considered in [11, 27].

A discretization in Lagrangian variables is also employed in Paper 2, but instead of being based on a special type of solution such as (0.2.1), the discretization is founded upon variational principles. We will here indicate how this semidiscrete system is derived, and to reduce the amount of terms we will consider the discretization of the CH equation only. Defining the discrete "labels" $\xi_i = i\Delta\xi$ for $\Delta\xi > 0$, $i \in \mathbb{Z}$, and the difference operators

$$\mathrm{D}_{\pm} f_i = \pm \frac{f_{i\pm1} - f_i}{\Delta\xi},$$

we introduce the discrete set of characteristics $y_i(t)$ satisfying the initial condition $y_i(0) = y(0, \xi_i)$, or equivalently $(y_0)_i = y_0(\xi_i)$. Then we introduce the discrete Lagrangian velocity $U_i(t)$ such that $U_i(0) = U(0, y(0, \xi_i))$, or $(U_0)_i = U_0((y_0)_i)$. Based on these quantities we introduce the discrete energy

$$\Delta\xi \sum_i \left[(U_i)^2 \mathrm{D}_+ y_i + \frac{(\mathrm{D}_+ U_i)^2}{\mathrm{D}_+ y_i} \right], \tag{0.2.4}$$

which is a discretization of (0.1.5) in Lagrangian variables

$$\frac{1}{2} \int_{\mathbb{R}} \left(U^2 y_\xi + \frac{U_\xi^2}{y_\xi} \right) d\xi. \tag{0.2.5}$$

Following the derivation of the paper, we combine the discrete characteristic equation $\dot{y}_i = U_i$ and the Euler–Lagrange equation resulting from a first variation of (0.2.4) to obtain the infinite-dimensional ODE system

$$\begin{aligned} \dot{y}_i &= U_i, \\ (\mathrm{A}[\mathrm{D}_+ y]\dot{U})_i &= -U_i \mathrm{D}_+ U_i - \frac{1}{2} \mathrm{D}_- \left(U_i^2 + \left(\frac{\mathrm{D}_+ U_i}{\mathrm{D}_+ y_i} \right)^2 \right), \end{aligned} \tag{0.2.6}$$

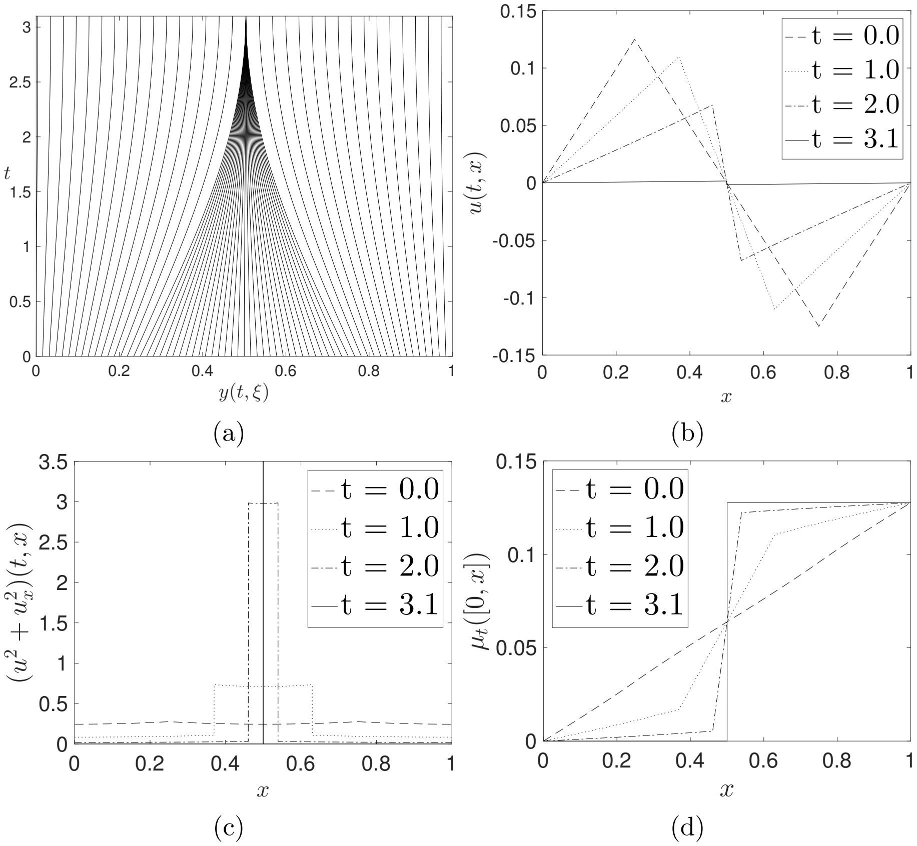
where for a grid function $v = \{v_i\}_{i \in \mathbb{Z}}$ we have defined the operator

$$(\mathrm{A}[\mathrm{D}_+ y]f)_i := (\mathrm{D}_+ y_i) v_i - \mathrm{D}_- \left(\frac{\mathrm{D}_+ v_i}{\mathrm{D}_+ y_i} \right). \tag{0.2.7}$$

Now, an alternative would be to stop at this point and call this our discrete scheme. However, unless one can guarantee $\mathrm{D}_+ y_i(t) \geq \delta$ for some constant $\delta > 0$, the division by $\mathrm{D}_+ y$ in (0.2.7) makes the analysis of existence and uniqueness of solutions for (0.2.6) difficult. This also causes trouble when applying (0.2.6) directly as a numerical method for solutions with wave breaking. Figure 0.1 displays the numerical results obtained with this method for a peakon-antipeakon example with periodic boundary conditions. Here we were able to run the scheme until collision time, around $t \approx 3.12$, when the ODE solver broke down. However, up to this time the plots clearly show the development of a delta distribution in the energy density at $x = 0.5$.

Omitting the dependence on t for the moment, we observe that the operator equation $(\mathrm{A}[\mathrm{D}_+ y]v)_i = f_i$ is a discrete version of the Sturm–Liouville equation

$$y_\xi(\xi) v(\xi) - \left(\frac{v_\xi(\xi)}{y_\xi(\xi)} \right)_\xi = f(\xi).$$



(a)

(b)

(c)

(d)

Figure 0.1: Peakon-antipeakon interaction computed with the prototype scheme (0.2.6). The 64 characteristics (a), wave profile (b), pointwise energy (c), and cumulative energy (d) before collision time at $t \approx 3.12$. We underline that the height of the profile at $t = 3.1$ in subfigure (c) is about 1.4×10^4.

The above equation can be solved for v through

$$v(\xi) = \frac{1}{2} \int_{\mathbb{R}} e^{-|y(\xi)-y(\eta)|} f(\eta)\, d\eta, \tag{0.2.8}$$

where we recognize the integration kernel as the Green's function for $\mathrm{Id} - \partial_x^2$, i.e., $\frac{1}{2} e^{-|x-x'|}$, evaluated in Lagrangian coordinates. This is exactly the sort of integral which defines the variables $P(t, \xi)$ and $Q(t, \xi)$ in [29, Eq. (2.10)], where the evolution equation for the Lagrangian velocity is explicitly given by $U_t = -Q$. Due to this similarity, it seemed natural to follow the approach of [29, 24] for the CH equation and 2CH system in the analysis of (0.2.6). However, then we would have to invert the operator $\mathrm{A}[\mathrm{D}_+y]$, and unlike the continuous case we cannot use composition of functions to obtain an explicit inverse such as (0.2.8) from the Eulerian Green's function. Instead, we prove the existence of discrete integral kernels, which correspond to $\frac{1}{2}e^{-|y(\xi)-y(\eta)|}$, by applying results from the Poincaré–Perron theory on difference equations. From these kernels we may then define an inverse for (0.2.7), which remains well-defined even as waves break, i.e., when $\mathrm{D}_+y_i = 0$. After proving existence and uniqueness of solutions for the new system, which is equivalent to (0.2.6) for $\mathrm{D}_+y_i > 0$, we construct sequences of interpolated functions which are shown to converge to solutions of the Lagrangian system considered in [29]. Hence, the variational discretization leads to conservative solutions of (0.1.4).

In paper 3 we study the corresponding variational discretization for the CH equation and 2CH system with periodic boundary conditions. Since the convergence of the discretization can be proved using the arguments of Paper 2, we choose instead to illustrate how it can work as a numerical method. To this end, we use an explicit ODE solver to integrate in time, and compare with other existing methods over several numerical examples. In particular, we introduce a periodic version of the multipeakon method considered in [28] and [30].

0.3 Variational principles and control theory applied to shape growth problems

In [8] the authors introduce and study two classes of variational problems which concern the optimal configuration of respectively tree roots and branches. The aim of the paper is to introduce mathematical models which serve as a step toward understanding the biological shapes appearing in nature. Each of these variational problems consists of a functional to be maximized, which is expressed as the difference of a gain functional and a cost functional. In the case of roots, the gain is expressed through a harvest functional which accounts for water and nutrients gathered by the tree roots. The gain for the branches comes from a sunlight functional, which measures the amount of sunlight absorbed by the leaves of the branch. For both cases, the cost functional represents a ramified transportation cost, for transporting water and nutrients from the roots to the trunk, and from the trunk to the leaves on the branches. The central idea of such ramified transport problems is that it is less costly to transport commodities, in this case nutrients, along a common path, than transporting them along separate paths. That is, the cost of transporting a commodity of size s along a path of length l is assumed to take the form $l \times s^{\alpha}$ for some $0 \leq \alpha \leq 1$. The limiting cases of $\alpha = 0$ and $\alpha = 1$ are respectively connected to the Steiner and Monge–Kantorovitch problems in transportation network theory. The mathematical framework for such ramified transport is detailed in [2].

In [8], several important properties are established for the functionals involved, which are then used to deduce existence of optimal configurations. This work is expanded upon in [6], where existence of optimal solutions are proved under less restrictive assumptions. Paper 4 ([5]) considers the variational problem for branches applied to plant stems, and two submodels are studied in detail. In the first model, the density of leaves is constant along the stem, leaving only the shape of the stem to be decided. The second model generalizes the first by including also the density of leaves as part of the configuration. Bearing in mind the general functional presented in [8], a specialized optimization problem is derived for each of the aforementioned models. The existence of maximizing solutions for both models is proved by means of the direct method presented in Section 0.1, where the semicontinuity properties of the functionals established in [8] plays a central role. Uniqueness of such solutions is then proved under some additional assumptions, by studying the necessary conditions for optimality. In the first model this is established through a more general form of the first variation presented in Section 0.1, while in the second model a more general form of the Pontryagin maximum principle in Theorem 0.1.1 provides the necessary conditions. After establishing these results for a single stem, one analyzes the existence and uniqueness for a competitive equilibrium, where the configuration of each individual stem is optimal given the configuration of all other stems.

Paper 5 concerns a shape optimization problem in two dimensions. More specifically, the aim is to find the optimal configuration for a set of branches in the plane, in order to maximize the gain functional for branches described above. The main result is that for $\frac{1}{2} \leq \alpha < 1$ in the ramified transport cost, the optimal shape is uniquely determined to be a solar panel-like shape. The same holds for $0 < \alpha < \frac{1}{2}$ under some additional restrictions on the angle of the incoming sunlight. This result is connected to Paper 4 in the sense that the second model of the paper can be used to describe the optimal distribution of leaves along the rays constituting the “solar panel”.

0.4 Summary of papers

Here I give a brief description of the papers included in the thesis, and how they came to be. For a more detailed description of the scientific content, the reader is referred to the abstracts of the papers in the subsequent sections.

Paper 1

Paper 1 is written for the proceedings of the XVI International Conference on Hyperbolic Problems in Aachen, Germany 2016, were I gave a contributed talk on my Master's thesis. The results of the thesis were later published as [22], while Paper 1 is published as [21]. In the proceedings paper, I show theoretical best case convergence rates for the finite element scheme in [22] for sufficiently regular data, and has little in common with the rest of the thesis, other than being on the discretization of a wave equation. Most of Paper 1 was written during a research stay at Institut Mittag-Leffler, Sweden during the workshop *Nonlinear Partial Differential Equations and Functional Inequalities* in the Fall of 2016.

Paper 2

Paper 2 constitutes in many ways the main work of my thesis, as it is also the paper I have spent most time on. It concerns a discretization for the 2CH system derived by my coauthor Xavier Raynaud, and I spent the first time of my PhD implementing it numerically to see whether it had potential as a numerical method for approximating solutions of the 2CH system. After promising numerical results for the periodic problem, I started the attempt at establishing convergence of the discretization on the full line. In the end, this turned out to be a quite theoretical paper on existence and uniqueness for the associated semidiscrete system, and convergence of the discretization to conservative solutions of the 2CH system.

Paper 3

Paper 3 returns to the numerical results which sparked the ideas of the second paper. After the quite theoretical work in Paper 2, it seemed appropriate to complement it with some illustrations of the variational discretization. Furthermore, the theoretical analysis led to a method which improves upon the prototype scheme which I first implemented, in that the improved method can handle wave breaking, or singularity formation, in the solution. To obtain a computationally feasible problem, my coauthor Katrin Grunert and I considered the periodic versions of the CH equation and 2CH system, and established the corresponding discretizations in this setting. As my other advisors Helge Holden and Xavier Raynaud have introduced a multipeakon method on the real line which is structurally similar to our discrete scheme, I derived a periodic version of this method to compare with. In addition, inspired by works of Camassa and collaborators, I augmented these multipeakon methods with efficient computational algorithms.

Paper 4

Paper 4 is a step in a completely different direction compared to the topics of Papers 1–3, as it concerns techniques from the calculus of variations and control theory applied to biological shape models. This paper was written during my research stay at the Pennsylvania State University in the academic year 2018–2019, under supervision of Alberto Bressan. In the course of this work, I learned a lot about the mathematical theory of control, of which I had no prior knowledge.

Paper 5

At the end of my research stay at Penn State, Professor Bressan and I started working on Paper 5. This is related to Paper 4 in that it concerns the optimal configuration of branches, given the same underlying models as in the previous paper. The methods involved are however quite different from Paper 4, as the proofs are less reliant on general results from control theory, and more tailored to the problem at hand.

0.5 Bibliography

[1] T. B. Benjamin. Internal waves of permanent form in fluids of great depth. *J. Fluid Mech.*, 29(3):559–592, 1967.

[2] M. Bernot, V. Caselles, and J.-M. Morel. *Optimal transportation networks*, volume 1955 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2009. Models and theory.

[3] A. Bressan and A. Constantin. Global conservative solutions of the Camassa-Holm equation. *Arch. Ration. Mech. Anal.*, 183(2):215–239, 2007.

[4] A. Bressan and A. Constantin. Global dissipative solutions of the Camassa-Holm equation. *Anal. Appl. (Singap.)*, 5(1):1–27, 2007.

[5] A. Bressan, S. T. Galtung, A. Reigstad, and J. Ridder. Competition models for plant stems. *J. Differential Equations*, 269(2):1571–1611, 2020.

[6] A. Bressan, M. Palladino, and Q. Sun. Variational problems for tree roots and branches. *Calc. Var. Partial Differential Equations*, 59(1):Paper No. 7, 31, 2020.

[7] A. Bressan and B. Piccoli. *Introduction to the mathematical theory of control*, volume 2 of *AIMS Series on Applied Mathematics*. American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2007.

[8] A. Bressan and Q. Sun. On the optimal shape of tree roots and branches. *Math. Models Methods Appl. Sci.*, 28(14):2763–2801, 2018.

[9] R. Camassa and D. D. Holm. An integrable shallow water equation with peaked solitons. *Phys. Rev. Lett.*, 71(11):1661–1664, 1993.

[10] R. Camassa, D. D. Holm, and J. M. Hyman. A new integrable shallow water equation. volume 31 of *Advances in Applied Mechanics*, pages 1–33. Elsevier, 1994.

[11] R. Camassa, J. Huang, and L. Lee. On a completely integrable numerical scheme for a nonlinear shallow-water wave equation. *J. Nonlinear Math. Phys.*, 12(suppl. 1):146–162, 2005.

[12] A. Constantin and J. Escher. Global existence and blow-up for a shallow water equation. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 26(2):303–328, 1998.

[13] A. Constantin and J. Escher. Wave breaking for nonlinear nonlocal shallow water equations. *Acta Math.*, 181(2):229–243, 1998.

[14] A. Constantin and R. I. Ivanov. On an integrable two-component Camassa-Holm shallow water system. *Phys. Lett. A*, 372(48):7129–7132, 2008.

[15] A. Constantin and B. Kolev. Geodesic flow on the diffeomorphism group of the circle. *Comment. Math. Helv.*, 78(4):787–804, 2003.

[16] J. Eckhardt and A. Kostenko. An isospectral problem for global conservative multi-peakon solutions of the Camassa-Holm equation. *Comm. Math. Phys.*, 329(3):893–918, 2014.

[17] J. Escher, M. Kohlmann, and J. Lenells. The geometry of the two-component Camassa-Holm and Degasperis-Procesi equations. *J. Geom. Phys.*, 61(2):436–452, 2011.

[18] W. H. Fleming and R. W. Rishel. *Deterministic and stochastic optimal control.* Springer-Verlag, Berlin-New York, 1975. Applications of Mathematics, No. 1.

[19] B. Fuchssteiner. Some tricks from the symmetry-toolbox for nonlinear equations: generalizations of the Camassa-Holm equation. *Phys. D*, 95(3-4):229–243, 1996.

[20] B. Fuchssteiner and A. S. Fokas. Symplectic structures, their Bäcklund transformations and hereditary symmetries. *Phys. D*, 4(1):47–66, 1981.

[21] S. T. Galtung. Convergence rates of a fully discrete Galerkin scheme for the Benjamin-Ono equation. In *Theory, numerics and applications of hyperbolic problems. I*, volume 236 of *Springer Proc. Math. Stat.*, pages 589–601. Springer, Cham, 2018.

[22] S. T. Galtung. A convergent Crank-Nicolson Galerkin scheme for the Benjamin-Ono equation. *Discrete Contin. Dyn. Syst.*, 38(3):1243–1268, 2018.

[23] K. Grunert. Blow-up for the two-component Camassa-Holm system. *Discrete Contin. Dyn. Syst.*, 35(5):2041–2051, 2015.

[24] K. Grunert, H. Holden, and X. Raynaud. Global conservative solutions to the Camassa-Holm equation for initial data with nonvanishing asymptotics. *Discrete Contin. Dyn. Syst.*, 32(12):4209–4227, 2012.

[25] K. Grunert, H. Holden, and X. Raynaud. Global solutions for the two-component Camassa–Holm system. *Comm. Partial Differential Equations*, 37(12):2245–2271, 2012.

[26] K. Grunert, H. Holden, and X. Raynaud. Global dissipative solutions of the two-component Camassa-Holm system for initial data with nonvanishing asymptotics. *Nonlinear Anal. Real World Appl.*, 17:203–244, 2014.

[27] H. Holden and X. Raynaud. A convergent numerical scheme for the Camassa-Holm equation based on multipeakons. *Discrete Contin. Dyn. Syst.*, 14(3):505–523, 2006.

[28] H. Holden and X. Raynaud. Global conservative multipeakon solutions of the Camassa-Holm equation. *J. Hyperbolic Differ. Equ.*, 4(1):39–64, 2007.

[29] H. Holden and X. Raynaud. Global conservative solutions of the Camassa-Holm equation—a Lagrangian point of view. *Comm. Partial Differential Equations*, 32(10-12):1511–1549, 2007.

[30] H. Holden and X. Raynaud. A numerical scheme based on multipeakons for conservative solutions of the Camassa-Holm equation. In *Hyperbolic problems: theory, numerics, applications*, pages 873–881. Springer, Berlin, 2008.

[31] H. Holden and X. Raynaud. Dissipative solutions for the Camassa-Holm equation. *Discrete Contin. Dyn. Syst.*, 24(4):1047–1112, 2009.

[32] R. S. Johnson. Camassa-Holm, Korteweg-de Vries and related models for water waves. *J. Fluid Mech.*, 455:63–82, 2002.

[33] S. Kouranbaeva. The Camassa-Holm equation as a geodesic flow on the diffeomorphism group. *J. Math. Phys.*, 40(2):857–868, 1999.

[34] J. Lenells. Conservation laws of the Camassa-Holm equation. *J. Phys. A*, 38(4):869–880, 2005.

[35] P. J. Olver and P. Rosenau. Tri-Hamiltonian duality between solitons and solitary-wave solutions having compact support. *Phys. Rev. E (3)*, 53(2):1900–1906, 1996.

[36] H. Ono. Algebraic solitary waves in stratified fluids. *J. Phys. Soc. Japan*, 39(4):1082–1091, 1975.

[37] B. van Brunt. *The calculus of variations.* Universitext. Springer-Verlag, New York, 2004.

[38] E. Wahlén. The interaction of peakons and antipeakons. *Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal.*, 13(3-4):465–472, 2006.

Part II

Research Papers

Paper 1

Convergence rates of a fully discrete Galerkin scheme for the Benjamin–Ono equation

Sondre Tesdal Galtung

Published in the proceedings of the *16th International Conference on Hyperbolic Problems: Theory, Numerics and Applications*

Convergence Rates of a Fully Discrete Galerkin Scheme for the Benjamin–Ono Equation

Sondre Tesdal Galtung

Abstract We consider a recently proposed fully discrete Galerkin scheme for the Benjamin–Ono equation which has been found to be locally convergent in finite time for initial data in $L^2(\mathbb{R})$. By assuming that the initial data is sufficiently regular, we obtain theoretical convergence rates for the scheme both in the full line and periodic versions of the associated initial value problem. These rates are illustrated with some numerical examples.

Keywords Benjamin–Ono equation · Finite element method · Convergence rates

2010 Mathematics Subject Classification 65M12 · 65M15 · 65M60 · 35Q53

1 Background

We will in the following consider the Benjamin–Ono (BO) equation [2, 7] which serves as a generic model for weakly nonlinear long waves with nonlocal dispersion. Its initial value problem reads

$$\begin{cases} u_t + uu_x - \mathrm{H}u_{xx} = 0, & (t,x) \in (0,T] \times \mathbb{R}, \\ u(0,x) = u_0(x), & x \in \mathbb{R}, \end{cases} \tag{1}$$

where H denotes the Hilbert transform defined by

$$\mathrm{H}u(\cdot,x) := \text{p.v.}\frac{1}{\pi}\int_{\mathbb{R}} \frac{u(\cdot,x-y)}{y}\mathrm{d}y,$$

S. T. Galtung (✉)
Department of Mathematical Sciences, NTNU Norwegian University of Science and Technology, NO-7491, Trondheim, Norway
e-mail: sondre.galtung@ntnu.no

C. Klingenberg and M. Westdickenberg (eds.), *Theory, Numerics and Applications of Hyperbolic Problems I*, Springer Proceedings in Mathematics & Statistics 236,
https://doi.org/10.1007/978-3-319-91545-6_45

for which p.v. denotes the Cauchy principal value. We may also consider the $2L$-periodic IVP for the BO equation

$$\begin{cases} u_t + uu_x - \mathrm{H}_{\mathrm{per}} u_{xx} = 0, & (t,x) \in (0,T] \times \mathbb{T}, \\ u(0,x) = u_0(x), & x \in \mathbb{T}, \end{cases} \tag{2}$$

where $\mathbb{T} := \mathbb{R}/2L\mathbb{Z}$, and $\mathrm{H}_{\mathrm{per}}$ denotes the $2L$-periodic Hilbert transform defined by

$$\mathrm{H}_{\mathrm{per}} u(\cdot,x) := \text{p.v.} \frac{1}{2L} \int_{-L}^{L} u(\cdot, x-y) \cot\left(\frac{\pi}{2L} y\right) \mathrm{d}y.$$

Based on a method for the Korteweg–de Vries equation due to Dutta and Risebro [4], Galtung [5] proposed a fully discrete Crank–Nicolson Galerkin scheme for (1) where an inherent smoothing effect is used to prove convergence locally for initial data u_0 in $L^2(\mathbb{R})$ and a finite time T which depends on $\|u_0\|_{L^2}$.

The scheme for (1) is defined in the following way. First one discretizes a subset of the real line by dividing it in intervals of equal length Δx, $I_j = [x_{j-1}, x_j]$, where $x_j := j\Delta x$, $j \in \mathbb{Z}$. For the temporal discretization, one analogously has $t_n = n\Delta t$, $n \in \{0, 1, \dots, N\}$, for a discretization parameter Δt such that $T = (N + 1/2)\Delta t$. Let us also for convenience define $t_{n+1/2} := (t_n + t_{n+1})/2$. Consider now the following finite-dimensional subspace of the Sobolev space $H^2(\mathbb{R})$:

$$S_{\Delta x} = \{v \in H^2(\mathbb{R}) \mid v \in \mathbb{P}_r(I_j), \, j \in \mathbb{Z}\}, \tag{3}$$

where $r \geq 2$ is a fixed integer and $\mathbb{P}_r(I)$ denotes the space of polynomials on the interval I of degree less than or equal to r. Given $R > 0$, we define $\varphi \in C^\infty(\mathbb{R})$, for which the derivative is a cutoff function, satisfying the following conditions:

1. $1 \leq \varphi(x) \leq 2 + 2R$,
2. $\varphi'(x) = 1$ for $|x| < R$,
3. $\varphi'(x) = 0$ for $|x| \geq R + 1$, and
4. $0 \leq \varphi'(x) \leq 1$ for all x.

This function plays a key role in establishing the previously mentioned smoothing effect for the scheme, and it may be chosen to be point symmetric in $(0, \varphi(0))$.

We need a reasonable approximation of u_0 in (1) as initial data u^0 for our scheme, and so we set $u^0 = \mathrm{P}u_0$, where P is the L^2-projection on $S_{\Delta x}$. Now, we define a sequence of approximations $\{u^n\}_{n=0}^N$ of the exact solution at each t_n by the following procedure: find $u^{n+1} \in S_{\Delta x}$ such that

$$\left\langle u^{n+1}, \varphi v \right\rangle - \frac{\Delta t}{2} \left\langle \left(u^{n+1/2}\right)^2, (\varphi v)_x \right\rangle + \Delta t \left\langle \mathrm{H}\left(u^{n+1/2}\right)_x, (\varphi v)_x \right\rangle = \left\langle u^n, \varphi v \right\rangle, \tag{4}$$

for all $v \in S_{\Delta x}$, where u^0 is defined as before and $u^{n+1/2} := (u^n + u^{n+1})/2$. Here, $\langle \cdot, \cdot \rangle$ is the standard L^2-inner product. Note that the inner product $\langle \cdot, \cdot \varphi \rangle =: \langle \cdot, \cdot \rangle_\varphi$

defines a norm which we denote $\|\cdot\|_{2,\varphi}$. The nonlinearity appearing in the above implicit scheme calls for some form of iterative method to solve (4) for each time step, and in [5] the following linearized scheme is used:

$$\begin{cases} \left\langle w^{\ell+1}, \varphi v \right\rangle - \frac{\Delta t}{2} \left\langle \left(\frac{w^{\ell}+u^n}{2} \right)^2, (\varphi v)_x \right\rangle + \Delta t \left\langle \left(\mathrm{H} \frac{w^{\ell+1}+u^n}{2} \right)_x, (\varphi v)_x \right\rangle = \left\langle w^{\ell}, \varphi v \right\rangle, \\ w^0 = u^n, \end{cases} \tag{5}$$

which is to hold for all $v \in S_{\Delta x}$. By assuming a CFL condition of the type $\Delta t = \mathrm{O}(\Delta x^2)$, the above iteration is shown to converge to the solution u^{n+1} of (4). From this, one can show that there exists $T > 0$ such that $u^{\Delta x}$, which is a piecewise linear interpolation of each u^n, belongs to the space $L^2(0, T; H^{1/2}_{\mathrm{loc}}(\mathbb{R}))$. Then, compactness arguments yield the convergence result.

Because a monotone increasing cutoff function is incompatible with the periodicity of (2) one cannot use the same arguments to prove convergence for L^2-initial data in this case, and so other tools are called for when considering low regularity initial data for the periodic BO equation. However, in this study we will assume the initial data to be as regular as needed, and so we will consider the convergence rate of the method in best-case scenarios. The established well-posedness of the BO equation for these more regular spaces then guarantees that the exact solution at all times is at least as regular as the initial data. This will even make us able to consider the periodic IVP (2) using a slightly adapted scheme where we have simply replaced the cutoff function φ with 1 wherever it appears.

In the upcoming analysis, we need some preliminary estimates for polynomial approximations in finite element spaces. For a function $v \in S_{\Delta x}$, we have the following inverse inequalities:

$$|v|_{W^{k,\infty}(\mathbb{R})} \leq \frac{C}{(\Delta x)^{1/2}} |v|_{H^k(\mathbb{R})}, \qquad k = 0, 1, \tag{6}$$

$$|v|_{H^{k+1}(\mathbb{R})} \leq \frac{C}{\Delta x} |v|_{H^k(\mathbb{R})}, \qquad k = 0, 1, \tag{7}$$

where the constant C is independent of v and Δx. Both here and in the following, $|\cdot|_{W^{k,p}(\mathbb{R})}$ denotes the seminorm of the Sobolev space $W^{k,p}(\mathbb{R})$ for which $H^k(\mathbb{R}) := W^{k,2}(\mathbb{R})$. The reader is referred to [3, p. 142] for a proof of the above inequalities.

Let us now consider two projections $\mathrm{P} : L^2(\mathbb{R}) \to S_{\Delta x}$ and $\mathrm{P}_\varphi : L^2(\mathbb{R}) \to S_{\Delta x}$ defined, respectively, by

$$\int_{\mathbb{R}} (\mathrm{P}u - u)\, v \,\mathrm{d}x = 0, \quad v \in S_{\Delta x}, \tag{8}$$

and

$$\int_{\mathbb{R}} \left(\mathrm{P}_\varphi u - u \right) \varphi v \,\mathrm{d}x = 0, \quad v \in S_{\Delta x}. \tag{9}$$

For these projections applied to a function $u \in H^2(\mathbb{R})$, we have the bounds

$$\begin{aligned} \|\mathrm{P}_0 u\|_{L^2(\mathbb{R})} &\leq C\|u\|_{L^2(\mathbb{R})}, \\ \|\mathrm{P}_0 u\|_{H^1(\mathbb{R})} &\leq C\|u\|_{H^1(\mathbb{R})}, \\ \|\mathrm{P}_0 u\|_{H^2(\mathbb{R})} &\leq C\|u\|_{H^2(\mathbb{R})}, \end{aligned} \tag{10}$$

where P_0 denotes either of the two projections and C is a constant which is independent of Δx. These bounds can be derived from the norm equivalence in a finite-dimensional space and the definitions of these projections.

We also have the following polynomial approximation error estimate on the discretized domain Ω. Given $u \in H^{l+1}(\Omega)$, $0 \leq m \leq l$ and $s := \min\{l, r\}$, then

$$|\mathrm{P}_0 u - u|_{H^m(\Omega)} \leq C\Delta x^{s+1-m}|u|_{H^{s+1}(\Omega)}, \quad m = 0, 1, 2, \tag{11}$$

where again P_0 denotes either of the two projections and C is a constant not depending on Δx. For a proof of (11) for P, we refer to [8, p. 98], and the result for P_φ follows from an adaption of the same proof.

The following properties of the Hilbert transform, which can be found in [6, p. 317], are also useful:

$$\begin{aligned} \langle \mathrm{H}u, v\rangle &= -\langle u, \mathrm{H}v\rangle \text{ for } u, v \in L^2(\mathbb{R}), \\ (\mathrm{H}u)_x &= \mathrm{H}u_x, \\ \|\mathrm{H}u\|_{L^2(\mathbb{R})} &= \|u\|_{L^2(\mathbb{R})}. \end{aligned}$$

Note that these properties hold analogously for the $2L$-periodic Hilbert transform $\mathrm{H}_{\mathrm{per}}$ on $\mathbb{T}$ with $L^2(\mathbb{T}) = L^2([-L, L])$, except that $\|\mathrm{H}_{\mathrm{per}}u\|_{L^2(\mathbb{T})} \leq \|u\|_{L^2(\mathbb{T})}$.

2 Analysis of Convergence Rates

In the following, we want to consider the L^2-norm of the difference $u^n - u(t_n)$, and we will do so by decomposing the error as

$$u^n - u(t_n) = (u^n - \mathrm{P}_0 u(t_n)) + (\mathrm{P}_0 u(t_n) - u(t_n)) =: \tau^n + \rho^n,$$

and we will use the notation $w^n := \mathrm{P}_0 u(t_n)$ for the sake of brevity. Here, $\mathrm{P}_0 = \mathrm{P}_\varphi$ in the full line case, and $\mathrm{P}_0 = \mathrm{P}$ for the periodic case. For ρ^n, we already have estimates for the L^2-norm by virtue of (11), and so it remains to estimate the norm of τ^n. As the analysis is similar for the full line and periodic problems, we will give detailed estimates for the former case and only indicate the main differences between the two for the latter case. Note that in the following, C will denote a constant which exact value is of no importance. Similarly, $C(R)$ will denote such a constant which depends on R and so on. When we write, e.g., L^2 it is understood from context if we are referring to $L^2(\mathbb{R})$ or $L^2(\mathbb{T}) = L^2([-L, L])$. For both the full line and periodic case, we have the following result which is proved in the next subsections.

Theorem 1. *Given sufficiently regular initial data u_0, say $u_0 \in H^{\max\{r+1,6\}}$, for the IVP of the BO equation, we have the following convergence rate for the fully discrete Galerkin scheme described in the previous section:*

$$\|u^n - u(t_n)\|_{L^2} = O(\Delta x^{r-1} + \Delta t^2), \quad n = 0, \dots, N. \tag{12}$$

2.1 Full Line Problem

From multiplying (1) by φv, where $v \in H^2$, and integrating by parts, we get

$$\langle u_t(t), \varphi v\rangle - \frac{1}{2}\left\langle u(t)^2, (\varphi v)_x\right\rangle + \langle \mathrm{H}u_x(t), (\varphi v)_x\rangle = 0, \quad t \in (0, T]. \tag{13}$$

From (4), (13), and (9), we are able to write

$$\begin{aligned}\left\langle \frac{\tau^{n+1} - \tau^n}{\Delta t}, \varphi v\right\rangle &= \left\langle \frac{u^{n+1} - u^n}{\Delta t}, \varphi v\right\rangle - \left\langle \frac{w^{n+1} - w^n}{\Delta t}, \varphi v\right\rangle \\ &= \left\langle \frac{u^{n+1} - u^n}{\Delta t}, \varphi v\right\rangle - \left\langle u_t(t_{n+1/2}), \varphi v\right\rangle \\ &\quad + \Bigg\langle \underbrace{u_t(t_{n+1/2}) - \frac{u(t_{n+1}) - u(t_n)}{\Delta t}}_{\kappa^{n+1/2}}, \varphi v\Bigg\rangle \\ &= -\frac{1}{2}\left\langle (u^{n+1/2})^2 - u(t_{n+1/2})^2, (\varphi v)_x\right\rangle \\ &\quad + \left\langle \mathrm{H}(u_x^{n+1/2} - u_x(t_{n+1/2})), (\varphi v)_x\right\rangle + \left\langle \kappa^{n+1/2}, \varphi v\right\rangle,\end{aligned}$$

for $v \in S_{\Delta x}$. As we are now considering u evaluated at $t_{n+1/2}$, we cannot use the previous decomposition of the error directly, but we instead write

$$u^{n+1/2} - u(t_{n+1/2}) = \tau^{n+1/2} + \rho^{n+1/2} + \underbrace{\frac{u(t_{n+1}) + u(t_n)}{2} - u(t_{n+1/2})}_{\sigma^{n+1/2}}.$$

Then, we may rewrite part of the nonlinear term as

$$\begin{aligned}(u^{n+1/2})^2 - u(t_{n+1/2})^2 &= (\tau^{n+1/2})^2 + 2\tau^{n+1/2}w^{n+1/2} + (w^{n+1/2})^2 - u(t_{n+1/2})^2 \\ &= (\tau^{n+1/2})^2 + 2\tau^{n+1/2}w^{n+1/2} \\ &\quad + (w^{n+1/2} + u(t_{n+1/2}))(\rho^{n+1/2} + \sigma^{n+1/2}).\end{aligned}$$

In the following, we want to use $\tau^{n+1/2} \in S_{\Delta x}$ as test function, and from integrating by parts, we get the following relevant identities:

$$\begin{aligned} \left\langle (\tau^{n+1/2})^2, (\varphi\tau^{n+1/2})_x \right\rangle &= -\frac{1}{3}\left\langle (\tau^{n+1/2})^3, \varphi_x \right\rangle, \\ 2\left\langle \tau^{n+1/2} w^{n+1/2}, (\varphi\tau^{n+1/2})_x \right\rangle &= -\left\langle (\tau^{n+1/2})^2, \varphi w_x^{n+1/2} \right\rangle + \left\langle (\tau^{n+1/2})^2, \varphi_x w^{n+1/2} \right\rangle. \end{aligned}$$

Inserting this in the previous equations, we get

$$\begin{aligned} \frac{1}{2}\|\tau^{n+1}\|_{2,\varphi}^2 = \frac{1}{2}\|\tau^{n}\|_{2,\varphi}^2 + \Delta t \Bigg[&-\frac{1}{6}\left\langle (\tau^{n+1/2})^3, \varphi_x \right\rangle - \frac{1}{2}\left\langle (\tau^{n+1/2})^2, \varphi w_x^{n+1/2} \right\rangle \\ &+ \frac{1}{2}\left\langle (\tau^{n+1/2})^2, \varphi_x w^{n+1/2} \right\rangle \\ &+ \frac{1}{2}\left\langle (w^{n+1/2} + u(t_{n+1/2}))(\rho^{n+1/2} + \sigma^{n+1/2}), (\varphi\tau^{n+1/2})_x \right\rangle \\ &- \left\langle \mathrm{H}\tau_x^{n+1/2}, (\varphi\tau^{n+1/2})_x \right\rangle - \left\langle \mathrm{H}\rho_x^{n+1/2}, (\varphi\tau^{n+1/2})_x \right\rangle \\ &- \left\langle \mathrm{H}\sigma_x^{n+1/2}, (\varphi\tau^{n+1/2})_x \right\rangle + \left\langle \kappa^{n+1/2}, \varphi\tau^{n+1/2} \right\rangle \Bigg]. \end{aligned}$$

From the commutator estimates presented in [5], we have the inequalities

$$\langle \mathrm{H}w_x, (\varphi w)_x \rangle \geq \left\| \sqrt{\varphi_x} D^{1/2} w \right\|_{L^2}^2 - \widetilde{C}\|w\|_{L^2}^2,$$

and

$$\left\langle w^3, \varphi_x \right\rangle \leq \left\| \sqrt{\varphi_x} D^{1/2} w \right\|_{L^2}^2 + C(1 + \|w\|_{L^2}^2)\|w\|_{L^2}^2$$

for $w \in H^2$. By inserting these in the preceding identity and using the L^2-isometry of the Hilbert transform, we obtain

$$\begin{aligned} \frac{1}{2}\|\tau^{n+1}\|_{2,\varphi}^2 &+ \Delta t \left\| \sqrt{\varphi_x} D^{1/2}\tau^{n+1/2} \right\|_{L^2}^2 - \Delta t \widetilde{C}\|\tau^{n+1/2}\|_{2,\varphi}^2 \\ &\leq \frac{1}{2}\|\tau^{n}\|_{2,\varphi}^2 + \frac{\Delta t}{3}\left\| \sqrt{\varphi_x} D^{1/2}\tau^{n+1/2} \right\|_{L^2}^2 \\ &+ \Delta t \Big[C(1 + \|u^{n+1/2} - w^{n+1/2}\|_{L^2}^2)\|\tau^{n+1/2}\|_{2,\varphi}^2 \\ &+ \frac{1}{2}\|w_x^{n+1/2}\|_{L^\infty}\|\tau^{n+1/2}\|_{2,\varphi}^2 + \frac{1}{2}\|w^{n+1/2}\|_{L^\infty}\|\tau^{n+1/2}\|_{2,\varphi}^2 \\ &+ \frac{1}{2}\|w_x^{n+1/2} + u_x(t_{n+1/2})\|_{L^\infty}(\|\rho^{n+1/2}\|_{2,\varphi} + \|\sigma^{n+1/2}\|_{2,\varphi})\|\tau^{n+1/2}\|_{2,\varphi} \\ &+ \frac{1}{2}\|w^{n+1/2} + u(t_{n+1/2})\|_{L^\infty}(\|\rho_x^{n+1/2}\|_{2,\varphi} + \|\sigma_x^{n+1/2}\|_{2,\varphi})\|\tau^{n+1/2}\|_{2,\varphi} \\ &+ C_R\|\rho_x^{n+1/2}\|_{L^2}(\|\tau^{n+1/2}\|_{2,\varphi} + C_R\|\tau_x^{n+1/2}\|_{L^2}) \\ &+ C_R\|\sigma_{xx}^{n+1/2}\|_{L^2}\|\tau^{n+1/2}\|_{2,\varphi} + C_R\|\kappa^{n+1/2}\|_{L^2}\|\tau^{n+1/2}\|_{2,\varphi} \Big]. \end{aligned}$$

From the Sobolev inequality $\|w\|_{L^\infty(\mathbb{R})} \leq \|w\|_{H^1(\mathbb{R})}$, the Cauchy–Schwarz inequality, (6) and reordering we then obtain

$$
\begin{aligned}
&\frac{1}{2}\|\tau^{n+1}\|_{2,\varphi}^2 + \frac{2\Delta t}{3}\,\|\sqrt{\varphi_x}D^{1/2}\tau^{n+1/2}\|_{L^2}^2\\
&\quad\leq \frac{1}{2}\|\tau^n\|_{2,\varphi}^2 + \Delta t C_R\Big[(1+\|u^{n+1/2}\|_{L^2}^2+\|w^{n+1/2}\|_{L^2}^2)\|\tau^{n+1/2}\|_{2,\varphi}^2\\
&\quad+\|w^{n+1/2}\|_{H^2}\|\tau^{n+1/2}\|_{2,\varphi}^2+\frac{1}{2}\|w^{n+1/2}\|_{H^1}\|\tau^{n+1/2}\|_{2,\varphi}^2\\
&\quad+(\|w^{n+1/2}\|_{H^2}+\|u(t_{n+1/2})\|_{H^2})(\|\rho^{n+1/2}\|_{L^2}+\|\sigma^{n+1/2}\|_{L^2})\|\tau^{n+1/2}\|_{2,\varphi}\\
&\quad+(\|w^{n+1/2}\|_{H^1}+u(t_{n+1/2})\|_{H^1})(\|\rho_x^{n+1/2}\|_{L^2}+\|\sigma_x^{n+1/2}\|_{L^2})\|\tau^{n+1/2}\|_{2,\varphi}\\
&\quad+\|\rho_x^{n+1/2}\|_{L^2}(\|\tau^{n+1/2}\|_{2,\varphi}+\frac{1}{\Delta x}\|\tau^{n+1/2}\|_{2,\varphi})\\
&\quad+\|\sigma_{xx}^{n+1/2}\|_{L^2}\|\tau^{n+1/2}\|_{2,\varphi}+\|\kappa^{n+1/2}\|_{L^2}\|\tau^{n+1/2}\|_{2,\varphi}\Big].
\end{aligned}
$$

The following result is a part of Lemma 4.1 in [5] and will be of use.

Lemma 1. *Let u^n be the solution of* (4) *and assume furthermore that the scheme fulfills a CFL condition of the form $\Delta t^2/\Delta x^3 \leq \tilde{C}$, where $\tilde{C}$ is a constant depending on $\|u_0\|_{L^2}$. Then, $\|u^n\|_{L^2} \leq C(\|u_0\|_{L^2})$ for $n = 0, ..., N$.*

Using Lemma 1, (10), Cauchy's inequality, and dropping the second term on the left-hand side, we get

$$
\begin{aligned}
\|\tau^{n+1}\|_{2,\varphi}^2 \leq \|\tau^n\|_{2,\varphi}^2 + \Delta t C(u,R)\Big[&\|\tau^{n+1}\|_{2,\varphi}^2+\|\tau^n\|_{2,\varphi}^2+\|\rho^{n+1/2}\|_{L^2}^2+|\rho^{n+1/2}|_{H^1}^2\\
&+\frac{1}{\Delta x^2}|\rho^{n+1/2}|_{H^1}^2+\|\sigma^{n+1/2}\|_{H^2}^2+\|\kappa^{n+1/2}\|_{L^2}^2\Big],
\end{aligned}
$$

which implies

$$
(1-\Delta t C(u,R))\|\tau^{n+1}\|_{2,\varphi}^2 \leq (1+\Delta t C(u,R))\|\tau^n\|_{2,\varphi}^2 + \Delta t C(u,R) S_n,
$$

where we have the remainder term

$$
S_n = \|\rho^{n+1/2}\|_{L^2}^2 + |\rho^{n+1/2}|_{H^1}^2 + \frac{1}{\Delta x^2}|\rho^{n+1/2}|_{H^1}^2 + \|\sigma^{n+1/2}\|_{H^2}^2 + \|\kappa^{n+1/2}\|_{L^2}^2.
$$

We will assume Δt small enough that the left-hand side of the previous inequality is strictly positive, say $1-\Delta t C(u,R)) \geq 1/2$. From Taylor's formula with integral remainder, we can derive the following estimate for the seminorms of $\sigma^{n+1/2}$:

$$
|\sigma^{n+1/2}|_{H^k}^2 \leq C\Delta t^3 \int_{t_n}^{t_{n+1}} |u_{tt}(s)|_{H^k}^2\,\mathrm{d}s, \tag{14}
$$

and the L^2-norm of $\kappa^{n+1/2}$,

$$\|\kappa^{n+1/2}\|_{L^2}^2 \le C\Delta t^3 \int_{t_n}^{t_{n+1}} \|u_{ttt}(s)\|_{L^2}^2 \, \mathrm{d}s. \tag{15}$$

Then, we may estimate the remainder term using (11), (14) and (15),

$$\begin{aligned} S_n &\le C\Delta x^{2(r+1)}(|u(t_n)|_{H^{r+1}} + |u(t_{n+1})|_{H^{r+1}}) + \frac{C\Delta x^{2r}}{\Delta x^2}(|u(t_n)|_{H^{r+1}}^2 + |u(t_{n+1})|_{H^{r+1}}^2) \\ &\quad + C\Delta t^3 \int_{t_n}^{t_{n+1}} \|u_{tt}(s)\|_{H^2}^2 \, \mathrm{d}s + C\Delta t^3 \int_{t_n}^{t_{n+1}} \|u_{ttt}(s)\|_{L^2}^2 \, \mathrm{d}s \\ &= C\Delta x^{2(r-1)} \sup_{0\le t\le T} |u(t)|_{H^{r+1}}^2 + C\Delta t^3 \left(\int_{t_n}^{t_{n+1}} \|u_{tt}(s)\|_{H^2}^2 \, \mathrm{d}s + \int_{t_n}^{t_{n+1}} \|u_{ttt}(s)\|_{L^2}^2 \, \mathrm{d}s \right). \end{aligned}$$

This yields

$$\begin{aligned} \|\tau^n\|_{2,\varphi}^2 &\le \left(\frac{1+C\Delta t}{1-C\Delta t}\right)^n \|\tau^0\|_{2,\varphi}^2 + \Delta t C \sum_{j=0}^{n-1} \left(\frac{1+C\Delta t}{1-C\Delta t}\right)^{n-j} S_j \\ &\le \mathrm{e}^{4CT} \|\tau^0\|_{2,\varphi}^2 + \Delta t \mathrm{e}^{4CT} \sum_{j=0}^{n-1} S_j \\ &\le TC(u,R,T)\Delta x^{2(r-1)} + C(T)\Delta t^4 \left(\int_0^T \|u_{tt}(s)\|_{H^2}^2 \, \mathrm{d}s + \int_0^T \|u_{ttt}(s)\|_{L^2}^2 \, \mathrm{d}s \right) \\ &= C(u,R,T)(\Delta x^{2(r-1)} + \Delta t^4). \end{aligned}$$

To ensure that the above norms are bounded, we assume that $u_0 \in H^s(\mathbb{R})$, $s \ge \max\{r+1, 6\}$, see Theorems 5.3.1 and 9.1 in [1]. Then, we have

$$\|\tau^n\|_{L^2} \le \|\tau^n\|_{2,\varphi} \le C(u,R,T)(\Delta x^{r-1} + \Delta t^2),$$

where we have employed (11) to deduce

$$\|\tau^0\|_{L^2} \le \|\mathrm{P}u_0 - u_0\|_{L^2} + \|u_0 - \mathrm{P}_\varphi u_0\|_{L^2} \le C\Delta x^{r+1},$$

and if one in the original scheme instead had set $u^0 := \mathrm{P}_\varphi u_0$, then one would have $\tau^0 = 0$ directly. From this and (11), we get

$$\|u^n - u(t_n)\|_{L^2} \le \|\tau^n\|_{L^2} + \|\rho^n\|_{L^2} \le C(u,R,T)(\Delta x^{r-1} + \Delta t^2), \quad n = 1, \dots, N,$$

which proves Theorem 1 for the full line case.

2.2 Periodic Problem

For the $2L$-periodic case, we follow the steps made for the real line case, but without the cutoff function φ involved in the scheme and all inner products now act on $[-L, L]$. In this case, it is straightforward to check that the L^2-norm of the fully discrete solution u^n is conserved, simply by choosing $v = u^{n+1/2}$ in the adapted version of (4), integrating by parts and applying the skew-symmetry of the Hilbert transform and the periodicity of u^n. The existence and uniqueness of solutions the adapted version of the iterative scheme (5) can be done analogously to the original version. In this case, we do not have the commutator estimates which were used to bound the terms $\langle \mathrm{H}\tau_x^n, (\varphi\tau^n)_x\rangle$ and $\langle (\tau^n)^2, (\varphi\tau^n)_x\rangle$ by $\|\tau^n\|_{2,\varphi}^2$, but since these now appear as, respectively, $\langle \mathrm{H}_{\mathrm{per}}\tau_x^n, \tau_x^n\rangle$ and $\langle (\tau^n)^2, \tau_x^n\rangle$ we use the skew-symmetricity of $\mathrm{H}_{\mathrm{per}}$ and the periodicity of τ^n to conclude that they both vanish. Apart from this, one proceeds similarly to obtain the estimate (12) for the periodic problem. Note that by obtaining this estimate, we have proved the convergence of the scheme in the periodic case given sufficiently regular initial data using a stability and consistency argument.

3 Numerical Experiments

In order to verify the convergence rates numerically, we applied the fully discrete schemes to the problems (1) and (2). Inspired by [4] we define the subspace $S_{\Delta x}$ as follows. Let f and g be the functions

$$f(y) = \begin{cases} 1 + y^2(2|y| - 3), & |y| \le 1, \\ 0, & |y| > 1, \end{cases}$$

$$g(y) = \begin{cases} y(1 - |y|)^2, & |y| \le 1, \\ 0, & |y| > 1. \end{cases}$$

For $j \in \mathbb{Z}$, we define the basis functions

$$v_{2j}(x) = f\left(\frac{x - x_j}{\Delta x}\right), \quad v_{2j+1}(x) = g\left(\frac{x - x_j}{\Delta x}\right),$$

where $x_j = j\Delta x$. Then, $\{v_j\}_{-M}^{M}$ spans a $4M + 2$ dimensional subspace of $H^2(\mathbb{R})$. In the following, we define $N := 2M$, which is the number of elements used in the approximation. Note that for this choice we have $r = 3$ in (3), and so we expect convergence rates of order $O(\Delta x^2 + \Delta t^2)$.

To approximate the full line for (1), we have chosen to consider a finite interval with periodic boundary conditions, and we claim that this is a reasonable approximation as long as the approximate and exact solutions are close to zero at the endpoints, simulating the decay at infinity on the real line, which is the case for our examples. We have chosen to set $\Delta t = O(\Delta x)$, contrary to the assertion $\Delta t = O(\Delta x^2)$ from the theory, as smaller time steps did not lead to significant improvement in the accuracy of the approximations. In the iteration (5) to obtain u^{n+1}, we chose the stopping condition $\|w^{\ell+1} - w^{\ell}\|_{L^2} \leq 0.002\Delta x \|u^n\|_{L^2}$. The integrals involved in the Hilbert transforms were computed with seven and eight point Gauss–Legendre quadrature rules, respectively, for the inner Cauchy principal value integral and the outer integral appearing in the inner product. For $t = n\Delta t$, we set $u_{\Delta x}(x,t) = u^n(x,t) = \sum_{j=-M}^{M} u_j^n v_j(x)$. We have measured the relative error $E := \|u_{\Delta x} - u\|_{L^2}/\|u\|_{L^2}$ of the numerical approximation compared to the exact solution u, where the L^2-norms were computed with the trapezoidal rule in the grid points x_j of the finest grid considered.

3.1 Full Line Problem

A solution to this problem is the double soliton given by

$$u_{s2}(x,t) = \frac{4c_1c_2\left(c_1\lambda_1^2 + c_2\lambda_2^2 + (c_1+c_2)^2c_1^{-1}c_2^{-1}(c_1-c_2)^{-2}\right)}{\left(c_1c_2\lambda_1\lambda_2 - (c_1+c_2)^2(c_1-c_2)^{-2}\right)^2 + (c_1\lambda_1 + c_2\lambda_2)^2},$$

where $\lambda_1 := x - c_1t - d_1$ and $\lambda_2 := x - c_2t - d_2$. When $c_2 > c_1$ and $d_1 > d_2$, this equation represents a tall soliton overtaking a smaller one while moving to the right. We applied the fully discrete scheme with initial data $u_0(x) = u_{s2}(x,0)$ and parameters $c_1 = 0.3$, $c_2 = 0.6$, $d_1 = -30$, and $d_2 = -55$. The time step was set to $\Delta t = 0.5\Delta x/\|u_0\|_{L^\infty}$ and the numerical solutions were computed for $t = 90$ and $t = 180$, that is, during and after the taller soliton overtakes the smaller one. To approximate the full line problem, we set the domain to $[-100, 100]$ with the aforementioned periodic boundary condition, and based on this domain we chose the weight function $\varphi(x) = 120 + x$ for all experiments in this setting. The results are presented in Table 1 and a comparison between the approximation for $N = 256$ and the exact solution is shown in Fig. 1. These results for the full line problem are also presented as numerical examples for this scheme in [5].

The plot shows that the numerical approximation appears to be close to the exact solution and this is confirmed by the errors which are decreasing from $N = 256$ onwards, but not with a consistent rate. According to our analysis, we should expect a convergence rate of 2, but at $t = 180$ it varies from slightly below 1 to slightly above 2. As pointed out in [5], this is a complicated numerical example since one has to approximate the nonlinear interaction between two solitons. Moreover, approximating the full line by a periodic finite interval could also be contributing to the error, and thus, we are led to believe that the method applied to the periodic problem will yield results which are in better agreement with theory.

Table 1 Relative L^2-error at $t = 90$ and $t = 180$ for full line problem with initial data u_{s2} and periodic boundary conditions

N	$t = 90$		$t = 180$	
	E	rate	E	rate
128	0.01844	-1.45	0.11959	-1.32
256	0.05021	1.58	0.29755	1.75
512	0.01678	0.68	0.08869	0.74
1024	0.01044	1.16	0.05295	2.35
2048	0.00467	0.08	0.01040	0.89
4096	0.00442		0.00561	

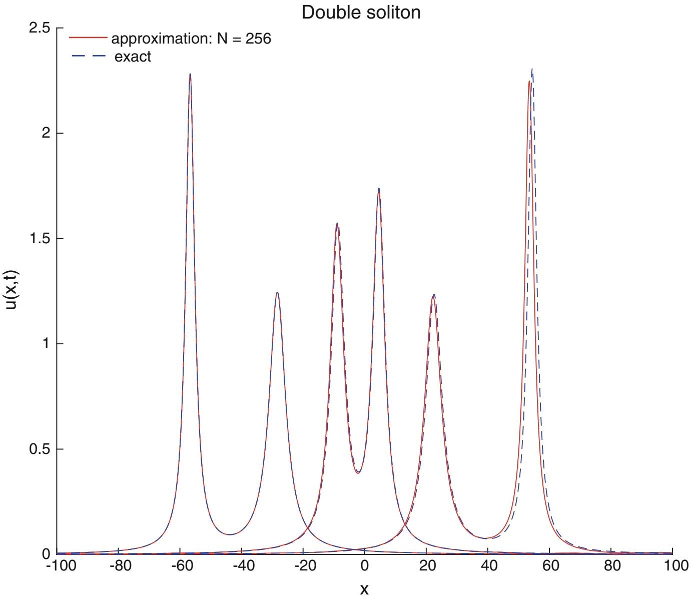


Fig. 1 Numerical approximation for $N = 256$ and exact solution for $t = 0$, 90, and 180, respectively, positioned from left to right in the plot, for full line problem with periodic boundary conditions. This figure is reproduced from [5]

3.2 Periodic Problem

In our second example, we consider the Cauchy problem for the $2L$-periodic BO equation (2). In this case, there exists a $2L$-periodic single wave solution that tends to a single soliton as the period goes to infinity, given by

$$u_{p1}(x, t) = \frac{2c\delta}{1 - \sqrt{1 - \delta^2} \cos\left(c\delta(x - ct)\right)}, \quad \delta = \frac{\pi}{cL}.$$

We applied the scheme with initial data $u_0(x) = u_{p1}(x, 0)$ with parameters $c = 0.25$ and $L = 15$. The time step was set to $\Delta t = 0.5\Delta x$ and the approximate solution was computed for $t = 480$, which is four periods for the exact solution. As previously mentioned, we do not have a weight function in this setting, which is equivalent to $\varphi = 1$. A visualization of the results for $N = 16$, 32, and 64 is given in Fig. 2.

Again, the plot indicates that the numerical approximation closes in on the exact solution, and this is confirmed by the errors in Table 2 which are decreasing with a rate of approximately 2, as predicted by theory. The reason for this better behavior compared to the previous example could also be its somewhat less complicated nature, where the exact solution is simply the translation of a single solitary wave.

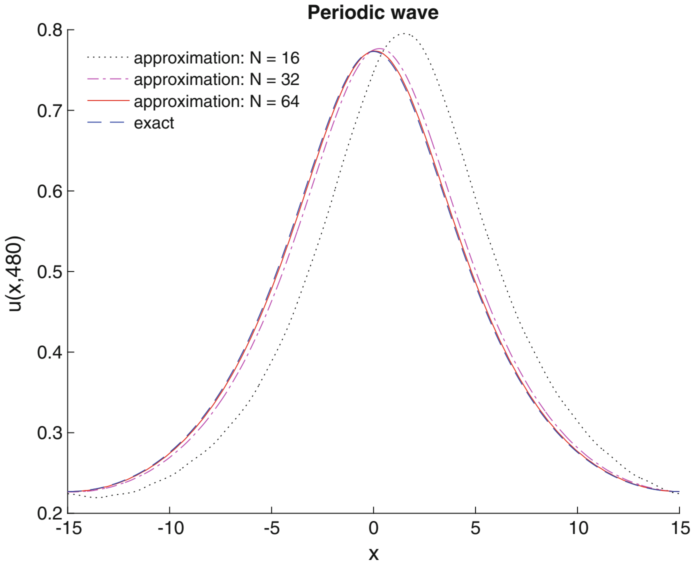


Fig. 2 Exact and numerical solutions of the $2L$-periodic problem at $t = 480$ for element numbers $N = 16, 32$, and 64, with $L = 15$ and initial data u_{p1}

Table 2 Relative L^2-error at $t = 480$ for $2L$-periodic problem with initial data u_{p1}

N	E	rate
16	0.14960222	2.41
32	0.02807195	2.28
64	0.00577740	2.16
128	0.00129088	2.07
256	0.00030683	1.97
512	0.00007805	1.85
1024	0.00002172	

Acknowledgements The author is grateful to Helge Holden for his encouragement and support to participate in the HYP2016 conference, and to Institut Mittag-Leffler for its hospitality during the Fall of 2016, providing an excellent working environment for this research.

References

1. L. Abdelouhab, J.L. Bona, M. Felland, J.-C. Saut, Nonlocal models for nonlinear, dispersive waves. Phys. D **40**(3), 360–392 (1989)
2. T.B. Benjamin, Internal waves of permanent form in fluids of great depth. J. Fluid Mech. **29**(3), 559–592 (1967)
3. P.G. Ciarlet, *The Finite Element Method for Elliptic Problems* (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002)
4. R. Dutta, N.H. Risebro, A note on the convergence of a Crank-Nicolson scheme for the KdV equation. Int. J. Numer. Anal. Model. **13**(5), 567–575 (2016)
5. S.T. Galtung, A convergent Crank–Nicolson Galerkin scheme for the Benjamin–Ono equation. Discrete Contin. Dyn. Syst. **38**(3), 1243–1268 (2018)
6. L. Grafakos, *Classical Fourier Analysis*, 3rd edn. (Springer, New York, 2014)
7. H. Ono, Algebraic solitary waves in stratified fluids. J. Phys. Soc. Jpn. **39**(4), 1082–1091 (1975)
8. A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations* (Springer, Heidelberg, 1994)

Paper 2

A semi-discrete scheme derived from variational principles for global conservative solutions of a Camassa–Holm system

Sondre Tesdal Galtung and Xavier Raynaud

Submitted for publication

A semi-discrete scheme derived from variational principles for global conservative solutions of a Camassa–Holm system

Sondre Tesdal Galtung[1] and Xavier Raynaud[1,2]

[1]Department of Mathematical Sciences, NTNU – Norwegian University of Science and Technology, Trondheim, Norway
[2]SINTEF Applied Mathematics and Cybernetics, Oslo, Norway

Abstract

We define kinetic and potential energies for which the principle of stationary action from Lagrangian mechanics yields a Camassa–Holm system (2CH) as the governing equations. After discretizing these energies, we use the same variational principle to derive a semi-discrete system of equations as an approximation of the 2CH system. The discretizaton is only available in Lagrangian coordinates and requires the inversion of a discrete Sturm–Liouville operator with time-varying coefficients. We show existence of fundamental solutions for this operator at initial time with appropriate decay. By propagating the fundamental solutions in time, we define an equivalent semi-discrete system for which we prove that there exists unique global solutions. Finally, we show how the solutions of the semi-discrete system can be used to construct a sequence of functions converging to the conservative solution of the 2CH system.

2.1 Introduction

The Camassa–Holm (CH) equation

$$u_t - u_{txx} + 3uu_x - 2u_xu_{xx} - uu_{xxx} = 0, \tag{2.1.1}$$

is first known to have appeared as a special case in a hierarchy of completely integrable partial differential equations presented in [25, Eqs. (26e) and (30)], although written in an alternative form. The equation gained prominence after it was derived in [8] as a limiting case in the shallow water regime of the Green–Naghdi equations from hydrodynamics, see also [18]. Since then. the Camassa–Holm equation has been widely studied due to its rich mathematical structure: It is for instance bi-Hamiltonian, admits a Lax pair and is completely integrable. The solutions may develop singularities in finite time even for smooth initial data, see, e.g., [12, 13].

The so-called two-component Camassa–Holm system (2CH)

$$u_t - u_{txx} + 3uu_x - 2u_x u_{xx} - uu_{xxx} + \rho\rho_x = 0, \tag{2.1.2a}$$

$$\rho_t + (\rho u)_x = 0 \tag{2.1.2b}$$

was first introduced in [39, Eq. (43)]. This is not the only two-component generalization which has been proposed for the CH equation. For instance, in [9, 23] the authors showed how similar systems are related to the AKNS hierarchy. However, we will here only consider (2.1.2), which similarly to (2.1.1) can be derived as a model for water waves. Indeed, the system was derived in [20] from the Euler equations in the case of constant vorticity, while different derivation based on the Green–Naghdi equations can be found in [14]. The 2CH system shares many properties with the CH equation: The equation is bi-hamiltonian [39], admits a Lax pair and is integrable [14]. Results on the well-posedness, blow-up and global existence of solutions to (2.1.2) are provided in [22, 33, 32, 21].

Both the CH equation and the 2CH system are geodesic equations, see [17, 15, 16, 21]. The CH equation is a geodesic on the group of diffeomorphisms for the right-invariant norm

$$E = \frac{1}{2}\left\|u\right\|^2_{\mathbf{H}^1} = \frac{1}{2}\int_{\mathbb{R}} (u^2 + u_x^2)\,dx. \tag{2.1.3}$$

To clarify this statement, we introduce the notation $y : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}$ for a path in the group of diffeomorphisms, meaning that $y(t,\xi)$ denotes the path of a particle initially at ξ, and the Eulerian velocity is given by $y_t(t,\xi)) = u(t, y(t,\xi))$. The geodesic equation is then obtained as an extremal solution for the action

$$\mathcal{A}(y) = \int_{t_0}^{t_1} E(t)\,dt = \frac{1}{2}\int_{t_0}^{t_1}\int_{\mathbb{R}} \left(y_t^2 y_\xi + \frac{y_{t\xi}^2}{y_\xi} \right) d\xi dt.$$

The momentum map, as defined in [2], is given by the Helmholtz transform $m(u) = u - u_{xx}$ in Eulerian coordinates. Then we may write the energy as $E = \frac{1}{2}\int_{\mathbb{R}} m(u)u\,dx$. For the 2CH system in [21], the diffeomorphism group is extended by a semi-direct product, which accounts for the variable ρ. Then the 2CH system is a geodesic for the norm $\frac{1}{2}\left\|u\right\|^2_{\mathbf{H}^1} + \frac{1}{2}\left\|\rho\right\|^2_{\mathbf{L}^2}$. We do not follow this purely geometrical approach here, but consider instead (2.1.2) from a Lagrangian mechanics perspective, see, e.g., [1]. Then we treat ρ as a density and define a potential energy given by

$$E^{\text{pot}} = \frac{1}{2}\int_{\mathbb{R}} (\rho - \rho_\infty)^2\,dx, \tag{2.1.4}$$

for a constant $\rho_\infty \geq 0$. Equation (2.1.4) can be interpreted as an elastic energy: It increases whenever the system deviates from the rest configuration given by $\rho \equiv \rho_\infty$. In a standard way, see, e.g., [1], the Lagrangian $\mathcal{L}$ is defined as the difference of a kinetic and potential energy

$$\mathcal{L} = E^{\text{kin}} - E^{\text{pot}}.$$

Here we set the kinetic energy equal to (2.1.3). The governing equation is then derived by the least action principle, also called principle of stationary action, on the group of diffeomorphisms.

To derive a discrete approximation of the CH and 2CH equations, we propose to follow the two steps of this variational approach. First, we discretize the path functions $y(t,\xi)$ by piecewise linear functions, $y_i(t) = y(t,\xi_i)$ for $\xi_i = i\varDelta\xi$, $i \in \mathbb{Z}$ and $\varDelta\xi > 0$. Then, we approximate the Lagrangian using these discretized variables. Finally, we obtain the governing equation for the discretized system from the principle of stationary action. The group structure of the diffeomorphisms is not carried over to the discrete setting, as the composition rule is not defined for the discrete functions. In practice, this means that our discretized equation will not have a purely Eulerian form and should be solved in Lagrangian variables. We retain two symmetries though, the time and space translation invariance. As a result, we have conservation of discrete counterparts of the integrals $\int_{\mathbb{R}} u^2 + u_x^2\,dx$ and $\int_{\mathbb{R}} u\,dx$, see Section 2.2.

We rewrite the 2CH system (2.1.2) in Lagrangian variables following [30]. We first apply the inverse of the Helmholtz operator $\mathrm{Id} - \partial_{xx}$ to obtain

$$u_t + uu_x = -P_x, \quad P - P_{xx} = u^2 + \frac{1}{2}u_x^2 + \frac{1}{2}\rho^2. \tag{2.1.5}$$

We rewrite the second equation above as a system of first-order equations,

$$\begin{bmatrix} -\partial_x & 1 \\ 1 & -\partial_x \end{bmatrix} \circ \begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix}, \tag{2.1.6}$$

for $Q = P_x$ and $f = u^2 + \frac{1}{2}u_x^2 + \frac{1}{2}\rho^2$. In Lagrangian variables we have $\bar{P}(\xi) = P(y(\xi))$, and the system (2.1.6) becomes

$$\begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} \bar{P} \\ \bar{Q} \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{f} \end{bmatrix}, \tag{2.1.7}$$

for $\bar{f} = f \circ y$. In (2.1.7), the operator denoted by y_ξ corresponds to pointwise multiplication by y_ξ. The matrix operator corresponds to the momentum map in Lagrangian coordinates and must be inverted to solve the system. In contrast to its Eulerian counterpart in (2.1.6), the operator evolves in time. This complicates the analysis significantly, especially in the discrete case. In Section 2.4, we introduce the operators $\mathcal{G}$ and $\mathcal{K}$ which define the fundamental solutions of the momentum operator,

$$\begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix}. \tag{2.1.8}$$

Note that the operator becomes singular when y_ξ vanishes. In the discrete case, the momentum operator and its fundamental solution are given by

$$\begin{bmatrix} -\mathrm{D}_- & \mathrm{D}_+ y \\ \mathrm{D}_+ y & -\mathrm{D}_+ \end{bmatrix} \circ \begin{bmatrix} \gamma & k \\ g & \kappa \end{bmatrix} = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix}, \tag{2.1.9}$$

where $\mathrm{D}_\pm$ denotes forward and backward difference operators, see Section 2.2. This is a form of *Jacobi difference equation*, cf. [41]. To establish solutions of (2.1.9), we shall invoke results from [24, 40], which generalize the Poincaré–Perron theory on difference equations. Section 2.3 is devoted to this analysis.

The CH equation and 2CH system can blow up in finite time, even for smooth initial data. Blow-up, also known as wave breaking, for the CH equation has been described in [11, 12, 19] and consists of a singularity where $\lim_{t\to t_c} u_x(t, x_c) = -\infty$ for some critical time t_c and location x_c. However, since the $\mathbf{H}^1$-norm of the solution is preserved, the solution remains continuous. In fact, the solution can be prolongated in two consistent ways: Conservative solutions will recover the total energy after the singularity, while dissipative solutions remove the energy that has been trapped in the singularity, see [5, 36, 30, 29, 6, 38, 31]. If $\rho > 0$ initially, no blow-up occurs and the 2CH system preserves the regularity of the initial data, see [30]. We can interpret this property as a regularization effect of the elastic energy: The particles cannot accumulate at a given location because of an elastic force that acts then as a repulsive force. The peakon-antipeakon collision is a good illustration of the dynamics of blow-up, see, e.g., [35]. We present this scenario in

Figures 2.1 and 2.2. In the peakon-antipeakon solution, which corresponds to $\rho_0 \equiv 0$, we observe the breakdown of the solution and the concentration of the energy distribution into a singular measure. At collision time, $u^2 + u_x^2 = 0$ and the energy reduces to a pure singular Dirac measure, which naturally cannot be plotted. In the case $\rho_0 \equiv 1$, the potential energy prevents the peaks from colliding, which is clear in the plot of the characteristics in Figure 2.1. The potential energy grows as the characteristics converge and results in an apparent force that divert them.

The global conservative solutions of the 2CH system are based on the following conservation law for the energy,

$$(\tfrac{1}{2}(u^2 + u_x^2 + (\rho - \rho_\infty)^2))_t + (u\tfrac{1}{2}(u^2 + u_x^2 + (\rho - \rho_\infty)^2))_x = -(uR)_x,$$

where $R = P - \frac{1}{2}u^2 - \frac{1}{2}\rho_\infty^2$ for P in (2.1.5). This equation enables us to compute the evolution of the cumulative energy defined from the energy distribution as

$$H(t,\xi) = \frac{1}{2}\int_{-\infty}^{y(t,\xi)} (u^2 + u_x^2 + (\rho - \rho_\infty)^2)\,dx$$

in Lagrangian coordinates, and we obtain $\frac{dH}{dt} = -(uR) \circ y$. This evolution equation is essential to keep track of the energy when the solution breaks down. To handle the blow-up of the solution, we need also to have a framework which allows the flow map $\xi \mapsto y(t,\xi)$ to become singular, that is where y_ξ can vanish and the momentum operator in Lagrangian coordinates become ill-posed. In [30], explicit expressions for P and Q are given. Here, we have to adopt a different approach where we propagate the fundamental solutions $\mathcal{K}$ and $\mathcal{G}$ from (2.1.8) in time. The equivalent system which is obtained for (2.1.2) is given by

$$y_t = U, \quad U_t = -Q, \quad H_t = -\mathcal{U}R, \tag{2.1.10a}$$

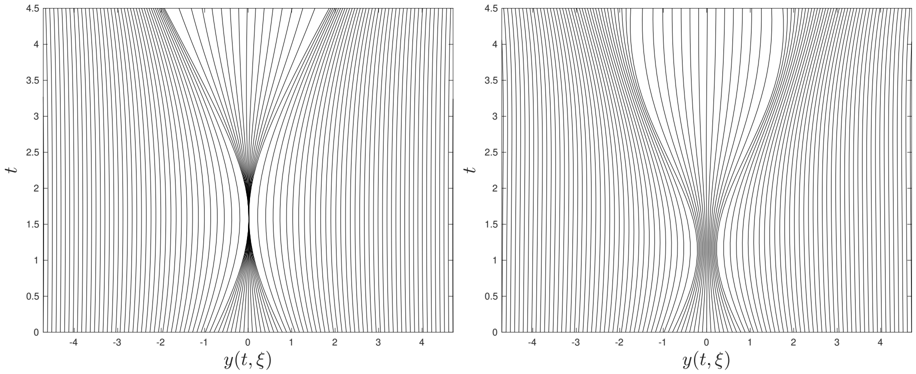
with the evolution equations for $\mathcal{K}$ and $\mathcal{G}$ given by

$$\frac{\partial}{\partial t}\mathcal{G} = [\mathcal{U}, \mathcal{K}], \quad \frac{\partial}{\partial t}\mathcal{K} = [\mathcal{U}, \mathcal{G}]. \tag{2.1.10b}$$

Here $[\mathcal{U}, \mathcal{K}]$ denotes the commutator of U and $\mathcal{K}$, see Section 2.4. In the case where $\rho_\infty = 0$, R and Q in (2.1.10a) are given by

$$\begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} \circ \begin{bmatrix} \frac{1}{2}U^2 \\ H \end{bmatrix}_\xi . \tag{2.1.10c}$$

The derivation of (2.1.10) can be carried over to our discrete system, and this is done in Section 2.4.



(a) $\rho_0 \equiv 0$

(b) $\rho_0 \equiv 1$

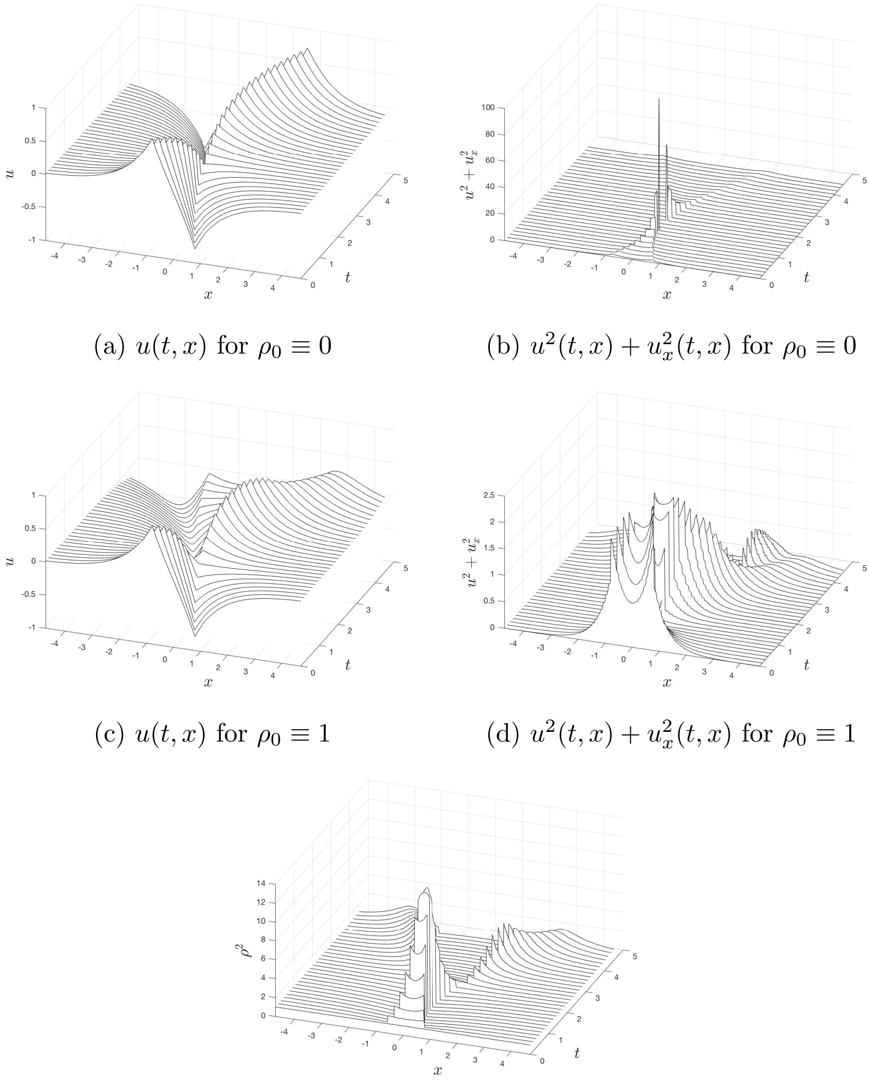
Figure 2.1: Plot of the characteristics for the peakon-antipeakon initial data for ρ_0 equal to 0 and 1. We observe the regularizing effect of $\rho_0 > 0$ which prevents the characteristics from colliding.

The short-time existence for the solution of the semi-discrete system relies on Lipschitz estimates. At this stage, one of the main ingredients in the proofs is the Young-type estimate for discrete operators presented in Proposition 2.4.2. For the global existence, we have to adapt the argument of the continuous case and complement it with *a priori* estimates of the fundamental solutions (g, k, γ, κ). These estimates follow from monotonicity properties of these operators, see Lemma 2.4.3. Section 2.5 is devoted to establishing the existence and unique of global solution to the discrete 2CH system. There we also present how one can construct initial data for the semi-discrete system. Finally, in Section 2.6, we explain how the solution of the semi-discrete system can be used to construct a sequence of functions that converge to the solution of the 2CH system (2.1.2).

2.2 Derivation of the semi-discrete CH system using a variational approach

The 2CH system can be derived as the Euler–Lagrange equation for the kinetic and potential energy given by (2.1.3) and (2.1.4) respectively. The variation is done with respect to the particle path, denoted by $y(t,\xi)$. This approach requires us to rewrite E^{kin} and E^{pot} in Lagrangian variables. For the kinetic energy, we obtain

$$E^{\mathrm{kin}}(t) = \frac{1}{2}\int_{\mathbb{R}} \left(y_t^2 y_\xi + \frac{y_{t\xi}^2}{y_\xi} \right)(t,\xi)\, d\xi. \tag{2.2.1}$$



(a) $u(t,x)$ for $\rho_0 \equiv 0$

(b) $u^2(t,x) + u_x^2(t,x)$ for $\rho_0 \equiv 0$

(c) $u(t,x)$ for $\rho_0 \equiv 1$

(d) $u^2(t,x) + u_x^2(t,x)$ for $\rho_0 \equiv 1$

(e) $(\rho(t,x) - \rho_\infty)^2$ for $\rho_0 \equiv 1$

Figure 2.2: Solutions for peakon-antipeakon initial data. For $\rho_0 \equiv 0$ we plot u in (a) and $u^2+u_x^2$ in (b). We observe the blow-up of u_x at $t_c \approx 1.5$ and the concentration of energy. For the same initial u_0, but $\rho_0 \equiv 1$, we plot the corresponding solution in (c) and (d), and observe that it does not blow-up. In (e) we plot the distribution of the potential energy given by $(\rho(t,x) - \rho_\infty)^2$, and observe how it grows when the peaks get closer to each other.

By definition, the density satisfies

$$\rho(t, y(t,\xi))y_\xi(t,\xi) = \rho(0, y(0,\xi))y_\xi(0,\xi), \tag{2.2.2}$$

which expresses that the mass is conserved. We rewrite the potential energy (2.1.4) in terms of y and obtain

$$E^{\mathrm{pot}}(t) = \frac{1}{2}\int_{\mathbb{R}} \left(\rho_0(y(0,\xi))\frac{y_\xi(0,\xi)}{y_\xi(t,\xi)} - \rho_\infty\right)^2 y_\xi(t,\xi)\,d\xi. \tag{2.2.3}$$

The definition of ρ given by (2.2.2) is equivalent to the conservation law (2.1.2b). We can check this statement directly:

$$\frac{d}{dt}(\rho(t,y)y_\xi) = (\rho_t(t,y) + \rho_x(t,y)u(t,y) + \rho(t,y)u_x(t,y))y_\xi = 0.$$

By introducing the Lagrangian density r, defined as

$$r(t,\xi) = \rho(t, y(t,\xi))y_\xi(t,\xi),$$

and requiring it to be preserved in time, we impose the definition of the density ρ in the system.

Next we shall discretize the kinetic and potential energies. First, we divide the line into a uniform grid by defining $\xi_j = j\mathit{\Delta}\xi$ for some discretization step $\mathit{\Delta}\xi > 0$ and $j \in \mathbb{Z}$. Then we approximate $y(t,\xi_j)$ with $y_j(t)$, and the spatial derivatives $y_\xi(t,\xi_j)$ with the finite difference $\mathrm{D}_+ y_j$. The finite difference operators D_+ and D_- are defined as

$$\mathrm{D}_\pm y_j := \pm\frac{y_{j\pm 1} - y_j}{\mathit{\Delta}\xi}, \tag{2.2.4}$$

and they satisfy the discrete product rule

$$\mathrm{D}_\pm(v_j w_j) = (\mathrm{D}_\pm v_j)w_{j\pm 1} + v_j(\mathrm{D}_\pm w_j). \tag{2.2.5}$$

When we later encounter grid functions with two indices, such as $g_{i,j}$ for $i, j \in \mathbb{Z}$, we will indicate partial differences by including the index in the difference operator, e.g., $\mathrm{D}_{j+} g_{i,j} = (g_{i,j+1} - g_{i,j})/\mathit{\Delta}\xi$. We use the notation $\boldsymbol{\ell}^p$ and $\boldsymbol{\ell}^\infty$ for the Banach spaces with norms

$$\|a\|_{\boldsymbol{\ell}^p} := \left(\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} |a_j|^p\right)^{\frac{1}{p}} \quad \text{and} \quad \|a\|_{\boldsymbol{\ell}^\infty} := \sup_{j\in\mathbb{Z}} |a_j|, \tag{2.2.6}$$

with $1 \le p < \infty$. For $p = 2$ we introduce a discrete analogue to the $\mathbf{L}^2$-inner product, namely

$$\langle v, w\rangle_{\boldsymbol{\ell}^2} := \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} v_j w_j, \quad v, w \in \boldsymbol{\ell}^2. \tag{2.2.7}$$

Then we may also define the adjoint, or transpose as we are working with real functions, of each difference operator in (2.2.4). We denote them by $(\mathrm{D}_\pm)^\top \colon \boldsymbol{\ell}^2 \to \boldsymbol{\ell}^2$, and they are defined through the relation

$$\sum_{j\in\mathbb{Z}} ((\mathrm{D}_\pm)^\top v_j)w_j = \sum_{j\in\mathbb{Z}} v_j(\mathrm{D}_\pm w_j), \quad v, w \in \boldsymbol{\ell}^2. \tag{2.2.8}$$

Summing by parts in the right-hand side of (2.2.8) we find that the operators in (2.2.4) satisfy $(\mathrm{D}_\pm)^\top = -\mathrm{D}_\mp$.

Turning back to the energy functionals, we discretize the kinetic energy (2.2.1) using finite differences to obtain

$$E^{\mathrm{kin}}_{\Delta\xi} := \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} \left((\dot{y}_j)^2(\mathrm{D}_+ y_j) + \frac{(\mathrm{D}_+\dot{y}_j)^2}{\mathrm{D}_+ y_j} \right). \tag{2.2.9}$$

The Lagrangian velocity is as usual defined as $U_i = \dot{y}_i$ and, using this notation, (2.2.9) becomes

$$E^{\mathrm{kin}}_{\Delta\xi} = \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} \left(U_j^2 \mathrm{D}_+ y_j + \frac{(\mathrm{D}_+ U_j)^2}{\mathrm{D}_+ y_j} \right).$$

The discrete counterpart of the potential energy (2.2.3) is similarly defined as

$$E^{\mathrm{pot}}_{\Delta\xi} := \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} \left(\frac{\mathrm{D}_+ y_{0,j}}{\mathrm{D}_+ y_j}\rho_{0,j} - \rho_\infty \right)^2 (\mathrm{D}_+ y_j),$$

where $y_{0,j} := y_0(\xi_j)$ and $\rho_{0,j} := \rho_0(y_{0,j})$. Now we define the Lagrangian as the difference between the kinetic and potential energy,

$$\mathcal{L}_{\mathrm{dis}} = E^{\mathrm{kin}}_{\Delta\xi} - E^{\mathrm{pot}}_{\Delta\xi}.$$

We compute the Fréchet derivatives of $\mathcal{L}_{\mathrm{dis}}$ with respect to y and $\dot{y}$. Formally, we have

$$\begin{aligned}
\delta E^{\mathrm{kin}}_{\Delta\xi} &= \Delta\xi \sum_{j\in\mathbb{Z}} \left(U_j(\delta U)_j(\mathrm{D}_+ y_j) + \frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\mathrm{D}_+(\delta U)_j \right) \\
&\quad + \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} \left((U_j)^2 \mathrm{D}_+(\delta y)_j - \left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right)^2 \mathrm{D}_+(\delta y)_j \right) \\
&= \Delta\xi \sum_{j\in\mathbb{Z}} \left(U_j(\mathrm{D}_+ y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right) \right)(\delta U)_j \\
&\quad - \Delta\xi \sum_{j\in\mathbb{Z}} \frac{1}{2}\mathrm{D}_- \left((U_j)^2 - \left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right)^2 \right)(\delta y)_j,
\end{aligned}$$

where in the final identity we have used (2.2.8) and $\mathrm{D}_+^\top = -\mathrm{D}_-$. This leads to the Fréchet derivatives

$$\left(\frac{\delta E^{\mathrm{kin}}_{\Delta\xi}}{\delta y}\right)_j = -\frac{1}{2}\mathrm{D}_-\left((U_j)^2 - \left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right)^2\right)$$

and

$$\left(\frac{\delta E^{\mathrm{kin}}_{\Delta\xi}}{\delta U}\right)_j = U_j(\mathrm{D}_+y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right) = \left(\frac{\delta \mathcal{L}_{\mathrm{dis}}}{\delta U}\right)_j, \tag{2.2.10}$$

where the rightmost equality in (2.2.10) is a consequence of $E^{\mathrm{pot}}_{\Delta\xi}$ being independent of U. Here the Fréchet derivative is given in $\boldsymbol{\ell}^2$, using the duality pairing defined by (2.2.7). For the potential term we find

$$\begin{aligned}\delta E^{\mathrm{pot}}_{\Delta\xi} &= \frac{\Delta\xi}{2}\sum_{j\in\mathbb{Z}}\Big(-2\left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j} - \rho_\infty\right)\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j}\mathrm{D}_+(\delta y)_j \\ &\quad + \left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j} - \rho_\infty\right)^2\mathrm{D}_+(\delta y)_j\Big) \\ &= \Delta\xi\sum_{j\in\mathbb{Z}}\frac{1}{2}\mathrm{D}_-\left(\left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j}\right)^2 - \rho_\infty^2\right)\delta y_j,\end{aligned}$$

which gives the Fréchet derivative

$$\left(\frac{\delta E^{\mathrm{pot}}_{\Delta\xi}}{\delta y}\right)_j = \frac{1}{2}\mathrm{D}_-\left(\left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j}\right)^2 - \rho_\infty^2\right).$$

The Euler–Lagrange equation gives us

$$\frac{\delta\mathcal{L}_{\mathrm{dis}}}{\delta y} - \frac{d}{dt}\frac{\delta\mathcal{L}_{\mathrm{dis}}}{\delta U} = 0, \tag{2.2.11}$$

see, e.g., [1]. Since

$$\begin{aligned}\frac{d}{dt}\left(\frac{\delta\mathcal{L}_{\mathrm{dis}}}{\delta U}\right)_j &= \frac{d}{dt}\left(U_j(\mathrm{D}_+y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right)\right) \\ &= \dot{U}_j(\mathrm{D}_+y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+\dot{U}_j}{\mathrm{D}_+y_j}\right) \\ &\quad + U_j(\mathrm{D}_+U_j) + \mathrm{D}_-\left(\left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right)^2\right),\end{aligned}$$

we obtain the following system of governing equations

$$\dot{y}_j = U_j \tag{2.2.12a}$$

and

$$\begin{aligned}(\mathrm{D}_+ y_j)\dot{U}_j - \mathrm{D}_-\left(\frac{\mathrm{D}_+\dot{U}_j}{\mathrm{D}_+ y_j}\right)\\ = -U_j(\mathrm{D}_+U_j) - \frac{1}{2}\mathrm{D}_-\left((U_j)^2 + \left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right)^2 + \left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j}\right)^2\right),\end{aligned} \tag{2.2.12b}$$

for $j \in \mathbb{Z}$. Note that we have omitted ρ_∞^2 on the right hand side in (2.2.12) as D_- maps constants to zero.

Next we show that energy, which coincides with the Hamiltonian, of the system is preserved in time. We can use the Legendre transform to define the Hamiltonian

$$\mathcal{H}_{\mathrm{dis}} = \left\langle \frac{\delta\mathcal{L}_{\mathrm{dis}}}{\delta U}, U\right\rangle_{\boldsymbol{\ell}^2} - \mathcal{L}_{\mathrm{dis}}. \tag{2.2.13}$$

Writing out the above Hamiltonian explicitly we have

$$\mathcal{H}_{\mathrm{dis}} = \frac{1}{2}\Delta\xi\sum_{j\in\mathbb{Z}}\left((U_j)^2 + \left(\frac{\mathrm{D}_+U_j}{\mathrm{D}_+y_j}\right)^2 + \left(\frac{\mathrm{D}_+y_{0,j}}{\mathrm{D}_+y_j}\rho_{0,j} - \rho_\infty\right)^2\right)(\mathrm{D}_+y_j). \tag{2.2.14}$$

We observe that the Lagrangian $\mathcal{L}_{\mathrm{dis}}$ does not depend explicitly on time. Then it is a classical result from mechanics, which follows from Noether's theorem, that $\mathcal{H}_{\mathrm{dis}}$ is time-invariant,

$$\frac{d\mathcal{H}_{\mathrm{dis}}}{dt} = 0.$$

The Lagrangian $\mathcal{L}_{\mathrm{dis}}$ is also invariant with respect to translation, which means that another time-invariant can be obtained. We denote by $\psi : \boldsymbol{\ell}^2 \times \mathbb{R} \to \boldsymbol{\ell}^2$ the transformation given by the uniform translation $(\psi(y,\varepsilon))_j = y_j + \varepsilon$. For simplicity, we write $y^\varepsilon(t) = \psi(y(t),\varepsilon)$. From the definition of ψ, we have

$$\dot{y}^\varepsilon(t) = \dot{y}(t) \quad \text{and} \quad \mathrm{D}_+y^\varepsilon(t) = \mathrm{D}_+y(t).$$

Hence, the Lagrangian $\mathcal{L}_{\mathrm{dis}}$ is invariant with respect to the transformation ψ. Then Noether's theorem gives us that the quantity $\left\langle \frac{\delta\mathcal{L}_{\mathrm{dis}}}{\delta U}, \frac{\delta y^\varepsilon}{\delta\varepsilon}\right\rangle_{\boldsymbol{\ell}^2}$

is preserved by the flow. In our case, $\left(\frac{\delta y^\varepsilon}{\delta \varepsilon}\right)_j = 1$ and $\left(\frac{\delta \mathcal{L}_{\mathrm{dis}}}{\delta U}\right)_j = U_j(\mathrm{D}_+ y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right)$, see (2.2.10). Thus, we obtain that the quantity

$$I = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \left(U_j(\mathrm{D}_+ y_j) - \mathrm{D}_-\left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right)\right) = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} U_j(\mathrm{D}_+ y_j), \quad (2.2.15)$$

is preserved. Note that I corresponds to a discretization of

$$\int_{\mathbb{R}} (u - u_{xx})\, dx = \int_{\mathbb{R}} u\, dx$$

in Eulerian coordinates, which is preserved by the 2CH system.

Let us return to (2.2.12), and in particular to the left-hand side which contains $\dot{U}_j$ but not in an explicit form. For a given sequence $a = \{a_j\}_{j\in\mathbb{Z}} \in \boldsymbol{\ell}^\infty$ and an arbitrary sequence $w = \{w_j\}_{j\in\mathbb{Z}} \in \boldsymbol{\ell}^2$, we define the operator $\mathrm{A}[a] : \boldsymbol{\ell}^2 \to \boldsymbol{\ell}^2$ by

$$(\mathrm{A}[a]w)_j \coloneqq a_j w_j - \mathrm{D}_-\left(\frac{\mathrm{D}_+ w_j}{a_j}\right), \quad j \in \mathbb{Z}. \quad (2.2.16)$$

The operator $\mathrm{A}[a]$ corresponds to the aforementioned momentum operator m when $a_j = \mathrm{D}_+ y_j$, and takes the form of a Sturm–Liouville operator. This operator is symmetric and positive definite for sequences a such that $a_j > 0$, as we can see from

$$\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} v_j (\mathrm{A}[a]w)_j = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \left(a_j w_j v_j + \frac{1}{a_j}(\mathrm{D}_+ w_j)(\mathrm{D}_+ v_j)\right),$$

where we once more have used (2.2.8). When $\mathrm{A}[\mathrm{D}_+ y]$ is positive definite, it is invertible and we may formally write (2.2.12) as a system of first order ordinary differential equations,

$$\begin{aligned} \dot{y}_j &= U_j, \\ \dot{U}_j &= -\mathrm{A}[\mathrm{D}_+ y]^{-1} \Bigg(U_j(\mathrm{D}_+ U_j) \\ &\quad + \frac{1}{2}\mathrm{D}_-\left((U_j)^2 + \left(\frac{\mathrm{D}_+ U_j}{\mathrm{D}_+ y_j}\right)^2 + \left(\frac{\mathrm{D}_+ y_{0,j}}{\mathrm{D}_+ y_j}\rho_{0,j}\right)^2\right)\Bigg). \end{aligned} \quad (2.2.17)$$

When solving the above system, we obtain approximations of the fluid velocity and density in Lagrangian variables, $U_j(t) \approx u(t, y(t,\xi_j))$ and $\rho_{0,j}/(\mathrm{D}_+ y_j(t)) \approx \rho(t, y(t,\xi_j))$.

We conclude this section with some comments on the Hamiltonian form of the equations. Hamiltonian equations in generalized position and momentum variables can be derived from the Lagrangian approach in classical mechanics, see, e.g., [1]. The generalized momentum is defined as $p = \frac{\delta \mathcal{L}_{\mathrm{dis}}}{\delta U}(y, U)$. When we express the Hamiltonian $\mathcal{H}_{\mathrm{dis}}$ given in (2.2.14) in term of y and p, the Hamiltonian equations are then given as usual by

$$\dot{y} = \frac{\delta \mathcal{H}_{\mathrm{dis}}}{\delta p}, \quad \dot{p} = -\frac{\delta \mathcal{H}_{\mathrm{dis}}}{\delta y}. \tag{2.2.18}$$

From (2.2.15), we get that the momentum is $p_j = (\mathrm{A}[\mathrm{D}_+ y]U)_j$. Hence, $U_j = (\mathrm{A}[\mathrm{D}_+ y]^{-1} p)_j$, and the Hamiltonian (2.2.14) is

$$\mathcal{H}_{\mathrm{dis}} = \frac{1}{2} \mathit{\Delta}\xi \sum_{j \in \mathbb{Z}} p_j (\mathrm{A}[\mathrm{D}_+ y]^{-1} p)_j + E^{\mathrm{pot}}_{\mathit{\Delta}\xi}.$$

Moreover, (2.2.18) are exactly $\dot{y} = U$ and (2.2.11). If we introduce the fundamental solution $g_{i,j}$ of the operator $\mathrm{A}[\mathrm{D}_+ y]$, see Section 2.3, the Hamiltonian can be rewritten as

$$\mathcal{H}_{\mathrm{dis}} = \frac{1}{2} \mathit{\Delta}\xi \sum_{j \in \mathbb{Z}} p_j \mathit{\Delta}\xi \sum_{i \in \mathbb{Z}} g_{i,j} p_i = \frac{1}{2} \sum_{i,j \in \mathbb{Z}} (\mathit{\Delta}\xi p_i)(\mathit{\Delta}\xi p_j) g_{i,j} + E^{\mathrm{pot}}_{\mathit{\Delta}\xi}.$$

In the case $\rho \equiv 0$, i.e., $E^{\mathrm{pot}}_{\mathit{\Delta}\xi} = 0$, we recognized the similarity of this expression to that of

$$\mathcal{H}_{\mathrm{mp}} = \frac{1}{2} \sum_{i,j=1}^{N} p_i p_j e^{-|y_i - y_j|}$$

given in [8]. The Hamiltonian $\mathcal{H}_{\mathrm{mp}}$ defines the multipeakon solutions, which can be seen as an other form of discretization for CH, see [37] for the global conservative case. Then, the two discretization appear as the results of two different choices of discretization for the inverse momentum operator: We have $g_{i,j}$ in the case of this paper and $\hat{g}_{i,j} = e^{-|y_i - y_j|}$ in [8]. We note that a numerical study of discretizations of the periodic CH equation considering both multipeakons and the variational method presented in this paper can be found in [26].

2.3 Construction of the fundamental solutions of the discrete momentum operator

In this section we will construct fundamental solutions for the momentum operator in (2.2.16). However, we will first present some results on sequences which will prove useful in the later analysis.

2.3.1 Some useful results for sequence spaces

Given a grid parameter $\Delta\xi > 0$, let $\boldsymbol{\ell}$ be the space of all bilaterally infinite sequences $a = \{a_j\}_{j\in\mathbb{Z}}$ defined on the lattice $\Delta\xi\mathbb{Z} =: \{j\Delta\xi \mid j \in \mathbb{Z}\}$. We have already encountered the Banach spaces $\boldsymbol{\ell}^\infty$ and $\boldsymbol{\ell}^p$ with norms defined in (2.2.6). Let us also define a discrete analogue of the $\mathbf{H}^1(\mathbb{R})$-inner product,

$$\langle a, b\rangle_{\mathbf{h}^1} =: \Delta\xi \sum_{j\in\mathbb{Z}} \left[a_j b_j + (\mathrm{D}_+ a_j)(\mathrm{D}_+ b_j)\right], \tag{2.3.1}$$

which induces a norm in the usual manner. Finally, we introduce the subspace of $\boldsymbol{\ell}^\infty$ defined as

$$\mathbf{V}_{\Delta\xi} \coloneqq \{a \in \boldsymbol{\ell}^\infty \mid \mathrm{D}_+ a \in \boldsymbol{\ell}^2\}, \qquad \|a\|_{\mathbf{V}_{\Delta\xi}} \coloneqq \|a\|_{\boldsymbol{\ell}^\infty} + \|\mathrm{D}_+ a\|_{\boldsymbol{\ell}^2}. \tag{2.3.2}$$

Now we are set to present some results for sequences which will be useful to us.

Proposition 2.3.1 (Useful results for sequences)**.** *We list some useful results for sequences $a\colon \mathbb{Z} \to \mathbb{R}$, $b\colon \mathbb{Z} \to \mathbb{R}$, and $f\colon \mathbb{Z}\times\mathbb{Z} \to \mathbb{R}$, where we use the convention $q/\infty = 0$ for $q < \infty$, and $\infty/\infty = 1$.*

The inverse inequalities

$$\|a\|_{\boldsymbol{\ell}^\infty} \le \frac{1}{\sqrt{\Delta\xi}}\|a\|_{\boldsymbol{\ell}^2} \le \frac{1}{\Delta\xi}\|a\|_{\boldsymbol{\ell}^1}, \tag{2.3.3}$$

the discrete Sobolev-type inequality

$$\|a\|_{\boldsymbol{\ell}^\infty} \le \frac{1}{\sqrt{2}}\|a\|_{\mathbf{h}^1}, \tag{2.3.4}$$

the summation by parts formula

$$\Delta\xi \sum_{j=m}^{n} (\mathrm{D}_+ a_j) b_j + \Delta\xi \sum_{j=m}^{n} a_j (\mathrm{D}_- b_j) = a_{n+1} b_n - a_m b_{m-1}, \tag{2.3.5}$$

and a discrete generalized Hölder inequality

$$\left\| \prod_{k=1}^{n} a_k \right\|_{\boldsymbol{\ell}^q} \le \prod_{k=1}^{n} \|a_k\|_{\boldsymbol{\ell}^{p_k}} \ \textit{for} \ \sum_{k=1}^{n} \frac{1}{p_k} = \frac{1}{q}, \quad q, p_k \in [1,\infty], \tag{2.3.6}$$

where in (2.3.6) the j-th component of a product of sequences is interpreted as

$$\left(\prod_{k=1}^{n} a_k\right)_j = \prod_{k=1}^{n} (a_k)_j.$$

Furthermore, any sequence a such that $\mathrm{D}_+a \in \boldsymbol{\ell}^2$ is bounded in the "discrete Hölder seminorm",

$$\sup_{j,k\in\mathbb{N}, j\neq k} \frac{|a_j - a_k|}{(\Delta\xi|j-k|)^{1/2}} \leq \|\mathrm{D}_+a\|_{\boldsymbol{\ell}^2}\,. \tag{2.3.7}$$

In addition, such sequences satisfy the asymptotic relation

$$\lim_{j\to\pm\infty} \sqrt{\Delta\xi}\,|\mathrm{D}_+a_j| = 0, \tag{2.3.8}$$

which in particular implies

$$\lim_{j\to\pm\infty} |a_{j+1}-a_j| = \lim_{j\to\pm\infty} \Delta\xi\,|\mathrm{D}_+a_j| = 0 \textit{ for } a \in \mathbf{V}_{\Delta\xi}. \tag{2.3.9}$$

For the proof of Proposition 2.3.1 we refer to Appendix 2.A.

2.3.2 Construction of fundamental solutions

In this section we construct a Green's function or fundamental solution for the operator defined in (2.2.16). Note that when $a = \mathrm{D}_+y$ coincides with the constant sequence $\mathbf{1} = \{1\}_{j\in\mathbb{Z}}$, we have from (2.2.16) that $\mathrm{A}[\mathbf{1}] = \mathrm{Id} - \mathrm{D}_-\mathrm{D}_+$, which corresponds to the operator used in the difference schemes studied in [10, 34]. As the coefficients are constant, the authors are able to find an explicit Green's function g which can be written as

$$g_j = \frac{1}{\sqrt{4+\Delta\xi^2}}\left(1+\frac{\Delta\xi^2}{2}+\frac{\Delta\xi}{2}\sqrt{4+\Delta\xi^2}\right)^{-|j|} \tag{2.3.10}$$

and fulfills $(\mathrm{Id}-\mathrm{D}_-\mathrm{D}_+)g = \delta_0$. Here $\delta_0 = \{\delta_{0,j}\}_{j\in\mathbb{Z}}$ for the Kronecker delta $\delta_{i,j}$, equal to one when the indices coincide and zero otherwise. In our case, the coefficients appearing in the definition of $\mathrm{A}[\mathrm{D}_+y]$ are varying with the grid index j, which significantly complicates the construction of the Green's function.

Let us consider the operator $\mathrm{A}[a]$ from (2.2.16) and the equation $(\mathrm{A}[a]g)_j = f_j$. We want to prove that there exists a solution which decreases exponentially as $j \to \pm\infty$. To this end, we want to find a Green's function for the operator $\mathrm{A}[a]$, and the first step is to realize that the homogeneous operator equation $(\mathrm{A}[a]g)_j = 0$ can be written as

$$\frac{\mathrm{D}_+g_j}{a_j} = \Delta\xi a_j g_j + \frac{\mathrm{D}_+g_{j-1}}{a_{j-1}}.$$

This can again be restated as a *Jacobi difference equation*, see [41, Eq. (1.19)],

$$-\frac{1}{a_j}g_{j+1}+\left(\frac{1}{a_j}+\frac{1}{a_{j-1}}+a_j(\varDelta\xi)^2\right)g_j-\frac{1}{a_{j-1}}g_{j-1}=0,$$

or equivalently in matrix form

$$\begin{bmatrix} g_j \\ g_{j+1}\end{bmatrix}=\begin{bmatrix} 0 & 1 \\ -\frac{a_j}{a_{j-1}} & 1+\frac{a_j}{a_{j-1}}+(a_j\varDelta\xi)^2\end{bmatrix}\begin{bmatrix} g_{j-1} \\ g_j\end{bmatrix}=:\tilde{A}_j\begin{bmatrix} g_{j-1} \\ g_j\end{bmatrix}. \quad (2.3.11)$$

Observe that $\tilde{A}_j$ is not symmetric and always contains positive, negative and zero entries under the assumption $a_j>0$. Moreover, $\tilde{A}_j$ is ill-defined when $a_{j-1}=0$. This case is of importance to us, as it corresponds to wave breaking for the system. We want to allow for $a_j=0$ in order to obtain solutions globally in time. If we go back to the first restatement of the operator equation and introduce the variable

$$\gamma_j:=\frac{\mathrm{D}_+g_j}{a_j}=\frac{g_{j+1}-g_j}{a_j\varDelta\xi}, \quad (2.3.12)$$

we get the following characterization of the homogeneous problem

$$\begin{bmatrix} -\mathrm{D}_+ & a_j \\ a_j & -\mathrm{D}_-\end{bmatrix}\begin{bmatrix} g_j \\ \gamma_j\end{bmatrix}=\begin{bmatrix} 0 \\ 0\end{bmatrix}, \quad (2.3.13)$$

or equivalently

$$\begin{bmatrix} g_{j+1} \\ \gamma_j\end{bmatrix}=\begin{bmatrix} 1+(a_j\varDelta\xi)^2 & a_j\varDelta\xi \\ a_j\varDelta\xi & 1\end{bmatrix}\begin{bmatrix} g_j \\ \gamma_{j-1}\end{bmatrix}=:A_j\begin{bmatrix} g_j \\ \gamma_{j-1}\end{bmatrix}. \quad (2.3.14)$$

Here A_j is a symmetric matrix with positive entries whenever $a_j>0$, and it reduces to the identity matrix when $a_j=0$. We will use (2.3.14) rather than (2.3.11) to construct our Green's function, and it will also significantly simplify the analysis of the asymptotic behavior for the solutions.

Lemma 2.3.2 (Properties of matrix A_j)**.** *Consider A_j from* (2.3.14) *and assume $a_j=1+\mathrm{D}_+b_j\geq 0$ where $\mathrm{D}_+b\in\boldsymbol{\ell}^2$. Then* $\det A_j=1$ *and there exist $M_b>m_b>0$ depending on $\|\mathrm{D}_+b\|_{\boldsymbol{\ell}^2}$ and $\varDelta\xi$ such that the eigenvalues $\lambda_j^{\pm}$ of A_j satisfy*

$$m_b\leq\lambda_j^-<1<\lambda_j^+\leq M_b \quad (2.3.15)$$

uniformly with respect to j when $a_j>0$. Moreover there is the obvious identity $\lambda_j^{\pm}=1$ when $a_j=0$. Asymptotically we have $\lim_{j\to\pm\infty}A_j=A$*,* *where A is given by A_j after setting $a_j = 1$, and the eigenvalues $\lambda^\pm$ of A satisfy*

$$m \le \lambda^- < 1 < \lambda^+ \le M \tag{2.3.16}$$

for $M > m > 0$ depending only on $\Delta\xi$. Moreover, as the eigenvalues are strictly positive it follows that the spectral radius of A_j, $\operatorname{spr}(A_j) := \max\{|\lambda_j^+|, |\lambda_j^-|\}$ satisfies $\|A_j\| = \operatorname{spr}(A_j) = \lambda_j^+$, $\|A\| = \operatorname{spr}(A) = \lambda^+$, and both matrices can be diagonalized: $A_j = R_j \Lambda_j R_j^\top$, $A = R \Lambda R^\top$.

Proof of Lemma 2.3.2. To see that $\det A_j = 1$ one can compute it directly, or see it from the eigenvalues

$$\begin{aligned} \lambda_j^\pm &:= 1 + \frac{(a_j \Delta\xi)^2}{2} \pm \frac{a_j \Delta\xi}{2} \sqrt{4 + (a_j \Delta\xi)^2} \\ &= \frac{1}{4} \left(\sqrt{4 + (a_j \Delta\xi)^2} \pm a_j \Delta\xi \right)^2, \end{aligned} \tag{2.3.17}$$

which shows that A_j is invertible irrespective of the value of a_j. As A_j is real and symmetric it can be diagonalized with orthonormal eigenvectors $r_j^\pm$ as follows

$$A_j = R_j \Lambda_j R_j^\top, \quad \Lambda_j = \begin{bmatrix} \lambda_j^- & 0 \\ 0 & \lambda_j^+ \end{bmatrix}, \quad R_j = \begin{bmatrix} \dfrac{1}{\sqrt{1+\lambda_j^+}} & \dfrac{1}{\sqrt{1+\lambda_j^-}} \\ -\dfrac{1}{\sqrt{1+\lambda_j^-}} & \dfrac{1}{\sqrt{1+\lambda_j^+}} \end{bmatrix}. \tag{2.3.18}$$

Since $\mathrm{D}_+ b \in \boldsymbol{\ell}^2$, for any $j \in \mathbb{Z}$ we have the bound

$$\sqrt{\Delta\xi}\, |\mathrm{D}_+ b_j| = \left(\Delta\xi\, |\mathrm{D}_+ b_j|^2 \right)^{1/2} \le \|\mathrm{D}_+ b\|_{\boldsymbol{\ell}^2}$$

which leads to

$$0 \le a_j \Delta\xi \le \Delta\xi + \sqrt{\Delta\xi}\, \|\mathrm{D}_+ b\|_{\boldsymbol{\ell}^2} =: K_b,$$

meaning a_j is bounded from above and below. Then it follows that

$$0 < \left(\frac{\sqrt{4 + K_b^2} - K_b}{2} \right)^2 \le \lambda_j^- \le 1 \le \lambda_j^+ \le \left(\frac{\sqrt{4 + K_b^2} + K_b}{2} \right)^2$$

corresponding to (2.3.15). Furthermore, we have $\lim_{j \to \pm\infty} a_j \Delta\xi = \Delta\xi$ by (2.3.9). We denote by A, Λ, R, and $\lambda^\pm$ the matrices and eigenvalues given by A_j, Λ_j, R_j, and $\lambda_j^{\pm}$ after replacing a_j by 1. From the preceding limit, (2.3.17) and (2.3.18) we obtain

$$\lim_{j\to\pm\infty}(A_j,\Lambda_j,R_j)=(A,\Lambda,R). \tag{2.3.19}$$

Bounds for $\lambda^{\pm}$ are obtained similarly to the bounds for $\lambda_j^{\pm}$. As A_j, A are symmetric and hence normal, their norms coincide with the spectral radius spr$(\cdot)$ which here coincides with the largest eigenvalue. □

Note that (2.3.14) corresponds to a transition from (g_j,γ_{j-1}) to (g_{j+1},γ_j), so that A_j can be considered as a transfer matrix between these two states. Thus, solving the homogeneous operator equation $(\mathrm{A}[a]g)_j=0$ bears clear resemblance to propagating a discrete dynamical system, and this is also the idea employed in the analysis of Jacobi difference equations in [41, Eq. (1.28)]. However, in making the change of variable to obtain (2.3.14) we lose the symmetry of the difference equation, and so the results in [41] are no longer directly applicable. On the other hand, our system can be regarded as a more general Poincaré difference system, and our idea is then to apply the results [24, Thm. 1.1] and [40, Thm. 1] to the matrix product

$$\Phi_{k,j}:=\begin{cases}A_{k-1}\cdots A_j, & k>j,\\ I, & k=j,\\ (A_k)^{-1}\cdots(A_{j-1})^{-1}, & k<j,\end{cases} \tag{2.3.20}$$

which is the transition matrix from (g_j,γ_{j-1}) to (g_k,γ_{k-1}). Note that in the lemma below, the norms can be taken to be the standard Euclidean norm, but one could use any vector norm.

Lemma 2.3.3 (Existence of exponentially decaying solutions)**.** *Consider the matrix equation*

$$v_n=(\Phi_{n,0})v_0, \qquad v_n=\begin{bmatrix}g_n\\ \gamma_{n-1}\end{bmatrix}, \tag{2.3.21}$$

coming from (2.3.14) *with* $\Phi_{n,0}$ *as defined in* (2.3.20)*. Then there exist initial vectors* $v_0=v_0^{\pm}$ *such that the corresponding solutions* $v_n^{\pm}$ *satisfy*

$$\lim_{n\to\mp\infty}\sqrt[n]{\|v_n^{\pm}\|}=\lambda^{-}. \tag{2.3.22}$$

That is, there exist solutions v_n *with exponential decay in either direction, owing to the Lyapunov exponent* $\lambda^-<1$*. Moreover, the initial vectors are unique up to a constant factor.*

Proof of Lemma 2.3.3. We begin with the case of increasing n, and we want to apply [24, Thm. 1.2] which states that for sequences of positive matrices $\{A_n\}$ satisfying $\lim_{n\to+\infty} A_n = A$ for some positive matrix A we have

$$\lim_{n\to+\infty} \frac{A_n A_{n-1}\dots A_1 A_0}{\|A_n A_{n-1}\dots A_1 A_0\|} = vw^\top \tag{2.3.23}$$

for some vectors v and w with positive entries such that $Av = \operatorname{spr}(A)v$. As mentioned in [3, Rem. 4], there is in general no easy way of determining the vector w explicitly.

We recall that our A_n has positive entries, unless $a_n = 0$ in which case we have $A_n = I$. Because of (2.3.19), there can only be finitely many $n \geq 0$ for which A_n reduces to the identity. If we instead consider the sequence of positive matrices consisting of our $\{A_n\}$ where we have omitted the finitely many identity matrices, they clearly still satisfy (2.3.19) and so (2.3.23) holds with $\operatorname{spr}(A) = \lambda^+$ and $v = r^+$ from (2.3.17) and (2.3.18). However, as the matrices we omitted were identities, it is clear that the limit in (2.3.23) for both sequences coincide. Hence, [24, Thm. 1.1] holds for our nonnegative sequence as well.

Now, as $A_n \geq I$ entrywise it follows that the entries of $\Phi_{n,0}$ are nondecreasing for $n \geq 0$, which means that $\|\Phi_{n,0}\|$ is also nondecreasing for such n. Therefore, by (2.3.23) we have that any initial vector v_0 such that $w^\top v_0 \neq 0$ leads to a solution v_n with nondecreasing norm, and which then by [40, Thm. 1] must satisfy

$$\varrho = \lim_{n\to+\infty} \sqrt[n]{\|v_n\|} \tag{2.3.24}$$

with $\varrho = \lambda^+ > 1$ due to (2.3.16), i.e., an asymptotically exponentially increasing solution. Indeed, the nondecreasing norm rules out the possibility of $v_n = 0$ for n large enough. It follows that (2.3.24) holds for ϱ equal to either λ^+ or λ^-, but if it were $\lambda^- < 1$, then $\|v_n\|$ could not be nondecreasing. However, choosing instead a nonzero v_0 such that $w^\top v_0 = 0$, we obtain an asymptotically exponentially decreasing solution v_n satisfying (2.3.24) with $\varrho = \lambda^- < 1$. This follows by once more excluding the scenario of $v_n = 0$ for large enough n, since v_0 is nonzero and each A_n has full rank. Then the only remaining possibility is v_n satisfying (2.3.24) with $\varrho = \lambda^-$. An obvious choice of v_0 given $w = [w_1 \;\, w_2]^\top$ is then $v_0 = [w_2 \;\, {-w_1}]^\top$.

For decreasing n, we will be able to reuse the arguments from above. From (2.3.20) we find that $\Phi_{n,0}$ is a product of inverses of A_n for $n < 0$, and by (2.3.14) we have

$$\begin{bmatrix} g_j \\ \gamma_{j-1} \end{bmatrix} = (A_j)^{-1} \begin{bmatrix} g_{j+1} \\ \gamma_j \end{bmatrix} = \begin{bmatrix} 1 & -a_j \varDelta\xi \\ -a_j\varDelta\xi & 1 + (a_j\varDelta\xi)^2 \end{bmatrix} \begin{bmatrix} g_{j+1} \\ \gamma_j \end{bmatrix}.$$

Since $(A_n)^{-1}$ contains entries of opposite sign, it would appear that we may not be able to use our previous argument. However, a change of variables will do the trick for us. First recall (2.3.12) which shows that γ_j corresponds to a rescaled forward difference for g_j, hence its sign indicates whether g is increasing or decreasing at index j. For an increasing solution in the direction of increasing n it is then necessary for g_n and γ_{n-1} to share the same sign as $n \to +\infty$. On the other hand, for an increasing solution in the direction of decreasing n, the forward difference for γ_{n-1} should have the opposite sign of g_n as $n \to -\infty$. Therefore, a change of variables allows us to rewrite the previous equations as

$$\begin{bmatrix} g_j \\ -\gamma_{j-1} \end{bmatrix} = \begin{bmatrix} 1 & a_j \varDelta\xi \\ a_j \varDelta\xi & 1 + (a_j \varDelta\xi)^2 \end{bmatrix} \begin{bmatrix} g_{j+1} \\ -\gamma_j \end{bmatrix} =: B_j \begin{bmatrix} g_{j+1} \\ -\gamma_j \end{bmatrix}, \qquad (2.3.25)$$

and

$$\begin{bmatrix} g_n \\ -\gamma_{n-1} \end{bmatrix} = B_n \dots B_{-1} \begin{bmatrix} g_0 \\ -\gamma_{-1} \end{bmatrix}, \quad n < 0$$

and for this system we may use the positive matrix technique from before. The eigenvalues of B_j in (2.3.25) are the same as those of A_j, but they switch positions in the corresponding eigenvectors $\tilde{r}_j^{\pm}$ compared to $r_j^{\pm}$ of A_j:

$$\tilde{r}_j^{\pm} = \begin{bmatrix} \dfrac{1}{\sqrt{1+\lambda_j^{\pm}}} \\ \pm\dfrac{1}{\sqrt{1+\lambda_j^{\mp}}} \end{bmatrix}, \qquad r_j^{\pm} = \begin{bmatrix} \dfrac{1}{\sqrt{1+\lambda_j^{\mp}}} \\ \pm\dfrac{1}{\sqrt{1+\lambda_j^{\pm}}} \end{bmatrix}.$$

The same argument as in the case of increasing n then proves the existence of v_0 giving exponentially decreasing solutions as $n \to -\infty$.

The uniqueness follows from the uniqueness of limits in (2.3.23), which for a given eigenvector v of A means that w is unique up to a constant factor. But then again, since we are in $\mathbb{R}^2$, the vector orthogonal to w is unique up to a constant factor. □

Remark 2.3.4 (Signs of the initial vectors). Here we underline that the form of $\Phi_{n,0}$ implies that the entries of $v_0^{\pm}$ in Lemma 2.3.3 must be nonzero, with opposite signs for v_0^- and same sign for v_0^+. Indeed, by (2.3.21) and (2.3.22) we have

$$\lim_{n\to+\infty} \left\| (\Phi_{n,0}) v_0^- \right\| = 0.$$

Let us then assume $v_0^- \neq 0$ with nonnegative entries of the same sign, namely $v_0^- \geq 0$ ($v_0^- \leq 0$) understood entrywise. From the definition (2.3.20) and $A_n \geq I$, it is clear that $(\Phi_{n,0})v_0^- \geq v_0^-$ ($(\Phi_{n,0})v_0^- \leq v_0^-$) for $n \geq 0$, and so it is impossible for the norm to tend to zero. Hence, the entries of v_0^- must be nonzero and of opposite sign. For $n \to -\infty$, we can use (2.3.25) and the same argument to arrive at the same conclusion for $[g_0 \ \ {-\gamma_{-1}}]^\top$, implying that $v_0^+ = [g_0 \ \ \gamma_{-1}]^\top$ has nonzero entries of equal sign.

Theorem 2.3.5 (Existence of a discrete Green's function). *Let $\{a_j\}_{j\in\mathbb{Z}}$ be a nonnegative sequence such that $a_j = 1+\mathrm{D}_+b_j$ with $\mathrm{D}_+b \in \boldsymbol{\ell}^2$. Then, for any given index i, there exists a unique sequence $g_i = \{g_{i,j}\}_{j\in\mathbb{Z}}$ such that*

$$(\mathrm{A}[a]g_i)_j = \frac{\delta_{i,j}}{\Delta\xi}. \tag{2.3.26}$$

Proof. Our strategy follows a standard approach for constructing Green's functions: We first construct solutions of the homogeneous version of (2.3.26) with exponential decay, and then we "glue" them together making sure we obtain a delta function at a given point. We start by constructing $g_{0,j}$ centered at $i = 0$.

Choosing $v_0^\pm$ from Lemma 2.3.3 we set

$$\begin{bmatrix} g_0^- \\ \gamma_{-1}^- \end{bmatrix} := v_0^-, \qquad \begin{bmatrix} g_0^+ \\ \gamma_{-1}^+ \end{bmatrix} := v_0^+, \tag{2.3.27}$$

and define the sequences

$$\begin{bmatrix} g_n^- \\ \gamma_{n-1}^- \end{bmatrix} := \Phi_{n,0}\begin{bmatrix} g_0^- \\ \gamma_{-1}^- \end{bmatrix}, \quad \begin{bmatrix} g_n^+ \\ \gamma_{n-1}^+ \end{bmatrix} := \Phi_{n,0}\begin{bmatrix} g_0^+ \\ \gamma_{-1}^+ \end{bmatrix}, \quad n \in \mathbb{Z}, \tag{2.3.28}$$

where by construction $g^\pm, \gamma^\pm$ are exponentially decreasing for $n \to \mp\infty$. Then, applying the operator $\mathrm{A}[a]$ to $g^\pm$ we find

$$(\mathrm{A}[a]g^\pm)_j = a_j g_j^\pm - \mathrm{D}_-\gamma_j^\pm = 0, \qquad j \in \mathbb{Z}$$

by construction of $g^\pm$ and $\gamma^\pm$. Let us then define

$$g_{0,j} := C\begin{cases} g_j^- g_0^+, & j \geq 0, \\ g_j^+ g_0^-, & j < 0, \end{cases} \qquad \gamma_{0,j} := C\begin{cases} \gamma_j^- g_0^+, & j \geq 0, \\ \gamma_j^+ g_0^-, & j < 0 \end{cases} \tag{2.3.29}$$

for some hitherto unspecified constant C, and observe from the homogeneous equation that $a_j g_{0,j} - \mathrm{D}_-\gamma_{0,j} = 0$ for $j \neq 0$. Moreover, we have $\mathrm{D}_+g_{0,j} = a_j\gamma_{0,j}$ for all j by construction. Now we would like to show that the constant C can be chosen to obtain $\mathrm{A}[a]g_{0,0} = 1/\varDelta\xi$. From (2.3.5), we get

$$\varDelta\xi \sum_{j=m}^{n} g_j^+ (\mathrm{A}[a]g^-)_j - \varDelta\xi \sum_{j=m}^{n} (\mathrm{A}[a]g^+)_j g_j^- = W_n(g^-, g^+) - W_{m-1}(g^-, g^+), \tag{2.3.30}$$

where we in the spirit of [41, Eq. (1.21)] have defined a discrete Wronskian

$$W_n(g^-, g^+) \coloneqq g_{n+1}^- \gamma_n^+ - g_{n+1}^+ \gamma_n^- = g_n^- \gamma_n^+ - g_n^+ \gamma_n^-, \tag{2.3.31}$$

and the last equality follows from the identity $g_{n+1}^{\pm} = g_n^{\pm} + \varDelta\xi a_n \gamma_n^{\pm}$. Since the left-hand side of (2.3.30) vanishes by definition of $g^{\pm}$, we have $W_n(g^-, g^+) = W_{m-1}(g^-, g^+)$ for any $n, m \in \mathbb{Z}$. That is, the Wronskian $W_n(g^-, g^+)$ is a constant $W(g^-, g^+)$ for the constructed sequences g^+ and g^-. An alternative way to see this can be found in Remark 2.3.6.

Next, we want to show that the Wronskian is nonzero. Considering

$$W(g^-, g^+) = W_{-1}(g^-, g^+) = g_0^- \gamma_{-1}^+ - g_0^+ \gamma_{-1}^- = g_0^+ \gamma_{-1}^- + g_0^- (-\gamma_{-1}^+)$$

and the definition (2.3.27), we use the sign properties stated in Remark 2.3.4 to conclude that the two terms in the final sum are always nonzero and of the same sign, implying $W(g^-, g^+) \neq 0$. Finally, we will determine the constant C by considering the backward difference

$$\begin{aligned}\mathrm{D}_{j-}\gamma_{0,0} &= C\frac{\gamma_0^- g_0^+ - \gamma_{-1}^+ g_0^-}{\varDelta\xi} = C\frac{\gamma_0^- g_0^+ - \gamma_{-1}^- g_0^+ + \gamma_{-1}^- g_0^+ - \gamma_{-1}^+ g_0^-}{\varDelta\xi} \\ &= C g_0^+ a_0 g_0^- - C\frac{W_{-1}(g^-, g^+)}{\varDelta\xi} = a_0 g_{0,0} - C\frac{W(g^-, g^+)}{\varDelta\xi},\end{aligned}$$

which leads to

$$(\mathrm{A}[a]g_0)_0 = a_0 g_{0,0} - \mathrm{D}_-\gamma_{0,0} = C\frac{W(g^-, g^+)}{\varDelta\xi}.$$

Consequently, setting $C^{-1} = W(g^-, g^+)$ in (2.3.29) gives the desired Green's function.

Note that there is nothing special about the index $i = 0$ where we centered the Green's function. We can simply use the sequences (2.3.28) from before and define

$$g_{i,j} = \frac{1}{W(g^-, g^+)} \begin{cases} g_j^+ g_i^-, & j \geq i, \\ g_j^- g_i^+, & j < i, \end{cases} \qquad \gamma_{i,j} = \frac{1}{W(g^-, g^+)} \begin{cases} \gamma_j^+ g_i^-, & j \geq i, \\ \gamma_j^- g_i^+, & j < i \end{cases} \tag{2.3.32}$$

to obtain a Green's function $g_{i,j}$ centered at an arbitrary i.

The uniqueness of $g_{i,j}$ follows from the vectors $v_0^{\pm}$ in Lemma 2.3.3 being uniquely defined up to constant factors. Indeed, when constructing the Green's function in (2.3.32) these factors disappear since we are dividing by the Wronskian $W(g^-, g^+)$, and so we have no degrees of freedom left in our construction of $g_{i,j}$, hence it is unique. □

Remark 2.3.6. The constancy of the Wronskian (2.3.31) can be derived in an alternative way using only (2.3.14). Observe that

$$W_{n-1}(g^-, g^+) = \begin{bmatrix} g_n^+ & \gamma_{n-1}^+ \end{bmatrix} \begin{bmatrix} -\gamma_{n-1}^- \\ g_n^- \end{bmatrix}.$$

Without loss of generality we may assume $n \geq k$, and transposing (2.3.14) we find

$$\begin{bmatrix} g_n^+ & \gamma_{n-1}^+ \end{bmatrix} = \begin{bmatrix} g_k^+ & \gamma_{k-1}^+ \end{bmatrix} (\Phi_{n,k})^\top.$$

On the other hand, by interchanging rows (2.3.14) can be written

$$\begin{bmatrix} -\gamma_{n-1}^- \\ g_n^- \end{bmatrix} = \begin{bmatrix} 1 & -a_{n-1}\mathit{\Delta}\xi \\ -a_{n-1}\mathit{\Delta}\xi & 1 + (a_{n-1}\mathit{\Delta}\xi)^2 \end{bmatrix} \begin{bmatrix} -\gamma_{n-2}^- \\ g_{n-1}^- \end{bmatrix} = (A_{n-1})^{-1} \begin{bmatrix} -\gamma_{n-2}^- \\ g_{n-1}^- \end{bmatrix},$$

which leads to

$$\begin{bmatrix} -\gamma_{n-1}^- \\ g_n^- \end{bmatrix} = (\Phi_{n,k})^{-\top} \begin{bmatrix} -\gamma_{k-1}^- \\ g_k^- \end{bmatrix}.$$

It is then clear that

$$W_{n-1}(g^-, g^+) = \begin{bmatrix} g_k^+ & \gamma_{k-1}^+ \end{bmatrix} (\Phi_{n,k})^\top (\Phi_{n,k})^{-\top} \begin{bmatrix} -\gamma_{k-1}^- \\ g_k^- \end{bmatrix} = W_{k-1}(g^-, g^+),$$

which is what we claimed.

Note that $\mathrm{A}[a]$ is not the only way to discretize the operator

$$a(\xi)\,\mathrm{Id} - \frac{\partial}{\partial\xi}\frac{1}{a(\xi)}\frac{\partial}{\partial\xi}$$

with first order differences, we may also consider

$$(\mathrm{B}[a]k)_j := a_j k_j - \mathrm{D}_+\left(\frac{\mathrm{D}_- k_j}{a_j}\right). \tag{2.3.33}$$

In fact, we will need the Green's function for this operator as well to close our upcoming system of differential equations. Fortunately, the existence of Green's function for (2.3.33) follows from the considerations already made in Theorem 2.3.5.

Corollary 2.3.7. *Under the same assumptions on* $\{a_j\}_{j\in\mathbb{Z}}$ *as in Theorem 2.3.5, for any given index* i *there exists a unique sequence* $k_i = \{k_{i,j}\}_{j\in\mathbb{Z}}$ *such that*

$$(\mathrm{B}[a]k_i)_j = \frac{\delta_{i,j}}{\varDelta\xi}. \tag{2.3.34}$$

Proof of Corollary 2.3.7. By manipulating the homogeneous version of (2.3.34) we find it to be equivalent to

$$\frac{\mathrm{D}_- k_{j+1}}{a_{j+1}} = \varDelta\xi a_j k_j + \frac{\mathrm{D}_- k_j}{a_j}.$$

Introducing

$$\kappa_j = \frac{\mathrm{D}_- k_j}{a_j} = \frac{k_j - k_{j-1}}{a_j\varDelta\xi}, \tag{2.3.35}$$

the previous equation can be written as

$$\begin{bmatrix}\kappa_{j+1}\\ k_j\end{bmatrix} = \begin{bmatrix}1+(a_j\varDelta\xi)^2 & a_j\varDelta\xi\\ a_j\varDelta\xi & 1\end{bmatrix}\begin{bmatrix}\kappa_j\\ k_{j-1}\end{bmatrix} = A_j\begin{bmatrix}\kappa_j\\ k_{j-1}\end{bmatrix},$$

where we recognize the matrix A_j from (2.3.14). Going backward we find

$$\begin{bmatrix}\kappa_j\\ k_{j-1}\end{bmatrix} = \begin{bmatrix}1 & -a_j\varDelta\xi\\ -a_j\varDelta\xi & 1+(a_j\varDelta\xi)^2\end{bmatrix}\begin{bmatrix}\kappa_{j+1}\\ k_j\end{bmatrix},$$

or equivalently

$$\begin{bmatrix}-\kappa_j\\ k_{j-1}\end{bmatrix} = \begin{bmatrix}1 & a_j\varDelta\xi\\ a_j\varDelta\xi & 1+(a_j\varDelta\xi)^2\end{bmatrix}\begin{bmatrix}-\kappa_{j+1}\\ k_j\end{bmatrix} = B_j\begin{bmatrix}-\kappa_{j+1}\\ k_j\end{bmatrix}$$

with B_j from (2.3.25). Hence, we get the solution for free from 2.3.5. Indeed, choosing

$$\begin{bmatrix}\kappa_n^-\\ k_{n-1}^-\end{bmatrix} = \begin{bmatrix}g_n^-\\ \gamma_{n-1}^-\end{bmatrix}, \qquad \begin{bmatrix}-\kappa_n^+\\ k_{n-1}^+\end{bmatrix} = \begin{bmatrix}g_n^+\\ -\gamma_{n-1}^+\end{bmatrix}$$

we know that these sequences have the correct decay at infinity. Defining

$$\begin{aligned}
k_{i,j} &= \frac{1}{W(g^-,g^+)}\begin{cases}k_j^-k_i^+, & j>i,\\ k_j^+k_i^-, & j\le i,\end{cases} = \frac{-1}{W(g^-,g^+)}\begin{cases}\gamma_j^-\gamma_i^+, & j>i,\\ \gamma_j^+\gamma_i^-, & j\le i,\end{cases}\\
\kappa_{i,j} &= \frac{1}{W(g^-,g^+)}\begin{cases}\kappa_j^-k_i^+, & j>i,\\ \kappa_j^+k_i^-, & j\le i,\end{cases} = \frac{-1}{W(g^-,g^+)}\begin{cases}g_j^-\gamma_i^+, & j>i,\\ g_j^+\gamma_i^-, & j\le i,\end{cases}
\end{aligned} \tag{2.3.36}$$

it follows from (2.3.13) that $(\mathrm{B}[a]k_i)_j = a_jk_{i,j} - \mathrm{D}_{j+}\kappa_{i,j} = 0$ for $j \neq i$. Moreover, by the constancy of (2.3.31) we find $(\mathrm{B}[a]k_i)_i = 1/\varDelta\xi$ in the same way as for $(\mathrm{A}[a]g_i)_i$. □

Remark 2.3.8. Note that we may observe directly from (2.3.32) and (2.3.36) that $g_{i,j} = g_{j,i}$, $k_{i,j} = k_{j,i}$, and $\kappa_{i,j} = -\gamma_{j,i}$. Moreover, the eigenvalues

$$\lambda^{\pm} = \frac{1}{2}\left(2 + \Delta\xi^2 \pm \Delta\xi\sqrt{4 + \Delta\xi^2}\right)$$

are exactly those found in (2.3.10) for the operator $\mathrm{Id} - \mathrm{D}_-\mathrm{D}_+$. In fact, for $a_j \equiv 1$ the sequences g and k coincide since $\mathrm{D}_-\mathrm{D}_+ = \mathrm{D}_+\mathrm{D}_-$, and their explicit expression (2.3.10) can be recovered from the columns of $\Lambda^n R^{-1}$ in the diagonalization $A^n = R\Lambda^n R^{-1}$.

Observe that by (2.2.16), (2.3.12), (2.3.33), and (2.3.35) we can rewrite (2.3.26) and (2.3.34) in the compact form

$$\begin{bmatrix} -\mathrm{D}_{j-} & a_j \\ a_j & -\mathrm{D}_{j+} \end{bmatrix} \begin{bmatrix} \gamma_{i,j} & k_{i,j} \\ g_{i,j} & \kappa_{i,j} \end{bmatrix} = \frac{1}{\Delta\xi} \begin{bmatrix} \delta_{i,j} & 0 \\ 0 & \delta_{i,j} \end{bmatrix}. \tag{2.3.37}$$

Lemma 2.3.9 (Sign properties of the discrete Green's functions)**.** *Assume $a_j \geq 0$ for $j \in \mathbb{Z}$, and let g, γ, k, and κ be solutions of* (2.3.37) *which decay to zero for $|j - i| \to +\infty$. Then the following sign properties hold,*

(i) $g_{i,j} > 0$ *and* $k_{i,j} > 0$ *for* $j \in \mathbb{Z}$,

(ii) $\operatorname{sgn}(\gamma_{i,j}) = \operatorname{sgn}(i - j - 1/2)$ *and* $\operatorname{sgn}(\kappa_{i,j}) = \operatorname{sgn}(i - j + 1/2)$.

In particular, this leads to the monotonicity properties

$$\begin{aligned} &\max_{j\in\mathbb{Z}} g_{i,j} = g_{i,i}, \quad \lim_{|j-i|\to+\infty} g_{i,j} \searrow 0, \\ &\max_{j\in\mathbb{Z}} k_{i,j} = k_{i,i}, \quad \lim_{|j-i|\to+\infty} k_{i,j} \searrow 0, \end{aligned} \tag{2.3.38}$$

where the arrows denote monotone decrease.

In Figure 2.3 we have included a sketch of $g_{i,n}$, $\gamma_{i,n}$, $k_{i,n}$, and $\kappa_{i,n}$ for $\Delta\xi = 0.2$, $i = 0, 4$ and $a_n = a(n\Delta\xi)$ given by

$$a(\xi) = \begin{cases} 2, & -1 < \xi \leq 0.5, \\ 0, & 0.5 < \xi \leq 1, \\ 4, & 1 < \xi \leq 1.5 \\ 1, & \text{otherwise.} \end{cases} \tag{2.3.39}$$

We say sketch, as they have been computed on a finite grid where $n \in \{-20, \ldots, 20\}$ with boundary conditions $\gamma_{i,-21} = g_{i,21} = 0$ and $k_{i,-21} = \kappa_{i,21} = 0$, and consequently we find that neither of $g_{i,-20}$, $\gamma_{i,20}$, $\kappa_{i,-20}$ or

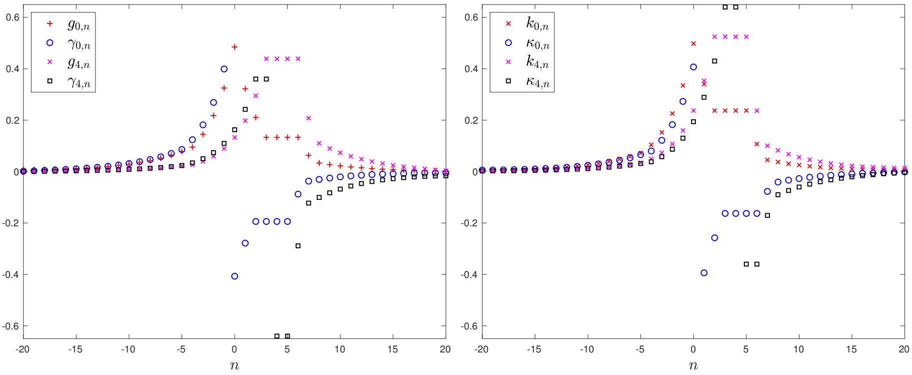


Figure 2.3: Sketch of $g_{i,n}$, $\gamma_{i,n}$, $k_{i,n}$, and $\kappa_{i,n}$ for $\Delta\xi = 0.2$, $i = 0, 4$ and $a_n = a(n\Delta\xi)$ for $a(\xi)$ defined in (2.3.39). Note the jump of size $-1 + \mathcal{O}(\Delta\xi)$ at $n = i$ for both γ and κ.

$k_{i,20}$ are exactly zero. However, the exponential decay makes them very small and the qualitative behavior indicated in Lemma 2.3.9 is still the same. Note how $a(\xi)$ being zero on the interval $(0.5, 1]$ leads to constant kernel values in that neighborhood, even at the peaks for the kernels centered at $\xi_4 = 0.8$.

Proof of Lemma 2.3.9. We prove this only for g and γ as the proof for k and κ is similar. The proof relies on the reasoning in Remark 2.3.4.

As a first step we want to show that the properties (i) and (ii) hold for $g_{i,i}$, $g_{i,i+1}$, $\gamma_{i,i-1}$, and $\gamma_{i,i}$. To this end, we recall from the proof of Theorem 2.3.5 that since $g_{i,j}$ and $\gamma_{i,j}$ satisfy (2.3.37), they must also satisfy

$$\begin{bmatrix} g_{i,j} \\ -\gamma_{i,j-1} \end{bmatrix} = B_j \cdots B_{i-1} \begin{bmatrix} g_{i,i} \\ -\gamma_{i,i-1} \end{bmatrix}, \qquad j \leq i-1$$

and

$$\begin{bmatrix} g_{i,j} \\ \gamma_{i,j-1} \end{bmatrix} = A_{j-1} \cdots A_{i+1} \begin{bmatrix} g_{i,i+1} \\ \gamma_{i,i} \end{bmatrix}, \qquad j \geq i+2,$$

with A_k and B_k as defined in (2.3.14) and (2.3.25). By our assumptions, the Green's functions must tend to zero asymptotically, and we recall from Remark 2.3.4 that a necessary condition for this is for the vectors $[g_{i,i}, -\gamma_{i,i-1}]^\top$ and $[g_{i,i+1}, \gamma_{i,i}]^\top$ to have entries of opposite sign. Hence, $g_{i,i}\gamma_{i,i-1} > 0$ and $g_{i,i+1}\gamma_{i,i} < 0$, where we stress the importance of $a_j \geq 0$ for this argument to hold. Using only (2.3.37) we calculate

$$\begin{aligned} 0 &> g_{i,i+1}\gamma_{i,i} - g_{i,i}\gamma_{i,i-1} \\ &= \Delta\xi \frac{(g_{i,i+1} - g_{i,i})\gamma_{i,i} + g_{i,i}(\gamma_{i,i} - \gamma_{i,i-1})}{\Delta\xi} \end{aligned}$$

$$
\begin{aligned}
&= \Delta\xi \left[a_i \gamma_{i,i}\gamma_{i,i} + g_{i,i} \left[a_i g_{i,i} - \frac{1}{\Delta\xi} \right] \right] \\
&= \Delta\xi a_i \left[(g_{i,i})^2 + (\gamma_{i,i})^2 \right] - g_{i,i}.
\end{aligned}
$$

Since $a_j \geq 0$, it follows that $g_{i,i} \geq 0$. Recalling that $g_{i,i}$ must be nonzero according to the sign requirements, we necessarily have $g_{i,i} > 0$, and then $\gamma_{i,i-1} > 0$ follows. Moreover, multiplying the identity $g_{i,i+1} - \Delta\xi a_i \gamma_{i,i} = g_{i,i}$ by $g_{i,i+1}$ and using $a_i \geq 0$, $g_{i,i} > 0$, and $g_{i,i+1}\gamma_{i,i} < 0$, we must have $g_{i,i+1} > 0$, which then implies $\gamma_{i,i} < 0$.

Next we must prove that (i) and (ii) hold for the remaining values of j, and this will be achieved with a contradiction argument. We define the vectors

$$
v_j^+ := \begin{bmatrix} g_{i,j} \\ \gamma_{i,j-1} \end{bmatrix}, \qquad v_j^- := \begin{bmatrix} g_{i,j+1} \\ -\gamma_{i,j} \end{bmatrix}
$$

such that v_{i+1}^+ and v_{i-1}^- both have positive first component and negative second component, and satisfy

$$
v_{j+1}^+ := A_j v_j^+ \text{ for } j \geq i+1, \qquad v_{j-1}^- := B_j v_j^- \text{ for } j \leq i-1.
$$

If we can prove that they retain the sign property under the above propagation, then we are done. Let us consider

$$
v_{j+1}^+ := A_j v_j^+, \qquad j \geq i+1.
$$

Assume that v_j^+ does not retain the sign property, then there is some $k \geq i+1$ which is the first index such that v_{k+1}^+ does not have a positive first component and negative second component. We consider the two possible cases.

The first case is $v_{k+1}^+ \geq 0$ ($v_{k+1}^+ \leq 0$) considered entrywise. First of all, v_{k+1}^+ cannot be the zero vector as A_k has full rank, since then v_k^+ would also have to be zero, which contradicts $k+1$ being the first problematic index. Otherwise, the entrywise inequality $A_{k+1} \geq I$ leads to $v_{k+2}^+ = A_{k+1}v_{k+1}^+ \geq v_{k+1}^+$ ($v_{k+2}^+ \leq v_{k+1}^+$), and thus $\lim_{n\to+\infty} v_n^+ \geq v_{k+1}^+$ ($\lim_{n\to+\infty} v_n^+ \leq v_{k+1}^+$). This is however impossible, as it contradicts the assumed decay of the Green's functions.

The remaining case is that the entries interchange sign from v_k^+ to v_{k+1}^+. However, then we would have

$$
v_k^+ = (A_k)^{-1} v_{k+1}^+ = \begin{bmatrix} 1 & -a_k\Delta\xi \\ -a_k\Delta\xi & 1 + (a_k\Delta\xi)^2 \end{bmatrix} v_{k+1}^+.
$$

Since $a_k \geq 0$, v_k^+ would also have negative first component and positive second component, which contradicts $k+1$ being the first problematic index. Hence, v_j^+ always has positive first component and negative second component for $j > i$, thus for $j \geq i$ it follows that $g_{i,j}$ is always positive, while $\gamma_{i,j}$ is always negative which shows that $g_{i,j}$ is decreasing in this direction.

A similar argument holds in the other direction when considering v_j^- and B_j. Then $-\gamma_{i,j}$ is always negative for $j < i$, which means that $g_{i,j}$ is increasing with j for these indices. Thus, (i) and (ii) hold for $\{g_{i,j}\}_{j\in\mathbb{Z}}$ and $\{\gamma_{i,j}\}_{j\in\mathbb{Z}}$. □

2.4 An equivalent semi-discrete system for global solutions in time

We now return to the initial value problem (2.1.2). We use the Lagrangian formulation introduced in earlier works, see [30], but reformulate the governing equations by including the fundamental solutions of the momentum operator in the solution.

2.4.1 Reformulation of the continuous problem using operator propagation

The 2CH system can be written as

$$u_t + uu_x + P_x = 0, \quad \rho_t + (u\rho)_x = 0$$

for P implicitly defined by

$$P - P_{xx} = u^2 + \frac{1}{2}u_x^2 + \frac{1}{2}\rho^2. \tag{2.4.1}$$

Let us introduce $\bar{\rho} \coloneqq \rho - \rho_\infty \in \mathbf{L}^2$ to ease notation. Note that most expressions simplify when we consider $\rho_\infty = 0$. We have chosen to cover the case of arbitrary ρ_∞, to allow for the initial condition $\rho(0,x) = \varepsilon$, for any $\varepsilon > 0$. Such initial data lead to solutions without blow-up, see [28]. In the case of the 2CH system, the conservation law for the energy is given by

$$(\tfrac{1}{2}(u^2 + u_x^2 + \bar{\rho}^2))_t + (u\tfrac{1}{2}(u^2 + u_x^2 + \bar{\rho}^2))_x + (uR)_x = 0, \tag{2.4.2}$$

where we have used P from (2.4.1) to define

$$R = P - \tfrac{1}{2}u^2 - \tfrac{1}{2}\rho_\infty^2.$$

We can check that the first order system

$$\begin{bmatrix} -\partial_x & 1 \\ 1 & -\partial_x \end{bmatrix} \circ \begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} uu_x \\ \tfrac{1}{2}(u^2 + u_x^2 + \bar{\rho}^2) + \rho_\infty\bar{\rho} \end{bmatrix} \tag{2.4.3}$$

is equivalent to (2.4.1). Hence,

$$u_t + uu_x + Q = 0, \tag{2.4.4a}$$
$$\rho_t + (u\rho)_x = 0 \tag{2.4.4b}$$

and (2.4.3) is yet another form of the 2CH system.

We introduce as before the Lagrangian position $y(t,\xi)$ and velocity $U(t,\xi)$. Moreover, we define the Lagrangian density

$$r(t,\xi) := \rho(t,y(t,\xi))y_\xi(t,\xi),$$

and the cumulative energy H given by

$$\begin{aligned} H(t,\xi) &= \frac{1}{2}\int_{-\infty}^{y(t,\xi)} (u^2 + u_x^2 + \bar{\rho}^2)(t,x)\,dx \\ &= \frac{1}{2}\int_{-\infty}^{\xi} ((u^2 + u_x^2 + \bar{\rho}^2)\circ y)y_\xi(t,\eta)\,d\eta, \end{aligned}$$

as well as the Lagrangian variables $\bar{Q} = Q \circ y$ and $\bar{R} = R \circ y$. From (2.4.4), we get $U_t = -\bar{Q}$ and $r_t = 0$, while the conservation of energy (2.4.2) yields $H_t = -U\bar{R}$. Finally, we also rewrite the system (2.4.3) in terms of the Lagrangian variables. To simplify the notation, we replace $\bar{Q}$ by Q, and similarly for $\bar{R}$. The equivalent system in Lagrangian variables is then given by

$$y_t = U, \tag{2.4.5a}$$
$$U_t = -Q, \tag{2.4.5b}$$
$$H_t = -UR, \tag{2.4.5c}$$
$$r_t = 0, \tag{2.4.5d}$$

with

$$\begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} UU_\xi \\ H_\xi + \rho_\infty(r - \rho_\infty y_\xi) \end{bmatrix}. \tag{2.4.6}$$

In (2.4.6) we use the same notation for the variable y_ξ and the operator for point-wise multiplication by y_ξ. We will use this convention for the rest of the paper. The equivalence between (2.4.3) and (2.4.6) holds only assuming the that $y_\xi \geq 0$ and all the functions are smooth enough to do the manipulation.

Note that we need to decompose the variables y and r in (2.4.5) to give them a decay which enables us to define them in a proper functional setting. We define ζ and $\bar{r}$ as

$$y(t,\xi) = \zeta(t,\xi) + \xi \quad \text{and} \quad r(t,\xi) = \bar{r}(t,\xi) + \rho_\infty y_\xi(t,\xi).$$

The Banach space which contains ζ and H is the subspace of bounded and continuous functions with derivative in $\mathbf{L}^2$,

$$\mathbf{V} \coloneqq \{f \in \mathbf{C}_{\mathrm{b}}(\mathbb{R}) \mid f_\xi \in \mathbf{L}^2(\mathbb{R})\}, \tag{2.4.7}$$

endowed with the norm $\|f\|_{\mathbf{V}} \coloneqq \|f\|_{\mathbf{L}^\infty} + \|f_\xi\|_{\mathbf{L}^2}$. Recall that we defined a discrete version of (2.4.7) in (2.3.2). Then we let

$$\mathbf{E} \coloneqq \mathbf{V} \times \mathbf{H}^1 \times \mathbf{V} \times \mathbf{L}^2 \tag{2.4.8}$$

be a Banach space tailored for the tuple $X = (\zeta, U, H, \bar{r})$ with norm

$$\|X\|_{\mathbf{E}} \coloneqq \|\zeta\|_{\mathbf{V}} + \|U\|_{\mathbf{H}^1} + \|H\|_{\mathbf{V}} + \|\bar{r}\|_{\mathbf{L}^2}. \tag{2.4.9}$$

The unique solution of (2.4.5), as studied in [30], is then completely described by this tuple.

Let us define the operators $\mathcal{G}$ and $\mathcal{K}$ as the fundamental solutions to the operator in (2.4.6), meaning that they satisfy

$$\begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix}. \tag{2.4.10}$$

As we mentioned in the introduction, the operators $\mathcal{K}$ and $\mathcal{G}$ can be computed explicitly using the fundamental solution of the Helmholtz operators in Eulerian coordinates. If we define

$$g(\eta, \xi) = \frac{1}{2} e^{-|y(\xi) - y(\eta)|} \tag{2.4.11a}$$

and

$$\kappa(\eta, \xi) = -\frac{1}{2} \operatorname{sgn}(\xi - \eta) e^{-|y(\xi) - y(\eta)|}, \tag{2.4.11b}$$

then we can check that the operators defined as $\mathcal{G}(f) = \int_{\mathbb{R}} g(\eta, \xi) f(\eta)\, d\eta$ and $\mathcal{K}(f) = \int_{\mathbb{R}} \kappa(\eta, \xi) f(\eta)\, d\eta$ are solutions to (2.4.10), again assuming y is monotone increasing in ξ. This means that we can obtain explicit expressions for R and Q given by

$$\begin{aligned} R = {} & \int_{\mathbb{R}} \kappa(\eta, \xi) U(\eta) U_\xi(\eta)\, d\eta \\ & + \int_{\mathbb{R}} g(\eta, \xi)(H_\xi(\eta) + \rho_\infty(r(\eta) - \rho_\infty y_\xi(\eta)))\, d\eta, \end{aligned} \tag{2.4.12a}$$

$$\begin{aligned} Q = {} & \int_{\mathbb{R}} g(\eta, \xi) U(\eta) U_\xi(\eta)\, d\eta \\ & + \int_{\mathbb{R}} \kappa(\eta, \xi)(H_\xi(\eta) + \rho_\infty(r(\eta) - \rho_\infty y_\xi(\eta)))\, d\eta. \end{aligned} \tag{2.4.12b}$$

In [36, 30], the authors prove that the right-hand side of their respective versions of (2.4.5) is locally Lipschitz, and consecutive contraction arguments yield the existence of a unique short-time solution. In the same manner, we would like to prove that there exists a unique short-time solution for our semi-discrete system, but the explicit forms for g and κ in (2.4.11) are not available in the discrete setting. As a remedy, we propagate the kernel operators corresponding to $\mathcal{K}$ and $\mathcal{G}$ by incorporating them in the governing equations. Given the evolution of y, that is, $y_t = U$, we can derive evolution equations for $\mathcal{G}$ and $\mathcal{K}$. Let us see how this can be done in the continuous case before dealing with the discrete case. Formally we have

$$\begin{aligned}\frac{\partial \mathcal{G}(f)}{\partial t} &= \frac{1}{2}\int_{\mathbb{R}} \frac{\partial}{\partial t} e^{-|y(t,\xi)-y(t,\eta)|} f(\eta)\,d\eta \\ &= \int_{\mathbb{R}} \frac{\operatorname{sgn}(y(t,\eta)-y(t,\xi))}{2}(y_t(t,\xi)-y_t(t,\eta))e^{-|y(t,\xi)-y(t,\eta)|}f(\eta)\,d\eta \\ &= \int_{\mathbb{R}} \frac{\operatorname{sgn}(y(t,\eta)-y(t,\xi))}{2}(U(t,\xi)-U(t,\eta))e^{-|y(t,\xi)-y(t,\eta)|}f(\eta)\,d\eta.\end{aligned}$$

Here we again assume that we know *a priori* that y remains a monotone function with respect to ξ. Then, we can rewrite the last equality as

$$\frac{\partial}{\partial t}\mathcal{G}(f) = -\frac{1}{2}\int_{\mathbb{R}} \operatorname{sgn}(\xi-\eta)(U(t,\xi)-U(t,\eta))e^{-|y(t,\xi)-y(t,\eta)|}f(\eta)\,d\eta. \tag{2.4.13}$$

For a function U, we can associate a pointwise multiplication operator, which we denote by $\mathcal{U}$. That is, we write $\mathcal{U}(f)(\xi) = U(\xi)f(\xi)$ for any function f and any point ξ. The integral kernel of $\mathcal{U}$ would be singular and equal to $U(\xi)\delta(\xi-\eta)$. Using this notation, we can rewrite (2.4.13) as

$$\frac{\partial}{\partial t}\mathcal{G}(f) = (\mathcal{U}\circ\mathcal{K})(f) - (\mathcal{K}\circ\mathcal{U})(f).$$

This can equivalently be stated as

$$\frac{\partial}{\partial t}\mathcal{G} = [\mathcal{U},\mathcal{K}], \qquad \frac{\partial}{\partial t}\mathcal{K} = [\mathcal{U},\mathcal{G}], \tag{2.4.14}$$

where the evolution equation for $\mathcal{K}$ is derived analogously. An equivalent system of equations for the 2CH system is then given by

$$y_t = U, \quad U_t = -Q, \quad H_t = UR, \quad r_t = 0, \tag{2.4.15a}$$

$$\frac{\partial}{\partial t}\mathcal{G} = [\mathcal{U},\mathcal{K}], \quad \frac{\partial}{\partial t}\mathcal{K} = [\mathcal{U},\mathcal{G}], \tag{2.4.15b}$$

with R and Q given as

$$\begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} \circ \begin{bmatrix} UU_\xi \\ H_\xi + \rho_\infty(r - \rho_\infty y_\xi) \end{bmatrix} \tag{2.4.16}$$

Then, the new system (2.4.15) and (2.4.16) gives rise to the same solutions as the one given by (2.4.5), (2.4.11) and (2.4.12).

We note that the evolution equation for $\mathcal{G}$ and $\mathcal{K}$ can be obtained directly from the product identity (2.4.10). Indeed, after differentiation with respect to time, we get

$$\begin{bmatrix} 0 & U_\xi \\ U_\xi & 0 \end{bmatrix} \circ \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} + \begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} \dot{\mathcal{K}} & \dot{\mathcal{G}} \\ \dot{\mathcal{G}} & \dot{\mathcal{K}} \end{bmatrix} = 0,$$

which implies

$$\begin{bmatrix} \dot{\mathcal{K}} & \dot{\mathcal{G}} \\ \dot{\mathcal{G}} & \dot{\mathcal{K}} \end{bmatrix} = - \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix} \circ \begin{bmatrix} 0 & U_\xi \\ U_\xi & 0 \end{bmatrix} \circ \begin{bmatrix} \mathcal{K} & \mathcal{G} \\ \mathcal{G} & \mathcal{K} \end{bmatrix}. \tag{2.4.17}$$

This expression, which corresponds to $\frac{dM^{-1}}{dt} = -M^{-1}\frac{dM}{dt}M^{-1}$ for a matrix M, can be simplified to (2.4.14) using integration by parts. Then it follows from (2.4.17) that the identity (2.4.10) is preserved by the evolution equation.

The following proposition establishes properties and *a priori* bounds for the fundamental solutions g and κ. Those bounds are obvious from the explicit expressions given in (2.4.11), but we prove them here using the relation

$$\begin{bmatrix} -\partial_\xi & y_\xi \\ y_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} g \\ \kappa \end{bmatrix} = \begin{bmatrix} 0 \\ \delta \end{bmatrix}$$

which define them. Notice that parts of the proof resembles a standard proof of the Sobolev inequality

$$\|f\|_{\mathbf{L}^\infty} \leq \frac{1}{\sqrt{2}} \|f\|_{\mathbf{H}^1}. \tag{2.4.18}$$

Proposition 2.4.1. *Assume we have functions $g, \kappa \colon \mathbb{R}^2 \to \mathbb{R}$ satisfying* (2.4.10) *for $y_\xi(\xi) \geq 0$. Furthermore, assume $g(\eta,\xi) \geq 0$ and* $\operatorname{sgn}(\kappa(\eta,\xi)) = \operatorname{sgn}(\eta - \xi)$. *Then, for a given η, we have $|g(\eta,\xi)| = |\kappa(\eta,\xi)|$ for a.e. ξ, and we have the upper bound $|g(\eta,\xi)|, |\kappa(\eta,\xi)| \leq g(\eta,\eta) = \frac{1}{2}$.*

Proof of Proposition 2.4.1. We start by observing

$$(g(\eta,\xi))^2 = \frac{1}{2}\left[\int_{-\infty}^{\xi} \left(g(\eta,s)^2\right)_s ds - \int_{\xi}^{+\infty} \left(g(\eta,s)^2\right)_s ds\right]$$

$$= \int_{-\infty}^{\xi} g(\eta, s) g_s(\eta, s)\, ds - \int_{\xi}^{+\infty} g(\eta, s) g_s(\eta, s)\, ds,$$

Then, we have

$$\begin{aligned}(g(\eta,\eta))^2 &= \int_{-\infty}^{\eta} g(\eta, s) g_s(\eta, s)\, ds - \int_{\eta}^{+\infty} g(\eta, s) g_s(\eta, s)\, ds \\ &= \int_{-\infty}^{\eta} y_\xi(s) g(\eta, s)\kappa(\eta, s)\, ds - \int_{\eta}^{+\infty} y_\xi(s) g(\eta, s)\kappa(\eta, s)\, ds,\end{aligned}$$

where we use $\partial_\xi g = y_\xi \kappa$ from (2.4.10). It follows that

$$\begin{aligned}(g(\eta,\eta))^2 &= \int_{-\infty}^{+\infty} y_\xi(s) g(\eta, s)|\kappa(\eta, s)|\, ds \\ &\leq \frac{1}{2}\int_{-\infty}^{+\infty} y_s(s)\left[g(\eta,s)^2 + \kappa(\eta,s)^2\right] ds \\ &= \frac{1}{2}\int_{-\infty}^{+\infty} g(\eta, s)\left[y_\xi(s) g(\eta, s) - (\kappa(\eta, s))_s\right] ds.\end{aligned}$$

Hence, by $y_\xi g - \partial_\xi \kappa = \delta$ from (2.4.10), we find

$$(g(\eta,\eta))^2 \leq \frac{1}{2}\int_{-\infty}^{+\infty} g(\eta, s)\delta(\eta - s) ds = \frac{1}{2} g(\eta,\eta),$$

and the result $g(\eta,\eta) \leq \frac{1}{2}$ follows. Since $y_\xi(\xi) g(\eta,\xi) - \kappa_\xi(\eta,\xi) = \delta(\eta,\xi)$, we find that for $\xi < \eta$,

$$\begin{aligned}(\kappa(\eta,\xi))^2 &= 2\int_{-\infty}^{\xi} \kappa(\eta, s)\kappa_s(\eta, s)\, ds \\ &= 2\int_{-\infty}^{\xi} \kappa(\eta, s) y_\xi(s) g(\eta, s)\, ds \\ &= 2\int_{-\infty}^{\xi} g_s(\eta, s) g(\eta, s)\, ds \\ &= (g(\eta,\xi))^2,\end{aligned}$$

and consequently $\kappa(\eta,\xi) = g(\eta,\xi)$ for $\xi < \eta$, where we have used the sign properties of g and κ. In a similar way we find $\kappa(\eta,\xi) = -g(\eta,\xi)$ for $\xi > \eta$, obtaining $\lim_{\xi\to\eta\mp} \kappa(\eta,\xi) = \pm g(\eta,\eta)$. Moreover, since $\kappa_\xi(\eta,\xi) = y_\xi(\xi) g(\eta,\xi) \geq 0$ for $\xi \neq \eta$ we have

$$\sup_\xi |\kappa(\eta,\xi)| = \lim_{\xi\to\eta\mp} |\kappa(\eta,\xi)| \leq \frac{1}{2}.$$

In fact, we find that we must have equality, as the above limits show

$$\lim_{\xi\to\eta+}\kappa(\eta,\xi)-\lim_{\xi\to\eta-}\kappa(\eta,\xi)\geq 1.$$

Then, the jump condition for κ required at $\eta=\xi$ to obtain a delta implies that the difference above is exactly one, which can only happen if

$$\lim_{\xi\to\eta\mp}\pm\kappa(\eta,\xi)=g(\eta,\eta)=\frac{1}{2}.$$

Since $g_\xi(\eta,\xi)=y_\xi(\xi)\kappa(\eta,\xi)$ is positive for $\xi<\eta$, and negative for $\xi>\eta$, it follows that $g(\eta,\xi)\leq g(\eta,\eta)$. □

Due to our lack of explicit formulae for the discrete counterparts of g and κ, an argument similar to that of Proposition 2.4.1 will help us to establish bounds also in the discrete case.

2.4.2 Reformulation of the semi-discrete system using operator propagation

Turning back to the formal expression (2.2.17), we use the the Green's functions from Theorem 2.3.5 and Corollary 2.3.7 to write out the right-hand side explicitly. Considering (2.3.37) where we now have $a_j=\mathrm{D}_+y_j$, we observe that they correspond to the discrete versions of (2.4.10). Indeed, we have the following identity

$$\begin{bmatrix}-\mathrm{D}_{j-} & (\mathrm{D}_+y_j)\\ (\mathrm{D}_+y_j) & -\mathrm{D}_{j+}\end{bmatrix}\circ\begin{bmatrix}\gamma_{i,j} & k_{i,j}\\ g_{i,j} & \kappa_{i,j}\end{bmatrix}=\frac{1}{\varDelta\xi}\begin{bmatrix}\delta_{i,j} & 0\\ 0 & \delta_{i,j}\end{bmatrix}\tag{2.4.19}$$

which has to be compared with (2.4.10) in the continuous case. Thus, (2.2.17) can be rewritten as

$$\dot{U}_j=-\varDelta\xi\sum_{i\in\mathbb{Z}}g_{i,j}\left(U_i(\mathrm{D}_+U_i)+\mathrm{D}_-\left(\frac{h_i}{\mathrm{D}_+y_i}+\rho_\infty\frac{\bar{r}_i}{\mathrm{D}_+y_i}\right)\right),\tag{2.4.20}$$

where we have defined

$$\bar{r}_i:=\rho_{0,i}-\rho_\infty(\mathrm{D}_+y_i)\tag{2.4.21}$$

and

$$h_i:=\frac{1}{2}(U_i)^2(\mathrm{D}_+y_i)+\frac{1}{2}\frac{(\mathrm{D}_+U_i)^2}{\mathrm{D}_+y_i}+\frac{1}{2}\frac{\bar{r}_i^2}{\mathrm{D}_+y_i}.\tag{2.4.22}$$

From the expressions in (2.4.20) and (2.4.22), it seems that, if D_+y_i goes to zero for some index i and time t, then $\dot{U}_j$ and h_i blow up. However, it turns out that these quantities remain bounded, which allows us to extend the solution globally in time. To obtain a well-defined system, we are going to remove the explicit dependence on $1/\mathrm{D}_+y_i$ by adding h to the set of variables of the system.

With the discrete kernels g, k, γ, and κ, we are able to express $\mathrm{A}[\mathrm{D}_+y]^{-1}$ in (2.2.17) to obtain (2.4.20). However, since we do not know their explicit form as functions of D_+y_j, we derive a system analogous to (2.4.15) by introducing g, k, γ, and κ as variables. To compute the evolution of g, k, γ, and κ, we repeat the procedure from the continuous case. By differentiating (2.4.19) and using the fact that $\dot{y}_i = U_i$, we get

$$\begin{bmatrix} \dot{\gamma} & \dot{k} \\ \dot{g} & \dot{\kappa} \end{bmatrix} = - \begin{bmatrix} \gamma & k \\ g & \kappa \end{bmatrix} * \begin{bmatrix} 0 & \mathrm{D}_+U \\ \mathrm{D}_+U & 0 \end{bmatrix} \begin{bmatrix} \gamma & k \\ g & \kappa \end{bmatrix}$$

which in explicit form yields

$$\begin{aligned} \dot{g}_{i,j} &= -\kappa_{m,j} * ((\mathrm{D}_+U_m)\gamma_{i,m}) - g_{m,j} * ((\mathrm{D}_+U_m)g_{i,m}), \\ \dot{\gamma}_{i,j} &= -k_{m,j} * ((\mathrm{D}_+U_m)\gamma_{i,m}) - \gamma_{m,j} * ((\mathrm{D}_+U_m)g_{i,m}), \end{aligned} \tag{2.4.23}$$

and

$$\begin{aligned} \dot{k}_{i,j} &= -k_{m,j} * ((\mathrm{D}_+U_m)k_{i,m}) - \gamma_{m,j} * ((\mathrm{D}_+U_m)\kappa_{i,m}), \\ \dot{\kappa}_{i,j} &= -\kappa_{m,j} * ((\mathrm{D}_+U_m)k_{i,m}) - g_{m,j} * ((\mathrm{D}_+U_m)\kappa_{i,m}). \end{aligned}$$

Here we denote by $(g*f)_j$ the action of the operator $g_{i,j}$ as a summation kernel on a sequence f_i, defined as

$$(g*f)_j = \varDelta\xi \sum_{i\in\mathbb{Z}} g_{i,j} f_i.$$

For the operators, we introduce the following norms

$$\begin{aligned} \|g\|_{\boldsymbol{\ell}^p} &= \sup_i \|g_i\|_{\boldsymbol{\ell}^p} = \sup_i \left(\varDelta\xi \sum_{j\in\mathbb{Z}} |g_{i,j}|^p \right)^{\frac{1}{p}}, \\ \|g\|_{\boldsymbol{\ell}^\infty} &= \sup_i \left(\sup_j |g_{i,j}| \right). \end{aligned}$$

Moreover, for the kernel operator g we have that the transpose $g^\top$ of g is given by $(g^\top)_{i,j} = g_{j,i}$. Then, the following result, reminiscent of Young's convolution inequality, will prove useful.

Proposition 2.4.2 (Young's inequality for general operators)**.**

$$\|g*f\|_{\boldsymbol{\ell}^r} \leq \|g\|_{\boldsymbol{\ell}^q}^{\frac{q}{r}} \left\|g^\top\right\|_{\boldsymbol{\ell}^q}^{1-\frac{q}{r}} \|f\|_{\boldsymbol{\ell}^p}, \tag{2.4.24}$$

for

$$1+\frac{1}{r}=\frac{1}{p}+\frac{1}{q}, \qquad p,q,r\in[1,\infty],$$

with the convention $q/\infty=0$ for $q<\infty$, and $\infty/\infty=1$.

We refer to Appendix 2.A for a proof of Proposition 2.4.2 . Note that the standard Young's inequality is usually given for a translation invariant kernel where g takes the form $g_{i,j}=\hat{g}_{i-j}$ for some sequence $\hat{g}$. For an operator of this form, we can check that $g^{\top}=\tau\circ g\circ\tau$, where the operator τ inverts the indexing, that is $\tau(f)_j=f_{-j}$. Since the operator τ is an isometry in all $\boldsymbol{\ell}^q$-spaces, the expression (2.4.24) simplifies to

$$\|g*f\|_{\boldsymbol{\ell}^r}\leq\|g\|_{\boldsymbol{\ell}^q}\|f\|_{\boldsymbol{\ell}^p}.$$

Now that we have evolution equations and norms for the kernels, we are set to prove some fundamental properties in the next lemma.

Lemma 2.4.3 (Preservation of identities)**.** *Let $T>0$, and assume that, for $t\in[0,T]$, $(\mathrm{D}_+y_j(t))_t=\mathrm{D}_+U_j(t)$ for $j\in\mathbb{Z}$, and that g,k,γ,κ and D_+U are bounded in $\boldsymbol{\ell}^2$-norm in $[0,T]$. Then, for $t\in[0,T]$ the sequences $g_{i,j}(t)$, $k_{i,j}(t)$, $\gamma_{i,j}(t)$, $\kappa_{i,j}(t)$ satisfy the following identities:*

(i) The Green's function identities (2.4.19)*,*

(ii) The symmetry identities

$$g_{j,i}=g_{i,j} \quad \text{and} \quad k_{j,i}=k_{i,j}, \tag{2.4.25}$$

and the antisymmetry identity

$$\gamma_{j,i}=-\kappa_{i,j}. \tag{2.4.26}$$

Proof of Lemma 2.4.3. Recall from Remark 2.3.8 that these identities are satisfied for $t=0$ by construction. The rest of the proof then relies on Grönwall's inequality. *(i)*: We introduce the four operators z_l for $l=1,2,3,4$ defined as

$$z_{1,i,j}=(\mathrm{D}_+y_i)g_{i,j}-\mathrm{D}_{j-}\gamma_{i,j}-\frac{\delta_{i,j}}{\Delta\xi}, \quad z_{2,i,j}=(\mathrm{D}_+y_j)k_{i,j}-\mathrm{D}_{j+}\kappa_{i,j}-\frac{\delta_{i,j}}{\Delta\xi},$$
$$z_{3,i,j}=(\mathrm{D}_+y_j)\gamma_{i,j}-\mathrm{D}_{j+}g_{i,j}, \qquad z_{4,i,j}=(\mathrm{D}_+y_j)\kappa_{i,j}-\mathrm{D}_{j-}k_{i,j}.$$

Using $(\mathrm{D}_+y_j(t))_t=\mathrm{D}_+U_j(t)$ and (2.4.23) we find that

$$(z_{1,i,j})_t=(\mathrm{D}_+y_j)_tg_{i,j}+(\mathrm{D}_+y_j)\dot{g}_{i,j}-\mathrm{D}_{j-}\dot{\gamma}_{i,j}$$

$$
\begin{aligned}
&= (\mathrm{D}_+U_j)g_{i,j} - (\mathrm{D}_+y_j)\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)\left(g_{i,m}g_{m,j} + \gamma_{i,m}\kappa_{m,j}\right) \\
&\quad + \mathrm{D}_{j-}\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)\left(g_{i,m}\gamma_{m,j} + \gamma_{i,m}k_{m,j}\right) \\
&= (\mathrm{D}_+U_j)g_{i,j} - \Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)g_{i,m}\left((\mathrm{D}_+y_j)g_{m,j} - \mathrm{D}_{j-}\gamma_{m,j}\right) \\
&\quad - \Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)\gamma_{i,m}\left((\mathrm{D}_+y_j)\kappa_{m,j} - \mathrm{D}_{j-}k_{m,j}\right) \\
&= -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)(g_{i,m}z_{1,m,j} + \gamma_{i,m}z_{4,m,j}).
\end{aligned}
$$

Similarly, one shows that

$$
\begin{aligned}
(z_{2,i,j})_t &= -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)(k_{i,m}z_{2,m,j} + \kappa_{i,m}z_{3,m,j}), \\
(z_{3,i,j})_t &= -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)(g_{i,m}z_{3,m,j} + \gamma_{i,m}z_{2,m,j})
\end{aligned}
$$

and

$$
(z_{4,i,j})_t = -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+U_m)(k_{i,m}z_{4,m,j} + \kappa_{i,m}z_{1,m.j}).
$$

Integrating the first of these, taking absolute values, applying Hölder's inequality and taking supremum over i we obtain

$$
\begin{aligned}
\sup_i |z_{1,i,j}(t)| \le \sup_i |z_{1,i,j}(0)| &+ \int_0^t \|\mathrm{D}_+U(s)\|_{\boldsymbol{\ell}^2} \|g(s)\|_{\boldsymbol{\ell}^2} \sup_m |z_{1,m,j}(s)| \, ds \\
&+ \int_0^t \|\mathrm{D}_+U(s)\|_{\boldsymbol{\ell}^2} \|\gamma(s)\|_{\boldsymbol{\ell}^2} \sup_m |z_{4,m,j}(s)| \, ds
\end{aligned}
$$

Treating the three other relations similarly and defining

$$
Z(t) = \sum_{l=1}^{4} \|z_l(t)\|_{\boldsymbol{\ell}^\infty},
$$

we may add the four inequalities to obtain an inequality of the form

$$
Z(t) \le Z(0) + \int_0^t C(s)Z(s)\, ds,
$$

where

$$
C(s) = 2\,\|\mathrm{D}_+U\|_{\boldsymbol{\ell}^2} \left(\|g\|_{\boldsymbol{\ell}^2} + \|k\|_{\boldsymbol{\ell}^2} + \|\gamma\|_{\boldsymbol{\ell}^2} + \|\kappa\|_{\boldsymbol{\ell}^2}\right)(s)
$$

is bounded by assumption. Since $Z(0)=0$, Grönwall's inequality yields $Z(t)=0$ for $t\in[0,T]$, which proves the result.

(ii): We prove the symmetry of g. From (2.4.19) we have

$$(\mathrm{D}_+y_m)g_{i,m}-\mathrm{D}_{m-}\gamma_{i,m}=\frac{\delta_{i,m}}{\Delta\xi},$$

such that a summation by parts shows

$$\begin{aligned} g_{j,i}&=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_+y_m)g_{i,m}-\mathrm{D}_{m-}\gamma_{i,m}\right]g_{j,m}\\ &=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_+y_m)g_{i,m}g_{j,m}+\gamma_{i,m}\mathrm{D}_{m+}g_{j,m}\right]. \end{aligned}$$

Then we use the identity $\mathrm{D}_{m+}g_{j,m}=(\mathrm{D}_+y_m)\gamma_{j,m}$ from (2.4.19) twice, first for j and then for i, to obtain

$$\begin{aligned} g_{j,i}&=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_+y_m)g_{i,m}g_{j,m}+\gamma_{i,m}(\mathrm{D}_+y_m)\gamma_{j,m}\right]\\ &=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_+y_m)g_{i,m}g_{j,m}+(\mathrm{D}_{m+}g_{i,m})\gamma_{j,m}\right]. \end{aligned}$$

After summing by parts and using (2.4.19) once more, we end up with

$$g_{j,i}=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}g_{i,m}\left[(\mathrm{D}_+y_m)g_{j,m}+\mathrm{D}_{m-}\gamma_{j,m}\right]=g_{i,j},$$

and the symmetry of g is proved. A similar procedure shows the symmetry of $k_{i,j}$. For the antisymmetry we also use (2.4.19) and the same techniques to compute

$$\begin{aligned} \gamma_{j,i}&=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_+y_m)k_{i,m}-\mathrm{D}_{m+}\kappa_{i,m}\right]\gamma_{j,m}\\ &=\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[k_{i,m}\mathrm{D}_{m+}g_{j,m}+\kappa_{i,m}\mathrm{D}_{m-}\gamma_{j,m}\right]\\ &=-\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\left[(\mathrm{D}_{m-}k_{i,m})g_{j,m}-\kappa_{i,m}\mathrm{D}_{m-}\gamma_{j,m}\right]\\ &=-\mathit{\Delta}\xi\sum_{m\in\mathbb{Z}}\kappa_{i,m}\left[(\mathrm{D}_+y_m)g_{j,m}-\mathrm{D}_{m-}\gamma_{j,m}\right]\\ &=-\kappa_{i,j}. \end{aligned}$$

□

Returning to (2.4.20), the second term in the right-hand side can be simplified as follows,

$$\begin{aligned}
-\Delta\xi \sum_{i\in\mathbb{Z}} g_{i,j}\mathrm{D}_-\left(\frac{h_i}{\mathrm{D}_+ y_i} + \rho_\infty \frac{\bar{r}_i}{\mathrm{D}_+ y_i}\right) &= \Delta\xi \sum_{i\in\mathbb{Z}} \frac{\mathrm{D}_{i+} g_{j,i}}{\mathrm{D}_+ y_i}\left(h_i + \rho_\infty \bar{r}_i\right) \\
&= \Delta\xi \sum_{i\in\mathbb{Z}} \gamma_{j,i}\left(h_i + \rho_\infty \bar{r}_i\right) \\
&= -\Delta\xi \sum_{i\in\mathbb{Z}} \kappa_{i,j}\left(h_i + \rho_\infty \bar{r}_i\right),
\end{aligned}$$

where we have used (2.4.19) and (2.4.26). We define

$$Q_j := \Delta\xi \sum_{i\in\mathbb{Z}} g_{i,j} U_i(\mathrm{D}_+ U_i) + \Delta\xi \sum_{i\in\mathbb{Z}} \kappa_{i,j}\left(h_i + \rho_\infty \bar{r}_i\right).$$

Then, the evolution of U is given by

$$\dot{U}_j = -Q_j \tag{2.4.27}$$

The form of Q also motivates the definition

$$R_j := \Delta\xi \sum_{i\in\mathbb{Z}} \gamma_{i,j} U_i(\mathrm{D}_+ U_i) + \Delta\xi \sum_{i\in\mathbb{Z}} k_{i,j}\left(h_i + \rho_\infty \bar{r}_i\right).$$

Indeed, with these definitions we have

$$\begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} \gamma & k \\ g & \kappa \end{bmatrix} * \begin{bmatrix} U(\mathrm{D}_+ U) \\ h + \rho_\infty \bar{r} \end{bmatrix},$$

meaning R and Q satisfies

$$\begin{bmatrix} -\mathrm{D}_- & (\mathrm{D}_+ y_j) \\ (\mathrm{D}_+ y_j) & -\mathrm{D}_+ \end{bmatrix} \circ \begin{bmatrix} R_j \\ Q_j \end{bmatrix} = \begin{bmatrix} U_j(\mathrm{D}_+ U_j) \\ h_j + \rho_\infty \bar{r}_j \end{bmatrix}. \tag{2.4.28}$$

We recognize this as the discrete version of (2.4.6).

The relation $\dot{U}_j = -Q_j$ shows that we have a differential equation for U in the variables y, U, H, $\bar{r}$, g, and κ. From (2.4.21) we obtain

$$\dot{\bar{r}}_j = \dot{r}_j - \rho_\infty \mathrm{D}_+ \dot{y}_j = -\rho_\infty \mathrm{D}_+ U_j. \tag{2.4.29}$$

Next, we introduce the cumulative energy H_j as

$$H_j = \Delta\xi \sum_{i=-\infty}^{j-1} h_i,$$

so that $h_j = \mathrm{D}_+ H_j$. To obtain the evolution equation of H, we first multiply (2.4.22) by $\mathrm{D}_+ y_i$ and differentiate the result with respect to time to obtain

$$\begin{aligned}\frac{d}{dt}\left((\mathrm{D}_+ y_i) h_i\right) = -U_i Q_i (\mathrm{D}_+ y_i)^2 + U_i^2 (\mathrm{D}_+ y_i)(\mathrm{D}_+ U_i)\\ - (\mathrm{D}_+ U_i)(\mathrm{D}_+ Q_i) - \rho_\infty \bar{r}_i \mathrm{D}_+ U_i,\end{aligned}$$

after using (2.4.27) and (2.4.29). Then, we use the relation between Q and R given in (2.4.28) to obtain

$$\frac{d}{dt}\left((\mathrm{D}_+ y_i) h_i\right) = (\mathrm{D}_+ U_i) h_i - (\mathrm{D}_+ y_i)[U_i(\mathrm{D}_- R_i) + R_i(\mathrm{D}_+ U_i)].$$

Simplifying further, we obtain

$$\dot{h}_i = -\left[U_i(\mathrm{D}_- R_i) + R_i(\mathrm{D}_+ U_i)\right].$$

This leads to

$$\dot{H}_j = -\Delta\xi \sum_{i=-\infty}^{j-1} [U_i(\mathrm{D}_- R_i) + R_i(\mathrm{D}_+ U_i)] = -U_j R_{j-1},$$

where in the last equality we have used the decay at infinity together with (2.3.5).

Collecting all the equations and applying the relations (2.4.25) and (2.4.26) we obtain the closed system

$$\dot{\zeta}_j = U_j, \tag{2.4.30a}$$

$$\dot{U}_j = -Q_j \tag{2.4.30b}$$

$$\dot{H}_j = -U_j R_{j-1}, \tag{2.4.30c}$$

$$\dot{\bar{r}}_j = -\rho_\infty \mathrm{D}_+ U_j, \tag{2.4.30d}$$

$$\dot{g}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ U_m)\left(g_{i,m} g_{m,j} + \gamma_{i,m}\kappa_{m,j}\right), \tag{2.4.30e}$$

$$\dot{k}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ U_m)\left(k_{i,m} k_{m,j} + \kappa_{i,m}\gamma_{m,j}\right), \tag{2.4.30f}$$

$$\dot{\gamma}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ U_m)\left(\gamma_{i,m} k_{m,j} + g_{i,m}\gamma_{m,j}\right), \tag{2.4.30g}$$

$$\dot{\kappa}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ U_m)\left(\kappa_{i,m} g_{m,j} + k_{i,m}\kappa_{m,j}\right), \tag{2.4.30h}$$

where $y_j = j\Delta\xi + \zeta_j$, and we recall

$$R_j = \Delta\xi \sum_{i\in\mathbb{Z}} \gamma_{i,j} U_i (\mathrm{D}_+ U_i) + \Delta\xi \sum_{i\in\mathbb{Z}} k_{i,j}\left(h_i + \rho_\infty \bar{r}_i\right),$$

$$Q_j = \mathit{\Delta}\xi \sum_{i\in\mathbb{Z}} g_{i,j} U_i(\mathrm{D}_+ U_i) + \mathit{\Delta}\xi \sum_{i\in\mathbb{Z}} \kappa_{i,j} \left(h_i + \rho_\infty \bar{r}_i\right).$$

2.5 Existence and uniqueness of the solution to the semi-discrete 2CH system

In this section, we show that the semi-discrete system (2.4.30) has a unique, globally defined solution. Let us first introduce the functional setting for the analysis. We define the discrete versions of the spaces used in the continuous setting, namely

$$\mathbf{E}_{\mathit{\Delta}\xi} := \mathbf{V}_{\mathit{\Delta}\xi} \times \mathbf{h}^1 \times \mathbf{V}_{\mathit{\Delta}\xi} \times \boldsymbol{\ell}^2,$$

with norm

$$\|(\zeta, U, H, \bar{r})\|_{\mathbf{E}_{\mathit{\Delta}\xi}} := \|\zeta\|_{\mathbf{V}_{\mathit{\Delta}\xi}} + \|U\|_{\mathbf{h}^1} + \|H\|_{\mathbf{V}_{\mathit{\Delta}\xi}} + \|\bar{r}\|_{\boldsymbol{\ell}^2}.$$

Since we have included the operator kernels as solution variables in (2.4.30), we have to introduce a space for them as well. To account for that the kernels are well-behaved, we choose their space to be $\boldsymbol{\ell}^* := \boldsymbol{\ell}^1 \cap \boldsymbol{\ell}^\infty$ with norm $\|\cdot\|_{\boldsymbol{\ell}^*} = \|\cdot\|_{\boldsymbol{\ell}^1} + \|\cdot\|_{\boldsymbol{\ell}^\infty}$, and this will be sufficient for our purposes. We note that $\boldsymbol{\ell}^* \subset \boldsymbol{\ell}^2$, since we have the inequality

$$\|g\|_{\boldsymbol{\ell}^2} \leq \|g\|_{\boldsymbol{\ell}^\infty}^{1/2} \|g\|_{\boldsymbol{\ell}^1}^{1/2} \leq \frac{1}{2}(\|g\|_{\boldsymbol{\ell}^\infty} + \|g\|_{\boldsymbol{\ell}^1}). \tag{2.5.1}$$

Thus, we will consider solution tuples of the form

$$X = (\zeta, U, H, \bar{r}, g, k, \gamma, \kappa) \in \mathbf{E}_{\mathit{\Delta}\xi} \times (\boldsymbol{\ell}^*)^4 =: \mathbf{E}^{\mathrm{ker}}_{\mathit{\Delta}\xi},$$

where $\mathbf{E}^{\mathrm{ker}}_{\mathit{\Delta}\xi}$ denotes the space $\mathbf{E}_{\mathit{\Delta}\xi}$ augmented with the space for the kernel operators $\boldsymbol{\ell}^*$.

2.5.1 Existence and uniqueness of the solution to the semi-discrete system

To prove the short-time existence of (2.4.30), we consider an auxiliary system which corresponds to (2.4.30), except that we have decoupled ζ, U and H from their discrete derivatives $\mathrm{D}_+\zeta$, D_+U and D_+H by introducing the sequences α, β and h. The reason for this is that we cannot take for granted that the kernels satisfy (2.4.19) for $t > 0$, and then we cannot use (2.4.28) when estimating the right-hand side of (2.4.30b) in $\mathbf{h}^1$-norm. Once the short-time existence of solutions to the auxiliary system is established, we will prove that the coupling between y, U, H

and their discrete derivatives is indeed preserved if it holds initially. The auxiliary system reads

$$\dot{\zeta}_j = U_j, \qquad \dot{U}_j = -Q_j, \qquad \dot{H}_j = -U_j R_{j-1}, \tag{2.5.2a}$$
$$\dot{r}_j = -\rho_\infty \beta_j, \qquad \dot{\alpha}_j = \beta_j, \tag{2.5.2b}$$

$$\dot{\beta}_j = -R_j(1+\alpha_j) + h_j + \rho_\infty r_j, \tag{2.5.2c}$$
$$\dot{h}_j = \left((U_j)^2 - R_j\right)\beta_j - U_j Q_j (1+\alpha_j), \tag{2.5.2d}$$

and

$$\dot{g}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \left(g_{i,m} g_{j,m} - \gamma_{i,m}\gamma_{j,m}\right), \tag{2.5.2e}$$
$$\dot{k}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \left(k_{i,m} k_{j,m} - \kappa_{i,m}\kappa_{j,m}\right), \tag{2.5.2f}$$
$$\dot{\gamma}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \left(\gamma_{i,m} k_{j,m} - g_{i,m}\kappa_{j,m}\right), \tag{2.5.2g}$$
$$\dot{\kappa}_{i,j} = -\Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \left(\kappa_{i,m} g_{j,m} - k_{i,m}\gamma_{j,m}\right), \tag{2.5.2h}$$

where we have momentarily redefined R and Q as

$$\begin{bmatrix} R \\ Q \end{bmatrix} = \begin{bmatrix} \gamma & k \\ g & \kappa \end{bmatrix} * \begin{bmatrix} U\beta \\ h + \rho_\infty \bar{r} \end{bmatrix}.$$

Equations (2.5.2c), (2.5.2d), and the second equation of (2.5.2b), have been obtained formally by applying D_+ to (2.4.30a), (2.4.30b), and (2.4.30c), in combination with (2.4.28). We collect all the variables in a tuple

$$Y = (\zeta, U, H, r, \alpha, \beta, h, g, k, \gamma, \kappa) \in \mathbf{E}^{\mathrm{aux}}_{\Delta\xi}$$

for

$$\mathbf{E}^{\mathrm{aux}}_{\Delta\xi} := \boldsymbol{\ell}^\infty \times \left(\boldsymbol{\ell}^2 \cap \boldsymbol{\ell}^\infty\right) \times \boldsymbol{\ell}^\infty \times (\boldsymbol{\ell}^2)^4 \times (\boldsymbol{\ell}^*)^4,$$

and introduce the corresponding norm

$$\begin{aligned}\|Y\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}} := {}& \|\zeta\|_{\boldsymbol{\ell}^\infty} + \|U\|_{\boldsymbol{\ell}^2} + \|U\|_{\boldsymbol{\ell}^\infty} + \|H\|_{\boldsymbol{\ell}^\infty} + \|r\|_{\boldsymbol{\ell}^2} \\ & + \|\alpha\|_{\boldsymbol{\ell}^2} + \|\beta\|_{\boldsymbol{\ell}^2} + \|h\|_{\boldsymbol{\ell}^2} + \|g\|_{\boldsymbol{\ell}^*} + \|k\|_{\boldsymbol{\ell}^*} + \|\gamma\|_{\boldsymbol{\ell}^*} + \|\kappa\|_{\boldsymbol{\ell}^*}.\end{aligned}$$

Note how we require $U \in \boldsymbol{\ell}^\infty$ to account for the fact that the decoupling of U and D_+U deprives us of the continuous inclusion $\mathbf{h}^1 \subset \boldsymbol{\ell}^\infty$.

Lemma 2.5.1 (Short-time solution for (2.5.2))**.** *Let $Y_0 \in \mathbf{E}^{\mathrm{aux}}_{\Delta\xi}$ be such that $1+\alpha_j \geq 0$ for all j, and with initial auxiliary variables $g_0, k_0, \gamma_0, \kappa_0$ constructed according to Theorem 2.3.5 and Corollary 2.3.7 with $a_j = 1+\alpha_j$. Then, there exists a time $T>0$ depending only on $\|Y_0\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}}$ such that* (2.5.2) *has a unique solution $Y \in \mathbf{C}^1([0,T], \mathbf{E}^{\mathrm{aux}}_{\Delta\xi})$ with initial datum Y_0.*

Proof of Lemma 2.5.1. We will use the symmetry and anti-symmetry identities (2.4.25) and (2.4.26) in our estimates and we explain now why it can be done. First, we note that these identities hold initially by the construction of (2.3.32) and (2.3.36). Then, from the evolution equations (2.5.2e)–(2.5.2h) one can check that the symmetry identities are preserved by the Picard fixed-point operator which we will use here to prove the short-time existence of (2.5.2). Then, by establishing local Lipschitz regularity of the right-hand side, we can prove the existence of a short-time solution in the closed subset of $\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}$ where (2.4.25) and (2.4.26) hold.

Let us consider two functions in $\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}$,

$$Y = (\zeta, U, H, r, \alpha, \beta, h, g, k, \gamma, \kappa)$$

and

$$\tilde{Y} = \left(\tilde{\zeta}, \tilde{U}, \tilde{H}, \tilde{r}, \tilde{\alpha}, \tilde{\beta}, \tilde{h}, \tilde{g}, \tilde{k}, \tilde{\gamma}, \tilde{\kappa}\right).$$

For the Lipschitz estimates, we first treat the right-hand sides of (2.5.2e)–(2.5.2h). We only provide details for (2.5.2e) as (2.5.2f)–(2.5.2h) can be treated similarly.

We start by considering the $\boldsymbol{\ell}^\infty$-norm using the following splitting,

$$\begin{aligned}
&\left| -\Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \left(g_{j,m}g_{i,m} - \gamma_{i,m}\gamma_{j,m}\right) + \Delta\xi \sum_{m\in\mathbb{Z}} \tilde{\beta}_m \left(\tilde{g}_{j,m}\tilde{g}_{i,m} - \tilde{\gamma}_{i,m}\tilde{\gamma}_{j,m}\right)\right| \\
&\qquad \leq \left| \Delta\xi \sum_{m\in\mathbb{Z}} \beta_m g_{j,m}g_{i,m} - \Delta\xi \sum_{m\in\mathbb{Z}} \tilde{\beta}_m \tilde{g}_{j,m}\tilde{g}_{i,m}\right| \\
&\qquad\qquad + \left| \Delta\xi \sum_{m\in\mathbb{Z}} \beta_m \gamma_{i,m}\gamma_{j,m} - \Delta\xi \sum_{m\in\mathbb{Z}} \tilde{\beta}_m \tilde{\gamma}_{i,m}\tilde{\gamma}_{j,m}\right|
\end{aligned}$$

We estimate the first term as follows

$$\left| \Delta\xi \sum_{m\in\mathbb{Z}} \beta_m g_{j,m}g_{i,m} - \Delta\xi \sum_{m\in\mathbb{Z}} \tilde{\beta}_m \tilde{g}_{j,m}\tilde{g}_{i,m}\right|$$

$$\leq \|g\|_{\ell^\infty}\,\|g\|_{\ell^2}\,\|\beta-\tilde{\beta}\|_{\ell^2}+\|g\|_{\ell^\infty}\,\|\tilde{\beta}\|_{\ell^2}\,\|g-\tilde{g}\|_{\ell^2}+\|\tilde{\beta}\|_{\ell^2}\,\|\tilde{g}\|_{\ell^2}\,\|g-\tilde{g}\|_{\ell^\infty}$$

and the second term has a similar estimate. For the ℓ^1-norm, use the same splitting and consider again only the first term. We make use of the symmetry properties of the kernel operators, as given in Lemma (2.4.3), to switch between indices and obtain

$$\begin{aligned}&\Delta\xi\sum_{i\in\mathbb{Z}}\left|\Delta\xi\sum_{m\in\mathbb{Z}}\beta_m g_{j,m}g_{i,m}-\Delta\xi\sum_{m\in\mathbb{Z}}\tilde{\beta}_m\tilde{g}_{j,m}\tilde{g}_{i,m}\right|\\&\leq \|g\|_{\ell^1}\,\|g\|_{\ell^2}\,\|\beta-\tilde{\beta}\|_{\ell^2}+\|g\|_{\ell^1}\,\|\tilde{\beta}\|_{\ell^2}\,\|g-\tilde{g}\|_{\ell^2}+\|\tilde{\beta}\|_{\ell^2}\,\|\tilde{g}\|_{\ell^2}\,\|g-\tilde{g}\|_{\ell^1}\,,\end{aligned}$$

From (2.5.1) we get $\|g-\tilde{g}\|_{\ell^2}\leq\frac{1}{2}(\|g-\tilde{g}\|_{\ell^\infty}+\|g-\tilde{g}\|_{\ell^1})$ and therefore we can conclude that the right-hand side in (2.5.2e) is Lipschitz-continuous with respect to the $\mathbf{E}^{\text{aux}}_{\Delta\xi}$-norm.

Let us consider Lipschitz properties of R and Q. We decompose Q in Q_1+Q_2 where

$$(Q_1)_j := \Delta\xi\sum_{i\in\mathbb{Z}} g_{i,j}U_i(\mathrm{D}_+U_i), \tag{2.5.3a}$$

$$(Q_2)_j := \Delta\xi\sum_{i\in\mathbb{Z}} \kappa_{i,j}\left(h_i+\rho_\infty\bar{r}_i\right). \tag{2.5.3b}$$

Similarly, we decompose R in R_1+R_2 where

$$(R_1)_j := \Delta\xi\sum_{i\in\mathbb{Z}} \gamma_{i,j}U_i(\mathrm{D}_+U_i), \tag{2.5.4a}$$

$$(R_2)_j := \Delta\xi\sum_{i\in\mathbb{Z}} k_{i,j}\left(h_i+\rho_\infty\bar{r}_i\right). \tag{2.5.4b}$$

We have $Q_2=\kappa * f$ for $f=h+\rho_\infty r$ so that

$$\|f\|_{\ell^2}=\|h+\rho_\infty r\|_{\ell^2}\leq\|h\|_{\ell^2}+\rho_\infty\|r\|_{\ell^2}.$$

Starting with Q_2, we have

$$\|Q_2-\tilde{Q}_2\|_{\ell^2}=\|\kappa * f-\tilde{\kappa} * \tilde{f}\|_{\ell^2}\leq\|(\kappa-\tilde{\kappa}) * f\|_{\ell^2}+\|\tilde{\kappa} * (f-\tilde{f})\|_{\ell^2}$$

For the first term above, applying the Young's inequality (2.4.24) with $r=p=2$ and $q=1$, we get

$$\|(\kappa-\tilde{\kappa}) * f\|_{\ell^2}\leq\|\kappa-\tilde{\kappa}\|_{\ell^1}^{\frac{1}{2}}\|(\kappa-\tilde{\kappa})^\top\|_{\ell^1}^{\frac{1}{2}}\|f\|_{\ell^2}$$

Using the antisymmetry property (2.4.26) of κ and $\tilde{\kappa}$, namely $\kappa^\top = -\gamma$ and $\tilde{\kappa}^\top = -\tilde{\gamma}$, we get

$$\|(\kappa - \tilde{\kappa}) * f\|_{\boldsymbol{\ell}^2} \leq \|\kappa - \tilde{\kappa}\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}} \|\gamma - \tilde{\gamma}\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}} \|f\|_{\boldsymbol{\ell}^2}$$

Hence, we obtain the following estimate in $\boldsymbol{\ell}^2$-norm,

$$\|Q_2 - \tilde{Q}_2\|_{\boldsymbol{\ell}^2} \leq \frac{\|\gamma - \tilde{\gamma}\|_{\boldsymbol{\ell}^1} + \|\kappa - \tilde{\kappa}\|_{\boldsymbol{\ell}^1}}{2} \|f\|_{\boldsymbol{\ell}^2} + \frac{\|\tilde{\gamma}\|_{\boldsymbol{\ell}^1} + \|\tilde{\kappa}\|_{\boldsymbol{\ell}^1}}{2} \|f - \tilde{f}\|_{\boldsymbol{\ell}^2}.$$

For the $\boldsymbol{\ell}^\infty$-norm, we use the same splitting

$$\|Q_2 - \tilde{Q}_2\|_{\boldsymbol{\ell}^\infty} \leq \|(\kappa - \tilde{\kappa}) * f\|_{\boldsymbol{\ell}^\infty} + \|\kappa * (f - \tilde{f})\|_{\boldsymbol{\ell}^\infty}.$$

Applying Young's inequality (2.4.24), for $r = \infty$ and $p = q = 2$, and the symmetry property of κ, we obtain in a similar way as before that

$$\|Q_2 - \tilde{Q}_2\|_{\boldsymbol{\ell}^\infty} \leq \|\gamma - \tilde{\gamma}\|_{\boldsymbol{\ell}^2} \|f\|_{\boldsymbol{\ell}^2} + \|\tilde{\gamma}\|_{\boldsymbol{\ell}^2} \|f - \tilde{f}\|_{\boldsymbol{\ell}^2}.$$

In a similar fashion as for Q_2 we find

$$\begin{aligned}
\|R_2 - \tilde{R}_2\|_{\boldsymbol{\ell}^2} &\leq \|k - \tilde{k}\|_{\boldsymbol{\ell}^1} \, \|f\|_{\boldsymbol{\ell}^2} + \|\tilde{k}\|_{\boldsymbol{\ell}^1} \|f - \tilde{f}\|_{\boldsymbol{\ell}^2}, \\
\|R_2 - \tilde{R}_2\|_{\boldsymbol{\ell}^\infty} &\leq \|k - \tilde{k}\|_{\boldsymbol{\ell}^2} \, \|f\|_{\boldsymbol{\ell}^2} + \|\tilde{k}\|_{\boldsymbol{\ell}^2} \|f - \tilde{f}\|_{\boldsymbol{\ell}^2}.
\end{aligned}$$

Furthermore, analogous applications of (2.4.24) and (2.4.26) produce

$$\begin{aligned}
\|Q_1 - \tilde{Q}_1\|_{\boldsymbol{\ell}^2} &\leq \|g - \tilde{g}\|_{\boldsymbol{\ell}^2} \, \|U\beta\|_{\boldsymbol{\ell}^1} + \|\tilde{g}\|_{\boldsymbol{\ell}^2} \|U\beta - \tilde{U}\tilde{\beta}\|_{\boldsymbol{\ell}^1}, \\
\|Q_1 - \tilde{Q}_1\|_{\boldsymbol{\ell}^\infty} &\leq \|g - \tilde{g}\|_{\boldsymbol{\ell}^\infty} \, \|U\beta\|_{\boldsymbol{\ell}^1} + \|\tilde{g}\|_{\boldsymbol{\ell}^\infty} \|U\beta - \tilde{U}\tilde{\beta}\|_{\boldsymbol{\ell}^1}, \\
\|R_1 - \tilde{R}_1\|_{\boldsymbol{\ell}^2} &\leq \|\gamma - \tilde{\gamma}\|_{\boldsymbol{\ell}^2} \, \|U\beta\|_{\boldsymbol{\ell}^1} + \|\tilde{\gamma}\|_{\boldsymbol{\ell}^2} \|U\beta - \tilde{U}\tilde{\beta}\|_{\boldsymbol{\ell}^1}, \\
\|R_1 - \tilde{R}_1\|_{\boldsymbol{\ell}^\infty} &\leq \|\gamma - \tilde{\gamma}\|_{\boldsymbol{\ell}^\infty} \, \|U\beta\|_{\boldsymbol{\ell}^1} + \|\tilde{\gamma}\|_{\boldsymbol{\ell}^\infty} \|U\beta - \tilde{U}\tilde{\beta}\|_{\boldsymbol{\ell}^1}.
\end{aligned}$$

For the $\boldsymbol{\ell}^1$-norms above we then apply the Cauchy–Schwarz inequality to obtain

$$\|U\beta\|_{\boldsymbol{\ell}^1} \leq \|U\|_{\boldsymbol{\ell}^2} \, \|\beta\|_{\boldsymbol{\ell}^2}\,, \quad \|U\beta - \tilde{U}\tilde{\beta}\|_{\boldsymbol{\ell}^1} \leq \|U\|_{\boldsymbol{\ell}^2} \, \|\beta - \tilde{\beta}\|_{\boldsymbol{\ell}^2} + \|\tilde{U}\|_{\boldsymbol{\ell}^2} \|\beta - \tilde{\beta}\|_{\boldsymbol{\ell}^2},$$

which contain the relevant norms.

From the preceding estimates on Q_1 and Q_2 the local Lipschitz property of the right-hand side of the second equation in (2.5.2a) in the $\boldsymbol{\ell}^2 \cap \boldsymbol{\ell}^\infty$-norm is clear. Furthermore, since $U \in \boldsymbol{\ell}^\infty$, the previous $\boldsymbol{\ell}^\infty$-estimates on R and Q also show that the right-hand sides of (2.5.2c) and (2.5.2d) are locally Lipschitz in the $\boldsymbol{\ell}^2$-norm. For the last equation in (2.5.2a), we introduce the right-shift operator $(\tau R)_j = R_{j-1}$ and we have

$$\begin{aligned}\|U(\tau R) - \tilde{U}(\tau\tilde{R})\|_{\ell^\infty} &\leq \|U - \tilde{U}\|_{\ell^\infty}\,\|\tau R\|_{\ell^\infty} + \|\tilde{U}\|_{\ell^\infty}\|\tau(R-\tilde{R})\|_{\ell^\infty} \\ &\leq \|U - \tilde{U}\|_{\ell^\infty}\,\|R\|_{\ell^\infty} + \|\tilde{U}\|_{\ell^\infty}\|R-\tilde{R}\|_{\ell^\infty},\end{aligned}$$

The remaining right-hand sides in (2.5.2a) and (2.5.2b) are linear in the solution variables, and thus Lipschitz in their respective norms. Hence, for (2.5.2) written as $\dot{Y} = \hat{F}(Y)$ we have

$$\|\hat{F}(Y) - \hat{F}(\tilde{Y})\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}} \leq C(\|Y\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}}, \|\tilde{Y}\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}})\|Y - \tilde{Y}\|_{\mathbf{E}^{\mathrm{aux}}_{\Delta\xi}},$$

which is what we set out to prove. □

The final step in obtaining short-time existence for (2.4.30) from the auxiliary system, is to show that if the initial data for (2.5.2) satisfy

$$\begin{bmatrix} -\mathrm{D}_{j-} & (1+\alpha_j) \\ (1+\alpha_j) & -\mathrm{D}_{j+}\end{bmatrix} \circ \begin{bmatrix}\gamma_{i,j} & k_{i,j} \\ g_{i,j} & \kappa_{i,j}\end{bmatrix} = \frac{1}{\Delta\xi}\begin{bmatrix}\delta_{i,j} & 0 \\ 0 & \delta_{i,j}\end{bmatrix} \tag{2.5.5a}$$

$$\alpha = \mathrm{D}_+\zeta, \quad \beta = \mathrm{D}_+U, \text{ and } \quad h = \mathrm{D}_+H, \tag{2.5.5b}$$

then these identities are preserved in time by the solution. The result for (2.5.5a) has been proved in Lemma 2.4.3, as it only depends on the identity $(\mathrm{D}_+y)_t = \mathrm{D}_+U$, which is replaced here by $\alpha_t = \beta$. Using (2.5.5a), we infer from (2.4.28) that

$$\begin{bmatrix} -\mathrm{D}_- & (1+\alpha_j) \\ (1+\alpha_j) & -\mathrm{D}_+\end{bmatrix} \circ \begin{bmatrix} R_j \\ Q_j\end{bmatrix} = \begin{bmatrix} U_j\beta_j \\ h_j + \rho_\infty\bar{r}_j\end{bmatrix}. \tag{2.5.6}$$

From the definition of (2.5.2) we get

$$\frac{d}{dt}(\alpha_j - \mathrm{D}_+\zeta_j) = \beta_j - \mathrm{D}_+U_j, \tag{2.5.7a}$$

while the expression for D_+Q_j from (2.5.6) yields

$$\frac{d}{dt}(\beta_j - \mathrm{D}_+U_j) = 0, \tag{2.5.7b}$$

and from the expression for D_-R_j we obtain

$$\frac{d}{dt}(h_j - \mathrm{D}_+H_j) = -R_j(\beta_j - \mathrm{D}_+U_j). \tag{2.5.7c}$$

Hence, the equations (2.5.7) give us that (2.5.5b) holds for all time if it holds initially. Then we have proved the following theorem.

Theorem 2.5.2 (Short-time solution for (2.4.30))**.** *Given $X_0 \in \mathbf{E}^{\mathrm{ker}}_{\Delta\xi}$ such that $1 + \mathrm{D}_+\zeta_j \geq 0$ and g_0, k_0, γ_0, and κ_0 are constructed according to Theorem 2.3.5 and Corollary 2.3.7 with $a_j = 1 + \mathrm{D}_+\zeta_j$. Then, there exists a time T depending only on $\|X_0\|_{\mathbf{E}^{\mathrm{ker}}_{\Delta\xi}}$ such that* (2.4.30) *has a unique solution $X \in C^1([0,T], \mathbf{E}^{\mathrm{ker}}_{\Delta\xi})$ with initial datum X_0.*

The next step is to prove that there exists a subset, denoted by $\mathcal{B}$, of $\mathbf{E}^{\mathrm{ker}}_{\Delta\xi}$ which is preserved by the evolution equation. For this subset, the solution exists globally in time. The subset $\mathcal{B}$ is defined as follows.

Definition 2.5.3. The set $\mathcal{B}$ is composed of all $(\zeta, U, H, \bar{r}, g, k, \gamma, \kappa) \in \mathbf{E}^{\mathrm{ker}}_{\Delta\xi}$ such that

(a) g, k, γ, κ satisfy the properties listed in Lemma 2.4.3 for $a = \mathrm{D}_+y$,

(b) $(\mathrm{D}_+y, \mathrm{D}_+U, \mathrm{D}_+H, \bar{r}) \in (\boldsymbol{\ell}^\infty)^4$,

(c) $2(\mathrm{D}_+y_j)(\mathrm{D}_+H_j) = (U_j)^2(\mathrm{D}_+y_j)^2 + (\mathrm{D}_+U_j)^2 + \bar{r}_j^2$ for all j,

(d) $\mathrm{D}_+y_j \geq 0$, $\mathrm{D}_+H_j \geq 0$, $\mathrm{D}_+y_j + \mathrm{D}_+H_j > 0$ for all j.

Lemma 2.5.4 (Properties preserved by the flow)**.** *Given initial datum $X_0 \in \mathcal{B}$, let $X(t) \in C^1([0,T], \mathbf{E}^{\mathrm{ker}}_{\Delta\xi})$ be the corresponding short-time solution given by Theorem 2.5.2. Then $X(t) \in \mathcal{B}$ for all $t \in [0,T]$.*

Proof of Lemma 2.5.4. Property (a) follows from Lemma 2.4.3, since the solution variables in $X(t)$ satisfy $\mathrm{D}_+\dot{y}_j = \mathrm{D}_+U_j$ and $\mathrm{D}_+U \in \boldsymbol{\ell}^2$, where we as usual have $\mathrm{D}_+y_j = 1 + \mathrm{D}_+\zeta_j$.

The proof of property (b) essentially follows [30, Lem. 3.3], which again is based on [36, Lem. 2.4], and the argument is as follows. Consider U, R, Q defined in (2.5.3a), (2.5.3b), (2.5.4a) and (2.5.4b) as given functions for $t \in [0,T]$ based on the solution variables in $X(t)$. Then we can read off from (2.5.2) that the variables $(\mathrm{D}_+y, \mathrm{D}_+U, \mathrm{D}_+H, \bar{r})(t)$ coming from $X(t)$ satisfy the following affine system,

$$\begin{aligned}
\dot{\alpha}_j &= \beta_j \\
\dot{\beta}_j &= -R_j\alpha_j + h_j + \rho_\infty r_j \\
\dot{h}_j &= \left((U_j)^2 - R_j\right)\beta_j - U_jQ_j\alpha_j, \\
\dot{r}_j &= -\rho_\infty\beta_j,
\end{aligned} \tag{2.5.8}$$

in the respective variables (α, β, h, r). We know that U, R, and Q are bounded in the $\boldsymbol{\ell}^\infty$-norm, and so the affine system (2.5.8) has bounded coefficients. Then we may take any norm we like, in particular the $\boldsymbol{\ell}^\infty$-norm, and right-hand side of (2.5.8) will be locally Lipschitz in that norm. Hence, the result follows from a standard contraction argument.

To prove (c) we simply differentiate the identity with respect to time while applying (2.4.28) and (2.4.30), or (2.5.2) if you will, to find

$$\begin{aligned}\frac{d}{dt}&\left[(\mathrm{D}_+y_j)h_j-\frac{1}{2}(U_j)^2(\mathrm{D}_+y_j)^2-\frac{1}{2}(\mathrm{D}_+U_j)^2-\frac{1}{2}\bar{r}_j^2\right]\\ &=(\mathrm{D}_+y_j)\dot{h}_j+(\mathrm{D}_+\dot{y}_j)h_j-U_j\dot{U}_j(\mathrm{D}_+y_j)^2-(U_j)^2(\mathrm{D}_+y_j)(\mathrm{D}_+\dot{y}_j)\\ &\quad-(\mathrm{D}_+U_j)(\mathrm{D}_+\dot{U}_j)-\bar{r}_j\dot{\bar{r}}_j\\ &=(\mathrm{D}_+y_j)\left[\left((U_j)^2-R_j\right)(\mathrm{D}_+U_j)-U_jQ_j(\mathrm{D}_+y_j)\right]\\ &\quad+(\mathrm{D}_+U_j)h_j+U_jQ_j(\mathrm{D}_+y_j)^2-(U_j)^2(\mathrm{D}_+y_j)(\mathrm{D}_+U_j)\\ &\quad+(\mathrm{D}_+U_j)\left[R_j(\mathrm{D}_+y_j)-h_j-\rho_\infty\bar{r}_j\right]+\bar{r}_j\rho_\infty(\mathrm{D}_+U_j)\\ &=0,\end{aligned}$$

where we identify h_j and D_+H_j. Consequently, if (c) holds for $t=0$, then it will hold for all $t\in(0,T]$.

To prove (d) we fix $j\in\mathbb{Z}$ and define

$$t^*:=\sup\{t\in[0,T]\ :\ \mathrm{D}_+y_j(t')\geq 0, t'\in[0,t]\},$$

and assume $t^*<T$. Since D_+y_j is continuous with respect to time we have $\mathrm{D}_+y_j(t^*)=0$, which by (c) and $h\in\boldsymbol{\ell}^\infty$ from (b) implies

$$\mathrm{D}_+\dot{y}_j(t^*)=\mathrm{D}_+U_j(t^*)=\bar{r}_j(t^*)=0.$$

From (2.4.28) and (2.4.30) we get

$$\mathrm{D}_+\ddot{y}_j=-\mathrm{D}_+Q_j=-R_j(\mathrm{D}_+y_j)+h_j+\rho_\infty\bar{r}_j,$$

implying $\mathrm{D}_+\ddot{y}_j(t^*)=h_j(t^*)$. Assume first $h_j(t^*)=0$ which implies

$$(\mathrm{D}_+y_j,\mathrm{D}_+U_j,h_j,\bar{r}_j)(t^*)=(0,0,0,0).$$

Uniqueness of solutions for (2.5.8) then yields

$$(\mathrm{D}_+y_j,\mathrm{D}_+U_j,h_j,\bar{r}_j)(0)=(0,0,0,0),$$

which contradicts $X_0\in\mathcal{B}$. Assume then $h_j(t^*)<0$. This contradicts the definition of t^* as there would then be a neighborhood of t^* where $\mathrm{D}_+y_j<0$. Therefore we must have $h_j(t^*)>0$ and so there must be a neighborhood of t^* where $\mathrm{D}_+y_j>0$ contradicting the definition of t^*. Hence $t^*=T$ and we have proved $\mathrm{D}_+y_j(t)\geq 0$ for $t\in[0,T]$.

When $\mathrm{D}_+y_j > 0$ it follows from (2.4.22) that $h_j \geq 0$. On the other hand, if $\mathrm{D}_+y_j(t) = 0$ we have just seen that $h_j(t) < 0$ would imply that $\mathrm{D}_+y_j < 0$ in a punctured neighborhood of t, which is impossible. Thus we must have $h_j(t) \geq 0$ for $t \in [0,T]$. For the last inequality, assume that $\mathrm{D}_+y_j + h_j > 0$ does not hold for $t \in [0,T]$. Then by continuity there is a t such that $(\mathrm{D}_+y_j + h_j)(t) = 0$, but this would again by uniqueness of solutions for (2.5.8) mean that $(\mathrm{D}_+y_j + h_j)(0) = 0$ which contradicts $X_0 \in \mathcal{B}$. □

For the rest of the paper we will only consider $X \in \mathcal{B} \cap \mathbf{E}^{\mathrm{ker}}_{\Delta\xi}$, as solutions in this set contains all the relevant solutions to the original 2CH system (2.1.2). Lemma 2.5.4, and in particular the preservation of the identity

$$2(\mathrm{D}_+y_j)h_j = U_j^2(\mathrm{D}_+y_j)^2 + (\mathrm{D}_+U_j)^2 + \bar{r}_j^2, \tag{2.5.9}$$

allows us to prove useful estimates for the solutions in $\mathcal{B}$. We have

$$\Delta\xi \sum_{j\in\mathbb{Z}} |U_j|\,|\mathrm{D}_+U_j| \leq H_\infty(t), \tag{2.5.10}$$

where $H_\infty(t) = \lim_{n\to+\infty} H_n$ is the total energy of the discrete system. This quantity corresponds to $\mathcal{H}_{\mathrm{dis}}$ in (2.2.13). Indeed, the Hamiltonian (2.2.14) is conserved for $t \in [0,T]$, that is $H_\infty(t) = H_\infty(0) < \infty$ for $t \in [0,T]$. We denote the preserved total energy $H_\infty(t)$ by H_∞. Turning back to the inequality (2.5.10), it can be proved as follows,

$$\begin{aligned}
\Delta\xi \sum_{j\in\mathbb{Z}} |U_j|\,|\mathrm{D}_+U_j| &\leq \Delta\xi \sum_{j\in\mathbb{Z}} |U_j| \sqrt{(\mathrm{D}_+y_j)[2h_j - U_j^2(\mathrm{D}_+y_j)]} \\
&\leq \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} U_j^2(\mathrm{D}_+y_j) + \frac{1}{2}\Delta\xi \sum_{j\in\mathbb{Z}} [2h_j - U_j^2(\mathrm{D}_+y_j)] \\
&= H_\infty,
\end{aligned}$$

where in the first inequality we have used (2.5.9), and in the second inequality we have used $\mathrm{D}_+y_j \geq 0$ together with the Cauchy–Schwarz inequality. An immediate consequence of (2.5.10) is that $\|U\|_{\boldsymbol{\ell}^\infty}$ can be uniformly bounded by a constant depending only on H_∞. To show this, we note that by adding and subtracting in (2.2.5) we have the identity

$$\mathrm{D}_\pm(U_i)^2 = 2U_i(\mathrm{D}_\pm U_i) \pm \Delta\xi(\mathrm{D}_\pm U_i)^2.$$

Taking advantage of the decay of U at infinity, we may then write

$$(U_j)^2 = -2\Delta\xi \sum_{i=j}^{\infty} U_i(\mathrm{D}_+U_i) - (\Delta\xi)^2 \sum_{i=j}^{\infty} (\mathrm{D}_+U_i)^2 \leq 2\Delta\xi \sum_{i\in\mathbb{Z}} |U_i|\,|\mathrm{D}_+U_i|,$$

from which the bound

$$\sup_{0\leq t\leq T} \|U(t)\|_{\boldsymbol{\ell}^\infty} \leq \sqrt{2H_\infty} \tag{2.5.11}$$

follows. From (2.5.11) and (2.4.30a) we then obtain the estimate

$$\|\zeta(t)\|_{\boldsymbol{\ell}^\infty} \leq \|\zeta(0)\|_{\boldsymbol{\ell}^\infty} + \sqrt{2H_\infty}t. \tag{2.5.12}$$

Another useful estimate coming from (2.5.9) is

$$|\bar{r}_j| \leq \sqrt{2(\mathrm{D}_+ y_j)h_j}. \tag{2.5.13}$$

Now that Lemma 2.5.4 has established $\mathrm{D}_+ y_j(t) \geq 0$ in the short-time solution for $t \in [0, T]$, we can apply Lemma 2.3.9 with $a_j = \mathrm{D}_+ y_j$. Indeed, the sequences g, γ, k, and κ solve (2.4.19) and belong to $\boldsymbol{\ell}^*$ for $t \in [0, T]$, and so they correspond to the unique decaying solution. These properties contained in Lemmas 2.3.9 and 2.4.3 are essential to establish the *a priori* estimates contained in the next lemma.

Lemma 2.5.5 (A priori relations and inequalities for the kernels)**.** *As a consequence of establishing the preservation of the summation kernels and their sign properties over time, we have the identities*

$$\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{D}_+ y_j)|\gamma_{i,j}| = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} |\mathrm{D}_{j+} g_{i,j}| = 2\,\|g\|_{\boldsymbol{\ell}^\infty}\,, \tag{2.5.14a}$$

$$\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{D}_+ y_j)|\kappa_{i,j}| = \mathit{\Delta}\xi \sum_{i\in\mathbb{Z}} |\mathrm{D}_{j-} k_{i,j}| = 2\,\|k\|_{\boldsymbol{\ell}^\infty}\,, \tag{2.5.14b}$$

as well as

$$\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{D}_+ y_j) g_{i,j} = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{A}[\mathrm{D}_+ y] g_i)_j = 1, \tag{2.5.15a}$$

$$\mathit{\Delta}\xi \sum_{i\in\mathbb{Z}} (\mathrm{D}_+ y_j) k_{i,j} = \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{B}[\mathrm{D}_+ y] k_i)_j = 1, \tag{2.5.15b}$$

and the bounds

$$\|g\|_{\boldsymbol{\ell}^\infty}\,, \|k\|_{\boldsymbol{\ell}^\infty}\,, \|\gamma\|_{\boldsymbol{\ell}^\infty}\,, \|\kappa\|_{\boldsymbol{\ell}^\infty} \leq 1, \tag{2.5.16}$$

$$\begin{aligned} &\|g\|_{\boldsymbol{\ell}^1} \leq 1 + 2\,\|\zeta\|_{\boldsymbol{\ell}^\infty}\,, &&\|k\|_{\boldsymbol{\ell}^1} \leq 1 + 2\,\|\zeta\|_{\boldsymbol{\ell}^\infty}\,, \\ &\|\gamma\|_{\boldsymbol{\ell}^1} \leq 2\left[1 + \|\zeta\|_{\boldsymbol{\ell}^\infty}\right], &&\|\kappa\|_{\boldsymbol{\ell}^1} \leq 2\left[1 + \|\zeta\|_{\boldsymbol{\ell}^\infty}\right]. \end{aligned} \tag{2.5.17}$$

Proof of Lemma 2.5.5. To prove (2.5.14a) we use $\mathrm{D}_+ y_j \geq 0$ and (2.4.19) for the leftmost equalities, while for the rightmost equalities we use the monotonicity properties of (2.3.38) to write

$$\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} |\mathrm{D}_{j+} g_{i,j}| = \mathit{\Delta}\xi \sum_{j=-\infty}^{i-1} \mathrm{D}_{j+} g_{i,j} - \mathit{\Delta}\xi \sum_{j=i}^{\infty} \mathrm{D}_{j+} g_{i,j} = 2 g_{i,i} = 2\left\| g \right\|_{\boldsymbol{\ell}^\infty}.$$

We obtain (2.5.14a) in the same way. To obtain (2.5.15), we use the definitions of the operators A in (2.2.16) and B in (2.3.33), and apply telescopic cancellation to the differences $\mathrm{D}_{j+}\gamma_{i,j}$ and $\mathrm{D}_{j-}\kappa_{i,j}$ in the identities (2.4.19). In the same manner, telescopic cancellation applied to (2.4.19) yields

$$\gamma_{i,j} = \begin{cases} \mathit{\Delta}\xi \sum_{m=-\infty}^{j} (\mathrm{D}_+ y_m) g_{i,m}, & j \leq i-1, \\ -\mathit{\Delta}\xi \sum_{m=j+1}^{\infty} (\mathrm{D}_+ y_m) g_{i,m}, & j \geq i, \end{cases}$$

Using the fact that $\mathrm{D}_+ y_j \geq 0$ and $g_{i,j} \geq 0$, the triangle inequality and (2.5.15) yield (2.5.16) for γ. We proceed similarly for κ. For g, observe that, using (2.4.19), we can rewrite them as

$$\begin{aligned} g_{i,j} &= \sum_{m\in\mathbb{Z}} g_{i,m}\delta_{j,m} \\ &= \mathit{\Delta}\xi \sum_{m\in\mathbb{Z}} g_{i,m}\left[(\mathrm{D}_+ y_m) g_{j,m} - \mathrm{D}_{m-}\gamma_{j,m}\right] \\ &= \mathit{\Delta}\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ y_m)\left[g_{i,m} g_{j,m} + \gamma_{i,m}\gamma_{j,m}\right]. \end{aligned} \tag{2.5.18}$$

Using the decay at infinity we can then write

$$\begin{aligned} (g_{i,i})^2 &= \sum_{m=i}^{+\infty} \left[(g_{i,m+1})^2 - (g_{i,m})^2\right] \\ &= \mathit{\Delta}\xi \sum_{m=i}^{+\infty} \left[g_{i,m+1} + g_{i,m}\right] \mathrm{D}_{m+} g_{i,m} \\ &= \mathit{\Delta}\xi \sum_{m=i}^{+\infty} \left[g_{i,m+1} + g_{i,m}\right] (\mathrm{D}_+ y_m)|\gamma_{i,m}| \\ &\leq 2\mathit{\Delta}\xi \sum_{m=i}^{+\infty} g_{i,m}(\mathrm{D}_+ y_m)|\gamma_{i,m}| \\ &\leq \mathit{\Delta}\xi \sum_{m=i}^{+\infty} (\mathrm{D}_+ y_m)\left[(g_{i,m})^2 + (\gamma_{i,m})^2\right] \end{aligned}$$

$$\begin{aligned}&\leq \mathit{\Delta}\xi \sum_{m\in\mathbb{Z}} (\mathrm{D}_+ y_m) \left[(g_{i,m})^2 + (\gamma_{i,m})^2\right] \\ &= g_{i,i},\end{aligned}$$

where we have used (2.3.38) for the first inequality, and (2.5.18) for the final identity. The bound $g_{i,i} \leq 1$ follows, and note how the above estimates align nicely with those in the proof of Proposition 2.4.1. We then use $0 \leq g_{i,j} \leq g_{i,i}$ from (2.3.38) to conclude. A similar procedure can be applied to prove that $k_{i,j} \leq 1$. Furthermore, we have

$$\begin{aligned}\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} g_{i,j} &= \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \left[\mathrm{D}_+ y_j - \mathrm{D}_+ \zeta_j\right] g_{i,j} \\ &= 1 + \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \zeta_{j+1} (\mathrm{D}_{j+} g_{i,j}), \text{ from (2.5.15)}, \\ &= 1 + \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \zeta_{j+1} (\mathrm{D}_+ y_j) \gamma_{i,j}, \text{ from (2.4.19)}, \\ &\leq 1 + \|\zeta\|_{\boldsymbol{\ell}^\infty} \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{D}_+ y_j) |\gamma_{i,j}|,\end{aligned}$$

and the result on the $\boldsymbol{\ell}^1$ bound of g follows from (2.5.14) and (2.5.16). A similar procedure proves the bound on $\|k\|_{\boldsymbol{\ell}^1}$. For the bound on $\|\gamma\|_{\boldsymbol{\ell}^1}$ we find

$$\begin{aligned}&\mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} |\gamma_{i,j}| \\ &= \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \left[\mathrm{D}_+ y_j - \mathrm{D}_+ \zeta_j\right] |\gamma_{i,j}| \\ &= 2 g_{i,i} - \mathit{\Delta}\xi \sum_{j=-\infty}^{i-1} (\mathrm{D}_+ \zeta_j) \gamma_{i,j} + \mathit{\Delta}\xi \sum_{j=i}^{+\infty} (\mathrm{D}_+ \zeta_j) \gamma_{i,j} \\ &= 2\, \|g\|_{\boldsymbol{\ell}^\infty} - 2 \zeta_i \gamma_{i,i-1} + \mathit{\Delta}\xi \sum_{j=-\infty}^{i-1} \zeta_j (\mathrm{D}_{j-} \gamma_{i,j}) - \mathit{\Delta}\xi \sum_{j=i}^{+\infty} \zeta_j (\mathrm{D}_{j-} \gamma_{i,j}) \\ &= 2\, \|g\|_{\boldsymbol{\ell}^\infty} + (1 - 2\gamma_{i,i-1}) \zeta_i + \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} \operatorname{sgn}\left(i - j - \frac{1}{2}\right) \zeta_i (\mathrm{D}_+ y_j) g_{i,j} \\ &\leq 2\, \|g\|_{\boldsymbol{\ell}^\infty} + \|\zeta\|_{\boldsymbol{\ell}^\infty} \left[|1 - 2\gamma_{i,i-1}| + \mathit{\Delta}\xi \sum_{j\in\mathbb{Z}} (\mathrm{D}_+ y_j) g_{i,j} \right],\end{aligned}$$

where in the second equality we use Lemma 2.3.9, the third equality uses summation by parts (2.3.5), and the fourth is due to the kernel definition property (2.4.19). Then the result follows from (2.5.15), (2.5.16), and $0 \leq \gamma_{i,i-1} \leq 1$. A similar procedure proves the bound on $\|\kappa\|_{\boldsymbol{\ell}^1}$. □

A direct consequence of (2.5.16) is that the $\boldsymbol{\ell}^\infty$-norms of the kernels remain bounded by 1 for all time. Moreover, combining (2.5.17) with (2.5.12) we find that the $\boldsymbol{\ell}^1$-norms remain bounded for any finite t, namely

$$\begin{aligned} \|g(t)\|_{\boldsymbol{\ell}^1}, \|k(t)\|_{\boldsymbol{\ell}^1} &\leq 1 + 2\left[\|\zeta(0)\|_{\boldsymbol{\ell}^\infty} + \sqrt{2H_\infty}t\right], \\ \|\gamma(t)\|_{\boldsymbol{\ell}^1}, \|\kappa(t)\|_{\boldsymbol{\ell}^1} &\leq 2\left[1 + \|\zeta(0)\|_{\boldsymbol{\ell}^\infty} + \sqrt{2H_\infty}t\right]. \end{aligned} \tag{2.5.19}$$

Furthermore, Lemma 2.5.5 allows us to find a bound similar to (2.5.11) for $\|R\|_{\boldsymbol{\ell}^\infty}$ and $\|Q\|_{\boldsymbol{\ell}^\infty}$. For Q we find

$$\begin{aligned} \|Q\|_{\boldsymbol{\ell}^\infty} &\leq \|g * (U(\mathrm{D}_+U))\|_{\boldsymbol{\ell}^\infty} + \|\kappa * (h + \rho_\infty \bar{r})\|_{\boldsymbol{\ell}^\infty} \\ &\leq \|g\|_{\boldsymbol{\ell}^\infty}\|U(\mathrm{D}_+U)\|_{\boldsymbol{\ell}^1} + \|\kappa\|_{\boldsymbol{\ell}^\infty}\|h\|_{\boldsymbol{\ell}^1} + \rho_\infty\|\kappa * |\bar{r}|\,\|_{\boldsymbol{\ell}^\infty}. \end{aligned} \tag{2.5.20}$$

Using (2.5.13) and the Cauchy–Schwarz inequality, we have

$$\rho_\infty\|\kappa * |\bar{r}|\,\|_{\boldsymbol{\ell}^\infty} \leq \frac{1}{2}\rho_\infty^2\|\,|\kappa| * (\mathrm{D}_+y)\|_{\boldsymbol{\ell}^\infty} + \frac{1}{2}\|\,|\kappa| * (2h)\|_{\boldsymbol{\ell}^\infty}$$

which by (2.5.14) and (2.5.16) simplifies to

$$\rho_\infty\|\kappa * |\bar{r}|\,\|_{\boldsymbol{\ell}^\infty} \leq \frac{1}{2}\rho_\infty^2(2\|k\|_{\boldsymbol{\ell}^\infty}) + \|\kappa\|_{\boldsymbol{\ell}^\infty}\|h\|_{\boldsymbol{\ell}^1} \leq \rho_\infty^2 + H_\infty.$$

Using (2.5.10), we get $\|UD_+U\|_{\boldsymbol{\ell}^1} \leq H_\infty$. Hence, from (2.5.20), we get

$$\|Q\|_{\boldsymbol{\ell}^\infty} \leq 3H_\infty + \rho_\infty^2.$$

An analogous estimate for R can be obtained so that we can conclude with the bounds

$$\sup_{0\leq t\leq T}\|R(t)\|_{\boldsymbol{\ell}^\infty} \leq 3H_\infty + \frac{1}{2}\rho_\infty^2, \qquad \sup_{0\leq t\leq T}\|Q(t)\|_{\boldsymbol{\ell}^\infty} \leq 3H_\infty + \rho_\infty^2. \tag{2.5.21}$$

Now we are set to prove global existence for solutions of (2.4.30).

Theorem 2.5.6 (Global existence)**.** *Given initial datum X_0 in the set $\mathcal{B}$ from Definition 2.5.3, the system* (2.4.30) *admits a unique global solution $X \in \mathbf{C}^1([0,\infty), \mathbf{E}_{\Delta\xi})$, such that $X \in \mathcal{B}$ for all times. In particular, for $t > 0$, the norm $\|X(t)\|_{\mathbf{E}_{\Delta\xi}}$ is bounded by $C\,\|X(0)\|_{\mathbf{E}_{\Delta\xi}}$ for a constant C depending only on t, the total energy H_∞, the asymptotic density ρ_∞, and $\|\zeta(0)\|_{\boldsymbol{\ell}^\infty}$.*

Proof. The solution has a finite time of existence T only if

$$\|X\|_{\mathbf{E}_{\Delta\xi}} = \|\zeta\|_{\mathbf{V}_{\Delta\xi}} + \|U\|_{\mathbf{h}^1} + \|H\|_{\mathbf{V}_{\Delta\xi}} + \|\bar{r}\|_{\boldsymbol{\ell}^2}$$

blows up as t approaches T. Otherwise the solution can be prolonged by a small time interval by Theorem 2.5.2. Let X be the short-time solution given by 2.5.2 for initial datum X_0. We will prove $\sup_{0\leq t\leq T}\|X\|_{\mathbf{E}_{\Delta\xi}} < \infty$.

From the definition of the $\mathbf{h}^1$-norm and (2.3.4) we find that the right-hand side of (2.4.30a) is bounded in the $\mathbf{V}_{\Delta\xi}$-norm by $\frac{2+\sqrt{2}}{2}\|U\|_{\mathbf{h}^1}$, while the right-hand side of (2.4.30d) is bounded in $\boldsymbol{\ell}^2$-norm by $\rho_\infty\|U\|_{\mathbf{h}^1}$. Next, we estimate the right hand side of (2.4.30b),

$$\begin{aligned}\|Q\|_{\mathbf{h}^1} &\leq \|Q\|_{\boldsymbol{\ell}^2} + \|\mathrm{D}_+Q\|_{\boldsymbol{\ell}^2} \leq \|Q\|_{\boldsymbol{\ell}^2} + \|R(1+\mathrm{D}_+\zeta) - h - \rho_\infty\bar{r}\|_{\boldsymbol{\ell}^2}\\ &\leq \|Q\|_{\boldsymbol{\ell}^2} + \|R\|_{\boldsymbol{\ell}^2} + \|R\|_{\boldsymbol{\ell}^\infty}\|\mathrm{D}_+\zeta\|_{\boldsymbol{\ell}^2} + \|h+\rho_\infty\bar{r}\|_{\boldsymbol{\ell}^2},\end{aligned}$$

where we have used the definition of the $\mathbf{h}^1$-norm, (2.4.28) and the decomposition $\mathrm{D}_+y_j = 1 + \mathrm{D}_+\zeta_j$. Then, recalling the definitions (2.5.3a) and (2.5.3b) and applying the Young inequality (2.4.24) to the final expression above we see that it is bounded by

$$\begin{aligned}&\|g\|_{\boldsymbol{\ell}^1}\|U(\mathrm{D}_+U)\|_{\boldsymbol{\ell}^2} + \|\gamma\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|\kappa\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|h+\rho_\infty\bar{r}\|_{\boldsymbol{\ell}^2} + \|\gamma\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|\kappa\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|U(\mathrm{D}_+U)\|_{\boldsymbol{\ell}^2}\\ &\qquad + \|R\|_{\boldsymbol{\ell}^\infty}\|\mathrm{D}_+\zeta\|_{\boldsymbol{\ell}^2} + \|k\|_{\boldsymbol{\ell}^1}\|h+\rho_\infty\bar{r}\|_{\boldsymbol{\ell}^2} + \|h+\rho_\infty\bar{r}\|_{\boldsymbol{\ell}^2}\\ &\leq \left[\|g\|_{\boldsymbol{\ell}^1} + \|\gamma\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|\kappa\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\right]\|U\|_{\boldsymbol{\ell}^\infty}\|\mathrm{D}_+U\|_{\boldsymbol{\ell}^2} + \|R\|_{\boldsymbol{\ell}^\infty}\|\mathrm{D}_+\zeta\|_{\boldsymbol{\ell}^2}\\ &\qquad + \left[\|k\|_{\boldsymbol{\ell}^1} + \|\gamma\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\|\kappa\|_{\boldsymbol{\ell}^1}^{\frac{1}{2}}\right][\|h\|_{\boldsymbol{\ell}^2} + \rho_\infty\|\bar{r}\|_{\boldsymbol{\ell}^2}].\end{aligned}$$

Then, applying (2.5.10), (2.5.11), (2.5.19), (2.5.21) and the definitions of the $\mathbf{V}_{\Delta\xi}$- and $\mathbf{h}^1$-norms we obtain

$$\begin{aligned}\|Q\|_{\mathbf{h}^1} &\leq \left(3 + 4[\|\zeta(0)\|_{\boldsymbol{\ell}^\infty} + \sqrt{2H_\infty}t]\right)[\|U\|_{\mathbf{h}^1} + \|H\|_{\mathbf{V}_{\Delta\xi}} + \rho_\infty\|\bar{r}\|_{\boldsymbol{\ell}^2}]\\ &\qquad + \left(3H_\infty + \frac{1}{2}\rho_\infty^2\right)\|\zeta\|_{\mathbf{V}_{\Delta\xi}}.\end{aligned}$$

Finally, the $\mathbf{V}_{\Delta\xi}$-norm of the right-hand side of (2.4.30c) can be estimated as

$$\begin{aligned}\|U(\tau R)\|_{\mathbf{V}_{\Delta\xi}} &= \|U(\tau R)\|_{\boldsymbol{\ell}^\infty} + \left\|[U^2 - R](\mathrm{D}_+U) - UQ[1+\mathrm{D}_+\zeta]\right\|_{\boldsymbol{\ell}^2}\\ &\leq \|R\|_{\boldsymbol{\ell}^\infty}\|U\|_{\boldsymbol{\ell}^\infty} + [\|U\|_{\boldsymbol{\ell}^\infty}^2 + \|R\|_{\boldsymbol{\ell}^\infty}]\|\mathrm{D}_+U\|_{\boldsymbol{\ell}^2}\\ &\qquad + \|Q\|_{\boldsymbol{\ell}^\infty}\|U\|_{\boldsymbol{\ell}^2} + \|Q\|_{\boldsymbol{\ell}^\infty}\|U\|_{\boldsymbol{\ell}^\infty}\|\mathrm{D}_+\zeta\|_{\boldsymbol{\ell}^2}\end{aligned}$$

$$\leq \left(\frac{2+\sqrt{2}}{2}(3H_\infty + \rho_\infty^2) + 2H_\infty\right) \|U\|_{\mathbf{h}^1}$$
$$+ \sqrt{2H_\infty}\left(3H_\infty + \frac{1}{2}\rho_\infty^2\right) \|\zeta\|_{\mathbf{V}_{\Delta\xi}},$$

where we again use the notation $(\tau R)_j = R_{j-1}$. In the first identity above we have employed (2.4.28), while in the final line we have used the definitions of the $\mathbf{V}_{\Delta\xi}$- and $\mathbf{h}^1$-norms together with (2.3.4), (2.5.11) and (2.5.21).

Gathering all the above estimates of the right-hand sides, writing (2.4.30) in integral form, and taking norms we obtain the following inequality for $X(t) = (\zeta, U, H, \bar{r})(t)$,

$$\|X(t)\|_{\mathbf{E}_{\Delta\xi}} \leq \|X(0)\|_{\mathbf{E}_{\Delta\xi}} + C(H_\infty, \|\zeta(0)\|_{\boldsymbol{\ell}^\infty}, \rho_\infty) \int_0^t (1+s) \|X(s)\|_{\mathbf{E}_{\Delta\xi}} \, ds,$$

for $t \in [0,T]$ and some constant $C(H_\infty, \|\zeta(0)\|_{\boldsymbol{\ell}^\infty}, \rho_\infty)$ depending only on H_∞, $\|\zeta(0)\|_{\boldsymbol{\ell}^\infty}$ and ρ_∞. Grönwall's inequality then yields

$$\|X(t)\|_{\mathbf{E}_{\Delta\xi}} \leq \|X(0)\|_{\mathbf{E}_{\Delta\xi}} \exp\left\{C(H_\infty, \|\zeta(0)\|_{\boldsymbol{\ell}^\infty}, \rho_\infty)\left[t + \frac{1}{2}t^2\right]\right\}$$

for $t \in [0,T]$, which shows that $\|X(T)\|_{\mathbf{E}_{\Delta\xi}}$ is bounded, and we may according to Theorem 2.5.2 extend our solution indefinitely.

In retrospect, with the estimates (2.5.11) and (2.5.21) in hand, we see that a Grönwall estimate applied to (2.5.8) shows that the $\boldsymbol{\ell}^\infty$-norm of D_+y, D_+U, h and $\bar{r}$ at time $t \in [0,T]$ is bounded by their $\boldsymbol{\ell}^\infty$-norm at time $t = 0$ times a factor $\exp\{C(H_\infty, \rho_\infty)t\}$, where the constant $C(H_\infty, \rho_\infty)$ depends only on H_∞ and ρ_∞. □

We also mention that if $\rho > 0$ initially, the smoothness of the initial data for the 2CH system (2.1.2) is preserved, see [30]. This is because the characteristics do not collide in this case, and y_ξ remains positive for all time. In the discrete setting, this property takes the form of a lower bound for D_+y. For any given time T, there exists a constant $C > 0$ depending on $\max_{t\in[0,T]} \|X(t)\|_{\mathbf{E}_{\Delta\xi}}$, ρ_∞ and T such that

$$(\mathrm{D}_+y)_j(t) \geq \frac{\rho_{0,j}^2}{C},$$

for all j and $t \in [0,T]$. This follows from (c) in Definition 2.5.3. Thus, if $\rho_{0,j} > 0$, we will have $y_j(t) < y_{j+1}(t)$ for all time.

2.5.2 The choice of initial data

In this subsection we will elaborate upon the choice of initial data for (2.4.30). Let us first consider (2.1.2) where we in addition to $u_0 \in \mathbf{H}^1$, assume $u_{0,x}, \rho_0 - \rho_\infty \in \mathbf{L}^2 \cap \mathbf{L}^\infty$. This allows us to choose $y_j = \xi_j$ as the initial positions of the characteristics, since there is no concentration of energy in any points. As a consequence, $\zeta_j(0) = 0$ and the initial conditions for (2.4.30) can be chosen as $U_j(0) = U_0(\xi_j)$ and $\rho_j(0) = \rho_0(\xi_j)$. Then we define initial values for the auxiliary variables through

$$\begin{aligned} \bar{r}_j(0) &= \rho_j(0) - \rho_\infty, \\ H_j(0) &= \varDelta\xi \sum_{m=-\infty}^{j-1} \left[(U_m(0))^2 + (\mathrm{D}_+ U_m(0))^2 + (\bar{r}_m(0))^2 \right]. \end{aligned}$$

Moreover, since in this case $g_{i,j}(0)$, $k_{i,j}(0)$ are Green's functions for $\mathrm{A}[\mathbf{1}] = \mathrm{B}[\mathbf{1}] = \mathrm{Id} - \mathrm{D}_- \mathrm{D}_+$, they can be computed explicitly. Indeed, for $\mathrm{D}_+ y_j = 1$ we have

$$g_{i,j}(0) = k_{i,j}(0) = \frac{(\lambda^+)^{-|j-i|}}{\sqrt{4 + \varDelta\xi^2}},$$

with λ^+ defined in (2.3.10). Thus, initially we have the Eulerian Green's sequences as computed in [34].

On the other hand, in our construction of the Green's functions we have allowed for $\mathrm{D}_+ y_j = 0$, and so our discretization should be able to handle singular initial data as well. To fix the ideas we consider the CH equation (2.1.1) only, that is, we set $\rho_0 \equiv 0, \rho_\infty = 0$ and thus $r, \bar{r} \equiv 0$. Moreover, we take $u_0 \in \mathbf{H}^1$ and the positive, finite Radon measure μ_0 to be given, where the absolutely continuous part of μ_0 satisfies $d\mu_{0,\mathrm{ac}} = (u_0^2 + u_{0,x}^2)\,dx$, and we allow its singular part to be atomic. This means that $\mu_0((-\infty, x))$ may contain jump discontinuities.

In the usual manner we introduce the equispaced grid on $\mathbb{R}$. Inspired by works on conservative solutions of (2.1.1) in Lagrangian coordinates, we then define

$$y_j(0) \coloneqq \sup\{x \ : \ \mu_0((-\infty, x)) + x < \xi_j\}.$$

This is of course the same as interpolating the function

$$y_0(\xi) \coloneqq \sup\{x \ : \ \mu_0((-\infty, x)) + x < \xi\} \tag{2.5.22}$$

in the gridpoints ξ_j. In fact, (2.5.22) is given in [36, Eq. (3.21a)], and we can use the results therein to show that our choice of initial data will satisfy Definition 2.5.3. We also adopt their definition of U_0, $U_0(\xi) = u_0 \circ y_0(\xi)$, in [36, Eq. (3.21)], but we will have to modify the definition of H to satisfy our discrete identity (2.5.9). In our endeavor we will use that the continuous-setting variables $(y_0, U_0, H_0) \in \mathbf{V} \times \mathbf{H}^1 \times \mathbf{V}$. From [36] we have $|y_0(\xi) - \xi| \leq \mu_0(\mathbb{R})$, and since the total energy $\mu_0(\mathbb{R})$ is bounded, we have $\|y_0 - \mathrm{Id}\|_{\mathbf{L}^\infty} \leq \mu_0(\mathbb{R})$. Since $y_j(0) = y_0(\xi_j)$, this carries directly over to our setting, $|y_j(0) - \xi_j| \leq \mu_0(\mathbb{R})$, meaning $\|\zeta(0)\|_{\boldsymbol{\ell}^\infty} \leq \mu_0(\mathbb{R})$. Moreover, they prove $\xi \mapsto y_0(\xi)$ to be 1-Lipschitz, which yields

$$|y_0(\xi_{j+1}) - y_0(\xi_j)| \leq |\xi_{j+1} - \xi_j| = \mathit{\Delta}\xi \implies |\mathrm{D}_+ y_j(0)| \leq 1,$$

thus $\mathrm{D}_+ y(0) \in \boldsymbol{\ell}^\infty$. They also prove $\xi \to \int_{-\infty}^{y(\xi)} u_x^2(x)dx$ to be 1-Lipschitz. Then we have

$$\begin{aligned} |U_0(\xi_{j+1}) - U_0(\xi_j)| &= \left| \int_{y(\xi_j)}^{y(\xi_{j+1})} u(x)\,dx \right| \\ &\leq \sqrt{y(\xi_{j+1}) - y(\xi_j)} \sqrt{\int_{y(\xi_j)}^{y(\xi_{j+1})} u_x^2(x)dx}. \end{aligned} \tag{2.5.23}$$

Using that both factors in the final expression of (2.5.23) are 1-Lipschitz we obtain $|U_{j+1}(0) - U_j(0)| \leq \mathit{\Delta}\xi$, implying $|\mathrm{D}_+ U_j(0)| \leq 1$ and $\mathrm{D}_+ U(0) \in \boldsymbol{\ell}^\infty$. In addition, as $u_0 \in \mathbf{L}^\infty$ it is clear from $U_j(0) = u(y_j(0))$ that $\|U(0)\|_{\boldsymbol{\ell}^\infty} \leq \|u_0\|_{\mathbf{L}^\infty}$.

Now we need to choose H_j in such a way as to satisfy property (c) in Definition 2.5.3, and we will separate two possible cases. If $\mathrm{D}_+ y_j(0) > 0$ we define $h_j(0) \geq 0$ such that it satisfies $2h_j(\mathrm{D}_+ y_j) = (U_j)^2(\mathrm{D}_+ y_j)^2 + (\mathrm{D}_+ U_j)^2$. On the other hand, if $\mathrm{D}_+ y_j(0) = 0$ we set $h_j(0) = \frac{1}{2}$. Then we define $H_j(0) = \mathit{\Delta}\xi \sum_{m=-\infty}^{j-1} h_m(0)$. Let us estimate $h_j(0)$ in the case $\mathrm{D}_+ y_j(0) > 0$, where we note that another takeaway from (2.5.23) is $|U_{j+1}(0) - U_j(0)| \leq \sqrt{\mathit{\Delta}\xi}\sqrt{y_{j+1}(0) - y_j(0)}$, or equivalently $|\mathrm{D}_+ U_j(0)| \leq \sqrt{\mathrm{D}_+ y_j(0)}$. Using this and $\mathrm{D}_+ y_j(0) \leq 1$ together with property (c) we find

$$2h_j = U_j^2 \mathrm{D}_+ y_j + \frac{(\mathrm{D}_+ U_j)^2}{\mathrm{D}_+ y_j} \leq U_j^2 + 1 \leq \|u_0\|_{\mathbf{L}^\infty}^2 + 1.$$

Thus, $h(0) \in \boldsymbol{\ell}^\infty$. Finally,

$$h_j(0) + \mathrm{D}_+ y_j(0) = \begin{cases} \frac{1}{2} > 0, & \mathrm{D}_+ y_j(0) = 0, \\ h_j(0) + \mathrm{D}_+ y(0) > 0, & \mathrm{D}_+ y_j(0) > 0, \end{cases}$$

and the requirements (a)–(d) in Definition 2.5.3 are satisfied.

It remains to verify $(\zeta(0), U(0), H(0), 0) \in \mathbf{E}_{\Delta\xi}$. We already know that $\zeta(0) \in \boldsymbol{\ell}^\infty$. We know y_0 is continuous, so it follows that $\zeta_0 = y_0 - \mathrm{Id}$ is bounded and continuous, and we may write

$$|\zeta_0(\xi_{j+1}) - \zeta_0(\xi_j)|^2 = \left| \int_{\xi_j}^{\xi_{j+1}} (\zeta_0)_\xi(\xi)\, d\xi \right|^2 \leq \Delta\xi \int_{\xi_j}^{\xi_{j+1}} |(\zeta_0)_\xi(\xi)|^2 \, d\xi,$$

or equivalently

$$\Delta\xi \, |\mathrm{D}_+\zeta_j(0)|^2 \leq \int_{\xi_j}^{\xi_{j+1}} |(\zeta_0)_\xi(\xi)|^2 \, d\xi.$$

Summing over j in the above equation we obtain $\|\mathrm{D}_+\zeta(0)\|_{\boldsymbol{\ell}^2} \leq \|(\zeta_0)_\xi\|_{\mathbf{L}^2}$, and so $\zeta(0) \in \mathbf{V}_{\Delta\xi}$. A completely analogous procedure shows that $\|\mathrm{D}_+U(0)\|_{\boldsymbol{\ell}^2} \leq \|(U_0)_\xi\|_{\mathbf{L}^2}$. For the $\mathbf{L}^2$-norm of U we estimate

$$\begin{aligned}
\Delta\xi \sum_{j\in\mathbb{Z}} |U_j(0)|^2 &= \sum_{j\in\mathbb{Z}} \int_{\xi_j}^{\xi_{j+1}} \left| U_0(\xi) - \int_{\xi_j}^{\xi} (U_0)_\xi(s)\, ds \right|^2 \\
&\leq 2 \sum_{j\in\mathbb{Z}} \int_{\xi_j}^{\xi_{j+1}} |U_0(\xi)|^2 \, d\xi \\
&\quad + 2 \sum_{j\in\mathbb{Z}} \int_{\xi_j}^{\xi_{j+1}} \left(\int_{\xi_j}^{\xi_{j+1}} |(U_0)_\xi(s)| \, ds \right)^2 d\xi \\
&\leq 2 \|U_0\|_{\mathbf{L}^2}^2 + 2 \sum_{j\in\mathbb{Z}} \Delta\xi^2 \int_{\xi_j}^{\xi_{j+1}} |(U_0)_\xi(s)|^2 \, ds,
\end{aligned}$$

which translates into $\|U(0)\|_{\boldsymbol{\ell}^2}^2 \leq 2\|U_0\|_{\mathbf{L}^2}^2 + 2\Delta\xi^2 \|(U_0)_\xi\|_{\mathbf{L}^2}$, and so $U(0) \in \mathbf{h}^1$. Then it remains to check that $H(0) \in \mathbf{V}_{\Delta\xi}$, and from (2.5.9) we estimate

$$\begin{aligned}
2h_j &= U_j^2 \mathrm{D}_+y_j + (\mathrm{D}_+U_j)^2 - 2h_j\mathrm{D}_+\zeta_j \\
&\leq U_j^2 + (\mathrm{D}_+U_j)^2 + 2h_j\, |\mathrm{D}_+\zeta_j| \\
&\leq U_j^2 + (\mathrm{D}_+U_j)^2 + h_j + h_j\, |\mathrm{D}_+\zeta_j|^2 .
\end{aligned}$$

Now, summing over j we find $\|h(0)\|_{\boldsymbol{\ell}^1} \leq \|U(0)\|_{\mathbf{h}^1}^2 + \|h(0)\|_{\boldsymbol{\ell}^\infty} \|\mathrm{D}_+\zeta(0)\|_{\boldsymbol{\ell}^2}^2$, where the right-hand side is bounded by our previous estimates. Since $h_j(0) > 0$, it follows from our definition of $H_j(0)$ that $H_j(0) < H_{j+1}(0)$ and $H_j < \|h(0)\|_{\boldsymbol{\ell}^1}$, which yields $\|H(0)\|_{\boldsymbol{\ell}^\infty} = \|h(0)\|_{\boldsymbol{\ell}^1}$. Finally, we have $\|h(0)\|_{\boldsymbol{\ell}^2} \leq \|h(0)\|_{\boldsymbol{\ell}^\infty} \|h(0)\|_{\boldsymbol{\ell}^1}$, so $H(0) \in \mathbf{V}_{\Delta\xi}$. In conclusion we have the following theorem, where the functions involved should be compared to Definition 3.1 in [36].

Theorem 2.5.7. *We consider initial data of the Camassa–Holm equation given by a pair $u_0 \in \mathbf{H}^1$ and μ_0, where μ_0 is a positive, finite Radon measure whose absolutely continuous part satisfies $d\mu_{0,ac} = (u^2 + u_x^2)\,dx$, while its singular part is atomic. Then we can construct sequences of initial data of the type $(\zeta_0, U_0, H_0, g_0, k_0, \gamma_0, \kappa_0) \in \mathbf{V}_{\Delta\xi} \times \mathbf{h}^1 \times \mathbf{V}_{\Delta\xi} \times (\boldsymbol{\ell}^*)^4$ for the semi-discrete system* (2.4.30)*, which belongs to the set $\mathcal{B}$ in Definition 2.5.3 (for $\bar{r} \equiv 0$).*

The drawback of not having $y_j = \xi_j$ is that we do not have an explicit expression for the initial Green's functions. However, Theorem 2.3.5 guarantees their existence, so the semi-discrete scheme can still be used. In [26], the discretization (2.4.30) as a numerical method for the periodic version of (2.1.2). As the problem then is finite-dimensional, computing the Green's functions amounts to inverting a matrix and we are able to find them for any $\mathrm{D}_+ y_j \geq 0$. An interesting feature is then that we may allow for singular initial data from the very beginning in our numerical experiments. For instance, we could let $\mu_0((-\infty, x))$ be a pure step-function, meaning that all initial energy is concentrated in separated points on the domain, and then our scheme would yield the conservative solutions for this system.

2.6 Convergence of the scheme

In this section we interpolate the solutions of the semi-discrete scheme analyzed in Section 2.5. We shall then show that these interpolated functions converge to the solution of the 2CH system as written in (2.4.5) and (2.4.6). Let us in this section use $Y_{\Delta\xi}$ to denote the tuple of grid functions obtained in Theorem 2.5.6 for $t \in [0, T]$,

$$Y_{\Delta\xi}(t) = (\zeta, U, H, \bar{r})(t) \in \mathbf{E}_{\Delta\xi} = \mathbf{V}_{\Delta\xi} \times \mathbf{h}^1 \times \mathbf{V}_{\Delta\xi} \times \boldsymbol{\ell}^2, \tag{2.6.1}$$

to avoid confusing the sequence $\{U_j\}_{j\in\mathbb{Z}} \in \mathbf{h}^1$ with the reference solution $U \in \mathbf{H}^1$, etc. In order to ease notation below, we will write $\|Y_{\Delta\xi}\|$ for $\sup_{0\leq t\leq T} \|Y_{\Delta\xi}(t)\|$. We define the interpolated functions as follows

$$\begin{aligned}
V_\Delta(t,\xi) &= \sum_{j\in\mathbb{Z}} \left[V_j(t) + (\xi - \xi_j)(\mathrm{D}_+ V_j(t))\right] \chi_j(\xi), \\
\bar{r}_\Delta(t,\xi) &= \sum_{j\in\mathbb{Z}} \bar{r}_j(t)\chi_j(\xi), \\
R_\Delta(t,\xi) &= \sum_{j\in\mathbb{Z}} \left[R_j(t) + (\xi - \xi_{j+1})(\mathrm{D}_- R_j(t))\right] \chi_j(\xi),
\end{aligned} \tag{2.6.2}$$

where V is a placeholder for ζ, U, H, and Q, while $\chi_j(\xi)$ denotes the indicator function for the interval $[\xi_j, \xi_{j+1})$. We also introduce the functions

$$y_\Delta(t,\xi) := \xi + \zeta_\Delta(t,\xi), \qquad r_\Delta(t,\xi) := \bar{r}_\Delta(t,\xi) + \rho_\infty \frac{\partial y_\Delta(t,\xi)}{\partial \xi}. \tag{2.6.3}$$

Observe that the interpolated functions above are piecewise linear and continuous, except for $r_\Delta, \bar{r}_\Delta$ which are piecewise constant. In particular we note the identity

$$R_j + (\xi - \xi_{j+1})(\mathrm{D}_- R_j) = R_{j-1} + (\xi - \xi_j)(\mathrm{D}_- R_j), \quad \xi \in [\xi_j, \xi_{j+1}],$$

which shows $R_\Delta(t, \xi_j) = R_{j-1}$. Let us also recall the definition of the space $\mathbf{E}$ in (2.4.8). A consequence of Theorem 2.5.6 is that the tuple of interpolated functions

$$X_\Delta(t) := (\zeta_\Delta(t,\cdot), U_\Delta(t,\cdot), H_\Delta(t,\cdot), \bar{r}_\Delta(t,\cdot)) \tag{2.6.4}$$

satisfies $X_\Delta(t) \in \mathbf{C}^1([0,T], \mathbf{E})$ for any fixed $T > 0$ and $\Delta\xi > 0$. Let us now consider a given initial datum $X_0 = (\zeta_0, U_0, H_0, \bar{r}_0) \in \mathbf{E}$ for the equivalent 2CH system (2.4.5). Assume we have a sequence of initial data $Y_{\Delta\xi,0} \in \mathbf{E}_{\Delta\xi}$ such that the interpolation of $Y_{\Delta\xi,0}$, denoted by $X_{\Delta,0}$, converges to X_0, i.e.,

$$\lim_{\Delta\xi \to 0} \|X_{\Delta,0} - X_0\|_{\mathbf{E}} = 0. \tag{2.6.5}$$

For $T > 0$ and each $Y_{\Delta\xi,0}$, let $Y_{\Delta\xi}$ be the corresponding solution given by Theorem 2.5.6. Furthermore, we denote by $X \in \mathbf{C}([0,T], \mathbf{E})$ the solution to (2.4.5) with initial datum X_0, while $X_\Delta \in \mathbf{C}([0,T], \mathbf{E})$ is the function interpolated from $Y_{\Delta\xi}$ using (2.6.2). Then we have the following convergence result.

Theorem 2.6.1 (Convergence)**.** *The approximation X_Δ in* (2.6.4) *converges to the solution X of the 2CH system* (2.4.5) *in* $\mathbf{C}([0,T], \mathbf{E})$.

Proof of Theorem 2.6.1. The strategy of the proof is to show that our interpolated functions $(y_\Delta, U_\Delta, H_\Delta, r_\Delta)$ satisfy (2.4.5) and (2.4.6), where we allow for a small error of order $\mathcal{O}(\Delta\xi)$. For (2.4.5a), (2.4.5b), and (2.4.5d), we observe that, by construction, we have

$$\frac{\partial y_\Delta}{\partial t} = U_\Delta, \quad \frac{\partial U_\Delta}{\partial t} = -Q_\Delta, \quad \frac{\partial r_\Delta}{\partial t} = 0$$

due to (2.4.28), (2.4.30a), (2.4.30b), and (2.4.30d). Thus, the three linear equations in (2.4.5) are satisfied exactly by our interpolants. The next step is to check how well H_Δ satisfies (2.4.5c), and we find

$$\begin{aligned}\frac{\partial H_\Delta}{\partial t} &= -\sum_{j\in\mathbb{Z}} \left[U_j R_{j-1} + (\xi-\xi_j)[U_j(\mathrm{D}_- R_j) + R_j(\mathrm{D}_+ U_j)]\right]\chi_j \\ &= -\sum_{j\in\mathbb{Z}} \left[U_j\left[R_{j-1} + (\xi-\xi_j)(\mathrm{D}_- R_j)\right] + R_j(\xi-\xi_j)(\mathrm{D}_+ U_j)\right]\chi_j \\ &= -\sum_{j\in\mathbb{Z}} \left[U_j + (\xi-\xi_j)(\mathrm{D}_+ U_j)\right]\left[R_{j-1} + (\xi-\xi_j)(\mathrm{D}_- R_j)\right]\chi_j \\ &\quad + \sum_{j\in\mathbb{Z}} (\xi-\xi_j)(\xi-\xi_{j+1})(\mathrm{D}_+ U_j)(\mathrm{D}_- R_j)\chi_j \\ &= -U_\Delta R_\Delta + \sum_{j\in\mathbb{Z}} (\xi-\xi_j)(\xi-\xi_{j+1})(\mathrm{D}_+ U_j)(\mathrm{D}_- R_j)\chi_j .\end{aligned}$$

This identity then implies

$$\left(\frac{\partial H_\Delta}{\partial t} + U_\Delta R_\Delta\right)_\xi = \sum_{j\in\mathbb{Z}} (2\xi - \xi_j - \xi_{j+1})(\mathrm{D}_+ U_j)(\mathrm{D}_- R_j)\chi_j ,$$

almost everywhere. Combining the above identities we can estimate the error in the $\mathbf{V}$-norm as follows,

$$\begin{aligned}&\left\|\frac{\partial H_\Delta}{\partial t} + U_\Delta R_\Delta\right\|_{\mathbf{V}} \\ &\leq \Delta\xi^2 \sum_{j\in\mathbb{Z}} |\mathrm{D}_+ U_j|\,|\mathrm{D}_- R_j| + \left(\Delta\xi \sum_{j\in\mathbb{Z}} \Delta\xi^2\, |\mathrm{D}_+ U_j|^2\, |\mathrm{D}_- R_j|^2\right)^{\frac{1}{2}} \\ &\leq \Delta\xi\, \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^2}\, \|\mathrm{D}_- R\|_{\boldsymbol{\ell}^2} + \Delta\xi\, \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^\infty}\, \|\mathrm{D}_- R\|_{\boldsymbol{\ell}^2} \\ &\leq \Delta\xi \left(\|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^2} + \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^\infty}\right) \|(\mathrm{D}_+ y)Q - U(\mathrm{D}_+ U)\|_{\boldsymbol{\ell}^2} \\ &\leq \Delta\xi \left(\|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^2} + \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^\infty}\right)\left(\|\mathrm{D}_+ y\|_{\boldsymbol{\ell}^\infty}\, \|Q\|_{\boldsymbol{\ell}^2} + \|U\|_{\boldsymbol{\ell}^\infty}\, \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^2}\right).\end{aligned} \tag{2.6.6}$$

Now, for the relations (2.4.6), we measure the error in $\mathbf{L}^2$-norm. From (2.4.28), we obtain the relation

$$\frac{\partial y_\Delta}{\partial\xi} Q_\Delta - \frac{\partial R_\Delta}{\partial\xi} - U_\Delta \frac{\partial U_\Delta}{\partial\xi} = \sum_{j\in\mathbb{Z}} (\xi-\xi_j)\left[(\mathrm{D}_+ y_j)(\mathrm{D}_+ Q_j) - (\mathrm{D}_+ U_j)^2\right]\chi_j ,$$

and find

$$\begin{aligned}\left\| \frac{\partial y_\Delta}{\partial \xi} Q_\Delta - \frac{\partial R_\Delta}{\partial \xi} - U_\Delta \frac{\partial U_\Delta}{\partial \xi} \right\|_{\mathbf{L}^2} \\ \leq \Delta\xi \left(\|\mathrm{D}_+ y\|_{\boldsymbol{\ell}^\infty} \|\mathrm{D}_+ Q\|_{\boldsymbol{\ell}^2} + \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^\infty} \|\mathrm{D}_+ U\|_{\boldsymbol{\ell}^2} \right).\end{aligned} \tag{2.6.7}$$

Finally, using (2.4.28) once more, we have

$$\frac{\partial y_\Delta}{\partial \xi} R_\Delta - \frac{\partial S_\Delta}{\partial \xi} - \frac{\partial H_\Delta}{\partial \xi} - \rho_\infty \bar{r}_\Delta = \sum_{j\in\mathbb{Z}} (\xi - \xi_{j+1})(\mathrm{D}_- R_j)(\mathrm{D}_+ y_j)\chi_j$$

which can be estimated as

$$\left\| \frac{\partial y_\Delta}{\partial \xi} R_\Delta - \frac{\partial S_\Delta}{\partial \xi} - \frac{\partial H_\Delta}{\partial \xi} - \rho_\infty \bar{r}_\Delta \right\|_{\mathbf{L}^2} \leq \Delta\xi \|\mathrm{D}_+ y\|_{\boldsymbol{\ell}^\infty} \|\mathrm{D}_- R\|_{\boldsymbol{\ell}^2}. \tag{2.6.8}$$

The estimate (2.6.6) is exactly as we want it, (2.4.5c) is satisfied in the appropriate norm up to some small remainder. However, the estimates (2.6.7) and (2.6.8) require some more work, as we shall see next.

Let us estimate the **E**-norm of the difference between $X_\Delta(T)$ and the exact solution $X(T) := (\zeta, U, H, \bar{r})(T)$. From the above estimates and (2.4.5) we find

$$\begin{aligned}
\|(\zeta_\Delta - \zeta)(T,\cdot)\|_{\mathbf{V}} &\leq \|(\zeta_\Delta - \zeta)(0,\cdot)\|_{\mathbf{V}} + \int_0^T \|(U_\Delta - U)(t,\cdot)\|_{\mathbf{V}}\, dt \\
\|(U_\Delta - U)(T,\cdot)\|_{\mathbf{H}^1} &\leq \|(U_\Delta - U)(0,\cdot)\|_{\mathbf{H}^1} + \int_0^T \|(Q_\Delta - Q)(t,\cdot)\|_{\mathbf{H}^1}\, dt \\
\|(H_\Delta - H)(T,\cdot)\|_{\mathbf{V}} &\leq \|(H_\Delta - H)(0,\cdot)\|_{\mathbf{V}} \\
&\quad + \int_0^T \|(U_\Delta R_\Delta - UR)(t,\cdot)\|_{\mathbf{V}}\, dt \\
&\quad + \Delta\xi C_H(\|Y_{\Delta\xi}\|)T \\
\|(\bar{r}_\Delta - \bar{r})(T,\cdot)\|_{\mathbf{L}^2} &\leq \|(\bar{r}_\Delta - \bar{r})(0,\cdot)\|_{\mathbf{L}^2} \\
&\quad + \rho_\infty \int_0^T \left\| \frac{\partial (U_\Delta - U)(t,\cdot)}{\partial \xi} \right\|_{\mathbf{L}^2} dt,
\end{aligned} \tag{2.6.9}$$

where we have used that the final expression in (2.6.6) can be bounded by $\Delta\xi C_H(\|Y_{\Delta\xi}\|)$ for some constant C_H depending only on $\|Y_{\Delta\xi}\|$.

From (2.6.9), it is clear that we need estimates of $\|Q_\Delta - Q\|_{\mathbf{H}^1}$, $\|R_\Delta - R\|_{\mathbf{L}^\infty}$, and $\|(R_\Delta - R)_\xi\|_{\mathbf{L}^2}$ in terms of

$$\|X_\Delta - X\|_{\mathbf{E}} = \|\zeta_\Delta - \zeta\|_{\mathbf{V}} + \|U_\Delta - U\|_{\mathbf{H}^1} + \|H_\Delta - H\|_{\mathbf{V}} + \|\bar{r}_\Delta - \bar{r}\|_{\mathbf{L}^2},$$

and by definition of the $\mathbf{H}^1$-norm and the Sobolev inequality (2.4.18), it will be sufficient to bound $\|Q_\Delta - Q\|_{\mathbf{H}^1}$ and $\|R_\Delta - R\|_{\mathbf{H}^1}$. To this end, we note that by the estimates (2.6.7) and (2.6.8), it follows that

$$\begin{bmatrix} -\partial_\xi & (y_\Delta)_\xi \\ (y_\Delta)_\xi & -\partial_\xi \end{bmatrix} \circ \begin{bmatrix} R_\Delta \\ Q_\Delta \end{bmatrix} = \begin{bmatrix} U_\Delta (U_\Delta)_\xi \\ (H_\Delta)_\xi + \rho_\infty \bar{r}_\Delta \end{bmatrix} + \Delta\xi \begin{bmatrix} v_\Delta \\ w_\Delta \end{bmatrix} \tag{2.6.10}$$

for some functions $v_\Delta, w_\Delta \in \mathbf{L}^2$ which are bounded by a constant depending only on the norm $\|Y_{\Delta\xi}\|$ of (2.6.1). Recalling (2.4.12) and the operators defined in (2.4.10) we know that $R(t,\xi)$ and $Q(t,\xi)$ can be written as

$$\begin{aligned} R(t,\xi) &= \int_{\mathbb{R}} \kappa[t](\eta,\xi) U U_\xi(t,\eta)\, d\eta + \int_{\mathbb{R}} g[t](\eta,\xi)[H_\xi + \rho_\infty \bar{r}](t,\eta)\, d\eta \\ &= \mathcal{K}(UU_\xi) + \mathcal{G}(H_\xi + \rho_\infty \bar{r}), \\ Q(t,\xi) &= \int_{\mathbb{R}} g[t](\eta,\xi) U U_\xi(t,\eta)\, d\eta + \int_{\mathbb{R}} \kappa[t](\eta,\xi)[H_\xi + \rho_\infty \bar{r}](t,\eta)\, d\eta \\ &= \mathcal{G}(UU_\xi) + \mathcal{K}(H_\xi + \rho_\infty \bar{r}) \end{aligned}$$

with kernels

$$g[t](\eta,\xi) := \tfrac{1}{2} e^{-|y(t,\xi)-y(t,\eta)|}, \qquad \kappa[t](\eta,\xi) := -\operatorname{sgn}(\xi-\eta) g[t](\eta,\xi).$$

Due to the obvious similarities between (2.6.10) and (2.4.16) we would like to generalize the operator identity (2.4.10) by replacing $y(t,\xi)$ with any function $b(t,\xi)$ such that $b(t,\cdot) - \mathrm{Id} \in \mathbf{V}$ and $b_\xi(t,\xi) \geq 0$, in particular this holds for our $y_\Delta(t,\xi)$ in (2.6.3) by virtue of Lemma 2.5.4. This is can be done, and the unique $\mathbf{H}^1$-solution of

$$\begin{bmatrix} -\partial_\xi & b_\xi(t,\xi) \\ b_\xi(t,\xi) & -\partial_\xi \end{bmatrix} \begin{bmatrix} \phi(t,\xi) \\ \psi(t,\xi) \end{bmatrix} = \begin{bmatrix} v(t,\xi) \\ w(t,\xi) \end{bmatrix}$$

for $v(t,\cdot), w(t,\cdot) \in \mathbf{L}^2$ is then

$$\begin{aligned} \phi(t,\xi) &= \int_{\mathbb{R}} \frac{1}{2} e^{-|b(t,\xi)-b(t,\eta)|} \left[w(t,\eta) - \operatorname{sgn}(\xi-\eta) v(t,\eta) \right] d\eta, \\ \psi(t,\xi) &= \int_{\mathbb{R}} \frac{1}{2} e^{-|b(t,\xi)-b(t,\eta)|} \left[v(t,\eta) - \operatorname{sgn}(\xi-\eta) w(t,\eta) \right] d\eta. \end{aligned}$$

Consequently, we can generalize $\mathcal{G}$ and $\mathcal{K}$ from (2.4.10) to be operators from $\mathbf{V} \times \mathbf{L}^2$ to $\mathbf{H}^1$ as follows,

$$\mathcal{G}[t,\xi](b - \mathrm{Id}, f) := \int_{\mathbb{R}} \frac{1}{2} e^{-|b(t,\xi)-b(t,\eta)|} f(\eta)\, d\eta,$$

$$\mathcal{K}[t,\xi](b-\mathrm{Id},f) := -\int_{\mathbb{R}} \operatorname{sgn}(\xi-\eta)\frac{1}{2}e^{-|b(t,\xi)-b(t,\eta)|}f(\eta)\,d\eta.$$

Using these operators, we may write the general solutions $\phi(t,\xi)$, $\psi(t,\xi)$ as

$$\begin{aligned}\phi(t,\xi) &= \mathcal{K}[t,\xi](b-\mathrm{Id},v)+\mathcal{G}[t,\xi](b-\mathrm{Id},w),\\ \psi(t,\xi) &= \mathcal{G}[t,\xi](b-\mathrm{Id},v)+\mathcal{K}[t,\xi](b-\mathrm{Id},w).\end{aligned}$$

An argument analogous to [30, Lem. 3.1] then proves that the operators

$$\mathcal{R}_1[t,\cdot]\colon (\zeta,U,H,\bar{r}) \mapsto \mathcal{K}[t,\cdot](\zeta,UU_\xi)+\mathcal{G}[t,\cdot](\zeta,H_\xi+\rho_\infty\bar{r})$$

and

$$\mathcal{R}_2[t,\cdot]\colon (\zeta,v,w) \mapsto \mathcal{K}[t,\cdot](\zeta,v)+\mathcal{G}[t,\cdot](\zeta,w)$$

are locally Lipschitz as operators from $\mathbf{E}\to\mathbf{H}^1$ and $\mathbf{V}\times(\mathbf{L}^2)^2\to\mathbf{H}^1$ respectively, and the same is true for

$$\mathcal{Q}_1[t,\cdot]\colon (\zeta,U,H,\bar{r}) \mapsto \mathcal{G}[t,\cdot](\zeta,UU_\xi)+\mathcal{K}[t,\cdot](\zeta,H_\xi+\rho_\infty\bar{r})$$

and

$$\mathcal{Q}_2[t,\cdot]\colon (\zeta,v,w) \mapsto \mathcal{G}[t,\cdot](\zeta,v)+\mathcal{K}[t,\cdot](\zeta,w).$$

Finally turning back to the functions we are interested in, we note that, since our interpolants R_Δ and Q_Δ are solutions of (2.6.10), they can be written as

$$\begin{aligned}R_\Delta(t,\xi) &= \mathcal{R}_1[t,\xi]\,(\zeta_\Delta,U_\Delta,H_\Delta,\bar{r}_\Delta)+\Delta\xi\,\mathcal{R}_2[t,\xi](\zeta_\Delta,v_\Delta,w_\Delta),\\ Q_\Delta(t,\xi) &= \mathcal{Q}_1[t,\xi]\,(\zeta_\Delta,U_\Delta,H_\Delta,\bar{r}_\Delta)+\Delta\xi\,\mathcal{Q}_2[t,\xi](\zeta_\Delta,v_\Delta,w_\Delta).\end{aligned}$$

These should then be compared to R and Q for the exact solution, which now can be written as

$$\begin{aligned}R(t,\xi) &= \mathcal{R}_1[t,\xi](\zeta,U,H,\bar{r}),\\ Q(t,\xi) &= \mathcal{Q}_1[t,\xi](\zeta,U,H,\bar{r}).\end{aligned}$$

Then, we write

$$\begin{aligned}Q_\Delta(t,\xi)-Q(t,\xi) = \mathcal{Q}_1\,&(\zeta_\Delta,U_\Delta,H_\Delta,\bar{r}_\Delta)-\mathcal{Q}_1\,(\zeta,U,H,\bar{r})\\ &+\Delta\xi\,\mathcal{Q}_2[t,\xi](\zeta_\Delta,v_\Delta,w_\Delta)\end{aligned}$$

and it follows from the Lipschitz property that

$$\|Q_\Delta(t,\cdot)-Q(t,\cdot)\|_{\mathbf{H}^1} \le C_{Q,1}(\|X_\Delta(t)\|_{\mathbf{E}},\|X(t)\|_{\mathbf{E}})\,\|X_\Delta(t)-X(t)\|_{\mathbf{E}}$$

$$+ \varDelta\xi C_{Q,2}(\|Y_{\varDelta\xi}\|)$$

for constants $C_{Q,1}, C_{Q,2}$, and we can derive an analogous estimate for $\|R_{\varDelta}(t,\cdot) - R(t,\cdot)\|_{\mathbf{H}^1}$.

From the above estimates, the obvious inequality $\|f_\xi\|_{\mathbf{L}^2} \leq \|f\|_{\mathbf{H}^1}$, and $\|f\|_{\mathbf{V}} \leq \frac{2+\sqrt{2}}{2}\|f\|_{\mathbf{H}^1}$ coming from (2.4.18), we may add the equations in (2.6.9) to obtain

$$\begin{aligned}\|X_{\varDelta}(T) - X(T)\|_{\mathbf{E}} \leq \|X_{\varDelta}(0) - X(0)\|_{\mathbf{E}} + \varDelta\xi C_1(\|Y_{\varDelta\xi}\|)T \\ + C_2(\|Y_{\varDelta\xi}\|, \|X\|)\int_0^T \|X_{\varDelta}(t) - X(t)\|_{\mathbf{E}}\, dt,\end{aligned}$$

where we have used that both $\sup_{0\leq t\leq T}\|X_{\varDelta}\|_{\mathbf{E}} \leq C(\|Y_{\varDelta\xi}\|)$ and $\|X\| := \sup_{0\leq t\leq T}\|X(t)\|_{\mathbf{E}}$ and are bounded by constants depending on T and the $\mathbf{E}$-norm of their initial data. In particular, by Theorem 2.5.6 we know $\|Y_{\varDelta\xi}\|$ is bounded by a constant depending only on T, H_∞, $\|\zeta(0)\|_{\boldsymbol{\ell}^\infty}$, and ρ_∞. Grönwall's inequality then yields the estimate

$$\begin{aligned}\|X_{\varDelta}(T) - X(T)\|_{\mathbf{E}} \leq C_3\left(\|Y_{\varDelta\xi}\|, \|X\|\right) \\ \times \left[\|X_{\varDelta}(0) - X(0)\|_{\mathbf{E}} + \varDelta\xi C_1(\|Y_{\varDelta\xi}\|)T\right].\end{aligned}$$

Combining this estimate with (2.6.5), we obtain the desired result. □

Since convergence in Lagrangian coordinates implies convergence in the corresponding Eulerian coordinates, see [27] for details, this shows that interpolated solutions of the discrete two-component Camassa–Holm system can be used to obtain conservative solutions of the 2CH system (2.1.2). In particular, as conservative solutions of (2.1.1) are unique according to [4], our discretization of the CH equation corresponds to the unique conservative solution of (2.1.1).

Acknowledgments:

S. T. G. is thoroughly grateful to Katrin Grunert and Helge Holden for their careful reading of the manuscript, and the many helpful discussions in the course of this work.

Appendix 2.A Proofs of Propositions 2.3.1 and 2.4.2

Proof of (2.3.3). This is a consequence of the following inequalities,

$$|a_j|^2 = \frac{1}{\varDelta\xi}\varDelta\xi|a_j|^2 \leq \frac{1}{\varDelta\xi}\varDelta\xi\sum_{j\in\mathbb{Z}}|a_j|^2,$$

and

$$\Delta\xi \sum_{j\in\mathbb{Z}} |a_j|^2 \leq \frac{1}{\Delta\xi} \Delta\xi \sum_{j\in\mathbb{Z}} |a_j| \Delta\xi \sum_{i\in\mathbb{Z}} |a_i| = \frac{1}{\Delta\xi} \left(\Delta\xi \sum_{j\in\mathbb{Z}} |a_j| \right)^2 ,$$

where we have used the definitions (2.2.6). □

Proof of (2.3.4). We rewrite $(a_j)^2$ as

$$\begin{aligned} (a_j)^2 &= \frac{1}{2} \sum_{i=-\infty}^{j-1} \left((a_{i+1})^2 - (a_i)^2 \right) - \frac{1}{2} \sum_{i=j}^{\infty} \left((a_{i+1})^2 - (a_i)^2 \right) \\ &= \frac{\Delta\xi}{2} \left[\sum_{i=-\infty}^{j-1} (a_{i+1} + a_i) \mathrm{D}_+ a_i - \sum_{i=j}^{\infty} (a_{i+1} + a_i) \mathrm{D}_+ a_i \right] \\ &\leq \frac{\Delta\xi}{4} \sum_{i\in\mathbb{Z}} \left(|a_{i+1}|^2 + |a_i|^2 + 2 \left| \mathrm{D}_+ a_i \right|^2 \right) \\ &= \frac{1}{2} \|a\|_{\mathbf{h}^1}^2, \end{aligned}$$

where we have applied (2.3.1). □

Proof of (2.3.5). Telescopic cancellations yield

$$\begin{aligned} \Delta\xi \sum_{j=m}^{n} (\mathrm{D}_+ a_j) b_j &= \sum_{j=m}^{n} (a_{j+1} - a_j) b_j \\ &= \sum_{j=m}^{n} a_{j+1} b_j - \sum_{j=m}^{n} a_j b_{j-1} - \sum_{j=m}^{n} a_j (b_j - b_{j-1}) \\ &= a_{n+1} b_n - a_m b_{m-1} - \Delta\xi \sum_{j=m}^{n} a_j (\mathrm{D}_- b_j). \end{aligned}$$

□

A proof of (2.3.6) follows that of the continuous case, see, e.g., [7, Ex. 4.4]. To be precise, it comes from applying induction to the standard Hölder inequality.

Proof of (2.3.7). Without loss of generality we may assume $k \leq j$ and compute

$$|a_j - a_k| = \left| \Delta\xi \sum_{m=k}^{j-1} \mathrm{D}_+ a_m \right|$$

$$\begin{aligned}&\leq \left(\Delta\xi\sum_{m=k}^{j-1}|\mathrm{D}_+a_m|^2\right)^{1/2}\left(\Delta\xi\sum_{m=k}^{j-1}1\right)^{1/2}\\ &\leq \|\mathrm{D}_+a\|_{\boldsymbol{\ell}^2}\,|\Delta\xi(j-k)|^{1/2}.\end{aligned}$$

The result follows from taking supremum over j and k. □

Proof of (2.3.8). We first note

$$\|\mathrm{D}_+a\|_{\boldsymbol{\ell}^2}^2=\Delta\xi\sum_{|j|<n}|\mathrm{D}_+a_j|^2+\Delta\xi\sum_{|j|\geq n}|\mathrm{D}_+a_j|^2=:l_n+u_n,\quad n\in\mathbb{N},$$

where $l_n \nearrow \|\mathrm{D}_+a\|_{\boldsymbol{\ell}^2}^2$ and $u_n \searrow 0$ as $n\to+\infty$ by Bolzano–Weierstraß. Furthermore,

$$\begin{aligned}\sqrt{\Delta\xi}|\mathrm{D}_+a_j| &= \left(\Delta\xi|\mathrm{D}_+a_j|^2\right)^{1/2}\\ &\leq \min\left\{\left(\Delta\xi\sum_{k=j}^{+\infty}|\mathrm{D}_+a_k|^2\right)^{1/2},\left(\Delta\xi\sum_{k=-\infty}^{j}|\mathrm{D}_+a_k|^2\right)^{1/2}\right\}\end{aligned}$$

so that

$$\begin{aligned}\lim_{j\to\pm\infty}\sqrt{\Delta\xi}\,|\mathrm{D}_+a_j| &\leq \lim_{j\to\pm\infty}\left(\Delta\xi\sum_{|k|\geq|j|}|\mathrm{D}_+a_k|^2\right)^{1/2}\\ &= \lim_{j\to\pm\infty}(u_{|j|})^{1/2}=0.\end{aligned}$$

□

Proof (2.4.24). Let us denote $h=g*f$. Note that $r<\infty \implies p,q<\infty$, which shows that some configurations are impossible and can be excluded. We deal with the three remaining cases:

(i) $r<\infty$: From the generalized Hölder inequality we obtain

$$\begin{aligned}|h_j| &\leq \Delta\xi\sum_{i\in\mathbb{Z}}\left(|f_i|^{\frac{p}{r}}|g_{i,j}|^{\frac{q}{r}}\right)|f_i|^{1-\frac{p}{r}}|g_{i,j}|^{1-\frac{q}{r}}\\ &\leq \left[\Delta\xi\sum_{i\in\mathbb{Z}}\left(|f_i|^{\frac{p}{r}}|g_{i,j}|^{\frac{q}{r}}\right)^r\right]^{\frac{1}{r}}\left[\Delta\xi\sum_{i\in\mathbb{Z}}\left(|f_i|^{1-\frac{p}{r}}\right)^{\frac{rp}{r-p}}\right]^{\frac{r-p}{rp}}\\ &\quad\times\left[\Delta\xi\sum_{i\in\mathbb{Z}}\left(|g_{i,j}|^{1-\frac{q}{r}}\right)^{\frac{rq}{r-q}}\right]^{\frac{r-q}{rq}}\end{aligned}$$

$$\leq \left[\varDelta\xi \sum_{i\in\mathbb{Z}} |f_i|^p |g_{i,j}|^q\right]^{\frac{1}{r}} \left[\varDelta\xi \sum_{i\in\mathbb{Z}} |f_i|^p\right]^{\frac{r-p}{rp}} \times \left[\sup_{j\in\mathbb{Z}} \left(\varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}}\right]^{\frac{r-q}{r}}$$

which implies

$$\begin{aligned}
&\varDelta\xi \sum_{j\in\mathbb{Z}} |h_j|^r \\
&\quad \leq \|f\|_{\boldsymbol{\ell}^p}^{r-p} \left[\sup_{j\in\mathbb{Z}} \left(\varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}}\right]^{r-q} \varDelta\xi \sum_{j\in\mathbb{Z}} \varDelta\xi \sum_{i\in\mathbb{Z}} |f_i|^p |g_{i,j}|^q \\
&\quad \leq \|f\|_{\boldsymbol{\ell}^p}^{r-p} \left[\sup_{j\in\mathbb{Z}} \left(\varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}}\right]^{r-q} \varDelta\xi \sum_{i\in\mathbb{Z}} |f_i|^p \varDelta\xi \sum_{j\in\mathbb{Z}} |g_{i,j}|^q \\
&\quad \leq \|f\|_{\boldsymbol{\ell}^p}^{r} \left[\sup_{j\in\mathbb{Z}} \left(\varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}}\right]^{r-q} \left[\sup_{i\in\mathbb{Z}} \left(\varDelta\xi \sum_{j\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}}\right]^{q},
\end{aligned}$$

where we have used Fubini's theorem in the second inequality. Taking r-th roots we obtain the result.

(ii) $r = \infty$, $q < \infty$: We find

$$|h_j| \leq \varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}||f_i| \leq \|f\|_{\boldsymbol{\ell}^p} \left(\varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}|^q\right)^{\frac{1}{q}},$$

and taking supremum over j this corresponds to (2.4.24) where $q/\infty = 0$.

(iii) $r = q = \infty$: We find

$$\begin{aligned}
|h_j| &\leq \varDelta\xi \sum_{i\in\mathbb{Z}} |g_{i,j}||f_i| \\
&\leq \varDelta\xi \sum_{i\in\mathbb{Z}} |f_i| \left(\sup_{j\in\mathbb{Z}} |g_{i,j}|\right) \\
&\leq \sup_{i\in\mathbb{Z}} \left(\sup_{j\in\mathbb{Z}} |g_{i,j}|\right) \varDelta\xi \sum_{i\in\mathbb{Z}} |f_i|,
\end{aligned}$$

and taking supremum over j this corresponds to (2.4.24) where $\infty/\infty = 1$. □

Bibliography

[1] V. I. Arnold. *Mathematical methods of classical mechanics*, volume 60 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, second edition, 1989. Translated from the Russian by K. Vogtmann and A. Weinstein.

[2] V. I. Arnold and B. A. Khesin. *Topological methods in hydrodynamics*, volume 125 of *Applied Mathematical Sciences.* Springer-Verlag, New York, 1998.

[3] J. Borcea, S. Friedland, and B. Shapiro. Parametric Poincaré-Perron theorem with applications. *J. Anal. Math.*, 113:197–225, 2011.

[4] A. Bressan, G. Chen, and Q. Zhang. Uniqueness of conservative solutions to the Camassa-Holm equation via characteristics. *Discrete Contin. Dyn. Syst.*, 35(1):25–42, 2015.

[5] A. Bressan and A. Constantin. Global conservative solutions of the Camassa-Holm equation. *Arch. Ration. Mech. Anal.*, 183(2):215–239, 2007.

[6] A. Bressan and A. Constantin. Global dissipative solutions of the Camassa-Holm equation. *Anal. Appl. (Singap.)*, 5(1):1–27, 2007.

[7] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, New York, 2011.

[8] R. Camassa and D. D. Holm. An integrable shallow water equation with peaked solitons. *Phys. Rev. Lett.*, 71(11):1661–1664, 1993.

[9] M. Chen, Y. Zhang, et al. A two-component generalization of the Camassa-Holm equation and its solutions. *Letters in Mathematical Physics*, 75(1):1–15, 2006.

[10] G. M. Coclite, K. H. Karlsen, and N. H. Risebro. A convergent finite difference scheme for the Camassa-Holm equation with general H^1 initial data. *SIAM J. Numer. Anal.*, 46(3):1554–1579, 2008.

[11] A. Constantin and J. Escher. Global existence and blow-up for a shallow water equation. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 26(2):303–328, 1998.

[12] A. Constantin and J. Escher. Wave breaking for nonlinear nonlocal shallow water equations. *Acta Math.*, 181(2):229–243, 1998.

[13] A. Constantin and J. Escher. On the blow-up rate and the blow-up set of breaking waves for a shallow water equation. *Mathematische Zeitschrift*, 233(1):75–91, 2000.

[14] A. Constantin and R. I. Ivanov. On an integrable two-component Camassa-Holm shallow water system. *Phys. Lett. A*, 372(48):7129–7132, 2008.

[15] A. Constantin and B. Kolev. Least action principle for an integrable shallow water equation. *J. Nonlinear Math. Phys.*, 8(4):471–474, 2001.

[16] A. Constantin and B. Kolev. On the geometric approach to the motion of inertial mechanical systems. *J. Phys. A*, 35(32):R51–R79, 2002.

[17] A. Constantin and B. Kolev. Geodesic flow on the diffeomorphism group of the circle. *Comment. Math. Helv.*, 78(4):787–804, 2003.

[18] A. Constantin and D. Lannes. The hydrodynamical relevance of the Camassa–Holm and Degasperis–Procesi equations. *Archive for Rational Mechanics and Analysis*, 192(1):165–186, 2009.

[19] A. Constantin and L. Molinet. Global weak solutions for a shallow water equation. *Comm. Math. Phys.*, 211(1):45–61, 2000.

[20] J. Escher, D. Henry, B. Kolev, and T. Lyons. Two-component equations modelling water waves with constant vorticity. *Annali di Matematica Pura ed Applicata (1923-)*, 195(1):249–271, 2016.

[21] J. Escher, M. Kohlmann, and J. Lenells. The geometry of the two-component Camassa–Holm and Degasperis–Procesi equations. *Journal of Geometry and Physics*, 61(2):436–452, 2011.

[22] J. Escher, O. Lechtenfeld, and Z. Yin. Well-posedness and blow-up phenomena for the 2-component Camassa-Holm equation. *Discrete and continuous dynamical systems*, 19(3):493, 2007.

[23] G. Falqui. On a Camassa–Holm type equation with two dependent variables. *Journal of Physics A: Mathematical and General*, 39(2):327, 2005.

[24] S. Friedland. Convergence of products of matrices in projective spaces. *Linear Algebra Appl.*, 413(2-3):247–263, 2006.

[25] B. Fuchssteiner and A. S. Fokas. Symplectic structures, their Bäcklund transformations and hereditary symmetries. *Phys. D*, 4(1):47–66, 1981.

[26] S. T. Galtung and K. Grunert. A numerical study of variational discretizations of the Camassa–Holm equation, 2020. arXiv:2006.15562.

[27] M. Grasmair, K. Grunert, and H. Holden. On the equivalence of Eulerian and Lagrangian variables for the two-component Camassa-Holm system. In *Current research in nonlinear analysis*, volume 135 of *Springer Optim. Appl.*, pages 157–201. Springer, Cham, 2018.

[28] K. Grunert. Blow-up for the two-component Camassa-Holm system. *Discrete Contin. Dyn. Syst.*, 35(5):2041–2051, 2015.

[29] K. Grunert, H. Holden, and X. Raynaud. Global conservative solutions to the Camassa-Holm equation for initial data with nonvanishing asymptotics. *Discrete Contin. Dyn. Syst.*, 32(12):4209–4227, 2012.

[30] K. Grunert, H. Holden, and X. Raynaud. Global solutions for the two-component Camassa–Holm system. *Comm. Partial Differential Equations*, 37(12):2245–2271, 2012.

[31] K. Grunert, H. Holden, and X. Raynaud. Global dissipative solutions of the two-component Camassa-Holm system for initial data with nonvanishing asymptotics. *Nonlinear Anal. Real World Appl.*, 17:203–244, 2014.

[32] C. Guan, H. He, and Z. Yin. Well-posedness, blow-up phenomena and persistence properties for a two-component water wave system. *Nonlinear Analysis: Real World Applications*, 25:219–237, 2015.

[33] G. Gui and Y. Liu. On the global existence and wave-breaking criteria for the two-component Camassa–Holm system. *Journal of Functional Analysis*, 258(12):4251–4278, 2010.

[34] H. Holden and X. Raynaud. Convergence of a finite difference scheme for the Camassa-Holm equation. *SIAM J. Numer. Anal.*, 44(4):1655–1680, 2006.

[35] H. Holden and X. Raynaud. Global conservative multipeakon solutions of the Camassa-Holm equation. *J. Hyperbolic Differ. Equ.*, 4(1):39–64, 2007.

[36] H. Holden and X. Raynaud. Global conservative solutions of the Camassa-Holm equation—a Lagrangian point of view. *Comm. Partial Differential Equations*, 32(10-12):1511–1549, 2007.

[37] H. Holden and X. Raynaud. A numerical scheme based on multipeakons for conservative solutions of the Camassa-Holm equation. In *Hyperbolic problems: theory, numerics, applications*, pages 873–881. Springer, Berlin, 2008.

[38] H. Holden and X. Raynaud. Dissipative solutions for the Camassa-Holm equation. *Discrete Contin. Dyn. Syst.*, 24(4):1047–1112, 2009.

[39] P. J. Olver and P. Rosenau. Tri-Hamiltonian duality between solitons and solitary-wave solutions having compact support. *Phys. Rev. E (3)*, 53(2):1900–1906, 1996.

[40] M. Pituk. More on Poincaré's and Perron's theorems for difference equations. *J. Difference Equ. Appl.*, 8(3):201–216, 2002.

[41] G. Teschl. *Jacobi operators and completely integrable nonlinear lattices*, volume 72 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2000.

Paper 3

A numerical study of variational discretizations of the Camassa–Holm equation

Sondre Tesdal Galtung and Katrin Grunert

Submitted for publication

Paper 4

Competition models for plant stems

Alberto Bressan, Sondre Tesdal Galtung,
Audun Reigstad, and Johanna Ridder

Published in *Journal of Differential Equations*

Available online at www.sciencedirect.com

ScienceDirect

J. Differential Equations 269 (2020) 1571–1611

Journal of Differential Equations

www.elsevier.com/locate/jde

Competition models for plant stems

Alberto Bressan [a,*], Sondre T. Galtung [b], Audun Reigstad [b], Johanna Ridder [a]

[a] *Department of Mathematics, Penn State University, University Park, PA 16802, USA*
[b] *Department of Mathematical Sciences, NTNU – Norwegian University of Science and Technology, NO-7491 Trondheim, Norway*

Received 5 September 2019; revised 20 December 2019; accepted 11 January 2020
Available online 20 January 2020

Abstract

The models introduced in this paper describe a uniform distribution of plant stems competing for sunlight. The shape of each stem, and the density of leaves, are designed in order to maximize the captured sunlight, subject to a cost for transporting water and nutrients from the root to all the leaves. Given the intensity of light, depending on the height above ground, we first solve the optimization problem determining the best possible shape for a single stem. We then study a competitive equilibrium among a large number of similar plants, where the shape of each stem is optimal given the shade produced by all others. Uniqueness of equilibria is proved by analyzing the two-point boundary value problem for a system of ODEs derived from the necessary conditions for optimality.

MSC: 34B15; 49N90; 91A40; 92B05

Keywords: Optimal shape; Competitive equilibrium; Nonlinear boundary value problem

1. Introduction

Optimization problems for tree branches have recently been studied in [3,5]. In these models, optimal shapes maximize the total amount of sunlight gathered by the leaves, subject to a cost for

[*] Corresponding author.

E-mail addresses: axb62@psu.edu (A. Bressan), sondre.galtung@ntnu.no (S.T. Galtung), audun.reigstad@ntnu.no (A. Reigstad), johanna@jridder.de (J. Ridder).

https://doi.org/10.1016/j.jde.2020.01.013

building a network of branches that will bring water and nutrients from the root to all the leaves. Following [2,8,11,13,14], this cost is defined in terms of a ramified transport.

In the present paper we consider a competition model, where a large number of similar plants compete for sunlight. To make the problem tractable, instead of a tree-like structure we assume that each plant consists of a single stem. As a first step, assuming that the intensity of light $I(\cdot)$ depends only on the height above ground, we determine the corresponding optimal shape of the stem. This will be a curve $\gamma(\cdot)$ which can be found by classical techniques of the Calculus of Variations or optimal control [4,6,7]. In turn, given the density of plants (i.e., the average number of plants growing per unit area), if all stems have the same shape $\gamma(\cdot)$ one can compute the intensity of light $I(h)$ that reaches a point at height h.

An equilibrium configuration is now defined as a fixed point of the composition of the two maps $I(\cdot) \mapsto \gamma(\cdot)$ and $\gamma(\cdot) \mapsto I(\cdot)$. A major goal of this paper is to study the existence and properties of these equilibria, where the shape of each stem is optimal subject to the presence of all other competing plants.

In Section 2 we introduce our two basic models. In the first model, the length ℓ of the stems and the thickness (i.e., the density of leaves along each stem) are assigned a priori. The only function to optimize is thus the curve $\gamma : [0, \ell] \mapsto \mathbb{R}^2$ describing the shape of the stems. In the second model, also the length and the thickness of the stems are allowed to vary, and optimal values for these variables need to be determined.

In Section 3, given a light intensity function $I(\cdot)$, we study the optimization problem for Model 1, proving the existence of an optimal solution and deriving necessary conditions for optimality. We also give a condition which guarantees the uniqueness of the optimal solution. A counterexample shows that, in general, if this condition is not satisfied multiple solutions can exist. In Section 4 we consider the competition of a large number of stems, and prove the existence of an equilibrium solution. In this case, the common shape of the plant stems can be explicitly determined by solving a particular ODE.

The subsequent sections extend the analysis to a more general setting (Model 2), where both the length and the thickness of the stems are to be optimized. In Section 5 we prove the existence of optimal stem configurations, and derive necessary conditions for optimality, while in Section 6 we establish the existence of a unique equilibrium solution for the competitive game, assuming that the density (i.e., the average number of stems growing per unit area) is sufficiently small. The key step in the proof is the analysis of a two-point boundary value problem, for a system of ODEs derived from the necessary conditions.

In the above models, the density of stems was assumed to be uniform on the whole space. As a consequence, the light intensity $I(h)$ depends only of the height h above ground. Section 7, on the other hand, is concerned with a family of stems growing only on the positive half line. In this case the light intensity $I = I(h, x)$ depends also on the spatial location x, and the analysis becomes considerably more difficult. Here we only derive a set of equations describing the competitive equilibrium, and sketch what we conjecture should be the corresponding shape of stems.

The final section contains some concluding remarks. In particular, we discuss the issue of phototropism, i.e. the tendency of plant stems to bend in the direction of the light source. Devising a mathematical model, which demonstrates phototropism as an advantageous trait, remains a challenging open problem. For a biological perspective on plant growth we refer to [9]. A recent mathematical study of the stabilization problem for growing stems can be found in [1].

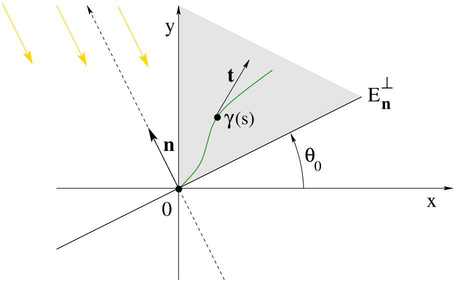


Fig. 1. By a reflection argument, it is not restrictive to assume that the tangent vector $\mathbf{t}(s)$ to the stem satisfies (2.4), i.e., it lies in the shaded cone.

2. Optimization problems for a single stem

We shall consider plant stems in the x-y plane, where y is the vertical coordinate. We assume that sunlight comes from the direction of the unit vector

$$\mathbf{n} = (n_1, n_2), \qquad n_2 < 0 < n_1.$$

As in Fig. 1, we denote by $\theta_0 \in]0, \pi/2[$ the angle such that

$$(-n_2, n_1) = (\cos\theta_0, \sin\theta_0). \tag{2.1}$$

Moreover, we assume that the light intensity $I(y) \in [0, 1]$ is a non-decreasing function of the height y. This is due to the presence of competing vegetation: close to the ground, less light can get through.

Model 1 (a stem with fixed length and constant thickness). We begin by studying a simple model, where each stem has a fixed length ℓ. Let $s \mapsto \gamma(s) = (x(s), y(s))$, $s \in [0, \ell]$, be an arc-length parameterization of the stem. As a first approximation, we assume that the leaves are uniformly distributed along the stem, with density κ. The total distribution of leaves in space is thus by a measure μ, with

$$\mu(A) = \kappa \cdot \text{meas}\Big(\big\{s \in [0, \ell]\,; \quad \gamma(s) \in A\big\}\Big) \tag{2.2}$$

for every Borel set $A \subseteq \mathbb{R}^2$.

Among all stems with given length ℓ, we seek the shape which will collect the most sunlight. This can be formulated as an optimal control problem. Since γ is parameterized by arc-length, the map $s \mapsto \gamma(s)$ is Lipschitz continuous with constant 1. Hence the tangent vector

$$\mathbf{t}(s) = \dot\gamma(s) = (\cos\theta(s), \sin\theta(s))$$

is well defined for a.e. $s \in [0, \ell]$. The map $s \mapsto \theta(s)$ will be regarded as a control function.

According to the model in [5], calling $\Phi(\cdot)$ the density of the projection of μ on the space $E_{\mathbf{n}}^{\perp}$ orthogonal to $\mathbf{n}$, the total sunlight captured by the stem is

$$\begin{aligned}\mathcal{S}(\gamma) &= \int \Big(1-\exp\{-\Phi(z)\}\Big)\,dz \\ &= \int_0^\ell I(y(s))\cdot\left(1-\exp\left\{\frac{-\kappa}{\cos(\theta(s)-\theta_0)}\right\}\right)\cos(\theta(s)-\theta_0)\,ds. \end{aligned} \tag{2.3}$$

In order to maximize (2.3), we claim that it is not restrictive to assume that the angle satisfies

$$\theta_0 \;\le\; \theta(s) \;\le\; \frac{\pi}{2} \qquad\qquad \text{for all } s\in[0,\ell]. \tag{2.4}$$

Indeed, for any measurable map $s\mapsto\theta(s)\in\,]-\pi,\pi]$, we can define a modified angle function $\theta^\sharp(\cdot)$ by setting

$$\theta^\sharp(s) \;=\; \begin{cases} \theta(s) & \text{if} \quad \theta(s)\in\,]0,\theta_0+\pi/2],\\ -\theta(s) & \text{if} \quad \theta(s)\in\,]-\pi,\,\theta_0-\pi/2],\\ 2\theta_0+\pi-\theta(s) & \text{if} \quad \theta(s)\in\,]\theta_0+\pi/2,\pi],\\ 2\theta_0-\theta(s) & \text{if} \quad \theta(s)\in\,]\theta_0-\pi/2,0]. \end{cases} \tag{2.5}$$

Calling $\gamma^\sharp:[0,\ell]\mapsto\mathbb{R}^2$ the curve whose tangent vector is $\dot\gamma^\sharp(s)=(\cos\theta^\sharp(s),\sin\theta^\sharp(s))$, since the light intensity function $y\mapsto I(y)$ is nondecreasing, we have $\mathcal{S}(\gamma^\sharp)\ge\mathcal{S}(\gamma)$.

By this first step, without loss of generality we can now assume $\theta(s)\in\,]0,\,\theta_0+\pi/2]$. To proceed further, consider the piecewise affine map

$$\varphi(\theta) \;=\; \begin{cases} \theta & \text{if} \quad \theta\in\,]\theta_0,\pi/2],\\ \pi-\theta & \text{if} \quad \theta\in[\pi/2,\,\theta_0+\pi/2],\\ 2\theta_0-\theta & \text{if} \quad \theta\in[0,\theta_0]. \end{cases} \tag{2.6}$$

Call γ^φ the curve whose tangent vector is $\dot\gamma^\varphi(s)=\Big(\cos(\varphi(\theta(s))),\,\sin(\varphi(\theta(s)))\Big)$. Since $I(\cdot)$ is nondecreasing, we again have $\mathcal{S}(\gamma^\varphi)\;\ge\;\mathcal{S}(\gamma)$. We now observe that, since $0<\theta_0<\pi/2$, there exists an integer $m\ge1$ such that the m-fold composition $\varphi^m\doteq\varphi\circ\cdots\circ\varphi$ maps $[0,\theta_0+\pi/2]$ into $[\theta_0,\pi/2]$. An inductive argument now yields $\mathcal{S}(\gamma^{\varphi^m})\;\ge\;\mathcal{S}(\gamma)$, completing the proof of our claim.

As shown in Fig. 2, left, we call z the coordinate along the space $E_{\mathbf{n}}^\perp$ perpendicular to $\mathbf{n}$, and let y be the vertical coordinate. Hence

$$dz(s) \;=\; \cos(\theta(s)-\theta_0)\,ds, \qquad\qquad dy(s) \;=\; \sin(\theta(s))\,ds. \tag{2.7}$$

In view of (2.4), one can express both γ and θ as functions of the variable y. Introducing the function

$$g(\theta) \;\doteq\; \left(1-\exp\left\{\frac{-\kappa}{\cos(\theta-\theta_0)}\right\}\right)\frac{\cos(\theta-\theta_0)}{\sin\theta}\,, \tag{2.8}$$

the problem can be equivalently formulated as follows.

(OP1) *Given a length $\ell > 0$, find $h > 0$ and a control function $y \mapsto \theta(y) \in [\theta_0, \pi/2]$ which maximizes the integral*

$$\int_0^h I(y)\, g(\theta(y))\, dy \tag{2.9}$$

subject to

$$\int_0^h \frac{1}{\sin\theta(y)}\, dy \;=\; \ell. \tag{2.10}$$

Model 2 (stems with variable length and thickness). Here we still assume that the plant consists of a single stem, parameterized by arc-length: $s \mapsto \gamma(s)$, $s \in [0, \ell]$. However, now we give no constraint on the length ℓ of the stem, and we allow the density of leaves to be variable along the stem.

Call $u(s)$ the density of leaves at the point $\gamma(s)$. In other words, μ is now the measure which is absolutely continuous w.r.t. arc-length measure on γ, with density u. Instead of (2.2) we thus have

$$\mu(A) \;=\; \int_{\{s\,;\ \gamma(s)\in A\}} u(s)\, ds\,. \tag{2.11}$$

Calling $I(y)$ the intensity of light at height y, the total sunlight gathered by the stem is now computed by

$$\mathcal{S}(\mu) \;=\; \int_0^\ell I(y(s)) \cdot \left(1 - \exp\left\{\frac{-u(s)}{\cos(\theta(s)-\theta_0)}\right\}\right) \cos(\theta(s)-\theta_0)\, ds. \tag{2.12}$$

As in [5], we consider a cost for transporting water and nutrients from the root to the leaves. This is measured by

$$\mathcal{I}^\alpha(\mu) \;=\; \int_0^\ell \left(\int_s^\ell u(t)\, dt\right)^\alpha ds, \tag{2.13}$$

for some $0 < \alpha < 1$. Notice that, in Model 1, this cost was the same for all stems and hence it did not play a role in the optimization.

For a given constant $c > 0$, we now consider a second optimization problem:

$$\text{maximize:} \quad \mathcal{S}(\mu) - c\mathcal{I}^\alpha(\mu), \tag{2.14}$$

subject to:

$$y(0) \;=\; 0, \qquad\qquad \dot{y}(s) \;=\; \sin\theta(s). \tag{2.15}$$

The maximum is sought over all controls $\theta : \mathbb{R}_+ \mapsto [0, \pi]$ and $u : \mathbb{R}_+ \mapsto \mathbb{R}_+$. Calling

$$z(t) \doteq \int_t^{+\infty} u(s)\,ds, \tag{2.16}$$

$$G(\theta, u) \doteq \left(1 - \exp\left\{\frac{-u}{\cos(\theta - \theta_0)}\right\}\right) \cos(\theta - \theta_0)\,, \tag{2.17}$$

this leads to an optimal control problem in a more standard form.

(OP2) *Given a sunlight intensity function $I(y)$, and constants $0 < \alpha < 1$, $c > 0$, find controls $\theta : \mathbb{R}_+ \mapsto [\theta_0, \pi/2]$ and $u : \mathbb{R}_+ \mapsto \mathbb{R}_+$ which maximize the integral*

$$\int_0^{+\infty} \Big[I(y)\, G(\theta, u) - c\, z^\alpha\Big]\, dt, \tag{2.18}$$

subject to

$$\begin{cases} \dot y(t) = \sin\theta, \\ \dot z(t) = -u, \end{cases} \qquad \begin{cases} y(0) = 0, \\ z(+\infty) = 0. \end{cases} \tag{2.19}$$

3. Optimal stems with fixed length and thickness

3.1. Existence of an optimal solution

Let $I(y)$ be the light intensity, which we assume is a non-decreasing function of the vertical component y. For a given $\kappa > 0$ (the thickness of the stem), we seek a curve $s \mapsto \gamma(s)$, starting at the origin and with a fixed length ℓ, which maximizes the sunlight functional defined at (2.9).

Theorem 3.1. *For any non-decreasing function $y \mapsto I(y) \in [0, 1]$ and any constants $\ell, \kappa > 0$ and $\theta_0 \in\,]0, \pi/2[$, the optimization problem* **(OP1)** *has at least one solution.*

Proof. 1. Let M be the supremum among all admissible payoffs in (2.9). By the analysis in [5] it follows that

$$0 \le M \le \kappa\, \mu(\mathbb{R}^2) = \kappa\, \ell.$$

Hence there exists a maximizing sequence of control functions $\theta_n : [0, h_n] \mapsto [\theta_0, \pi/2]$, so that

$$\int_0^{h_n} \frac{1}{\sin\theta_n(y)}\, dy = \ell \qquad \text{for all } n \ge 1, \tag{3.1}$$

$$\int_0^{h_n} I(y)g(\theta_n(y))\,dy \;\to\; M. \tag{3.2}$$

2. For each n, let $\theta_n^\sharp$ be the non-increasing rearrangement of the function θ_n. Namely, $\theta_n^\sharp$ is the unique (up to a set of zero measure) non-increasing function such that, for every $c \in \mathbb{R}$

$$\text{meas}\Big(\{s\,;\;\; \theta_n^\sharp(s) < c\}\Big) \;=\; \text{meas}\Big(\{s\,;\;\; \theta_n(s) < c\}\Big). \tag{3.3}$$

This can be explicitly defined as

$$\theta_n^\sharp(y) \;=\; \sup\Big\{\xi\,;\;\; \text{meas}\Big(\{\sigma \in [0,h_n]\,;\;\; \theta_n(\sigma) \ge \xi\}\Big) > y\Big\}.$$

For every $n \ge 1$ we claim that

$$\int_0^{h_n} \frac{1}{\sin\theta_n^\sharp(y)}\,dy \;=\; \int_0^{h_n} \frac{1}{\sin\theta_n(y)}\,dy \;=\; \ell, \tag{3.4}$$

$$\int_0^{h_n} I(y)g(\theta_n^\sharp(y))\,dy \;\ge\; \int_0^{h_n} I(y)g(\theta_n(y))\,dy. \tag{3.5}$$

Indeed, to prove the first identity we observe that, by (3.3), there exists a measure-preserving map $y \mapsto \zeta(y)$ from $[0,h_n]$ into itself such that $\theta_n^\sharp(y) = \theta_n(\zeta(y))$. Using ζ as new variable of integration, one immediately obtains (3.4).

To prove (3.5) we observe that the function g introduced at (2.8) is smooth and satisfies

$$g'(\theta) \;\le\; 0 \qquad \text{for all } \theta \in [\theta_0,\, \pi/2]. \tag{3.6}$$

Therefore, the map $y \mapsto g(\theta_n^\sharp(y))$ coincides with the non-decreasing rearrangement of $y \mapsto g(\theta_n(y))$. On the other hand, since $I(\cdot)$ is non-decreasing, it trivially coincides with the non-decreasing rearrangement of itself. Therefore, (3.5) is an immediate consequence of the Hardy-Littlewood inequality [10].

3. Since all functions $\theta_n^\sharp$ are non-increasing, they have bounded variation. Using Helly's compactness theorem, by possibly extracting a subsequence, we can find $h > 0$ and a non-increasing function $\theta^* : [0,h] \mapsto [\theta_0, \pi/2]$ such that

$$\lim_{n\to\infty} h_n \;=\; h\,, \qquad\qquad \lim_{n\to\infty} \theta_n^\sharp(y) \;=\; \theta^*(y) \qquad \text{for a.e. } y \in [0,h]. \tag{3.7}$$

This implies

$$\int_0^{h} \frac{1}{\sin\theta^*(y)}\,dy \;=\; \ell, \qquad\qquad \int_0^{h} I(y)g(\theta^*(y))\,dy \;=\; M,$$

proving the optimality of θ^*. □

3.2. Necessary conditions for optimality

Let $y \mapsto \theta^*(y)$ be an optimal solution. By the previous analysis we already know that the function $\theta^*(\cdot)$ is non-increasing. Otherwise, its non-increasing rearrangement achieves a better payoff. In particular, this implies that the left limit at the terminal point $y = h$ is well defined:

$$\theta^*(h) \;=\; \lim_{y\to h-} \theta^*(y). \tag{3.8}$$

Consider an arbitrary perturbation

$$\theta_\epsilon \;=\; \theta^* + \epsilon\Theta, \qquad\qquad h_\epsilon \;=\; h + \epsilon\eta.$$

The constraint (2.10) implies

$$\int_0^{h+\epsilon\eta} \frac{1}{\sin\theta_\epsilon(y)}\,dy \;=\; \ell. \tag{3.9}$$

Differentiating (3.9) w.r.t. ϵ one obtains

$$\frac{1}{\sin\theta^*(h)}\,\eta - \int_0^h \frac{\cos\theta^*(y)}{\sin^2\theta^*(y)}\,\Theta(y)\,dy \;=\; 0. \tag{3.10}$$

Next, calling

$$J_\epsilon \;\doteq\; \int_0^{h_\epsilon} I(y)g(\theta_\epsilon(y))dy$$

and assuming that $I(\cdot)$ is continuous at least at $y = h$, by (3.10) we obtain

$$\begin{aligned}0 \;=\; \frac{d}{d\epsilon}J_\epsilon\bigg|_{\epsilon=0} \;=\; &\int_0^h I(y)g'(\theta^*(y))\Theta(y)\,dy\\ &+ I(h)g(\theta^*(h))\cdot\sin\theta^*(h)\int_0^h \frac{\cos\theta^*(y)}{\sin^2\theta^*(y)}\,\Theta(y)\,dy.\end{aligned} \tag{3.11}$$

Since (3.11) holds for arbitrary perturbations $\Theta(\cdot)$, the optimal control $\theta^*(\cdot)$ should satisfy the identity

$$I(y)g'\big(\theta^*(y)\big) + \lambda\cdot\frac{\cos\theta^*(y)}{\sin^2\theta^*(y)} \;=\; 0, \qquad \text{for a.e. } y\in[0,h], \tag{3.12}$$

where

$$\lambda \;=\; I(h)g(\theta^*(h))\cdot\sin\theta^*(h). \tag{3.13}$$

It will be convenient to write

$$g(\theta) \;=\; \frac{G(\theta)}{\sin\theta}, \qquad\qquad G(\theta) \;\doteq\; \left(1-\exp\left\{\frac{-\kappa}{\cos(\theta-\theta_0)}\right\}\right)\cos(\theta-\theta_0). \tag{3.14}$$

Inserting (3.14) in (3.12) one obtains the pointwise identities

$$I(y)\Big(G'(\theta^*(y))\sin\theta^*(y)-G(\theta^*(y))\cos\theta^*(y)\Big)+\lambda\cdot\cos\theta^*(y) \;=\; 0. \tag{3.15}$$

At $y=h$, the identities (3.13) and (3.15) yield

$$G'(\theta^*(h))\,\tan\theta^*(h)-G(\theta^*(h)) \;=\; -\frac{I(h)G(\theta^*(h))}{I(h)}.$$

Hence

$$G'(\theta^*(h))\,\tan\theta^*(h) \;=\; 0,$$

which implies

$$\theta^*(h) \;=\; \theta_0\,, \qquad\qquad \lambda \;=\; I(h)g(\theta_0)\sin\theta_0 \;=\; \left(1-e^{-\kappa}\right)I(h)\,. \tag{3.16}$$

Notice that (3.15) corresponds to

$$\theta^*(y) \;=\; \arg\max_{\theta\in[0,\pi]}\left\{I(y)\frac{G(\theta)}{\sin\theta}-\frac{\lambda}{\sin\theta}\right\}. \tag{3.17}$$

Equivalently, $\theta=\theta^*(y)$ is the solution to

$$G'(\theta)\tan\theta-G(\theta) \;=\; -\frac{\lambda}{I(y)}, \tag{3.18}$$

where G is the function at (3.14).

Lemma 3.2. *Let G be the function at (3.14). Then for every $z\in\,]-\infty,\ e^{-\kappa}-1]$ the equation*

$$F(\theta) \;\doteq\; G'(\theta)\tan\theta-G(\theta) \;=\; z \tag{3.19}$$

has a unique solution $\theta=\varphi(z)\in[\theta_0,\pi/2[$.

Proof. Observing that

$$\begin{cases} G(\theta_0) \;=\; 1-e^{-\kappa}, \\ G'(\theta_0) \;=\; 0, \end{cases} \qquad\qquad \begin{cases} G'(\theta) \;<\; 0 \\ G''(\theta) \;<\; 0 \end{cases} \qquad \text{for }\ \theta\in\,]\theta_0,\pi/2[\,, \tag{3.20}$$

we obtain $F(\theta_0) = e^{-\kappa} - 1$ and

$$F'(\theta) = G''(\theta)\tan\theta + G'(\theta)\tan^2\theta < 0 \qquad \text{for } \theta \in [\theta_0, \pi/2[\,.$$

Therefore, for $\theta \in [\theta_0, \pi/2[$, the left hand side of (3.19) is monotonically decreasing from $e^{-\kappa} - 1$ to $-\infty$. We conclude that (3.19) has a unique solution $\theta = \varphi(z)$ for any $z \in]-\infty, e^{-\kappa} - 1]$. □

The optimal control $\theta^*(\cdot)$ determined by the necessary condition (3.18) is thus recovered by

$$\theta^*(y) = \varphi\left(\frac{-\lambda}{I(y)}\right) = \varphi\left(\frac{(e^{-\kappa} - 1)I(h)}{I(y)}\right). \tag{3.21}$$

Next, we need to determine h so that the constraint

$$L(h) \doteq \int_0^h \frac{1}{\sin(\theta^*(y))}\,dy = \ell \tag{3.22}$$

is satisfied. As shown by Example 3.4 below, the solution of (3.21)-(3.22) may not be unique.

In the following, we seek a condition on I which implies that L is monotone, i.e.,

$$L'(h) = \frac{1}{\sin(\theta_0)} + \int_0^h \frac{\cos\theta^*(y)}{\sin^2\theta^*(y)}\,\frac{1}{F'(\theta^*(y))}\,\frac{I'(h)}{I(y)}G(\theta_0)\,dy > 0\,. \tag{3.23}$$

This will guarantee that (3.22) has a unique solution. To get an upper bound for $F'(\theta)$, observe that, for $\theta \in [\theta_0, \pi/2[$,

$$\begin{aligned} F'(\theta) &\le \tan(\theta)G''(\theta) \\ &= -\tan(\theta)\Bigg[\cos(\theta - \theta_0)\left(1 - \left(\frac{\kappa}{\cos(\theta - \theta_0)} + 1\right)\exp\left\{\frac{-\kappa}{\cos(\theta - \theta_0)}\right\}\right) \\ &\quad + \tfrac{\tan^2(\theta-\theta_0)}{\cos(\theta-\theta_0)}\kappa^2 \exp\left\{\tfrac{-\kappa}{\cos(\theta-\theta_0)}\right\}\Bigg] \\ &= -\tan(\theta)\,\cos(\pi/2 - \theta_0)\big(1 - (\kappa + 1)e^{-\kappa}\big). \end{aligned}$$

Since $\theta^*(y) \in [\theta_0, \pi/2]$ and $G(\theta_0) = 1 - e^{-\kappa}$, using the above inequality one obtains

$$\begin{aligned} &\int_0^h \frac{\cos\theta^*(y)}{\sin^2\theta^*(y)} \cdot \frac{1}{|F'(\theta^*(y))|}\,\frac{I'(h)}{I(y)}G(\theta_0)\,dy \\ &\qquad \le \frac{\cos^2\theta_0}{\sin^3\theta_0} \cdot \frac{1 - e^{-\kappa}}{\cos(\pi/2 - \theta_0)\big(1 - (\kappa + 1)e^{-\kappa}\big)} \int_0^h \frac{I'(h)}{I(y)}\,dy\,. \end{aligned}$$

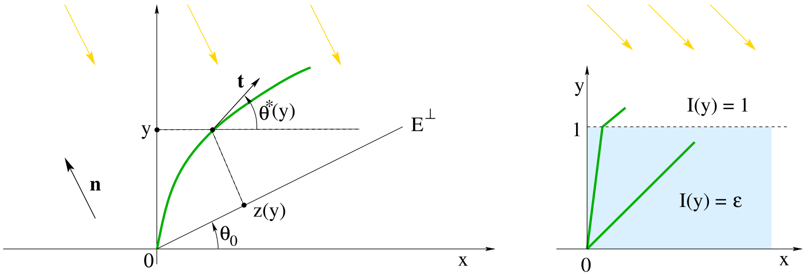


Fig. 2. Left: the optimal shape of a stem, as described in Theorem 3.3. Right: if the light intensity I changes abruptly as a function of the hight, the optimal shape may not be unique, as shown in Example 3.4.

Hence (3.23) is satisfied provided that

$$\int_0^h \frac{I'(h)}{I(y)}\,dy \;<\; \tan^2\theta_0 \cdot \frac{\cos(\pi/2-\theta_0)\big(1-(\kappa+1)e^{-\kappa}\big)}{1-e^{-\kappa}}\,. \tag{3.24}$$

From the above analysis, we conclude

Theorem 3.3. *Assume that the light intensity function I is Lipschitz continuous and satisfies the strict inequality (3.24) for a.e. $h \in [0,\ell]$. Then the optimization problem* **(OP1)** *has a unique optimal solution $\theta^* : [0,h^*] \mapsto [\theta_0, \pi/2]$. The function θ^* is non-increasing, and satisfies*

$$\theta^*(y) \;=\; \varphi\left((e^{-\kappa}-1)\frac{I(h^*)}{I(y)}\right), \tag{3.25}$$

where $z \mapsto \varphi(z) = \theta$ is the function implicitly defined by (3.19).

The following example shows that, without the bound (3.24) on the sunlight intensity function $I(\cdot)$, the conclusion of Theorem 3.3 can fail.

Example 3.4 *(non-uniqueness).* Choose $\mathbf{n} = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$, $\ell = 6/5 < \sqrt{2}$, $\kappa = 1$,

$$I(y) \;=\; \begin{cases} \varepsilon & \text{if} \quad y \in [0,1], \\ 1 & \text{if} \quad y > 1, \end{cases}$$

with $\varepsilon > 0$l.

By Theorem 3.1 at least one optimal solution exists. By the previous analysis, any optimal solution $\theta^* : [0,h^*] \mapsto [\theta_0, \pi/2]$ satisfies the necessary conditions (3.25). In this particular case, this implies that $\theta^*(y)$ is constant separately for $y < 1$ and for $y > 1$. As shown in Fig. 2, right, these necessary conditions can have two solutions.

Solution 1. If $h^* < 1$, then $I(y) = \varepsilon$ for all $y \in [0,h^*]$ and the necessary conditions (3.25) yield

$$\theta_1^*(y) \;=\; \varphi(e^{-1}-1) \;=\; \theta_0 \;=\; \pi/4 \qquad \text{for all } y\in[0,h^*].$$

The total sunlight collected is

$$\mathcal{S}_\varepsilon(\theta_1^*) \;=\; \frac{6}{5}(1-e^{-1})\,. \tag{3.26}$$

Solution 2. If $h^*>1$, then $I(h^*)=1$ and the necessary conditions (3.25) yield

$$\theta_2^*(y) \;=\; \varphi\left((e^{-1}-1)\frac{I(h^*)}{I(y)}\right) \;=\; \begin{cases} \varphi\left((e^{-1}-1)\varepsilon^{-1}\right) & \text{if} \quad y\in[0,1],\\ \pi/4 & \text{if} \quad y>1.\end{cases}$$

Calling $\alpha=\alpha(\varepsilon)\doteq\;\varphi\left((e^{-1}-1)\varepsilon^{-1}\right)$, the total sunlight collected in this case is

$$\mathcal{S}_\varepsilon(\theta_2^*) \;=\; \left(1-\exp\left\{-\frac{1}{\cos(\alpha-\pi/4)}\right\}\right)\cos(\alpha-\pi/4)\,\varepsilon+\left(\frac{6}{5}-\frac{1}{\sin\alpha}\right)\left(1-e^{-1}\right). \tag{3.27}$$

We claim that, for a suitable choice of $\varepsilon\in\,]0,1[$, the two quantities in (3.26) and (3.27) become equal. Indeed, as $\varepsilon\to0+$ we have

$$\alpha(\varepsilon)\;\doteq\;\varphi\left(\frac{e^{-1}-1}{\varepsilon}\right)\;\to\;\frac{\pi}{2}\,,$$

$$\mathcal{S}_\varepsilon(\theta_1^*)\;\to\;0,\qquad\qquad \mathcal{S}_\varepsilon(\theta_2^*)\;\to\;\frac{1-e^{-1}}{5}\,. \tag{3.28}$$

On the other hand, as $\varepsilon\to1$ we have $\alpha(\varepsilon)\to\pi/4$. By continuity, there exists $\varepsilon_1\in\,]0,1[$ such that

$$\sin\alpha(\varepsilon_1)\;=\;\frac{5}{6}\,.$$

As $\varepsilon\to\varepsilon_1+$, we have

$$\mathcal{S}_\varepsilon(\theta_2^*)\;\to\;\left(1-\exp\left\{-\frac{1}{\cos(\alpha(\varepsilon_1)-\pi/4)}\right\}\right)\cos(\alpha(\varepsilon_1)-\pi/4)\,\varepsilon_1\;<\;\mathcal{S}_{\varepsilon_1}(\theta_1^*). \tag{3.29}$$

Comparing (3.28) with (3.29), by continuity we conclude that there exists some $\widehat{\varepsilon}\in\,]0,\varepsilon_1[$ such that $\mathcal{S}_{\widehat{\varepsilon}}(\theta_1^*)\;=\;\mathcal{S}_{\widehat{\varepsilon}}(\theta_2^*)$. Hence for $\varepsilon=\widehat{\varepsilon}$ the optimization problem has two distinct solutions.

We remark that in this example the light intensity $I(y)$ is discontinuous at $y=1$. However, by a mollification one can still construct a similar example with two optimal configurations, also for $I(\cdot)$ smooth. Of course, in this case the derivative $I'(h)$ will be extremely large for $h\approx1$, so that the assumption (3.24) fails.

4. A competition model

In the previous analysis, the light intensity function $I(\cdot)$ was a priori given. We now consider a continuous distribution of stems, and determine the average sunlight $I(y)$ available at height y above ground, depending on the density of vegetation above y.

Let the constants $\ell, \kappa > 0$ be given, specifying the length and thickness of each stem. We now introduce another constant $\rho > 0$ describing the density of stems, i.e. how many stems grow per unit area. Assume that all stems have the same height and shape, described by the function $\theta : [0, h] \mapsto [\theta_0, \pi/2]$. For any $y \in [0, h]$, the total amount of vegetation at height $\geq y$, per unit length, is then measured by

$$\rho \cdot \int_y^h \frac{\kappa}{\sin\theta(y)}\, dy.$$

The corresponding light intensity function is defined as

$$I(y) \doteq \exp\left\{-\rho \cdot \int_y^h \frac{\kappa}{\sin\theta(y)}\, dy\right\} \qquad \text{for} \quad y \in [0, h], \tag{4.1}$$

while $I(y) = 1$ for $y \geq h$. We are interested in equilibrium configurations, where the shape of the stems is optimal for the light intensity $I(\cdot)$. We recall that θ_0 is the angle of incoming light rays, as in (2.1), while the constants $\ell, \kappa > 0$ denote the length and thickness of the stems.

Definition 4.1. Given an angle $\theta_0 \in]0, \pi/2]$ and constants $\ell, \kappa, \rho > 0$, we say that a light intensity function $I^* : \mathbb{R}_+ \mapsto [0, 1]$ and a stem shape function $\theta^* : [0, h^*] \mapsto [\theta_0, \pi/2]$ yield a **competitive equilibrium** if the following holds.

(i) The stem shape function $\theta^* : [0, h^*] \mapsto [\theta_0, \pi/2]$ provides an optimal solution to the optimization problem **(OP1)**, with light intensity function $I = I^*$.
(ii) For all $y \geq 0$, the light intensity at height y satisfies

$$I^*(y) = \exp\left\{-\rho \cdot \int_{\min\{y,h^*\}}^{h^*} \frac{\kappa}{\sin\theta^*(y)}\, dy\right\}. \tag{4.2}$$

If the density of vegetation is sufficiently small, we now show that an equilibrium configuration exists.

Theorem 4.2. *Let the light angle $\theta_0 \in]0, \pi/2]$ be given, together with the constants $\ell, \kappa > 0$ determining the common length and thickness of all the stems. Then there exists a constant $c_0 > 0$ such that, for all $0 < \rho \leq c_0$, an equilibrium configuration exists.*

Proof. 1. Consider the set of stem configurations

$$\mathcal{K} \doteq \Big\{\Theta : [0, \ell] \mapsto [\theta_0, \pi/2], \quad \Theta \text{ is nonincreasing}\Big\}, \tag{4.3}$$

and the set of light intensity functions

$$\mathcal{J} \;\doteq\; \Big\{ I : [0,+\infty[\,\mapsto [0,1]\,;\;\; I \text{ is nondecreasing}, \;\; I(y)=1 \;\text{ for } y\geq \ell, \\ I \text{ is Lipschitz continuous with constant } \frac{\rho\kappa}{\sin\theta_0} \Big\}. \tag{4.4}$$

We observe that $\mathcal{K}$ is a compact, convex subset of $\mathbf{L}^1([0,\ell])$, while $\mathcal{J}$ is a compact, convex subset of $\mathcal{C}^0([0,+\infty[)$.

If $\Theta(\cdot)\in\mathcal{K}$ describes the common configuration of all stems, we denote by $I^\Theta(\cdot)$ the corresponding light intensity function. Moreover, for a given function $I(\cdot)$, we denote by $\Theta^*(I)$ the corresponding optimal configuration of plant stems.

In the following steps we shall prove that:

(i) The map $\Theta\mapsto I^\Theta$ is continuous from $\mathcal{K}$ into $\mathcal{J}$.
(ii) The map $I\mapsto\Theta^*(I)$ is continuous from $\mathcal{J}$ into $\mathcal{K}$.

As a consequence, the composed map $\Theta\mapsto\Theta^*(I^\Theta)$ is continuous from $\mathcal{K}$ into itself. By Schauder's theorem, it has a fixed point, which provides an equilibrium solution.

2. Given $\Theta\in\mathcal{K}$, define the constant

$$\bar h \;\doteq\; \int_0^\ell \sin\Theta(t)\,dt\,. \tag{4.5}$$

More generally, for $s\in[0,\ell]$, set

$$y(s) \;\doteq\; \int_0^s \sin\Theta(t)\,dt \;\in\; [0,\bar h]. \tag{4.6}$$

We observe that, since $\Theta(t)\in[\theta_0,\pi/2]$, the inverse function $y\mapsto s(y)$ from $[0,\bar h]$ into $[0,\ell]$ is a strictly increasing bijection, with Lipschitz constant $L=\frac{1}{\sin\theta_0}$. The corresponding light intensity function is determined by

$$I^\Theta(y) \;=\; \begin{cases} \exp\big\{-\rho\kappa(\ell-s(y))\big\} & \text{if } \; y\in[0,\bar h], \\ \qquad 1 & \text{if } \; y>\ell. \end{cases} \tag{4.7}$$

From the above definitions it follows that $\Theta\mapsto I^\Theta$ is continuous from $\mathcal{K}$ into $\mathcal{J}$.

3. Next, let $I\in\mathcal{J}$. Given the constants ℓ,κ, by choosing $\rho>0$ small enough, any Lipschitz continuous function $I:[0,\ell]\mapsto[0,1]$ with Lipschitz constant $L=\frac{\rho\kappa}{\sin\theta_0}$ will satisfy the inequality (3.24). Hence, by Theorem 3.3, the optimization problem **(OP1)** has a unique optimal solution $\theta^*:[0,h^*]\mapsto[\theta_0,\pi/2]$.

Notice that in Theorem 3.3 this solution is written in terms of the variable $y\in[0,h^*]$, and satisfies the optimality condition (3.25). In terms of the arc-length parameter $s\in[0,\ell]$, this corresponds to

$$\Theta^*(s) \;=\; \theta^*(h(s))$$

where the variable $y(s) \in [0, h^*]$ is implicitly defined by

$$\int_0^{y(s)} \frac{1}{\sin\theta^*(z)}\,dz \;=\; s.$$

In view of (2.3), given $I \in \mathcal{J}$ and $\Theta \in \mathcal{K}$, the total sunlight collected by the stem is computed by

$$\mathcal{S}(I,\Theta) \;=\; = \int_0^{\ell} I(y(s)) \cdot \left(1 - \exp\left\{\frac{-\kappa}{\cos(\Theta(s)-\theta_0)}\right\}\right)\cos(\Theta(s)-\theta_0)\,ds, \tag{4.8}$$

where

$$y(s) \;\doteq\; \int_0^{s} \sin\Theta(s)\,ds.$$

From the above formulas it follows that the map $(I,\Theta) \mapsto \mathcal{S}(I,\Theta)$ is continuous on the compact set $\mathcal{J} \times \mathcal{K}$. In particular, the function

$$I \;\mapsto\; \max_{\Theta\in\mathcal{K}} \mathcal{S}(I,\Theta) \tag{4.9}$$

is continuous on the compact set $\mathcal{J}$.

Given a light intensity function $I \in \mathcal{J}$, call $\Theta^*(I) \in \mathcal{K}$ the unique optimal stem shape. We claim that the map $I \mapsto \Theta^*(I)$ is continuous.

Indeed, this is a straightforward consequence of continuity and compactness. If continuity fails, there exists a convergent sequence $I_n \to I$ such that $\Theta(I_n)$ does not converge to $\Theta(I)$. By the compactness of $\mathcal{K}$, we can extract a subsequence such that

$$\Theta_{n_k} \;\to\; \Theta^\sharp \;\neq\; \Theta(I).$$

By continuity, one obtains

$$\begin{aligned}\mathcal{S}(I,\Theta(I)) = \sup_{\Theta\in\mathcal{K}} \mathcal{S}(I,\Theta) \;&=\; \lim_{k\to\infty}\sup_{\Theta\in\mathcal{K}} \mathcal{S}(I_{n_k},\Theta)\\ &= \lim_{k\to\infty} \mathcal{S}(I_{n_k},\Theta(I_{n_k}))) \;=\; \mathcal{S}(I,\Theta^\sharp).\end{aligned}$$

This contradicts the uniqueness of the optimal stem configuration, stated in Theorem 3.3. We thus conclude that the map $I \mapsto \Theta^*(I)$ is continuous, completing the proof. □

4.1. Uniqueness and representation of equilibrium solutions

By (3.21) and (4.2), this equilibrium configuration (h^*, θ^*) must satisfy the necessary condition

$$\theta^*(y) \;=\; \varphi\left((e^{-\kappa}-1)\exp\left\{\int_y^{h^*} \frac{\rho\kappa}{\sin\theta^*(y)}\,dy\right\}\right), \qquad\qquad y\in[0,h^*], \tag{4.10}$$

where φ is the function defined in Lemma 3.2. Here the constant h^* must be determined so that

$$\int_0^{h^*} \frac{1}{\sin\theta^*(y)}\,dy \;=\; \ell. \tag{4.11}$$

Based on (4.10), one obtains a simple representation of all equilibrium configurations, for any length $\ell > 0$. Indeed, for $t \in\,]-\infty, 0]$, let $t \mapsto \widehat{\zeta}(t)$ be the solution of the Cauchy problem

$$\zeta' \;=\; -\frac{\rho\kappa}{\sin\theta}, \qquad \text{where} \qquad \theta \;=\; \varphi\Big((e^{-\kappa}-1)\,e^{\zeta}\Big),$$

with terminal condition $\zeta(0) = 0$.

Notice that the corresponding function $t \mapsto \widehat{\theta}(t) = \varphi\Big((e^{-\kappa}-1)\,e^{\widehat{\zeta}(t)}\Big)$ satisfies

$$\widehat{\theta}(0) \;=\; \varphi(e^{-\kappa}-1) \;=\; \theta_0\,.$$

For any length ℓ of the stem, choose $h^* = h^*(\ell)$ so that

$$\int_{-h^*}^{0} \frac{1}{\sin\widehat{\theta}(t)}\,dt \;=\; \ell\,. \tag{4.12}$$

The shape of the stem that achieves the competitive equilibrium is then provided by

$$\theta^*(y) \;=\; \widehat{\theta}(y-h^*)\,, \qquad\qquad y\in[0,h^*]. \tag{4.13}$$

Since the backward Cauchy problem

$$\zeta' \;=\; -\frac{\rho\kappa}{\sin\Big(\varphi\big((e^{-\kappa}-1)\,e^{\zeta}\big)\Big)}\,, \qquad\qquad \zeta(0) = 0, \tag{4.14}$$

has a unique solution, we conclude that, if an equilibrium solution exists, by the representation (4.13) it must be unique. (See Fig. 3.)

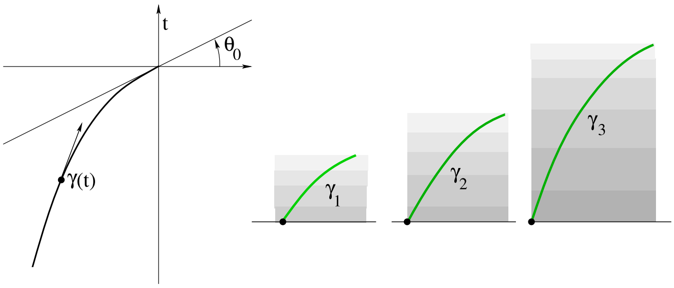


Fig. 3. Left: the curve γ, parameterized by the coordinate t. For $t < 0$, the tangent vector is $\frac{d\gamma}{dt} = (\tan\theta(t), 1)$, where $\theta(t)$ is obtained by solving the Cauchy problem (4.14). Right: for different lengths $0 < \ell_1 < \ell_2 < \ell_3$, the equilibrium configuration is obtained by taking the upper portion of the same curve γ, up to the length ℓ_i, $i = 1, 2, 3$.

5. Stems with variable length and thickness

We now consider the optimization problem **(OP2)**, allowing for stems of different lengths and with variable density of leaves.

5.1. Existence of an optimal solution

Theorem 5.1. *For any bounded, non-decreasing function $y \mapsto I(y) \in [0, 1]$ and any constants $0 < \alpha < 1$, $c > 0$ and $\theta_0 \in]0, \pi/2[$, the optimization problem* **(OP2)** *has at least one solution.*

Proof. 1. Consider a maximizing sequence of couples $(\theta_k, u_k) : \mathbb{R}_+ \mapsto [\theta_0, \pi/2] \times \mathbb{R}_+$. For $k \geq 1$, let

$$s \;\mapsto\; \gamma_k(s) \;=\; \left(\int_0^s \cos\theta_k(s)\,ds\,,\; \int_0^s \sin\theta_k(s)\,ds\right)$$

be the arc-length parameterization of the stem γ_k. Call μ_k the Radon measure on $\mathbb{R}^2$ describing the distribution of leaves along γ_k. For every Borel set $A \subseteq \mathbb{R}^n$, we thus have

$$\mu_k(A) \;=\; \int_{\{s\,;\;\gamma_k(s)\in A\}} u_k(s)\,ds. \tag{5.1}$$

For a given radius $\rho > 0$, we have the decomposition

$$\mu_k \;=\; \mu_k^\flat + \mu_k^\sharp\,,$$

where $\mu_k^\flat$ is the restriction of μ_k to the ball $B(0, \rho)$, while $\mu_k^\sharp$ is the restriction of μ_k to the complement $\mathbb{R}^2 \setminus B(0, \rho)$. By the same arguments used in steps **1-2** of the proof of Theorem 3.1 in [3], if the radius ρ is sufficiently large, then

$$\mathcal{S}(\mu_k^\flat) - c\mathcal{I}^\alpha(\mu_k^\flat) \;\geq\; \mathcal{S}(\mu_k) - c\mathcal{I}^\alpha(\mu_k) \tag{5.2}$$

for all $k \geq 1$. Here $\mathcal{S}$ and $\mathcal{I}^\alpha$ are the functionals defined at (2.12)-(2.13). According to (5.2), we can replace the measure μ_k with $\mu_k^\flat$ without decreasing the objective functional.

Without loss of generality we can thus choose $\ell > 0$ sufficiently large and assume that

$$u_k(s) = 0 \qquad \text{for all } s > \ell, \quad k \geq 1.$$

In turn, since $\mathcal{S}(\mu_k) - c\mathcal{I}^\alpha(\mu_k) \geq 0$, we obtain the uniform bound

$$\mathcal{I}^\alpha(\mu_k) \;\leq\; \kappa_1 \;\doteq\; \frac{1}{c}\mathcal{S}(\mu_k) \;\leq\; \frac{\ell}{c}. \tag{5.3}$$

2. In this step we show that the measures μ_k can be taken with uniformly bounded mass. Consider a measure μ_k for which (5.3) holds. By (2.13), for every $r \in [0, \ell]$ one has

$$\mathcal{I}^\alpha(\mu_k) \;\geq\; r \cdot \left(\int_r^\ell u_k(t)\,dt\right)^\alpha .$$

In view of (5.3), this implies

$$\int_r^\ell u_k(s)\,ds \;\leq\; \left(\frac{\kappa_1}{r}\right)^{1/\alpha}. \tag{5.4}$$

It thus remains to prove that, in our maximizing sequence, the functions u_k can be replaced with functions $\tilde{u}_k$ having a uniformly bounded integral over $[0, r]$, for some fixed $r > 0$.

Toward this goal we fix $0 < \varepsilon < \beta < 1$, and, for $j \geq 1$, we define $r_j = 2^{-j}$, and the interval $V_j =]r_{j+1}, r_j]$. Given $u = u_k$, if $\int_{V_j} u(s)\,ds > r_j^\varepsilon$, we introduce the functions

$$u_j(s) \;\doteq\; \chi_{V_j}(s)u(s), \qquad\qquad \tilde{u}_j(s) \;\doteq\; \min\{u_j(s), c_j\}, \tag{5.5}$$

choosing the constant $c_j \geq 2r_j^{\beta-1}$ so that

$$\int_{V_j} \tilde{u}_j(s)\,ds \;=\; r_j^\beta. \tag{5.6}$$

We then let $\mu_j = u_j\mu$ and $\tilde{\mu}_j = \tilde{u}_j\mu$ be the measures supported on V_j, corresponding to these densities.

For a fixed integer j^*, whose precise value will be chosen later, consider the set of indices

$$J \;\doteq\; \left\{ j \geq j^* \;\middle|\; \int_{V_j} u(s)\,ds > r_j^\varepsilon \right\} \tag{5.7}$$

and the modified density

$$\tilde u(s) \doteq u(s) + \sum_{j\in J} (\tilde u_j(s) - u_j(s)). \tag{5.8}$$

Moreover, call $\tilde\mu$ the measure obtained by replacing u with $\tilde u$ in (2.11). By (5.4) and (5.5) the total mass of $\tilde\mu$ is bounded. Indeed

$$\tilde\mu(\mathbb{R}^2) = \int_{r_{j^*}}^{\ell} \tilde u(s)\,ds + \int_0^{r_{j^*}} \tilde u(s)\,ds \le \left(\frac{\kappa_1}{r_{j^*}}\right)^{1/\alpha} + \sum_{j\ge j^*} r_j^{\varepsilon} \le \left(\frac{\kappa_1}{r_{j^*}}\right)^{1/\alpha} + \sum_{j\ge 1} 2^{-j\varepsilon} < +\infty. \tag{5.9}$$

We now claim that

$$\mathcal{S}(\tilde\mu) - c\mathcal{I}^\alpha(\tilde\mu) \;\ge\; \mathcal{S}(\mu) - c\mathcal{I}^\alpha(\mu). \tag{5.10}$$

Toward a proof of (5.10), we estimate

$$\begin{aligned}\mathcal{S}(\mu) - \mathcal{S}(\tilde\mu) &\le \sum_{j\in J}\Bigg(\int_{V_j} I(y(t))\cos(\theta(t)-\theta_0)\,dt \\ &\qquad - \int_{V_j} I(y(t))\left(1-\exp\left\{-\frac{\tilde u_j(t)}{\cos(\theta(t)-\theta_0)}\right\}\right)\cos(\theta(t)-\theta_0)\,dt\Bigg) \\ &\le \sum_{j\in J}\int_{r_{j+1}}^{r_j} \exp\{-\tilde u_j(t)\}dt \;\le\; \sum_{j\in J} r_{j+1}\exp\left\{-2r_j^{\beta-1}\right\}. \end{aligned} \tag{5.11}$$

To estimate the difference in the irrigation cost, we first observe that the inequality

$$\left(\int_r^\ell u(t)\,dt\right)^\alpha \;\le\; \frac{1}{r}\mathcal{I}^\alpha(\mu) \;=\; \frac{\kappa_1}{r}$$

implies

$$\left(\int_r^\ell u(t)\,dt\right)^{\alpha-1} \;\ge\; \left(\frac{\kappa_1}{r}\right)^{\frac{\alpha-1}{\alpha}}. \tag{5.12}$$

Since $\tilde u(s) \le u(s)$ for every $s\in[0,\ell]$, using (5.12) we now obtain

$$\begin{aligned}\mathcal{I}^\alpha(\mu) - \mathcal{I}^\alpha(\tilde\mu) &= \int_0^1 \frac{d}{d\lambda}\mathcal{I}^\alpha\big(\lambda\mu + (1-\lambda)\tilde\mu\big)\,d\lambda \\ &= \int_0^1\int_0^\ell \frac{d}{d\lambda}\left(\int_s^\ell [\lambda u(t) + (1-\lambda)\tilde u(t)]\,dt\right)^\alpha ds\,d\lambda\end{aligned}$$

$$\begin{aligned}
&= \int_0^1 \int_0^\ell \left\{ \alpha \left(\int_s^\ell [\lambda u(t) + (1-\lambda)\tilde{u}(t)]\, dt \right)^{\alpha-1} \int_s^\ell [u(t) - \tilde{u}(t)]\, dt \right\} ds\, d\lambda \\
&\geq \int_0^\ell \left\{ \alpha \left(\int_s^\ell u(t)\, dt \right)^{\alpha-1} \int_s^\ell [u(t) - \tilde{u}(t)]\, dt \right\} ds \\
&\geq \sum_{j\in J} \int_{r_{j+2}}^{r_{j+1}} \left[\alpha \left(\int_s^\ell u(t)\, dt \right)^{\alpha-1} \int_{r_{j+1}}^{r_j} (u_j(t) - \tilde{u}_j(t))\, dt \right] ds \\
&\geq \sum_{j\in J} \alpha \left(\frac{\kappa_1}{r_{j+2}} \right)^{\frac{\alpha-1}{\alpha}} \cdot (r_j^\varepsilon - r_j^\beta) \cdot r_{j+2} \\
&= \sum_{j\in J} \kappa_2 r_j^{1/\alpha} (r_j^\varepsilon - r_j^\beta),
\end{aligned} \tag{5.13}$$

where $\kappa_2 = \alpha(4\kappa_1)^{\frac{\alpha-1}{\alpha}}$. Combining (5.11) with (5.13) we obtain

$$c[\mathcal{I}^\alpha(\mu) - \mathcal{I}^\alpha(\tilde{\mu})] - [\mathcal{S}(\mu) - \mathcal{S}(\tilde{\mu})] \;\geq\; \sum_{j\in J} \Big(c\kappa_2 r_j^{1/\alpha}(r_j^\varepsilon - r_j^\beta) - r_{j+1} \exp\Big\{-2r_j^{\beta-1}\Big\} \Big). \tag{5.14}$$

By choosing the integer j^* large enough in (5.7), for $j \geq j^*$ all terms in the summation on the right hand side of (5.14) are ≥ 0. This implies (5.10).

3. By the two previous steps, w.l.o.g. we can assume that the measures μ_k have uniformly bounded support and uniformly bounded total mass. Otherwise, we can replace the sequence $(u_k)_{k\geq 1}$ with a new maximizing sequence $(\tilde{u}_k)_{k\geq 1}$ having these properties.

By taking a subsequence, we can thus assume the weak convergence $\mu_k \rightharpoonup \overline{\mu}$. The upper semicontinuity of the functional $\mathcal{S}$, proved in [5], yields

$$\mathcal{S}(\overline{\mu}) \;\geq\; \limsup_{k\to\infty} \; \mathcal{S}(\mu_k). \tag{5.15}$$

In addition, since all maps $s \mapsto \gamma_k(s)$ are 1-Lipschitz, by taking a further subsequence we can assume the convergence

$$\gamma_k(s) \;\to\; \overline{\gamma}(s) \tag{5.16}$$

for some limit function $\overline{\gamma}$, uniformly for $s \in [0, \ell]$.

Since each measure μ_k is supported on γ_k, the weak limit $\overline{\mu}$ is a measure supported on the curve $\overline{\gamma}$.

4. Since $\theta_k(s) \in [\theta_0, \pi/2]$, we can re-parameterize each stem γ_k in terms of the vertical variable

$$y_k(s) \;=\; \int_0^s \sin\theta_k(t)\, dt.$$

Calling $s = s_k(y)$ the inverse function, we thus obtain a maximizing sequence of couples

$$y \mapsto (\widehat{\theta}_k(y), \widehat{u}_k(y)) \doteq \Big(\theta_k(s_k(y)),\, u_k(s_k(y))\Big), \qquad\qquad y \in [0, h_k]\,.$$

Moreover, the stem γ_k can be described as the graph of the Lipschitz function

$$x = x_k(y) = \int_0^{s_k(y)} \cos\theta_k(s)\,ds.$$

Since all functions $x_k(\cdot)$ satisfy $x_k(0) = 0$ and are non-decreasing, uniformly continuous with Lipschitz constant $L = \cos\theta_0/\sin\theta_0$, by possibly extracting a further subsequence, we obtain the convergence $h_k \to \bar{h}$ and $x_k(\cdot) \to \bar{x}(\cdot)$. Here $\bar{x} : [0, \bar{h}] \mapsto \mathbb{R}$ is a nondecreasing continuous function with Lipschitz constant L, such that $\bar{x}(0) = 0$. More precisely, the convergence $x_k \to \bar{x}$ is uniform on every compact subinterval $[0, h]$ with $h < \bar{h}$.

5. We claim that the irrigation cost of $\overline{\mu}$ is no greater that the lim-inf of the irrigation costs for μ_k. Let $\sigma \mapsto \gamma(\sigma)$ be an arc-length parameterization of $\overline{\gamma}$. Since $s \mapsto \overline{\gamma}(s)$ is 1-Lipschitz, one has $d\sigma/ds \le 1$. We now compute

$$\begin{aligned}\mathcal{I}^\alpha(\overline{\mu}) &= \int_0^{\sigma(\ell)} \left(\int_\sigma^{\sigma(\ell)} \overline{u}(t)\,dt\right)^\alpha d\sigma = \int_0^{\sigma(\ell)} \left(\lim_{k\to\infty} \int_s^\ell u_k(t)\,dt\right)^\alpha d\sigma(s) \\ &\le \lim_{k\to\infty} \int_0^\ell \left(\int_s^\ell u_k(t)\,dt\right)^\alpha ds = \lim_{k\to\infty} \mathcal{I}^\alpha(\mu_k).\end{aligned} \tag{5.17}$$

6. Combining (5.15) with (5.17) we conclude that the measure $\overline{\mu}$, supported on the stem $\overline{\gamma}$, is optimal.

Let $\bar{u}$ be the density of the absolutely continuous part of $\overline{\mu}$ w.r.t. the arc-length measure on $\bar{\gamma}$, and call μ^* the measure that has density $\bar{u}$ w.r.t. arc-length measure. Since $\mathcal{S}(\mu^*) = \mathcal{S}(\overline{\mu})$, it follows that $\mu^* = \overline{\mu}$. Otherwise $\mathcal{I}^\alpha(\mu^*) < \mathcal{I}^\alpha(\overline{\mu})$ and $\overline{\mu}$ is not optimal. This argument shows that the optimal measure $\overline{\mu}$ is absolutely continuous w.r.t. the arc-length measure on $\overline{\gamma}$.

Calling $\sigma \mapsto \gamma(\sigma)$ the arc-length parameterization of $\overline{\gamma}$, the optimal solution to **(OP2)** is now provided by $\sigma \mapsto (\overline{\theta}(\sigma), \bar{u}(\sigma))$, where $\overline{\theta}$ is the orientation of the tangent vector:

$$\frac{d}{d\sigma}\overline{\gamma}(\sigma) = \big(\cos\overline{\theta}(\sigma),\, \sin\overline{\theta}(\sigma)\big). \quad \square$$

5.2. Necessary conditions for optimality

Let $t \mapsto (\theta^*(t), u^*(t))$ be an optimal solution to the problem **(OP2)**. The necessary conditions for optimality [4,6,7] yield the existence of dual variables p, q satisfying

$$\left\{\begin{aligned} \dot{p} &= -I'(y)\,G(\theta, u), \\ \dot{q} &= c\alpha\, z^{\alpha-1}, \end{aligned}\right. \qquad\qquad \left\{\begin{aligned} p(+\infty) &= 0, \\ q(0) &= 0, \end{aligned}\right. \tag{5.18}$$

and such that the maximality condition

$$(\theta^*(t), u^*(t)) = \arg\max_{\theta\in[0,\pi],\ u\geq 0} \Big\{ p(t)\sin\theta - q(t)u + I(y(t))\, G(\theta, u) - cz^\alpha \Big\}. \tag{5.19}$$

We recall that $G(\theta, u)$ is the function defined at (2.17). An intuitive interpretation of the quantities on the right-hand side of (5.19) goes as follows:

- $p(t)$ is the rate of increase in the gathered sunlight, if the upper portion of stem $\{\gamma(s)\,;\ s > t\}$ is raised higher.
- $q(t)$ is the rate at which the irrigation cost increases, adding mass at the point $\gamma(t)$.
- $I(y(t))\, G(\theta, u)$ is the sunlight captured by the leaves at the point $\gamma(t)$.

6. Uniqueness of the optimal stem configuration

Aim of this section is to show that, if the light intensity $I(y)$ remains sufficiently close to 1 for all $y \geq 0$, then the shape of the optimal stem is uniquely determined. This models a case where the density of external vegetation is small.

Theorem 6.1. *Let $h \mapsto I(h) \in [0, 1]$ be a non-decreasing, absolutely continuous function which satisfies*

$$I'(y) \leq Cy^{-\beta} \qquad \text{for a.e. } y > 0, \tag{6.1}$$

for some constants $C > 0$ and $0 < \beta < 1$. If

$$I(0) \geq 1 - \delta \tag{6.2}$$

for some $\delta > 0$ sufficiently small, then the optimal solution to **(OP2)** *is unique.*

Proof. We will show that the necessary conditions for optimality have a unique solution. This will be achieved in several steps. **1.** Given I, p, q, define the functions Θ, U by setting

$$\Big(\Theta(I, p, q),\ U(I, p, q)\Big) \doteq \arg\max_{\theta\in[0,\pi],\ u\geq 0} \Big\{ p\cdot\sin\theta - q\,u + I\cdot G(\theta, u) - cz^\alpha \Big\}. \tag{6.3}$$

We recall that G is the function defined at (2.17). Notice that one can write

$$G(\theta, u) = u\,\widetilde{G}\left(\frac{\cos(\theta - \theta_0)}{u}\right)$$

with

$$\widetilde{G}(x) \doteq \left(1 - \exp\left\{-\frac{1}{x}\right\}\right) x > 0, \qquad \widetilde{G}'(x) \leq 1, \qquad \widetilde{G}''(x) \leq 0, \qquad \text{for all } x > 0. \tag{6.4}$$

Denote by

$$\mathcal{H}(\theta, u) \doteq p\cdot\sin\theta - q\,u + I(y)\, G(\theta, u) - cz^\alpha \tag{6.5}$$

the quantity to be maximized in (6.3). Differentiating $\mathcal{H}$ w.r.t. θ and imposing that the derivative is zero, we obtain

$$\begin{aligned}\frac{p}{I} &= -\frac{G_\theta(\theta,u)}{\cos\theta}\\ &= \frac{\sin(\theta-\theta_0)}{\cos\theta}\left[1-\exp\left\{-\frac{u}{\cos(\theta-\theta_0)}\right\}-\frac{u}{\cos(\theta-\theta_0)}\exp\left\{-\frac{u}{\cos(\theta-\theta_0)}\right\}\right].\end{aligned} \tag{6.6}$$

Similarly, differentiating $\mathcal{H}$ w.r.t. u, we find

$$-q+IG_u(\theta,u) \;=\; -q+I\exp\left\{-\frac{u}{\cos(\theta-\theta_0)}\right\} \;=\; 0.$$

This yields

$$u \;=\; -\ln\left(\frac{q}{I}\right)\cos(\theta-\theta_0). \tag{6.7}$$

A lengthy but elementary computation shows that the Hessian matrix of second derivatives of $\mathcal{H}$ w.r.t. θ, u is negative definite, and the critical point is indeed the point where the global maximum is attained. By (6.7) it follows

$$U(I,p,q) \;=\; -\ln\left(\frac{q}{I}\right)\cos\big(\Theta(I,p,q)-\theta_0\big). \tag{6.8}$$

Inserting (6.8) in (6.6) and using the identity

$$\frac{\sin(\theta-\theta_0)}{\cos\theta} \;=\; \cos\theta_0\tan\theta-\sin\theta_0$$

we obtain

$$\Theta(I,p,q) \;=\; \arctan\left(\tan\theta_0+\frac{\frac{1}{\cos\theta_0}\,\frac{p}{I}}{1-\frac{q}{I}+\frac{q}{I}\ln\left(\frac{q}{I}\right)}\right). \tag{6.9}$$

Introducing the function

$$w(I,p,q) \;\doteq\; \frac{p/I}{1-\frac{q}{I}+\frac{q}{I}\ln\left(\frac{q}{I}\right)}, \tag{6.10}$$

by (6.9) one has the identities

$$\left\{\begin{aligned}\sin\big(\Theta(I,p,q)\big) &= \frac{\sin\theta_0+w}{\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}},\\ \cos\big(\Theta(I,p,q)-\theta_0\big) &= \frac{1+w\sin\theta_0}{\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}}.\end{aligned}\right. \tag{6.11}$$

Note that $w\geq 0$, because $p,q,I\geq 0$. In turn, from (6.11) it follows

$$\left\{\begin{aligned} \cos\big(\Theta(I,p,q)\big) &= \frac{\cos\theta_0}{\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}},\\ \sin\left(\Theta(I,p,q)-\theta_0\right) &= \frac{w\cos\theta_0}{\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}}.\end{aligned}\right. \tag{6.12}$$

2. The necessary conditions for the optimality of a solution to **(OP2)** yield the boundary value problem

$$\left\{\begin{aligned} \dot y(t) &= \sin\Theta,\\ \dot z(t) &= -U,\\ \dot p(t) &= -I'(y)G\big(\Theta,U\big),\\ \dot q(t) &= c\alpha z^{\alpha-1},\end{aligned}\right. \qquad \left\{\begin{aligned} y(0) &= 0,\\ z(T) &= 0,\\ p(T) &= 0,\\ q(T) &= I(y(T)),\\ q(0) &= 0.\end{aligned}\right. \tag{6.13}$$

Here $[0,T[$ is the interval where $u>0$, while

$$\Theta = \Theta(I(y),p,q), \qquad U = U(I(y),p,q) \tag{6.14}$$

are the functions introduced at (6.3), or more explicitly at (6.8)-(6.9). Notice that the length T of the stem is a quantity to be determined, using the boundary conditions in (6.13).

3. Since the control system (2.19) and the running cost (2.18) do not depend explicitly on time, the Hamiltonian function

$$H(y,z,p,q) \doteq \max_{\theta\in[0,\pi],\ u\geq 0}\Big\{p\cdot\sin\theta - q\,u + I(y)\,G(\theta,u) - cz^\alpha\Big\} \tag{6.15}$$

is constant along trajectories of (6.13). Observing that the terminal conditions in (6.13) imply $H(y(T),z(T),p(T),q(T))=0$, one has the first integral

$$H(y(t),z(t),p(t),q(t)) = 0 \qquad \text{for all } t\in[0,T]. \tag{6.16}$$

This yields

$$\begin{aligned} 0 &= p\sin\Theta + \left[I(y)-q+q\ln\left(\frac{q}{I(y)}\right)\right]\cos\left(\Theta-\theta_0\right) - cz^\alpha\\ &= \frac{p\left[\sin\theta_0+w\right]+\left[I(y)-q+q\ln\left(\frac{q}{I(y)}\right)\right]\left[1+w\sin\theta_0\right]}{\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}} - cz^\alpha\\ &= I(y)\left[1-\frac{q}{I(y)}+\frac{q}{I(y)}\ln\left(\frac{q}{I(y)}\right)\right]\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2} - cz^\alpha.\end{aligned}$$

We can use this identity to express z as a function of the other variables:

$$\begin{aligned} z\big(I(y),p,q\big) &= \left\{\frac{I(y)}{c}\left[1-\frac{q}{I(y)}+\frac{q}{I(y)}\ln\left(\frac{q}{I(y)}\right)\right]\sqrt{\cos^2\theta_0+(w+\sin\theta_0)^2}\right\}^{1/\alpha} \\ &= c^{-1/\alpha}\Bigg\{\left(\left[I(y)-q+q\ln\left(\frac{q}{I(y)}\right)\right]\cos\theta_0\right)^2 \\ &\qquad\qquad + \left(p+\left[I(y)-q+q\ln\left(\tfrac{q}{I(y)}\right)\right]\sin\theta_0\right)^2\Bigg\}^{1/2\alpha}. \end{aligned} \tag{6.17}$$

4. Since I is given as a function of the height y, it is convenient to rewrite the equations (6.13) using y as an independent variable. Using the identity (6.17), we obtain a system of two equations for the variables p, q:

$$\begin{aligned} \frac{d}{dy}p(y) &= -I'(y)\left[1-\frac{q(y)}{I(y)}\right]\frac{\cos\big(\Theta\big(I(y),p(y),q(y)\big)-\theta_0\big)}{\sin\Theta\big(I(y),p(y),q(y)\big)} \\ &= -I'(y)\left[1-\frac{q(y)}{I(y)}\right]\frac{1+w\sin\theta_0}{w+\sin\theta_0} \\ &\doteq -I'(y)\,f_1\big(I(y),p(y),q(y)\big), \end{aligned} \tag{6.18}$$

$$\begin{aligned} \frac{d}{dy}q(y) &= \frac{c\alpha\big[z\big(I(y),p(y),q(y)\big)\big]^{\alpha-1}}{\sin\Theta\big(I(y),p(y),q(y)\big)} \\ &= \frac{\alpha c^{1/\alpha}}{w+\sin\theta_0}\left[\cos^2\theta_0+(\sin\theta_0+w)^2\right]^{1-\frac{1}{2\alpha}} \\ &\qquad\times\left[I(y)\left(1-\frac{q}{I(y)}+\frac{q}{I(y)}\ln\left(\frac{q}{I(y)}\right)\right)\right]^{1-\frac{1}{\alpha}} \\ &\doteq f_2\big(I(y),p(y),q(y)\big), \end{aligned} \tag{6.19}$$

where $w = w(I,p,q)$ is the function introduced at (6.10). Note that under our assumptions, f_1 remains bounded, while f_2 diverges as $q(y)\to I(y)$. The system (6.13) can now be equivalently formulated as

$$\begin{cases} p'(y) = -I'(y)\,f_1\big(I(y),p,q\big), \\ q'(y) = f_2\big(I(y),p,q\big), \end{cases} \qquad \begin{cases} p(h) = 0, \\ q(h) = I(h), \end{cases} \qquad q(0)=0. \tag{6.20}$$

5. To prove uniqueness of the solution to the boundary value problem (6.13), it thus suffices to prove the following (see Fig. 4, right).

(U) *Call*

$$y \mapsto \big(p(y,h),\,q(y,h)\big) \tag{6.21}$$

the solution to the system (6.20), *with the two terminal conditions given at* $y=h$. *Then there is a unique choice of* $h>0$ *which satisfies also the third boundary condition*

$$q(0,h) = 0. \tag{6.22}$$

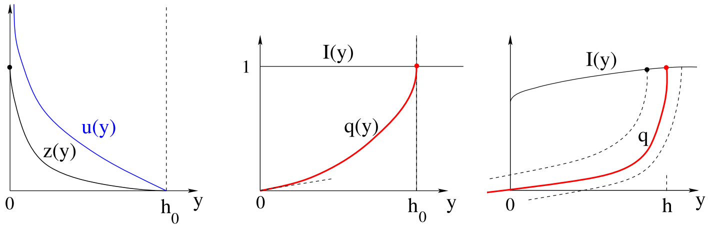


Fig. 4. Left and center: sketch of the solution of the system (5.18) in the case where $I(y) \equiv 1$. Left: the graphs of the functions z in (6.25) and $u = -\ln q$. Center: the graph of the function q at (6.26). The figure on the right shows the case where $I(\cdot)$ is not constant. As before, h must be determined so that $q(0, h) = 0$.

To make the argument more clear, the uniqueness property **(U)** will be proved in two steps.

(i) When $I(y) \equiv 1$, the map

$$h \mapsto q(0,h) \tag{6.23}$$

is strictly decreasing, hence it vanishes at a unique point h_0.

(ii) For all functions $I(\cdot)$ sufficiently close to the constant map $\equiv 1$, the map (6.23) is strictly decreasing in a neighborhood of h_0.

In the case $I(y) \equiv 1$, recalling (6.9) we obtain (see Fig. 4)

$$I'(y) = 0, \qquad p(y,h) = 0, \qquad \Theta(I,0,q) = \theta_0, \qquad G(\theta_0,U) = 1-e^{-U},$$
$$U(1,0,q) = \operatorname*{argmax}_u \left\{-qu + G(\theta_0,U)\right\} = \operatorname*{argmax}_u \{-qu+1-e^{-u}\} = -\ln q,$$

The system (6.13) can now be written as

$$\begin{cases} p'(y) = 0, \\ q'(y) = \dfrac{c\alpha z^{\alpha-1}}{\sin\theta_0}, \\ z'(y) = \dfrac{\ln q}{\sin\theta_0}, \end{cases} \qquad \begin{cases} p(h) = 0, \\ q(h) = 1, \\ z(h) = 0, \end{cases} \qquad q(0) = 0. \tag{6.24}$$

From (6.24) it follows $p(y) \equiv 0$, while

$$\frac{dz}{dq} = \frac{\ln q}{c\alpha z^{\alpha-1}}.$$

Integrating the above ODE with terminal conditions $q = 1$, $z = 0$, one obtains

$$z = c^{-1/\alpha}\Big[1 + q\ln q - q\Big]^{1/\alpha}. \tag{6.25}$$

The second equation in (6.24) thus becomes

$$q'(y) \;=\; \frac{\alpha c^{1/\alpha}}{\sin\theta_0}\Big[1+q\ln|q|-q\Big]^{\frac{\alpha-1}{\alpha}}. \tag{6.26}$$

Notice that here the right hand side is strictly positive for all $q \in\,]-1,1[\,$. Of course, only positive values of q are relevant for the optimization problem, but for the analysis it is convenient to extend the definition also to negative values of q. The solution of (6.26) with terminal condition $q(h) = 1$ is implicitly determined by

$$h - y \;=\; \frac{\sin\theta_0}{\alpha c^{1/\alpha}}\int_{q(y)}^{1}\Big[1+s\ln|s|-s\Big]^{\frac{1-\alpha}{\alpha}}\,ds\,. \tag{6.27}$$

The map $h \mapsto q(0,h)$ thus vanishes at the unique point

$$h_0 \;=\; \frac{\sin\theta_0}{\alpha c^{1/\alpha}}\int_{0}^{1}\Big[1+s\ln|s|-s\Big]^{\frac{1-\alpha}{\alpha}}\,ds. \tag{6.28}$$

As expected, the height h_0 of the optimal stem decreases as we increase the constant c in the transportation cost. A straightforward computation yields

$$\frac{\partial}{\partial h}q(0,h) \;=\; -\frac{\alpha c^{1/\alpha}}{\sin\theta_0}\Big[1+q(0,h)\ln|q(0,h)|-q(0,h)\Big]^{\frac{1-\alpha}{\alpha}}. \tag{6.29}$$

In particular, at $h = h_0$ we have $q^{(h_0)}(0) = 0$ and hence

$$\frac{d}{dh}q(0,h)\bigg|_{h=h_0} \;=\; -\frac{\alpha c^{1/\alpha}}{\sin\theta_0} \;<\; 0. \tag{6.30}$$

6. We will show that a strict inequality as in (6.30) remains valid for a more general function $I(\cdot)$, provided that the assumptions (6.1)-(6.2) hold.

Toward this goal, we need to determine how p and q vary w.r.t. the parameter h. Denoting by

$$P(y) \;\doteq\; \frac{\partial p(y,h)}{\partial h}, \qquad Q(y) \;\doteq\; \frac{\partial q(y,h)}{\partial h} \tag{6.31}$$

their partial derivatives, by (6.20) one obtains the linear system

$$\begin{pmatrix} P(y) \\ Q(y)\end{pmatrix}' \;=\; \begin{pmatrix} -I'(y)f_{1,p} & -I'(y)f_{1,q} \\ f_{2,p} & f_{2,q}\end{pmatrix}\begin{pmatrix} P(y) \\ Q(y)\end{pmatrix}. \tag{6.32}$$

The boundary conditions at $y = h$ require some careful consideration. As $y \to h-$, we expect $f_2(I(y), p(y), q(y)) \to +\infty$ and $Q(y) \to -\infty$. To cope with this singularity we introduce the new variable

$$\widetilde{Q}(y) \;\doteq\; \frac{Q(y)}{f_2\big(I(y), p(y), q(y)\big)}. \tag{6.33}$$

The system (6.32), together with the new boundary conditions for P, $\widetilde{Q}$, can now be written as

$$\begin{cases} P'(y) \;=\; -I'(y)\big[f_{1,p}P + f_{1,q}f_2\widetilde{Q}\big], \\[2mm] \widetilde{Q}'(y) \;=\; \dfrac{f_{2,p}}{f_2}P - \dfrac{I'(y)[f_{2,I} - f_{2,p}f_1]}{f_2}\widetilde{Q}, \end{cases} \qquad \begin{cases} P(h) \;=\; 0, \\ \widetilde{Q}(h) \;=\; -1. \end{cases} \tag{6.34}$$

To analyze this system we must compute the partial derivatives of f_1 and f_2. From the definition (6.10) it follows

$$\frac{\partial w}{\partial I} \;=\; \frac{w^2}{p}\left[1 - \frac{q}{I}\right], \qquad \frac{\partial w}{\partial p} \;=\; \frac{w}{p}, \qquad \frac{\partial w}{\partial q} \;=\; -\frac{w^2}{p}\ln\left(\frac{q}{I}\right). \tag{6.35}$$

Using (6.35), from (6.18), (6.19) we obtain

$$\begin{cases} f_{1,p}\big(I(y), p, q\big) \;=\; \dfrac{1 - \frac{q}{I(y)}}{I(y)\tan^2\Theta\left[1 - \frac{q}{I(y)} + \frac{q}{I(y)}\ln\left(\frac{q}{I(y)}\right)\right]}, \\[4mm] f_{1,q}\big(I(y), p, q\big) \;=\; \dfrac{1}{I(y)}\dfrac{\cos(\Theta - \theta_0)}{\sin\Theta} - \dfrac{\sin(\Theta - \theta_0)\cos\Theta\left[1 - \frac{q}{I(y)}\right]\ln\left(\frac{q}{I}\right)}{I(y)\sin^2\Theta\left[1 - \frac{q}{I(y)} + \frac{q}{I(y)}\ln\left(\frac{q}{I(y)}\right)\right]}, \\[4mm] f_{2,p}\big(I(y), p, q\big) \;=\; -\left[1 + \dfrac{\alpha}{\sin^2\Theta} - 2\alpha\right]\dfrac{1}{z\big(I(y), p, q\big)}, \\[4mm] f_{2,q}\big(I(y), p, q)\big) \;=\; -\left[\dfrac{(1-\alpha)\sin\theta_0}{\sin^2\Theta} - \dfrac{\sin(\Theta - \theta_0)}{\cos\Theta}\left(1 + \dfrac{\alpha}{\sin^2\Theta} - 2\alpha\right)\right]\dfrac{\ln\left(\frac{q}{I(y)}\right)}{z\big(I(y), p, q\big)}, \\[4mm] f_{2,I}\big(I(y), p, q\big) \;=\; -\left[\dfrac{(1-\alpha)\sin\theta_0}{\sin^2\Theta} + \dfrac{\sin(\Theta - \theta_0)}{\cos\Theta}\left(1 + \dfrac{\alpha}{\sin^2\Theta} - 2\alpha\right)\right]\dfrac{1 - \frac{q}{I(y)}}{z\big(I(y), p, q\big)}. \end{cases} \tag{6.36}$$

At this stage, the strategy of the proof is straightforward. When $I'(y) \equiv 0$, the solution to (6.34) is trivially given by $P(y) \equiv 0$, $\widetilde{Q}(y) \equiv -1$. This implies

$$\frac{\partial}{\partial h}q(0, h) \;=\; \widetilde{Q}(0) \cdot f_2(I(0), p(0), q(0)) \;<\; 0.$$

We need to show that the same strict inequality holds when $\delta > 0$ in (6.2) is small enough. Notice that, if the right hand sides of the equations in (6.34) were bounded, letting $\|I'\|_{\mathbf{L}^\infty} \to 0$ a continuity argument would imply the uniform convergence $P(y) \to 0$ and $\widetilde{Q}(y) \to -1$. The same conclusion can be achieved provided that the right hand sides in (6.34) are uniformly integrable. This is precisely what will be proved in the next two steps, relying on the identities (6.36).

7. In this step we prove an inequality of the form

$$0 \;<\; \theta_0 \;\leq\; \Theta(I, p, q) \;\leq\; \theta^+ \;<\; \frac{\pi}{2}. \tag{6.37}$$

As a consequence, this implies that all terms in (6.36) involving $\sin\Theta$ or $\cos\Theta$ remain uniformly positive.

The lower bound $\Theta \geq \theta_0$ is an immediate consequence of (6.9). To obtain an upper bound on Θ, we set

$$q^\sharp \doteq \frac{q(y)}{I(y)}.$$

By (6.13), a differentiation yields

$$\dot{q}^\sharp = \frac{c\alpha z^{\alpha-1} - q^\sharp I' \sin(\Theta)}{I}.$$

Next, we observe that, by (6.13), one has

$$\frac{dz}{dq^\sharp} = \ln q^\sharp \cdot \cos(\Theta - \theta_0) \cdot \frac{I}{c\alpha z^{\alpha-1} - q^\sharp I' \sin(\Theta)} = \varphi_1(q^\sharp) \cdot \ln q^\sharp \cdot \alpha z^{\alpha-1}, \qquad \begin{cases} z(h) = 0, \\ q^\sharp(h) = 1. \end{cases}$$

In (6.2) we can now choose $\delta \leq c\alpha M^{\alpha-1}$, where $M \geq z(0)$ is an a priori bound on the mass of the stem, derived in Section 5. This ensures that φ_1 is a bounded, uniformly positive function for y close enough to h, say

$$0 < c^- \leq \varphi_1 \leq c^+,$$

for some constants c^-, c^+. Integrating, we obtain

$$z^\alpha = \int_0^z \alpha\zeta^{\alpha-1}\, d\zeta = -\int_{q^\sharp}^1 \varphi_1(s) \ln s\, ds = -\varphi_2(q^\sharp) \int_{q^\sharp}^1 \ln s\, ds = \varphi_3(q^\sharp) \cdot (1 - q^\sharp)^2, \tag{6.38}$$

and

$$\begin{aligned} \frac{dq^\sharp}{dy} &= \frac{c\alpha}{\sin\Theta} \left(-\int_{q^\sharp}^1 \varphi_1(s) \ln s\, ds \right)^{\frac{\alpha-1}{\alpha}} = \varphi_4(q^\sharp) \cdot \left(-\int_{q^\sharp}^1 \ln s\, ds \right)^{\frac{\alpha-1}{\alpha}} \\ &= \varphi_5(q^\sharp) \cdot (1 - q^\sharp)^{\frac{2(\alpha-1)}{\alpha}}. \end{aligned} \tag{6.39}$$

Here the φ_k are uniformly positive, bounded functions. Integrating (6.39) we obtain

$$\int_{q^\sharp}^1 \frac{1}{\varphi_5(s)} (1 - s)^{\frac{2(1-\alpha)}{\alpha}}\, ds = h - y. \tag{6.40}$$

To fix the ideas, assume

$$0 < c_3 \leq \varphi_5(s) \leq C_3.$$

Then

$$\frac{1}{c_3}\int_{q^\sharp}^{1}(1-s)^{\frac{2(1-\alpha)}{\alpha}}\,ds \;=\; \frac{\alpha}{(2-\alpha)c_3}(1-q^\sharp)^{\frac{2-\alpha}{\alpha}}\,ds \;\geq\; h-y.$$

$$1-q^\sharp(y)\;\geq\;\left(\frac{(2-\alpha)c_3}{\alpha}\right)^{\frac{\alpha}{2-\alpha}}(h-y)^{\frac{\alpha}{2-\alpha}}. \tag{6.41}$$

A similar argument yields

$$1-q^\sharp(y)\;\leq\;\left(\frac{(2-\alpha)C_3}{\alpha}\right)^{\frac{\alpha}{2-\alpha}}(h-y)^{\frac{\alpha}{2-\alpha}}. \tag{6.42}$$

Using (6.1) and (6.42) in the equation (6.18) we obtain a bound of the form

$$-p'(y)\;\leq\;C_1(1-q(y))\;\leq\;C_2(h-y)^{\frac{\alpha}{2-\alpha}} \tag{6.43}$$

for y in a left neighborhood of h, which yields

$$p(y)\;\leq\;\frac{C_2}{\alpha+1}(h-y)^{\frac{2}{2-\alpha}}. \tag{6.44}$$

Since $\alpha<1$, using (6.41) and (6.44) in (6.9) we obtain the limit $\Theta(y)\to\theta_0$ as $y\to h-$.

On the other hand, when y is bounded away from h, the denominator in (6.10) is strictly positive and the quantity $w=w(I,p,q)$ remains uniformly bounded. By (6.9), we obtain the upper bound $\Theta\leq\theta^+$, for some $\theta^+<\pi/2$.

8. Relying on (6.36), in this step we prove that all terms on the right hand sides of the ODEs in (6.34) are uniformly integrable.

(i) We first consider the terms appearing in the ODE for $P(y)$. Concerning $f_{1,p}$, as $y\to h-$ one has

$$f_{1,p}\;=\;\mathcal{O}(1)\cdot\left(1-\frac{q}{I}\right)^{-1}\;=\;\mathcal{O}(1)\cdot(h-y)^{\frac{-\alpha}{2-\alpha}}, \tag{6.45}$$

because of (6.41). Since $\alpha<1$, this implies that $f_{1,p}$ is an integrable function of y.

(ii) By the second equation in (6.36), as $y\to h-$ one has

$$f_{1,q}\;=\;\mathcal{O}(1)\cdot\frac{(1-q^\sharp)\ln(q^\sharp)}{1-q^\sharp+q^\sharp\ln(q^\sharp)}\;=\;\mathcal{O}(1). \tag{6.46}$$

(iii) The term f_2 blows up as $y\to h-$, due to the factor $z^{\alpha-1}$. However, this factor is integrable in y because, by (6.38), (6.41) and (6.42)

$$z^\alpha\big(I(y),p(y),q(y)\big)\;=\;\mathcal{O}(1)\cdot(h-y)^{\frac{2\alpha}{2-\alpha}}. \tag{6.47}$$

This implies

$$\begin{aligned} f_2\big(I(y),p(y),q(y)\big) &= \mathcal{O}(1)\cdot z^{\alpha-1}\big(I(y),p(y),q(y)\big)\\ &= \mathcal{O}(1)\cdot (h-y)^{-1+\frac{\alpha}{2-\alpha}}, \end{aligned} \tag{6.48}$$

showing that f_2 is integrable, because $\alpha > 0$.

(iv) We now solve the linear ODE for P in (6.34) with terminal condition $P(h) = 0$. By the estimates (6.45)-(6.46) and (6.48) one obtains a bound of the form

$$P(y) \;=\; \mathcal{O}(1)\cdot (h-y)^{\frac{\alpha}{2-\alpha}}, \tag{6.49}$$

valid in a left neighborhood of $y = h$.

(v) In a neighborhood of the origin, the function $f_{1,q}$ contains a logarithm which blows up as $y \to 0+$. However, this is integrable because, for $y \approx 0$, we have

$$\frac{q(y)}{I(y)} \approx \left(\frac{d}{dy}\frac{q(y)}{I(y)}\right)\bigg|_{y=0}\cdot y \;=\; \frac{c\alpha}{(z(0))^{1-\alpha} I(0)\sin(\Theta(0))}\,y,$$

and $\ln y$ is integrable in y. Recalling (6.1), as y ranges in a right neighborhood of the origin, i.e. for $y > 0$, we conclude

$$\begin{cases} I'(y)\cdot f_{1,q} f_2 \;=\; \mathcal{O}(1)\cdot I'(y) f_{1,q} \;=\; \mathcal{O}(1)\cdot y^{-\beta}\ln y,\\[2ex] \quad I'(y)\cdot f_{1,p} \;=\; \mathcal{O}(1)\cdot I'(y) \;=\; \mathcal{O}(1)\cdot y^{-\beta}. \end{cases} \tag{6.50}$$

This shows that, in (6.34), the coefficients in first equation are uniformly integrable in a right neighborhood of the origin.

(vi) It remains to consider the terms appearing in the ODE for $\widetilde{Q}(y)$. We first observe that

$$\frac{f_{2,p}}{f_2} \;=\; -\,\frac{\sin\Theta}{c\alpha}\left[1+\frac{\alpha}{\sin^2\Theta}-2\alpha\right] z^{-\alpha}\big(I(y),p(y),q(y)\big).$$

As $y \to h-$, by (6.47) and (6.49) this implies

$$\frac{f_{2,p}}{f_2}\cdot P \;=\; \mathcal{O}(1)\cdot (h-y)^{\frac{-2\alpha}{2-\alpha}}\cdot (h-y)^{\frac{\alpha}{2-\alpha}}, \tag{6.51}$$

which is integrable for $\alpha < 1$.

(vii) Finally, as $y \to h-$, we consider

$$\begin{aligned} \frac{f_{2,I}}{f_2} \;=\; &-\,\frac{\sin\Theta}{c\alpha}\left[\frac{(1-\alpha)\sin\theta_0}{\sin^2\Theta}+\frac{\sin(\Theta-\theta_0)}{\cos\Theta}\left(1+\frac{\alpha}{\sin^2\Theta}-2\alpha\right)\right]\\ &\times\frac{1-\frac{q}{I(y)}}{z^{\alpha}\big(I(y),p(y),q(y)\big)}\\ =\; &\mathcal{O}(1)\cdot(1-q^{\sharp})z^{-\alpha}\big(I(y),p(y),q(y)\big) \;=\; \mathcal{O}(1)\cdot (h-y)^{\frac{\alpha}{2-\alpha}}\cdot (h-y)^{\frac{-2\alpha}{2-\alpha}}, \end{aligned} \tag{6.52}$$

which is integrable in y since $\alpha < 1$. Similarly, by (6.51), (6.18), and (6.42), it follows

$$\frac{f_{2,p}}{f_2}\cdot f_1 \;=\; \mathcal{O}(1)\cdot (h-y)^{\frac{-2\alpha}{2-\alpha}}\cdot (h-y)^{\frac{\alpha}{2-\alpha}}, \tag{6.53}$$

which is again integrable.

9. The proof can now be accomplished by a contradiction argument. If the conclusion of the theorem were not true, one could find a sequence of absolutely continuous, non-decreasing functions $I_n : \mathbb{R}_+ \mapsto [0,1]$, all satisfying (6.1), with $I_n(0) \to 1$, and such that, for each $n \geq 1$, the optimization problem **(OP2)** has two distinct solutions, say $(\check\theta_n, \check u_n)$ and $(\hat\theta_n, \hat u_n)$. As a consequence, for each $n \geq 1$ the system (6.13) has two solutions. To fix the ideas, let the first solution be defined on $[0, \check h_n]$ and the second on $[0, \hat h_n]$, with $\check h_n < \hat h_n$. These two solutions will be denoted by $(\check p_n, \check q_n, \check z_n)$ and $(\hat p_n, \hat q_n, \hat z_n)$. They both satisfy the boundary conditions

$$\check p_n(\check h_n) = \hat p_n(\hat h_n) = 0, \qquad \check q_n(\check h_n) = I(\check h_n), \qquad \hat q_n(\hat h_n) = I(\hat h_n), \qquad \check q_n(0) = \hat q_n(0) = 0. \tag{6.54}$$

As a preliminary, we observe that, for $\delta > 0$ small, the heights $\hat h$, $\check h$ of optimal stems must remain uniformly positive. Indeed, by (2.3) the sunlight gathered by a stem γ of length ℓ is bounded by

$$\mathcal{S}(\gamma) \;\leq\; \ell.$$

Hence, for a sequence of stems γ_n with heights $\hat h_n \to 0$, the total sunlight satisfies

$$\mathcal{S}(\gamma_n) \;\leq\; \ell_n \;\leq\; \frac{\hat h_n}{\sin\theta_0} \;\to\; 0.$$

Therefore, for n large, none of these stems can be optimal.

Thanks to the last identity in (6.54), by the mean value theorem there exists some intermediate point $k_n \in [\check h_n, \hat h_n]$ such that, with the notation introduced at (6.21),

$$\frac{\partial q_n}{\partial h}(0, k_n) \;=\; 0. \tag{6.55}$$

For each $n \geq 1$ consider the corresponding system

$$\begin{cases} P_n'(y) \;=\; -I_n'(y)\big[f_{1,p}P_n + f_{1,q}f_2\widetilde Q_n\big], \\[2mm] \widetilde Q'(y) \;=\; \dfrac{f_{2,p}}{f_2}P_n - \dfrac{I_n'(y)[f_{2,I} - f_{2,p}f_1]}{f_2}\widetilde Q_n, \end{cases} \qquad \begin{cases} P_n(k_n) \;=\; 0, \\[2mm] \widetilde Q_n(k_n) \;=\; -1. \end{cases} \tag{6.56}$$

Since $f_2\big(I_n(0), p_n(0,k_n), 0\big) > 0$, by (6.55) it follows

$$\widetilde Q_n(0) \;=\; \frac{1}{f_2\big(I_n(0), p_n(0,k_n), 0\big)}\cdot\frac{\partial q_n}{\partial h}(0, k_n) \;=\; 0. \tag{6.57}$$

Let

$$P_n(y) \doteq \frac{\partial p(y,k_n)}{\partial h}, \qquad \widetilde{Q}_n(y) \doteq \frac{1}{f_2\big(I_n(y), p_n(y,k_n), q_n(y,k_n)\big)} \cdot \frac{\partial q(y,k_n)}{\partial h},$$

be the solutions to (6.56). By the previous steps, their derivatives $\big(P_n', \widetilde{Q}_n'\big)_{n\geq 1}$ form a sequence of uniformly integrable functions defined on the intervals $[0,k_n]$. Note that the existence of an upper bound $\sup_n k_n \doteq h^+ < +\infty$ follows from the existence proof.

Thanks to the uniform integrability, by possibly taking a subsequence, we can assume the convergence $k_n \to \overline{h} \in [0,h^+]$, the weak convergence of derivatives $P_n' \rightharpoonup P'$, $\widetilde{Q}_n' \rightharpoonup \widetilde{Q}'$ in $\mathbf{L}^1$, and the convergence

$$P_n \to P, \qquad \widetilde{Q}_n \to \widetilde{Q},$$

uniformly on every subinterval $[0,h]$ with $h < \bar{h}$.

Recalling that every I_n' satisfies the uniform bounds (6.1), since $I_n(y) \to I(y) \equiv 1$ uniformly for all $y \geq 0$, we conclude that $(P, \widetilde{Q})$ provides a solution to the linear system (6.34) on $[0,\bar{h}]$, corresponding to the constant function $I(y) \equiv 1$. We now observe that, when $I(y) \equiv 1$, the solution to (6.34) is $P(y) \equiv 0$ and $\widetilde{Q}(y) \equiv -1$. On the other hand, our construction yields

$$\widetilde{Q}(0) = \lim_{n\to\infty} \widetilde{Q}_n(0) = 0.$$

This contradiction achieves the proof of Theorem 6.1. □

7. Existence of an equilibrium solution

Given a nondecreasing light intensity function $I : \mathbb{R}_+ \mapsto [0,1]$, in the previous section we proved the existence of an optimal solution (θ^*, u^*) for the maximization problem **(OP2)**.

Conversely, let $\rho_0 > 0$ be the constant density of stems, i.e. the number of stems growing per unit area. If all stems have the same configuration, described by the couple of functions $y \mapsto (\theta(y), u(y))$ as in (2.18), then the corresponding intensity of light at height y above ground is computed as

$$I^{(\theta,u)}(y) \doteq \exp\left\{ -\frac{\rho_0}{\cos\theta_0} \int_y^{+\infty} \frac{u(\zeta)}{\sin\theta(\zeta)}\, d\zeta \right\}. \tag{7.1}$$

The main goal of this section is to find a competitive equilibrium, i.e. a fixed point of the composition of the two maps $I \mapsto (\theta^*, u^*)$ and $(\theta, u) \mapsto I^{(\theta,u)}$.

Definition 7.1. Given an angle $\theta_0 \in]0, \pi/2[$ and a constant $\rho_0 > 0$, we say that the light intensity function $I^* : \mathbb{R}_+ \mapsto [0,1]$ and the stem configuration $(\theta^*, u^*) : \mathbb{R}_+ \mapsto [\theta_0, \pi/2] \times \mathbb{R}_+$ yield a **competitive equilibrium** if the following holds.

(i) The couple (θ^*, u^*) provides an optimal solution to the optimization problem **(OP2)**, with light intensity function $I = I^*$.
(ii) The identity $I^* = I^{(\theta^*,u^*)}$ holds.

The main result of this section provides the existence of a competitive equilibrium, assuming that the density ρ_0 of stems is sufficiently small.

Theorem 7.2. *Let an angle $\theta_0 \in]0, \pi/2[$ be given. Then, for all $\rho_0 > 0$ sufficiently small, a unique competitive equilibrium (I^*, θ^*, u^*) exists.*

Proof. 1. Setting $C = 1$ and $\beta = 1/2$ in (6.1), we define the family of functions

$$\begin{aligned}\mathcal{F} \doteq \Big\{ I : \mathbb{R}_+ \mapsto [1-\delta,\, 1]\,; \quad & I \text{ is absolutely continuous,} \\ & I'(y) \in \big[0,\, y^{-1/2}\big] \quad \text{for a.e. } y > 0 \Big\},\end{aligned} \tag{7.2}$$

where $\delta > 0$ is chosen small enough so that the conclusion of Theorem 6.1 holds.

2. For each $I \in \mathcal{F}$, let $(\theta^{(I)}, u^{(I)})$ describe the corresponding optimal stem. Calling

$$h^{(I)} \;=\; \sup\,\big\{ y \geq 0\,;\;\; u^{(I)}(y) > 0 \big\}$$

the height of this stem, by the a priori bounds proved in Section 6 we have a uniform bound

$$h^{(I)} \;\leq\; h^+$$

for all $I \in \mathcal{F}$. Let $p^{(I)}, q^{(I)} : [0, h^{(I)}] \mapsto \mathbb{R}_+$ be the corresponding solutions of (6.20). For convenience, we extend all these functions to the larger interval $[0, h^+]$ by setting

$$p^{(I)}(y) \doteq p^{(I)}\big(h^{(I)}\big), \qquad\qquad q^{(I)}(y) \doteq q^{(I)}\big(h^{(I)}\big), \qquad \text{for all } y \in [h^{(I)}, h^+].$$

3. By the analysis in Section 6, for any $I \in \mathcal{F}$, the solution to the system of optimality conditions (6.13) satisfies

$$\theta_0 \;\leq\; \Theta(I(y), p(y), q(y)) \;\leq\; \theta^+, \qquad\qquad c_0\, y \;\leq\; \frac{q(y)}{I(y)} \;\leq\; 1, \tag{7.3}$$

for some $\theta^+ < \pi/2$ and $c_0 > 0$ sufficiently small. In view of (6.8), this implies

$$U(I(y), p(y), q(y)) \;\doteq\; -\ln\left(\frac{q(I)}{I(y)}\right)\cos\big(\Theta(I(y), p(y), q(y)) - \theta_0\big) \;\leq\; -\ln(c_0 y). \tag{7.4}$$

Note that $\Theta(I(y), p^{(I)}(y), q^{(I)}(y)) = \theta^{(I)}(y)$ and $U(I(y), p^{(I)}(y), q^{(I)}(y)) = u^{(I)}(y)$. Thus, if we choose $\rho_0 > 0$ small enough, it follows that the corresponding light intensity function $I^{(\theta,u)}$ at (7.1) is again in $\mathcal{F}$. A competitive equilibrium will be obtained by constructing a fixed point of the composition of the two maps

$$\Lambda_1 : I \;\mapsto\; \big(\theta^{(I)},\, u^{(I)}\big), \qquad\qquad \Lambda_2 : (\theta, u) \;\mapsto\; I^{(\theta,u)}. \tag{7.5}$$

In order to use Schauder's theorem, we need to check the continuity of these maps, in a suitable topology.

We start by observing that $\mathcal{F} \subset \mathcal{C}^0([0,h^+])$ is a compact, convex set. Again by the analysis in Section 6, as I varies within the domain $\mathcal{F}$, the corresponding functions $\theta^{(I)}$ are uniformly bounded in $\mathbf{L}^\infty([0,h^+])$, while $u^{(I)}$ is uniformly bounded in $\mathbf{L}^1([0,h^+])$.

From the estimate (6.43) it follows that the functions $p^{(I)}$ are equicontinuous on $[0,h^+]$. Recalling that $q = q^\sharp \cdot I$, by (6.39) we conclude that the functions $q^{(I)}$ are equicontinuous as well.

4. Motivated by (7.3)-(7.4), we consider the set of functions

$$\mathcal{U} \doteq \Big\{(\theta,u) \in \mathbf{L}^1\big([0,h^+];\ \mathbb{R}^2\big), \quad \theta(y) \in [\theta_0,\theta^+],\ \ 0 \le u(y) \le -\ln(c_0 y)\Big\}. \tag{7.6}$$

Thanks to the uniform bounds imposed on θ and u in the definition (7.6), the continuity of the map $\Lambda_2 : \mathcal{U} \mapsto \mathcal{C}^0$, defined at (7.1) is now straightforward.

5. To prove the continuity of the map Λ_1, consider a sequence of functions $I_n \in \mathcal{F}$, with $I_n \to I$ uniformly on $[0,h^+]$. Let $(\theta_n,u_n) : [0,h^+] \mapsto \mathbb{R}^2$ be the corresponding unique optimal solutions.

We claim that $(\theta_n,u_n) \to (\theta,u)$ in $\mathbf{L}^1([0,h^+])$, where (θ,u) is the unique optimal solution, given the light intensity I.

To prove the claim, let (p_n,q_n) be the corresponding solutions of the system (6.20). By the estimates on p',q' proved in Section 6, the functions (p_n,q_n) are equicontinuous. From any subsequence we can thus extract a further subsequence and obtain the convergence

$$p_{n_j} \to \widehat{p}, \qquad q_{n_j} \to \widehat{q}, \qquad I_{n_j} \to I, \tag{7.7}$$

for some functions $\widehat{p},\widehat{q}$, uniformly on $[0,h^+]$.

For every $j \ge 1$ we now have

$$\theta_{n_j}(y) \;=\; \Theta\big(I_{n_j}(y), p_{n_j}(y), q_{n_j}(y)\big), \qquad\qquad u_{n_j}(y) \;=\; U\big(I_{n_j}(y), p_{n_j}(y), q_{n_j}(y)\big),$$

where U and Θ are the functions in (6.8)-(6.9). By the dominated convergence theorem, the convergence (7.7) together with the uniform integrability of θ_{n_j} and u_{n_j} yields the $\mathbf{L}^1$ convergence

$$\|\theta_{n_j} - \widehat{\theta}\|_{\mathbf{L}^1} \;\to\; 0, \qquad\qquad \|u_{n_j} - \widehat{u}\|_{\mathbf{L}^1} \;\to\; 0. \tag{7.8}$$

In turn this implies that $(\widehat{p},\widehat{q})$ provide a solution to the problem (6.20), in connection with the light intensity I. By uniqueness, $\widehat{p} = p$ and $\widehat{q} = q$. Therefore, $\widehat{\theta} = \theta$ and $\widehat{u} = u$ as well.

The above argument shows that, from any subsequence, one can extract a further subsequence so that the $\mathbf{L}^1$-convergence (7.8) holds. Therefore, the entire sequence $(\theta_n,u_n)_{n\ge1}$ converges to (θ,u) in $\mathbf{L}^1([0,h^+])$. This establishes the continuity of the map Λ_1.

6. The map $\Lambda_2 \circ \Lambda_1$ is now a continuous map of the compact, convex domain $\mathcal{F} \subset \mathcal{C}^0([0,h^+])$ into itself. By Schauder's theorem it admits a fixed point $I^*(\cdot)$. By construction, the optimal stem configuration $\big(\theta^{(I^*)}, u^{(I^*)}\big)$ yields a competitive equilibrium, in the sense of Definition 7.1.

7. To prove uniqueness, we derive a set of necessary conditions satisfied by the equilibrium solution, and show that this system has a unique solution.

Using (6.8) and (6.11), we can rewrite the light intensity function (7.1) as

$$I(y) \;=\; \exp\left\{\frac{\rho_0}{\cos\theta_0}\int_y^\infty \ln\left(\frac{q}{I}\right)\frac{1+w\sin\theta_0}{\sin\theta_0+w}\,d\zeta\right\},$$

where $w=w(I,p,q)$ is the function introduced at (6.10). Differentiating w.r.t. y one obtains

$$I'(y) \;=\; -\frac{\rho_0}{\cos\theta_0}\ln\left(\frac{q}{I}\right)\frac{1+w\sin\theta_0}{\sin\theta_0+w}\cdot I \;\doteq\; f_3(I,p,q). \tag{7.9}$$

Combining (7.9) with (6.20), we conclude that the competitive equilibrium satisfies the system of equations and boundary conditions

$$\begin{cases} p'(y) \;=-f_1\big(I(y),p(y),q(y)\big)\cdot f_3(I(y),p(y),q(y)),\\ q'(y) \;=\; f_2\big(I(y),p(y),q(y)\big),\\ I'(y) \;=\; f_3(I(y),p(y),q(y)),\end{cases}\qquad \begin{cases} p(h) \;=\; 0,\\ q(h) \;=\; 1,\\ I(h) \;=\; 1,\end{cases} \tag{7.10}$$

together with

$$q(0) \;=\; 0. \tag{7.11}$$

Here the common height of the stems $h>0$ is a constant to be determined.

8. The uniqueness of solutions to (7.10) will be achieved by a contradiction argument. Since this is very similar to the one used in the proof of Theorem 6.1, we only sketch the main steps.

In analogy with (6.31), (6.33), denote by $p(y,h), q(y,h), I(y,h)$ the unique solution to the Cauchy problem (7.10), with terminal conditions given at $y=h$. Consider the functions

$$P(y)\doteq\frac{\partial p(y,h)}{\partial h},\qquad \widetilde{Q}(y)\doteq\frac{1}{f_2(I,p,q)}\,\frac{\partial q(y,h)}{\partial h},\qquad J(y)\doteq\frac{\partial I(y,h)}{\partial h}.$$

By (7.10), these functions satisfy

$$\begin{cases} P'(y) \;=\; -\big[f_{3,I}f_1+f_3f_{1,I}\big]J-\big[f_{3,p}f_1+f_3f_{1,p}\big]P-\big[f_{3,q}f_1+f_3f_{1,q}\big]f_2\widetilde{Q},\\ \widetilde{Q}'(y) \;=\; \dfrac{f_{2,I}}{f_2}J+\dfrac{f_{2,p}}{f_2}P-\dfrac{f_3}{f_2}\big[f_{2,I}-f_{2,p}f_1\big]\widetilde{Q},\\ J'(y) \;=\; f_{3,I}J+f_{3,p}P+f_{3,q}f_2\widetilde{Q},\end{cases} \tag{7.12}$$

with boundary conditions

$$P(h)=0,\qquad \widetilde{Q}(h)=-1,\qquad J(h)=0.$$

Set $d_0=\frac{\rho_0}{\cos\theta_0}$. Several of the partial derivatives on the right-hand side of (7.12) were computed in (6.36). The remaining ones are

$$f_{1,I}(I,p,q) = \frac{q}{I^2}\cdot\frac{1+w\sin\theta_0}{\sin\theta_0+w} - \frac{\cos^2\theta_0}{(\sin\theta_0+w)^2}\frac{w^2}{p}\Big[1-\frac{q}{I}\Big],$$

$$f_{3,I}(I,p,q) = -d_0\bigg[\Big(\ln\Big(\frac{q}{I}\Big)-1\Big)\frac{1+w\sin\theta_0}{\sin\theta_0+w} - I\ln\Big(\frac{q}{I}\Big)\frac{\cos^2\theta_0}{(\sin\theta_0+w)^2}\frac{w^2}{p}\Big(1-\frac{q}{I}\Big)\bigg],$$

$$f_{3,p}(I,p,q) = d_0 I\ln\Big(\frac{q}{I}\Big)\frac{\cos^2\theta_0}{(\sin\theta_0+w)^2}\frac{w}{p},$$

$$f_{3,q}(I,p,q) = -d_0 I\bigg[\frac{1}{q}\cdot\frac{1+w\sin\theta_0}{\sin\theta_0+w} + \Big[\ln\Big(\frac{q}{I}\Big)\Big]^2\frac{\cos^2\theta_0}{(\sin\theta_0+w)^2}\frac{w^2}{p}\bigg].$$

By the same arguments used in step **8** of the proof of Theorem 6.1, we conclude that the right-hand side of (7.12) is uniformly integrable.

9. Let a density $\rho_0 > 0$ be given. Assume that the problem (7.10)-(7.11) has two distinct solutions $(\hat p, \hat q, \hat I)$ and $(\check p, \check q, \check I)$, defined on $[0, \hat h]$ and $[0, \check h]$ say with $\hat h < \check h$. Since $\hat q(0) = \check q(0) = 0$, by the mean value theorem there exists $k \in [\hat h, \check h]$ such that $\frac{\partial q}{\partial h}(0, k) = 0$.

Next, if multiple solutions exist for arbitrarily small values of the density ρ_0, we can find a decreasing sequence $\rho_{0,n} \downarrow 0$ and corresponding solutions P_n, Q_n, I_n of (7.12), defined for $y \in [0, k_n]$, such that

$$P_n(k_n) = 0, \qquad \widetilde Q_n(k_n) = -1, \qquad J_n(k_n) = 0, \qquad \widetilde Q_n(0) = 0. \tag{7.13}$$

Thanks to the uniform integrability of the right hand sides of (7.12), by possibly extracting a subsequence we can achieve the convergence $k_n \to \bar h \in [0, h^+]$, the weak convergence $P_n' \rightharpoonup P'$, $\widetilde Q_n' \rightharpoonup \widetilde Q'$, $J_n' \rightharpoonup J'$ in $\mathbf{L}^1$, and the strong convergence

$$P_n \to P, \qquad \widetilde Q_n \to \widetilde Q, \qquad J_n \to J,$$

uniformly on every subinterval $[0, h]$ with $h < \bar h$.

To reach a contradiction, we observe that

$$J_n(y) \;=\; -\int_y^{k_n} J_n'(z)\,dz$$

and the right-hand side of J_n' in (7.12) consists of uniformly integrable terms which are multiplied by $\rho_{0,n}$. This implies $J(y) \equiv 0$. This corresponds to the case of an intensity function $I(y) \equiv 1$. But in this case we know that $\widetilde Q(y) \equiv -1$, contradicting the fact that, by (7.13),

$$\widetilde Q(0) \;=\; \lim_{n\to\infty} \widetilde Q_n(0) \;=\; 0. \quad \square$$

8. Stem competition on a domain with boundary

We consider here the same model introduced in Section 2, where all stems have fixed length ℓ and constant thickness κ. But we now allow the sunlight intensity $I = I(x, y)$ to vary w.r.t. both variables x, y. As shown in Fig. 5, left, we denote by

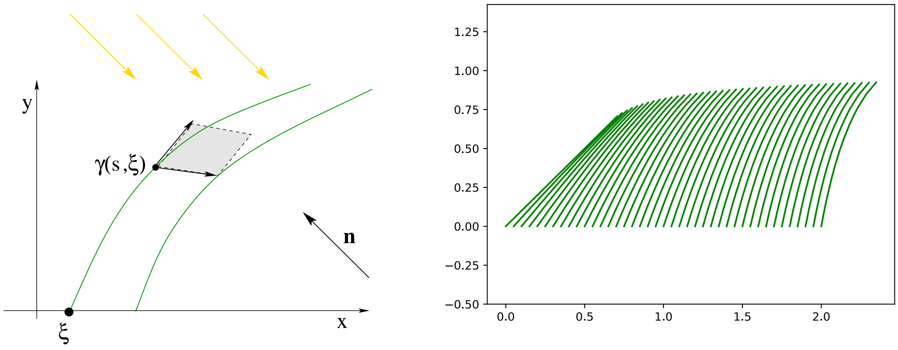


Fig. 5. Left: to leading order, the amount of vegetation in the shaded region is proportional to $\kappa\,\bar\rho(\xi)d\xi ds$. Since the area is computed in terms of the cross product $\frac{\partial\gamma}{\partial\xi}\times\frac{\partial\gamma}{\partial s}$, this motivates the formula (8.4). Right: a possible competitive equilibrium, where the light rays come from the direction $\mathbf{n}=(\frac{-1}{\sqrt2},\frac{1}{\sqrt2})$ and stems are distributed along the positive half line, with density as in (8.9). In this case, stems originating from points close to the origin have no incentive to grow upward, because they already receive a nearly maximum light intensity. Hence they bend to the right, almost perpendicularly to the light rays.

$$s\ \mapsto\ \gamma(s,\xi)\ =\ (x(s),y(s)),\qquad s\in[0,\ell],\tag{8.1}$$

the arc-length parameterization of the stem whose root is located at $(\xi,0)$, and write g for the function introduced at (2.8). This leads to the optimization problem

(OP3) *Given a light intensity function $I=I(x,y)$, find a control $s\mapsto\theta(s)\in[0,\pi]$ which maximizes the integral*

$$\int_0^\ell I(x(s),y(s))\,g(\theta(s))\,ds\tag{8.2}$$

subject to

$$\frac{d}{ds}(x(s),y(s))\ =\ (\cos\theta(s),\sin\theta(s)),\qquad\qquad (x(0),y(0))\ =\ (\xi,0).\tag{8.3}$$

Next, consider a function $\bar\rho(\xi)\geq0$ describing the density of stems which grow near $\xi\in\mathbb{R}$. At any point in space reached by a stem, i.e. such that

$$(x,y)=\gamma(s,\xi)\qquad\text{for some }\ \xi\in\mathbb{R},\ \ s\in[0,\ell],$$

the density of vegetation is

$$\rho(x,y)\ =\ \rho(\gamma(s,\xi))\ =\ \kappa\,\bar\rho(\xi)\cdot\left[\frac{\partial\gamma}{\partial\xi}\times\frac{\partial\gamma}{\partial s}\right]^{-1}.\tag{8.4}$$

The light intensity at a point $P = (x, y) \in \mathbb{R}^2$ is now given by

$$I(P) \;=\; \exp\left\{ -\int_0^{+\infty} \rho(P + t\mathbf{n})\,dt \right\}. \tag{8.5}$$

Definition 8.1. Given the constants ℓ, κ and the density $\bar\rho \in \mathbf{L}^\infty(\mathbb{R})$, we say that the maps $\gamma : [0,\ell] \times \mathbb{R}$ and $I : \mathbb{R} \times \mathbb{R}_+ \mapsto [0,1]$ yield a **competitive equilibrium** if the following holds:

(i) For each $\xi \in \mathbb{R}$, the stem $\gamma(\cdot, \xi)$ provides an optimal solution to **(OP3)**.
(ii) The function $I(\cdot)$ coincides with the light intensity determined by (8.4)-(8.5).

We shall not analyze the existence or uniqueness of the competitive equilibrium, in the case where the distribution of stem roots is not uniform. We only observe that, if the stem $\gamma(\cdot, \xi)$ in (8.1) is optimal, the necessary conditions yield the existence of a dual vector $s \mapsto \mathbf{p}(s)$ satisfying

$$\dot{\mathbf{p}}(s) \;=\; -\nabla I\big(x(s), y(s)\big)\, g(\theta(s)), \qquad\qquad \mathbf{p}(\ell) = (0,0), \tag{8.6}$$

and such that, for a.e. $s \in [0,\ell]$, the optimal angle $\theta^*(s)$ satisfies

$$\theta^*(s) \;=\; \underset{\theta}{\operatorname{argmax}} \Big\{ \mathbf{p}(s)\cdot(\cos\theta, \sin\theta) + I(x(s), y(s))g(\theta) \Big\}. \tag{8.7}$$

Differentiating the expression on the right hand side of (8.7) one obtains an implicit equation for $\theta^*(s)$, namely

$$I\big(x(s), y(s)\big)g'(\theta^*(s)) + \mathbf{p}(s)\cdot\mathbf{n}(s) \;=\; 0 \tag{8.8}$$

for a.e. $s \in [0,\ell]$. Here $\mathbf{n}(s) \doteq \big(-\sin\theta(s), \cos\theta(s)\big)$ is the unit vector perpendicular to the stem. Moreover, by (8.6) one has

$$\mathbf{p}(s) \;=\; \int_s^\ell \nabla I\big(x(\sigma), y(\sigma)\big)\, g\big(\theta^*(\sigma)\big)\, d\sigma.$$

An interesting case is where stems grow only on the half line $\{\xi \geq 0\}$. For example, one can take

$$\bar\rho(\xi) \;=\; \begin{cases} 0 & \text{if } \xi < 0, \\ b^{-1}\xi & \text{if } \xi \in [0,b], \\ 1 & \text{if } \xi > b. \end{cases} \tag{8.9}$$

In this case, we conjecture that the competitive equilibrium has the form illustrated in Fig. 5, right.

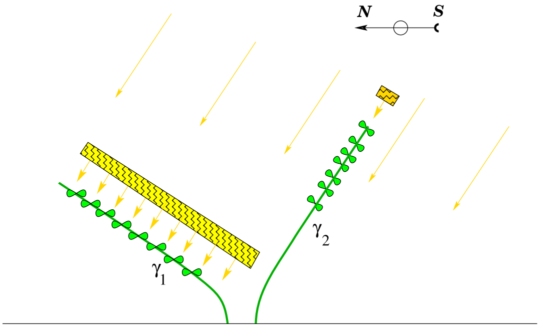


Fig. 6. The stem γ_1, oriented perpendicularly to the sun rays, collects much more sunlight than γ_2. Indeed, γ_1 would give the best orientation for solar panels. Notice that γ_2 minimizes the sunlight gathered because the upper leaves put the lower ones in shade.

9. Concluding remarks

A motivation for the present study was to understand whether competition for sunlight could explain phototropism, i.e. the tendency of plant stems to bend toward the light source. A naive approach may suggest that, if a stem bends in the direction of the light rays, the leaves will be closer to the sun and hence gather more light. However, since the average distance of the earth from the sun is approximately 90 million miles, getting a few inches closer cannot make a difference.

As shown in Fig. 6, if a single stem were present, to maximize the collected sunlight it should be perpendicular to the light rays, not parallel. In the presence of competition among several plant stems, our analysis shows that the best configuration is no longer perpendicular to light rays: the lower part of the stems should grow in a nearly vertical direction, while the upper part bends away from the sun.

Still, our competition models do not predict the tilting of stems in the direction of the sun rays. This may be due to the fact that these models are “static”, i.e., they do not describe how plants grow in time. This leaves open the possibility of introducing further models that can explain phototropism in a time-dependent framework. As suggested in [12], the preemptive conquering of space, in the direction of the light rays, can be an advantageous strategy. We leave these issues for future investigation.

Acknowledgments

The research of A. Bressan was partially supported by NSF, with grant DMS-1714237, “Models of controlled biological growth”. S.T. Galtung was supported in part by a grant from the U.S.-Norway Fulbright Foundation. A. Reigstad was supported by the grant “Waves and Nonlinear Phenomena” (250070) from the Research Council of Norway. S.T. Galtung and A. Reigstad are very grateful to the Department of Mathematics at Penn State University for the generous hospitality during the academic year 2018/2019.

References

[1] F. Ancona, A. Bressan, O. Glass, W. Shen, Feedback stabilization of stem growth, J. Dyn. Differ. Equ. 31 (2019) 1079–1106.

[2] M. Bernot, V. Caselles, J.M. Morel, Optimal Transportation Networks. Models and Theory, Springer Lecture Notes in Mathematics, vol. 1955, Berlin, 2009.

[3] A. Bressan, M. Palladino, Q. Sun, Variational problems for tree roots and branches, Calc. Var. Partial Differ. Equ. 59 (2020) 7.

[4] A. Bressan, B. Piccoli, Introduction to the Mathematical Theory of Control, AIMS Series in Applied Mathematics, AIMS, Springfield, MO, 2007.

[5] A. Bressan, Q. Sun, On the optimal shape of tree roots and branches, Math. Models Methods Appl. Sci. 28 (2018) 2763–2801.

[6] L. Cesari, Optimization - Theory and Applications, Springer-Verlag, 1983.

[7] W.H. Fleming, R.W. Rishel, Deterministic and Stochastic Optimal Control, Springer, 1975.

[8] E.N. Gilbert, Minimum cost communication networks, Bell Syst. Tech. J. 46 (1967) 2209–2227.

[9] O. Leyser, S. Day, Mechanisms in Plant Development, Blackwell Publishing, 2003.

[10] E. Lieb, M. Loss, Analysis, second edition, American Mathematical Society, Providence, 2001.

[11] F. Maddalena, J.M. Morel, S. Solimini, A variational model of irrigation patterns, Interfaces Free Bound. 5 (2003) 391–415.

[12] A. Runions, B. Lane, P. Prusinkiewicz, Modeling trees with a space colonization algorithm, in: Eurographics Workshop on Natural Phenomena, 2007.

[13] Q. Xia, Optimal paths related to transport problems, Commun. Contemp. Math. 5 (2003) 251–279.

[14] Q. Xia, Motivations, ideas and applications of ramified optimal transportation, ESAIM Math. Model. Numer. Anal. 49 (2015) 1791–1832.

Appendix 4.A Details on Paper 4

In the following we will provide some additional details to motivate and explain some results in Paper 4 ([2]) which for publication reasons had to be rather brief, leaving out several intermediate steps. We assume below that the reader has some familiarity with the papers [3, 2], and we will frequently refer to equations within these papers.

4.A.1 Details on Section 2 in Paper 4 [2]

Equation (2.3) in [2] gives the impression of being a straightforward result from [3], but this derivation is somewhat involved. The equation can be derived from Equations (2.22) and (2.23) in [3], which expresses the amount of sunlight absorbed by the measure μ in the presence of competing vegetation in form of a measure ν absolutely continuous with respect to the Lebesgue measure.

Moreover, the aforementioned (2.22) is simply presented as is, motivated by the less general equation (2.20) where both μ and ν are absolutely continuous with respect to the Lebesgue measure. We will first give a formal argument for (2.22) being the generalization of (2.20), before using (2.22) to justify Equation (2.3) in [2].

Following the notation in [3] we let $\mathbf{n}$ be the unit vector facing the incoming sunlight, $\pi_{\mathbf{n}}$ be the projection parallel to the light onto the perpendicular subspace $E_{\mathbf{n}}^{\perp}$ of dimension $d-1$, and $\mu^{\mathbf{n}}$ denote the projected measure

$$\mu^{\mathbf{n}}(A) = \mu\left(\{x \in \mathbb{R}^d \,;\, \pi_{\mathbf{n}}(x) \in A\}\right)$$

for every open set $A \in E_{\mathbf{n}}^{\perp}$.

Generalizations to non-absolutely continuous measures

In [3], they generalize the expression for the sunlight absorbed by a measure μ absolutely continuous with respect to the Lebesgue measure m_d in $\mathbb{R}^d$, which we denote by $\mu \ll m_d$, and for which the Radon–Nikodym derivative is $d\mu/dm_d = f$. To be precise, they take Equation (2.3) which reads

$$\mathcal{S}^{\mathbf{n}}(\mu) = \int_{E_{\mathbf{n}}^{\perp}} \left(1 - \exp\left\{-\int_{-\infty}^{\infty} f(y + t\mathbf{n})dt\right\}\right) dy$$

and generalize it to

$$\mathcal{S}^{\mathbf{n}}(\mu) = \int_{E_{\mathbf{n}}^{\perp}} \left(1 - \exp\left\{-\Phi^{\mathbf{n}}(y)\right\}\right) dy, \tag{4.A.1}$$

for μ with projection $\mu^{\mathbf{n}} \ll m_{d-1}$ onto $E_{\mathbf{n}}^{\perp}$, where m_{d-1} is the Lebesgue measure on $E_{\mathbf{n}}^{\perp}$ and $d\mu^{\mathbf{n}}/dm_{d-1} = \Phi^{\mathbf{n}}$. This generalization is simply claimed to be an easy generalization, and so we give a formal derivation here for the reader's benefit. Starting with $\mu^{\mathbf{n}}(A)$ we find

$$\mu^{\mathbf{n}}(A) = \mu(\pi_{\mathbf{n}}^{-1}(A)) \overset{\mu \ll m_d}{=} \int_{\mathbb{R}^d} \mathbf{1}_{\pi_{\mathbf{n}}^{-1}(A)} f \, dm_d = \int_{\mathbb{R}^d} \mathbf{1}_{A\times\mathbb{R}} f \, dm_d.$$

By the Fubini–Tonelli theorem, the rightmost expression above is equal to

$$\int_{E_{\mathbf{n}}^{\perp}} \mathbf{1}_A \int_{-\infty}^{\infty} f(y + t\mathbf{n}) \, dt \, dm_{d-1}(y) = \int_A \left(\int_{-\infty}^{\infty} f(y+t\mathbf{n}) \, dt \right) dm_{d-1}(y).$$

Next, in the expression for the function $\mathcal{S}_{\mu,\nu}^{\mathbf{n}}$ in (2.22) they generalize

$$\int_s^{\infty} f(y+t\mathbf{n}) \, dt$$

to $\Phi^{\mathbf{n}}(z)\mu^y([s,\infty))$. The explanation follows. First, observe that by the procedure above we may write

$$\begin{aligned}
\mu(A \times [s,\infty)) &= \int_{\mathbb{R}^d} \mathbf{1}_{A\times[s,\infty)} f \, dm_d \\
&= \int_{E_{\mathbf{n}}^{\perp}} \mathbf{1}_A \int_{-\infty}^{\infty} \mathbf{1}_{[s,\infty)} f(y+t\mathbf{n}) \, dt \, dm_{d-1}(y) \\
&= \int_A \left(\int_s^{\infty} f(y+t\mathbf{n}) \, dt \right) dm_{d-1}(y).
\end{aligned}$$

By the disintegration theorem, cf. [1, Thm. 2.28], we also have

$$\begin{aligned}
\mu(A \times [s,\infty)) &= \int_{\mathbb{R}^d} \mathbf{1}_{A\times[s,\infty)} \, d\mu \\
&= \int_{E_{\mathbf{n}}^{\perp}} \mathbf{1}_A \int_{-\infty}^{\infty} \mathbf{1}_{[s,\infty)} \, d\mu^y(t) \, d\mu^{\mathbf{n}}(y) \\
&= \int_A \mu^y\left([s,\infty)\right) \, d\mu^{\mathbf{n}}(y) \\
&= \int_A \mu^y\left([s,\infty)\right) \Phi^{\mathbf{n}}(y) \, dm_{d-1}(y).
\end{aligned}$$

Combining the previous expressions we obtain

$$\int_s^{\infty} f(z+t\mathbf{n}) \, dt = \Phi^{\mathbf{n}}(z)\mu^z\left([s,\infty)\right) \tag{4.A.2}$$

for m_{d-1}-a.e. $z \in E_{\mathbf{n}}^{\perp}$. As μ^z is a probability measure on the real line it follows that

$$\int_{-\infty}^{+\infty} f(z+t\mathbf{n})\,dt = \Phi^{\mathbf{n}}(z), \tag{4.A.3}$$

hence we obtain the generalization (4.A.1). Moreover, as (2.22) in [3] comes from (2.20) by replacing the left-hand sides of (4.A.2) and (4.A.3) with their respective right-hand sides, we have justified this generalization as well.

Equation for sunlight collected by single stem

Here we motivate Equation (2.3) in [2], the expression for sunlight absorbed by a single stem with background vegetation in the form of a light intensity function I. Let us assume that for m_1-a.e. $z \in E_{\mathbf{n}}^{\perp}$, the stem does not cross $E_{\mathbf{n}}(z) := \{z + s\mathbf{n} \,|\, s \in \mathbb{R}\}$ more than once. Indeed, to do otherwise would be suboptimal, as the stem would put itself in shade. For z with a single value s defining the non-empty intersection between the stem and $E_{\mathbf{n}}(z)$, we call this value $\sigma(z)$. Since μ^z is a probability measure which lives on the support of μ we must have $\mu^z = \delta_{\sigma(z)}$ for such z, and it follows that $\mu^z([s,\infty[) = \mathbf{1}_{(-\infty,\sigma(z)]}(s)$. Inserting this into the last term of $\mathcal{S}_{\mu,\nu}(z)$ as defined in [3, Eq. (2.22)] we obtain

$$\begin{aligned}
&\int_{-\infty}^{\infty} g(z+s\mathbf{n}) \exp\left\{-\int_s^{\infty} g(z+t\mathbf{n})\,dt\right\} \exp\left\{-\Phi^{\mathbf{n}}(z)\mu^z\left([s,\infty[\right)\right\} ds \\
&\quad= \exp\left\{-\Phi^{\mathbf{n}}(z)\right\} \int_{-\infty}^{\sigma(z)} g(z+s\mathbf{n}) \exp\left\{-\int_s^{\infty} g(z+t\mathbf{n})\,dt\right\} ds \\
&\qquad+ \int_{\sigma(z)}^{\infty} g(z+s\mathbf{n}) \exp\left\{-\int_s^{\infty} g(z+t\mathbf{n})\,dt\right\} ds \\
&\quad= \exp\left\{-\Phi^{\mathbf{n}}(z)\right\} \\
&\qquad\times \left(\exp\left\{-\int_{\sigma(z)}^{\infty} g(z+s\mathbf{n})\,ds\right\} - \exp\left\{-\int_{-\infty}^{\infty} g(z+s\mathbf{n})\,ds\right\}\right) \\
&\qquad+ 1 - \exp\left\{-\int_{\sigma(z)}^{\infty} g(z+s\mathbf{n})\,ds\right\}.
\end{aligned}$$

Using the above identity, the evaluation of $\mathcal{S}_{\mu,\nu}(z)$ yields two cancellations and the expression

$$\mathcal{S}_{\mu,\nu}(z) = \exp\left\{-\int_{\sigma(z)}^{\infty} g(z+s\mathbf{n})\,ds\right\} \left(1 - \exp\left\{-\Phi^{\mathbf{n}}(z)\right\}\right).$$

This must hold for a.e. $z \in E_{\mathbf{n}}^{\perp}$, since those z where the intersection of the stem and $E_{\mathbf{n}}(z)$ is a non-empty interval, i.e., the stem is parallel to $\mathbf{n}$, can at most be countable. It then follows that the total amount of sunlight absorbed by the stem is

$$\mathcal{S}_{\mu,\nu} = \int_{E_{\mathbf{n}}^{\perp}} \exp\left\{-\int_{\sigma(z)}^{\infty} g(z+s\mathbf{n})\,ds\right\}(1-\exp\{-\Phi^{\mathbf{n}}(z)\})\,dz$$

which is exactly the expression

$$\mathcal{S}(y) = \int_{E_{\mathbf{n}}^{\perp}} I(z)\,(1-\exp\{-\Phi(z)\})\,dz$$

from (2.3) in [2], where we identify the light intensity function I with the exponential function involving the density $g = d\nu/dm_d$,

$$I(z) \doteq \exp\left\{-\int_{\sigma(z)}^{\infty} g(z+s\mathbf{n})\,ds\right\}.$$

For the final expression, note that the projection of an infinitesimal segment $d\gamma(s)$ of the stem onto $E_{\mathbf{n}}^{\perp}$ equals $(\dot{\gamma}(s)\cdot\mathbf{n}^{\perp})\,ds$ to leading order. As $|dz|$ is the length of this projection it follows that

$$\left|\frac{ds}{dz}\right| = \left|\dot{\gamma}(s)\cdot\mathbf{n}^{\perp}\right| = \left|\cos(\theta(s)-\theta_0)\right|,$$

which leads to

$$\Phi(z(s)) = \kappa\left|\frac{dz}{ds}\right| = \frac{\kappa}{|\cos(\theta(s)-\theta_0)|}.$$

Finally, because of the assumption $\theta_0 \leq \theta(s) \leq \frac{\pi}{2}$ justified below, we may remove the absolute value to recover exactly (2.3) in [2].

Reflection argument

A reflection argument is presented below Equation (2.4) to prove that it is not restrictive to assume $\theta_0 \leq \theta(s) \leq \frac{\pi}{2}$. The argument with the m-fold composition coming from (2.5) and (2.6) contains a lot written between the lines, and we will give a more detailed exposition here.

The stem is above ground by definition, and so it should only be necessary to include the first and third cases in (2.5). If one is in the third case, the modified angle is a reflection about $\mathbf{n}$ which does not change the contribution from Φ, but such a reflection will elevate part of the stem and since I increases with height this increases the sunlight captured.

For the next step, we see that we are done if the new angle is mapped to the interval $[\theta_0, \pi/2]$. Observe that if $0 \leq \theta < \theta_0$ then the mapping $2\theta_0 - \theta$ is a reflection about $\mathbf{n}^\perp$, which does not change the value of $|\cos(\theta(s) - \theta_0)|$, but it will increase the contribution from I as part of the stem is lifted.

If $\pi/2 < \theta \leq \theta_0 + \pi/2$ then the mapping $\pi - \theta$ is a reflection about the y-axis, which does not change the elevation of the stem, but it decreases the angle between $\dot{\gamma}$ and $\mathbf{n}^\perp$, which increases the value of $|\cos(\theta(s) - \theta_0)|$. Since $(1 - e^{-1/x})x$ is monotone increasing in x, it follows that the amount of absorbed sunlight increases.

To see that this iteration must end after finitely many steps, observe that if we are in $[0, \theta_0)$ then the mapping $\theta \mapsto 2\theta_0 - \theta$ sends us to $(\theta_0, 2\theta_0] \subsetneq (\theta_0, \theta_0 + \pi/2]$. Likewise, if we are in $(\pi/2, \theta_0 + \pi/2]$, the mapping $\theta \mapsto \pi - \theta$ maps us to $[\pi/2 - \theta_0, \pi/2) \subsetneq [0, \pi/2)$. Let us assume that we start with angle $\phi_0 \in [0, \theta_0)$, and that we find the next iterate $\phi_1 \in (\pi/2, 2\theta_0]$. Then it follows

$$\phi_2 = \pi - \phi_1 = \pi - (2\theta_0 - \phi_0) = \phi_0 + 2(\pi/2 - \theta_0) > \phi_0.$$

This holds in general: if some angle $\phi_n \in [0, \theta_0)$ is not mapped into $[\theta_0, \pi/2]$ in the next iteration, we have $\phi_{n+2} = \phi_n + 2(\pi/2 - \theta_0) < \pi/2$. Similarly, if some angle $\phi_{n+1} \in (\pi/2, \theta_0 + \pi/2]$ is not mapped into $[\theta_0, \pi/2]$ in the next iteration we have $\phi_{n+3} = \phi_{n+1} - 2(\pi/2 - \theta_0) > \theta_0$. Since $\theta_0 < \pi/2$, there must be some finite m for which $\phi_m \in [\theta_0, \pi/2]$.

4.A.2 Details on Section 5 in Paper 4 [2]

Here we recall the optimization problem **(OP2)**. First we define the functions

$$z(t) := \int_t^{+\infty} u(s)\, ds, \quad G(\theta, u) := \left(1 - \exp\left\{\frac{-u}{\cos(\theta - \theta_0)}\right\}\right) \cos(\theta - \theta_0).$$

Then **(OP2)** is defined as follows. Given a sunlight intensity function $I(y)$, and constants $0 < \alpha < 1$ and $c > 0$, find controls $\theta : \mathbb{R}_+ \to [\theta_0, \frac{\pi}{2}]$ and $u : \mathbb{R}_+ \to \mathbb{R}_+$ which maximize the integral

$$\int_0^{+\infty} [I(y(t)) G(\theta(t), u(t)) - c z^\alpha(t)]\, dt \tag{4.A.4}$$

subject to

$$\begin{aligned} \dot{y}(t) &= \sin\theta(t), \qquad y(0) = 0, \\ \dot{z}(t) &= -u(t), \qquad z(+\infty) = 0. \end{aligned} \tag{4.A.5}$$

In Section 5.2 it is simply stated that the necessary conditions for optimality lead to the dual equations

$$\begin{aligned} \dot{p}(t) &= -I'(y(t))G(\theta(t), u(t)), & p(+\infty) &= 0, \\ \dot{q}(t) &= c\alpha z^{\alpha-1}(t), & q(0) &= 0, \end{aligned} \tag{4.A.6}$$

and the maximality condition

$$(\theta^*(t), u^*(t)) = \underset{\theta\in[0,\pi], u\geq 0}{\arg\max} \left\{ p(t)\sin\theta - q(t)u + I(y(t))G(\theta, u) - cz^{\alpha}(t) \right\}. \tag{4.A.7}$$

This is in fact an application of the Pontryagin maximum principle, and we will in the following provide details on how (4.A.6) and (4.A.7) are derived. First we rewrite **(OP2)** as the Mayer problem (0.1.11)–(0.1.14). To this end we introduce the auxiliary variable $w(t)$ satisfying

$$\dot{w}(t) = I(y(t))G(\theta(t), u(t)) - cz^{\alpha}(t), \qquad w(0) = 0,$$

for which we want to maximize $\phi_0(+\infty, (y, z, w)) = w(+\infty)$ subject to (4.A.5). Note that this Mayer problem does not correspond exactly to the one considered in Theorem 0.1.1, as one does not give an initial condition for the primal variable z, but only a terminal condition $z(+\infty) = 0$. We shall indicate below how this affects the derivation.

For this maximization of $w(+\infty)$, let (θ^*, u^*) be a pair of optimal controls with corresponding optimal trajectory (y^*, z^*, w^*). By virtue of (0.1.15) we have dual variables p, q, and r satisfying the evolution equation

$$\begin{aligned} \begin{bmatrix} \dot{p} & \dot{q} & \dot{r} \end{bmatrix} &= -\begin{bmatrix} p & q & r \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ I'(y)G(\theta, u) & -c\alpha z^{\alpha-1} & 0 \end{bmatrix} \\ &= \begin{bmatrix} -rI'(y)G(\theta, u) & rc\alpha z^{\alpha-1} & 0 \end{bmatrix}, \end{aligned} \tag{4.A.8}$$

where the matrix is the Jacobian corresponding to $D_x f$ in (0.1.15). For the terminal conditions of (p, q, r), we see that multiplying λ_i for $i = 0, \ldots, k$ with the same positive constant we may assume $\lambda_0 = 1$. Moreover, we think of the endpoint condition $z(+\infty) = 0$ as a constraint in (0.1.14) by setting $\phi_1((y, z, w)) = z$. Then it follows from (0.1.17) that $(p, q, r)(+\infty) = (0, \lambda_1, 1)$, where λ_1 is arbitrary. In consequence, we have $p(+\infty) = 0$ and $r(+\infty) = 1$, while $q(+\infty)$ is free. The dual variable q will instead be specified at the initial time $t = 0$, according to the Pontryagin principle presented in [5, Chap. 2], which is more general in the sense that the primal variables may be specified at either the initial or terminal time. Indeed, replacing the initial conditions $x(0) = \bar{x}$ in (0.1.12) by a set of initial constraints analogous to (0.1.14), there is an expression analogous to (0.1.17) which specifies the dual variables at $t = 0$. Since in our case we may regard $y(0) = 0$ and $w(0) = 0$ as initial constraints $\phi_i((y, z, w)) = 0$ for $\phi_2((y, z, w)) = y$ and $\phi_3((y, z, w)) = w$, this leads to $(p, q, r)(0) = (\lambda_2, 0, \lambda_3)$ where λ_2 and λ_3 are arbitrary. Hence, $q(0) = 0$, while $p(0)$ and $r(0)$ are free. Combining the initial and terminal conditions of (p, q, r) with (4.A.8) we obtain $r \equiv 1$, and consequently (4.A.6).

4.A.3 Details on Section 6 in Paper 4 [2]

Equation (6.32) comes from the differential equation satisfied by the derivative of a solution with respect to a parameter, details can be found in, e.g., [6]. In step 9 of the proof of Theorem 6.1 it is stated that due to uniform integrability one has weakly convergent subsequence in $\mathbf{L}^1$. This is can be seen as an application of the Dunford–Pettis theorem, cf. [4, Thm. 4.30].

Bibliography

[1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[2] A. Bressan, S. T. Galtung, A. Reigstad, and J. Ridder. Competition models for plant stems. *J. Differential Equations*, 269(2):1571–1611, 2020.

[3] A. Bressan and Q. Sun. On the optimal shape of tree roots and branches. *Math. Models Methods Appl. Sci.*, 28(14):2763–2801, 2018.

[4] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, New York, 2011.

[5] W. H. Fleming and R. W. Rishel. *Deterministic and stochastic optimal control.* Springer-Verlag, Berlin-New York, 1975. Applications of Mathematics, No. 1.

[6] T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Ann. of Math. (2)*, 20(4):292–296, 1919.

Paper 5

On a shape optimization problem for tree branches

Alberto Bressan and Sondre Tesdal Galtung

Submitted for publication

accepted for publication in Networks and Heterogeneous Media (NHM)
https://www.aimsciences.org/journal/1556-1801

On a Shape Optimization Problem for Tree Branches

Alberto Bressan(*) and Sondre T. Galtung(**)

(*) Department of Mathematics, Penn State University
University Park, Pa. 16802, USA.

(**) Department of Mathematical Sciences,
NTNU – Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

e-mails: axb62@psu.edu, sondre.galtung@ntnu.no.

Abstract

This paper is concerned with a shape optimization problem, where the functional to be maximized describes the total sunlight collected by a distribution of tree leaves, minus the cost for transporting water and nutrient from the base of the trunk to all the leaves. In the case of 2 space dimensions, the solution is proved to be unique, and explicitly determined.

Keywords: shape optimization, sunlight functional, branched transport.

MSC: 49Q10, 49Q20.

1 Introduction

In the recent papers [7, 9] two functionals were introduced, measuring the amount of light collected by the leaves, and the amount of water and nutrients collected by the roots of a tree. In connection with a ramified transportation cost [1, 14, 18], these lead to various optimization problems for tree shapes.

Quite often, optimal solutions to problems involving a ramified transportation cost exhibit a fractal structure [2, 3, 4, 12, 15, 16, 17]. In the present note we analyze in more detail the optimization problem for tree branches proposed in [7], in the 2-dimensional case. In this simple setting, the unique solution can be explicitly determined. Instead of being fractal, its shape reminds of a solar panel.

The present analysis was partially motivated by the goal of understanding phototropism, i.e., the tendency of plant stems to bend toward the source of light. Our results indicate that this behavior cannot be explained purely in terms of maximizing the amount of light collected by the leaves (Fig. 1). Apparently, other factors must have played a role in the evolution of this trait, such as the competition among different plants. See [6] for some results in this direction.

The remainder of this paper is organized as follows. In Section 2 we review the two functionals defining the shape optimization problem and state the main results. Proofs are then worked out in Sections 3 to 5.

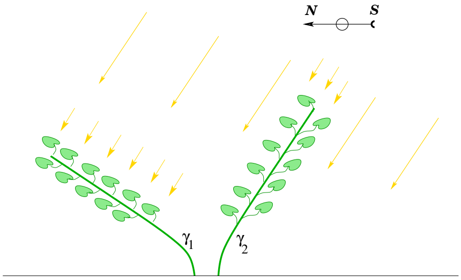


Figure 1: A stem γ_1 perpendicular to the sun rays is optimally shaped to collect the most light. For the stem γ_2 bending toward the light source, the upper leaves put the lower ones in shade.

2 Statement of the main results

We begin by reviewing the two functionals considered in [7, 9].

2.1 A sunlight functional

Let μ be a positive, bounded Radon measure on $\mathbb{R}^d_+ \doteq \{(x_1, x_2, \ldots, x_d)\,;\; x_d \geq 0\}$. Thinking of μ as the density of leaves on a tree, we seek a functional $\mathcal{S}(\mu)$ describing the total amount of sunlight absorbed by the leaves. Fix a unit vector

$$\mathbf{n} \;\in\; S^{d-1} \;\doteq\; \{x \in \mathbb{R}^d\,;\;\; |x| = 1\},$$

and assume that all light rays come parallel to $\mathbf{n}$. Call $E^\perp_\mathbf{n}$ the $(d-1)$-dimensional subspace perpendicular to $\mathbf{n}$ and let $\pi_\mathbf{n} : \mathbb{R}^d \mapsto E^\perp_\mathbf{n}$ be the perpendicular projection. Each point $\mathbf{x} \in \mathbb{R}^d$ can thus be expressed uniquely as

$$\mathbf{x} \;=\; \mathbf{y} + s\mathbf{n} \tag{2.1}$$

with $\mathbf{y} \in E^\perp_\mathbf{n}$ and $s \in \mathbb{R}$.

On the perpendicular subspace $E^\perp_\mathbf{n}$ consider the projected measure $\mu^\mathbf{n}$, defined by setting

$$\mu^\mathbf{n}(A) \;=\; \mu\Big(\big\{x \in \mathbb{R}^d\,;\;\; \pi_\mathbf{n}(x) \in A\big\}\Big). \tag{2.2}$$

Call $\Phi^\mathbf{n}$ the density of the absolutely continuous part of $\mu^\mathbf{n}$ w.r.t. the $(d-1)$-dimensional Lebesgue measure on $E^\perp_\mathbf{n}$.

Definition 2.1 *The total amount of sunlight from the direction* $\mathbf{n}$ *captured by a measure* μ *on* $\mathbb{R}^d$ *is defined as*

$$\mathcal{S}^\mathbf{n}(\mu) \;\doteq\; \int_{E^\perp_\mathbf{n}} \Big(1 - \exp\big\{-\Phi^\mathbf{n}(y)\big\}\Big)\,dy\,. \tag{2.3}$$

More generally, given an integrable function $\eta \in \mathbf{L}^1(S^{d-1})$*, the total sunshine absorbed by* μ *from all directions is defined as*

$$\mathcal{S}^\eta(\mu) \;\doteq\; \int_{S^{d-1}} \left(\int_{E^\perp_\mathbf{n}} \Big(1 - \exp\big\{-\Phi^\mathbf{n}(y)\big\}\Big)\,dy\right)\eta(\mathbf{n})\,d\mathbf{n}\,. \tag{2.4}$$

In the formula (2.4), $\eta(\mathbf{n})$ accounts for the intensity of light coming from the direction $\mathbf{n}$.

Remark 2.2 According to the above definition, the amount of sunlight $\mathcal{S}^{\mathbf{n}}(\mu)$ captured by the measure μ only depends on its projection $\mu^{\mathbf{n}}$ on the subspace perpendicular to $\mathbf{n}$. In particular, if a second measure $\widetilde{\mu}$ is obtained from μ by shifting some of the mass in a direction parallel to $\mathbf{n}$, then $\mathcal{S}(\widetilde{\mu}) = \mathcal{S}(\mu)$.

2.2 Optimal irrigation patterns

Consider a positive Radon measure μ on $\mathbb{R}^d$ with total mass $M = \mu(\mathbb{R}^d)$, and let $\Theta = [0, M]$. We think of $\xi \in \Theta$ as a Lagrangian variable, labeling a water particle.

Definition 2.3 *A measurable map*

$$\chi : \Theta \times \mathbb{R}_+ \ \mapsto \ \mathbb{R}^d \tag{2.5}$$

is called an **admissible irrigation plan** *if*

(i) For every $\xi \in \Theta$, the map $t \mapsto \chi(\xi, t)$ is Lipschitz continuous. More precisely, for each ξ there exists a stopping time $T(\xi)$ such that, calling

$$\dot{\chi}(\xi, t) \ = \ \frac{\partial}{\partial t}\, \chi(\xi, t)$$

the partial derivative w.r.t. time, one has

$$\big|\dot{\chi}(\xi, t)\big| \ = \ \begin{cases} 1 & \text{for a.e. } t \in \big[0, T(\xi)\big], \\ 0 & \text{for } t > T(\xi). \end{cases} \tag{2.6}$$

(ii) At time $t = 0$ all particles are at the origin: $\chi(\xi, 0) = \mathbf{0}$ for all $\xi \in \Theta$.

(iii) The push-forward of the Lebesgue measure on $[0, M]$ through the map $\xi \mapsto \chi(\xi, T(\xi))$ coincides with the measure μ. In other words, for every open set $A \subset \mathbb{R}^d$ there holds

$$\mu(A) \ = \ \text{meas}\Big(\{\xi \in \Theta\,; \quad \chi(\xi, T(\xi)) \in A\}\Big). \tag{2.7}$$

One may think of $\chi(\xi, t)$ as the position of the water particle ξ at time t.

To define the corresponding transportation cost, we first compute how many particles travel through a point $x \in \mathbb{R}^d$. This is described by

$$|x|_\chi \ \doteq \ \text{meas}\Big(\big\{\xi \in \Theta\,; \quad \chi(\xi, t) = x \quad \text{for some} \quad t \geq 0\big\}\Big). \tag{2.8}$$

We think of $|x|_\chi$ as the *total flux going through the point x*. Following [13, 14], we consider

Definition 2.4 (irrigation cost). *For a given $\alpha \in [0, 1]$, the total cost of the irrigation plan χ is*

$$\mathcal{E}^\alpha(\chi) \ \doteq \ \int_\Theta \left(\int_0^{T(\xi)} \big|\chi(\xi, t)\big|_\chi^{\alpha - 1}\, dt \right) d\xi. \tag{2.9}$$

The α-**irrigation cost** *of a measure μ is defined as*

$$\mathcal{I}^\alpha(\mu) \ \doteq \ \inf_\chi \mathcal{E}^\alpha(\chi), \tag{2.10}$$

where the infimum is taken over all admissible irrigation plans for the measure μ.

Remark 2.5 Sometimes it is convenient to consider more general irrigation plans where, in place of (2.6), for a.e. $t \in [0, T(\xi)]$ the speed satisfies $|\dot\chi(\xi,t)| \leq 1$. In this case, the cost (2.9) is replaced by

$$\mathcal{E}^\alpha(\chi) \doteq \int_\Theta \left(\int_0^{T(\xi)} |\chi(\xi,t)|_\chi^{\alpha-1} \, |\dot\chi(\xi,t)| \, dt \right) d\xi. \tag{2.11}$$

Of course, one can always re-parameterize each trajectory $t \mapsto \chi(\xi,t)$ by arc-length, so that (2.6) holds. This does not affect the cost (2.11).

Remark 2.6 In the case $\alpha = 1$, the expression (2.9) reduces to

$$\mathcal{E}^\alpha(\chi) \doteq \int_\Theta \left(\int_{\mathbb{R}_+} |\dot\chi_t(\xi,t)| \, dt \right) d\xi = \int_\Theta [\text{total length of the path } \chi(\xi,\cdot)] \, d\xi \,.$$

Of course, this length is minimal if every path $\chi(\cdot,\xi)$ is a straight line, joining the origin with $\chi(\xi, T(\xi))$. Hence

$$\mathcal{I}^\alpha(\mu) \doteq \inf_\chi \mathcal{E}^\alpha(\chi) = \int_\Theta |\chi(\xi, T(\xi))| \, d\xi = \int |x| \, d\mu \,.$$

On the other hand, when $\alpha < 1$, moving along a path which is traveled by few other particles comes at a high cost. Indeed, in this case the factor $\big|\chi(\xi,t)\big|_\chi^{\alpha-1}$ becomes large. To reduce the total cost, it is thus convenient that many particles travel along the same path.

For the basic theory of ramified transport we refer to the monograph [1]. For future use, we recall that optimal irrigation plans satisfy

Single Path Property: *If* $\chi(\xi,\tau) = \chi(\xi',\tau')$ *for some* $\xi, \xi' \in \Theta$ *and* $0 < \tau \leq \tau'$*, then*

$$\chi(\xi,t) = \chi(\xi',t) \qquad \textit{for all } t \in [0,\tau]. \tag{2.12}$$

2.3 The general optimization problem for branches.

Combining the two functionals (2.4) and (2.10), one can formulate an optimization problem for the shape of branches:

(OPB) Given a light intensity function $\eta \in \mathbf{L}^1(S^{d-1})$ and two constants $c > 0$, $\alpha \in [0,1]$, find a positive measure μ supported on R^d_+ that maximizes the payoff

$$\mathcal{S}^\eta(\mu) - c\,\mathcal{I}^\alpha(\mu). \tag{2.13}$$

2.4 Optimal branches in dimension $d = 2$.

We consider here the optimization problem for branches in the planar case $d = 2$. We assume that the sunlight comes from a single direction $\mathbf{n} = (\cos\theta_0, \sin\theta_0)$, so that the sunlight functional takes the form (2.3). Moreover, as irrigation cost we take (2.10), for some fixed $\alpha \in]0,1]$. For a given constant $c > 0$, this leads to the problem

$$\text{maximize:} \quad \mathcal{S}^{\mathbf{n}}(\mu) - c\mathcal{I}^{\alpha}(\mu), \tag{2.14}$$

over all positive measures μ supported on the half space $\mathbb{R}^2_+ \doteq \{x = (x_1, x_2)\,;\;\; x_2 \geq 0\}$. To fix the ideas, we shall assume that $0 < \theta_0 < \pi/2$. Our main goal is to prove that for this problem the "solar panel" configuration shown in Fig. 2 is optimal, namely:

Theorem 2.7 *Assume that $0 < \theta_0 \leq \pi/2$ and $1/2 \leq \alpha \leq 1$.Then the optimization problem (2.14) has a unique solution. The optimal measure is supported along two rays, namely*

$$\mathrm{Supp}(\mu) \;\subset\; \Big\{(r\cos\theta, r\sin\theta)\,;\quad r \geq 0, \quad \textit{either } \theta = 0 \textit{ or } \theta = \theta_0 + \frac{\pi}{2}\Big\} \;\doteq\; \Gamma_0 \cup \Gamma_1\,. \tag{2.15}$$

When $0 < \alpha < 1/2$, the same conclusion holds provided that the angle θ_0 satisfies

$$\cos\left(\frac{\pi}{2} - \theta_0\right) \;\geq\; 1 - 2^{2\alpha - 1}. \tag{2.16}$$

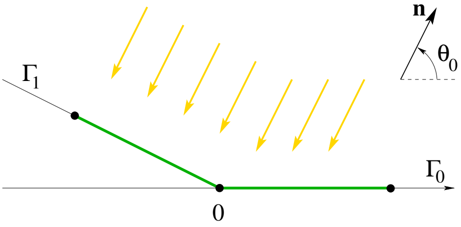


Figure 2: When the light rays impinge from a fixed direction $\mathbf{n}$, the optimal distribution of leaves is supported on the two rays Γ_0 and Γ_1.

In the case $\alpha = 1$ the result is straightforward. Indeed, for any measure μ we can consider its projection $\widetilde{\mu}$ on $\Gamma_0 \cup \Gamma_1$, obtained by shifting the mass in the direction parallel to the vector $\mathbf{n}$. In other words, for $x \in \mathbb{R}^2$ call $\phi^{\mathbf{n}}(x)$ the unique point in $\Gamma_0 \cup \Gamma_1$ such that $\phi^{\mathbf{n}}(x) - x$ is parallel to $\mathbf{n}$. Then let $\widetilde{\mu}$ be the push-forward of the measure μ w.r.t. $\phi^{\mathbf{n}}$. Since this projection satisfies $|\phi^{\mathbf{n}}(x)| \leq |x|$ for every $x \in \mathbb{R}^2_+$, the transportation cost decreases. On the other hand, by Remark 2.2 the sunlight captured remains the same. We conclude that

$$\mathcal{S}^{\mathbf{n}}(\widetilde{\mu}) - c\mathcal{I}^1(\widetilde{\mu}) \;\geq\; \mathcal{S}^{\mathbf{n}}(\mu) - c\mathcal{I}^1(\mu),$$

with strict inequality if μ is not supported on $\Gamma_0 \cup \Gamma_1$.

In the case $0 < \alpha < 1$, the result is not so obvious. Indeed, we do not expect that the conclusion holds if the hypothesis (2.16) is removed. A proof of Theorem 2.7 will be worked out in Sections 3 and 4.

Having proved that the optimal measure μ is supported on the two rays $\Gamma_0 \cup \Gamma_1$, the density of μ w.r.t. one-dimensional measure can then be determined using the necessary conditions derived in [6]. Indeed, the density u_1 of μ along the ray Γ_1 provides a solution to the scalar optimization problem

$$\text{maximize:} \;\; \mathcal{J}_1(u) \;\doteq\; \int_0^{+\infty} \left(1 - e^{-u(s)}\right) ds - c\int_0^{+\infty} \left(\int_s^{+\infty} u(r)\, dr\right)^{\alpha} ds\,, \tag{2.17}$$

among all non-negative functions $u : \mathbb{R}_+ \mapsto \mathbb{R}_+$. Here s is the arc-length variable along Γ_1. Similarly, the density u_0 of μ along the ray Γ_0 provides a solution to the problem

$$\text{maximize:} \quad \mathcal{J}_0(u) \doteq \int_0^{+\infty} \sin\theta_0 \left(1 - e^{-u(s)/\sin\theta_0}\right) ds - c\int_0^{+\infty} \left(\int_s^{+\infty} u(r)\, dr\right)^{\alpha} ds\,. \tag{2.18}$$

We write (2.17) in the form

$$\text{maximize:} \quad \mathcal{J}_1(u) \doteq \int_0^{+\infty} \Big[\left(1 - e^{-u(s)}\right) - cz^{\alpha}\Big] ds\,, \tag{2.19}$$

subject to

$$\dot{z} = -u, \qquad z(+\infty) = 0. \tag{2.20}$$

The necessary conditions for optimality (see for example [8, 11]) now yield

$$u(s) = \underset{\omega\geq 0}{\operatorname{argmax}} \Big\{ -e^{-\omega} - \omega q(s)\Big\} = -\ln q(s), \tag{2.21}$$

where the dual variable q satisfies

$$\dot{q} = c\alpha z^{\alpha-1}, \qquad q(0) = 0. \tag{2.22}$$

Notice that, by (2.21), $u > 0$ only if $q < 1$. Combining (2.20) with (2.22) one obtains an ODE for the function $q \mapsto z(q)$, with $q \in [0,1]$. Namely

$$\frac{dz(q)}{dq} = \frac{z^{1-\alpha}\ln q}{c\alpha}, \qquad z(1) = 0. \tag{2.23}$$

This equation admits the explicit solution

$$z(q) = c^{-1/\alpha}\left[1 + q\ln q - q\right]^{1/\alpha}. \tag{2.24}$$

Inserting (2.24) in (2.22), we obtain an implicit equation for $q(s)$:

$$s = \frac{1}{\alpha c^{1/\alpha}} \int_0^{q(s)} [1 + t\ln t - t]^{\frac{1-\alpha}{\alpha}}\, dt. \tag{2.25}$$

In turn, the density $u(s)$ of the optimal measure μ along Γ_1, as a function of the arc-length s, is recovered from (2.21). Notice that this measure is supported only on an initial interval $[0, \ell_1]$, determined by

$$\ell_1 = \frac{1}{\alpha c^{1/\alpha}} \int_0^1 [1 + s\ln s - s]^{\frac{1-\alpha}{\alpha}}\, ds.$$

The density of the optimal measure along the ray Γ_0 is computed in an entirely similar way. In this case, the equations (2.21) and (2.25) are replaced respectively by

$$u(s) = -(\sin\theta_0)\ln q(s),$$

$$s = \frac{(\sin\theta_0)^{\frac{1-\alpha}{\alpha}}}{\alpha c^{1/\alpha}} \int_0^{q(s)} [1 + t\ln t - t]^{\frac{1-\alpha}{\alpha}}\, dt.$$

Again, the condition $u(s) > 0$ implies $q(s) < 1$. Along Γ_0, the optimal measure μ is supported on an initial interval $[0, \ell_0]$, where

$$\ell_0 = \frac{(\sin\theta_0)^{\frac{1-\alpha}{\alpha}}}{\alpha c^{1/\alpha}} \int_0^1 [1 + s\ln s - s]^{\frac{1-\alpha}{\alpha}}\, ds.$$

2.5 The case $\alpha = 0$.

In the analysis of the optimization problem **(OPB)**, the case $\alpha = 0$ stands apart. Indeed, the general theorem on the existence of an optimal shape proved in [7] does not cover this case.

When $\alpha = 0$, a measure μ is irrigable only if it is concentrated on a set of dimension ≤ 1. When this happens, in any dimension $d \geq 3$ we have $\mathcal{S}^\eta(\mu) = 0$ and the optimization problem is trivial. The only case of interest occurs in dimension $d = 2$. In the following, $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathbb{R}^2$.

Theorem 2.8 *Let $\alpha = 0$, $d = 2$. Let $\eta \in \mathbf{L}^1(S^1)$ and define*

$$K \doteq \max_{|\mathbf{w}|=1} \int_{\mathbf{n} \in S^1} \Big| \langle \mathbf{w}, \mathbf{n} \rangle \Big| \, \eta(\mathbf{n}) \, d\mathbf{n}. \tag{2.26}$$

(i) If $K > c$, then the optimization problem **(OPB)** *has no solution, because the supremum of all possible payoffs is $+\infty$.*

(ii) If $K \leq c$, then the maximum payoff is zero, which is trivially achieved by the zero measure.

A proof will be given in Section 5.

3 Properties of optimal branch configurations

In this section we consider the optimization problem (2.14) in dimension $d = 2$. As a step toward the proof of Theorem 2.7, some properties of optimal branch configurations will be derived.

By the result in [7] we know that an optimal measure μ exists and has bounded support, contained in $\mathbb{R}^2_+ \doteq \{(x_1, x_2)\,; \quad x_2 \geq 0\}$. Call $M = \mu(\mathbb{R}^2_+)$ the total mass of μ and let $\chi : [0, M] \times \mathbb{R}_+ \mapsto \mathbb{R}^2_+$ be an optimal irrigation plan for μ.

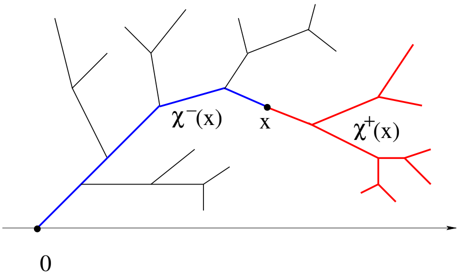


Figure 3: According to the definition (3.3), the set $\chi^-(x)$ is a curve joining the origin to the point x. The set $\chi^+(x)$ is a subtree, containing all paths that start from x.

Next, consider the set of all branches, namely

$$\mathcal{B} \doteq \{x \in \mathbb{R}^2_+\,; \quad |x|_\chi > 0\}. \tag{3.1}$$

By the single path property, we can introduce a partial ordering among points in $\mathcal{B}$. Namely, for any $x, y \in \mathcal{B}$ we say that $x \preceq y$ if for any $\xi \in [0, M]$ we have the implication

$$\chi(t, \xi) = y \qquad \Longrightarrow \qquad \chi(t', \xi) = x \qquad \text{for some} \;\; t' \in [0, t]. \tag{3.2}$$

This means that all particles that reach the point y pass through x before getting to y.

For a given $x \in \mathcal{B}$ the subsets of points $y \in \mathcal{B}$ that precede or follow x are defined as

$$\chi^-(x) \doteq \{y \in \mathcal{B};\ \ y \preceq x\}, \qquad\qquad \chi^+(x) \doteq \{y \in \mathcal{B};\ \ x \preceq y\}, \tag{3.3}$$

respectively (see Fig. 3).

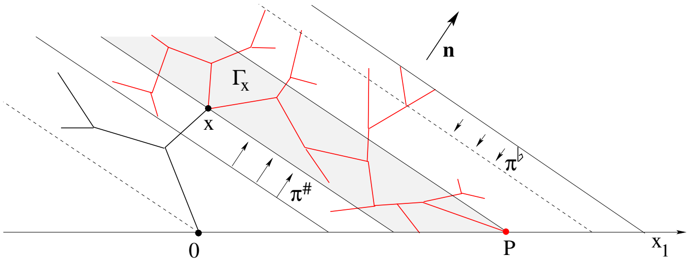


Figure 4: If the set $\chi^+(x)$ is not contained in the slab Γ_x (the shaded region), by taking the perpendicular projections $\pi^\sharp$ and $\pi^\flat$ we obtain another irrigation plan with strictly lower cost, which irrigates a new measure $\widetilde{\mu}$ gathering exactly the same amount of sunlight. Notice that here P is the point in the closed set $\overline{\chi^+(x)} \cap \mathbb{R}\mathbf{e}_1$ which has the largest inner product with $\mathbf{n}$.

We begin by deriving some properties of the sets $\chi^+(x)$. Introducing the unit vectors $\mathbf{e}_1 = (1,0)$, $\mathbf{e}_2 = (0,1)$, we denote by $\mathbb{R}\mathbf{e}_1$ the set of points on the x_1-axis. As before, $\mathbf{n} = (\cos\theta_0, \sin\theta_0)$ denotes the unit vector in the direction of the sunlight. Throughout the following, the closure of a set A is denoted by $\overline{A}$, while $\langle\cdot,\cdot\rangle$ denotes an inner product.

Lemma 3.1 *Let the measure μ provide an optimal solution to the problem (2.14), and let χ be an optimal irrigation plan for μ. Then, for every $x \in \mathcal{B}$, one has*

$$\chi^+(x) \ \subset\ \Gamma_x \ \doteq\ \Big\{y \in \mathbb{R}^2_+\,;\ \ \langle\mathbf{n}, y\rangle \ \in\ [a_x, b_x]\Big\}, \tag{3.4}$$

where $a_x \doteq \langle\mathbf{n}, x\rangle$, while b_x is defined as follows.

- *If $\overline{\chi^+(x)} \cap \mathbb{R}\mathbf{e}_1 = \emptyset$, then $b_x = a_x = \langle\mathbf{n}, x\rangle$.*
- *If $\overline{\chi^+(x)} \cap \mathbb{R}\mathbf{e}_1 \neq \emptyset$, then*

$$b_x \ =\ \max\ \{a_x, b'_x\}, \qquad\qquad b'_x \ \doteq\ \sup\Big\{\langle\mathbf{n}, z\rangle\,;\ \ z \in \overline{\chi^+(x)} \cap \mathbb{R}\mathbf{e}_1\Big\}.$$

Proof. The right-hand side of (3.4) is illustrated in Fig. 4. To prove the lemma, consider the set of all particles that pass through x, namely

$$\Theta_x \ \doteq\ \big\{\xi \in [0, M]\,;\ \ \chi(\tau,\xi) = x \ \text{ for some } \ \tau \geq 0\big\}.$$

1. We first show that, by the optimality of the solution,

$$\langle\mathbf{n}\,,\ \chi(\xi,t)\rangle \ \geq\ a_x \qquad\qquad \text{for all } \ \xi \in \Theta_x\,,\ t \geq \tau. \tag{3.5}$$

Indeed, consider the perpendicular projection on the half plane

$$\pi^\sharp : \mathbb{R}^2 \ \mapsto\ S^\sharp \ \doteq\ \{y \in \mathbb{R}^2\,;\ \langle\mathbf{n}, y\rangle \ \geq\ a_x\}.$$

Define the projected irrigation plan

$$\chi^\sharp(t,\xi) \doteq \begin{cases} \pi^\sharp\circ\chi(t,\xi) & \text{if } \xi\in\Theta_x\,, \quad t\geq\tau, \\ \chi(t,\xi) & \text{otherwise.}\end{cases}$$

Then the new measure $\mu^\sharp$ irrigated by $\chi^\sharp$ is still supported on $\mathbb{R}^2_+$ and has exactly the same projection on $E^\perp_{\mathbf{n}}$ as μ. Hence it gathers the same amount of sunlight. However, if the two irrigation plans do not coincide a.e., then the cost of $\chi^\sharp$ is strictly smaller than the cost of χ, contradicting the optimality assumption.

2. Next, we show that

$$\langle\mathbf{n}\,,\,\chi(\xi,t)\rangle \leq b_x \qquad\qquad \text{for all } \ \xi\in\Theta_x \ \ t\geq\tau. \tag{3.6}$$

Indeed, call

$$b'' \doteq \sup\Big\{\langle\mathbf{n},z\rangle\,; \quad z\in\chi^+(x)\Big\}.$$

If $b''\leq b_x$, we are done. In the opposite case, by a continuity and compactness argument we can find $\delta>0$ such that the following holds. Introducing the perpendicular projection on the half plane

$$\pi^\flat:\mathbb{R}^2 \mapsto S^\flat \doteq \{y\in\mathbb{R}^2\,;\ \langle\mathbf{n},y\rangle \leq b''-\delta\},$$

one has

$$\big\{\pi^\flat(y)\,; \quad y\in\chi^+(x)\big\} \subseteq \mathbb{R}^2_+\,. \tag{3.7}$$

Similarly as before, define the projected irrigation plan

$$\chi^\flat(t,\xi) \doteq \begin{cases} \pi^\flat\circ\chi(t,\xi) & \text{if } \xi\in\Theta_x\,, \quad t\geq\tau, \\ \chi(t,\xi) & \text{otherwise.}\end{cases}$$

Then the new measure $\mu^\flat$ irrigated by $\chi^\flat$ is supported on $\mathbb{R}^2_+\cap S^\flat$ and has exactly the same projection on $E^\perp_{\mathbf{n}}$ as μ. Hence it gathers the same amount of sunlight. However, if the two irrigation plans do not coincide a.e., then the cost of $\chi^\flat$ is strictly smaller than the cost of χ, contradicting the optimality assumption. This completes the proof of the Lemma. □

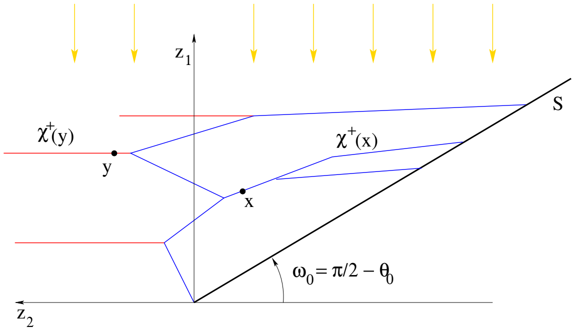


Figure 5: After a rotation of coordinates, the sunlight comes from the vertical direction. Here the blue lines correspond to the set $\mathcal{B}^*$ in (3.8).

Based on the previous lemma, we now consider the set

$$\mathcal{B}^* \doteq \{x\in\mathcal{B}\,;\ \ \overline{\chi^+(x)}\cap\mathbb{R}\mathbf{e}_1\neq\emptyset\}. \tag{3.8}$$

It will be convenient to rotate coordinates by an angle of $\pi/2-\theta_0$, and choose new coordinates (z_1, z_2) oriented as in Fig. 5. In these new coordinates, the direction of sunlight becomes vertical, while the positive x_1-axis corresponds to the line

$$\mathbf{S} \doteq \big\{(z_1, z_2)\,;\quad z_1 \geq 0\,,\quad z_2 = -\lambda z_1\big\}, \qquad \text{where} \quad \lambda = \tan\theta_0\,. \tag{3.9}$$

Calling $\big(z_1(\xi,t), z_2(\xi,t)\big)$ the corresponding coordinates of the point $\chi(\xi,t)$, from Lemma 3.1 we immediately obtain

Corollary 3.2 *Let χ be an optimal irrigation plan for a solution to (2.14). Then*

(i) For every $\xi \in [0, M]$, the map $t \mapsto z_1(\xi,t)$ is non-decreasing.

(ii) If $\bar z = (\bar z_1, \bar z_2) \notin \mathcal{B}^*$, then $\chi^+(\bar z)$ is contained in a horizontal line. Namely,

$$\chi^+(\bar z) \subset \{(\bar z_1, s)\,;\quad s \in \mathbb{R}\}. \tag{3.10}$$

To make further progress, we define

$$z_1^{\max} \;\doteq\; \sup\big\{z_1\,;\;(z_1,\, z_2) \in \mathcal{B}^*\big\}.$$

Moreover, on the interval $[0, z_1^{max}[$ we consider the function

$$\varphi(z_1) \;\doteq\; \sup\big\{s\,;\quad (z_1, s) \,\in\, \mathcal{B}^*\big\}. \tag{3.11}$$

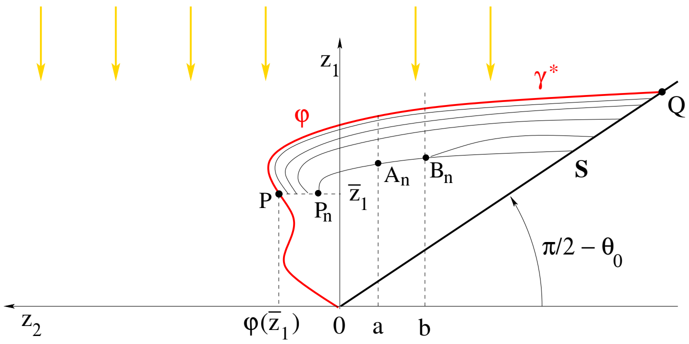


Figure 6: The construction used in the proof of Lemma 3.3.

Lemma 3.3 *For every $z_1 \in [0, z_1^{max}[$, the supremum $\varphi(z_1)$ is attained as a maximum.*

Proof. 1. Assume that, on the contrary, for some $\bar z_1$ the supremum is not a maximum. In this case, as shown in Fig. 6, there exist a sequence of points $P_n \to P$ with $P_n = (\bar z_1, s_n)$, $P = (\bar z_1, \bar z_2)$, $s_n \uparrow \bar z_2$. Here $P_n \in \mathcal{B}^*$ for every $n \geq 1$ but $P \notin \mathcal{B}^*$.

2. Choose two values a, b such that

$$-\lambda \bar z_1 \;<\; b \;<\; a \;<\; \varphi(\bar z_1).$$

By construction, for every $n \geq 1$ the set $\overline{\chi^+(P_n)}$ intersects $\mathbf{S}$. Therefore we can find points

$$P_n \prec A_n \prec B_n$$

all in $\mathcal{B}^*$, with

$$A_n = (t_n, a), \qquad B_n = (t'_n, b), \qquad\qquad \bar{z}_1 \leq t_n \leq t'_n \leq z_1^{max}.$$

3. Since the branches $\chi^+(A_n)$ are all disjoint, we have

$$\sum_{n \geq 1} |A_n|_\chi \leq M \doteq \mu(\mathbb{R}^2_+).$$

We can thus find N large enough so that

$$\varepsilon_N \doteq |A_N|_\chi < (a-b)^{\frac{1}{1-\alpha}}. \tag{3.12}$$

Consider the modified transport plan $\widetilde{\chi}$, obtained from χ by removing all particles that go through the point B_N. More precisely, $\widetilde{\chi}$ is the restriction of χ to the domain

$$\widetilde{\Theta} \doteq \Theta \setminus \{\xi\,;\ \ \chi(\xi, \tau) = B_N \quad \text{for some } \tau \geq 0\}.$$

Let $\widetilde{\mu}$ be the measure irrigated by $\widetilde{\chi}$.

Since $\widetilde{\mu} \leq \mu$, the total amount of sunlight gathered by the measure $\widetilde{\mu}$ satisfies

$$\mathcal{S}(\mu) - \mathcal{S}(\widetilde{\mu}) \leq (\mu - \widetilde{\mu})(\mathbb{R}^2). \tag{3.13}$$

We now estimate the reduction in the transportation cost, achieved by replacing μ with $\widetilde{\mu}$. Since all water particles reaching B_N must pass through A_N, they must cover a distance $\geq |B_N - A_N| \geq a - b$ traveling along a path whose maximum flux is $\leq \varepsilon_N$. The difference in the transportation costs can thus be estimated by

$$\mathcal{I}^\alpha(\mu) - \mathcal{I}^\alpha(\widetilde{\mu}) \geq (a-b) \cdot \alpha\varepsilon_N^{\alpha-1} \cdot (\mu - \widetilde{\mu})(\mathbb{R}^2). \tag{3.14}$$

If (3.12) holds, combining (3.13)-(3.14) we obtain

$$\mathcal{S}(\mu) - c\mathcal{I}^\alpha(\mu) < \mathcal{S}(\widetilde{\mu}) - c\mathcal{I}^\alpha(\widetilde{\mu}).$$

Hence the measure μ is not optimal. This contradiction proves the lemma. □

By the previous result, the graph of φ is contained in one single maximal trajectory of the transport plan χ. As in Figure 7, we let $s \mapsto \gamma(s)$ be the arc-length parameterization of this curve, which provides the left boundary of the set $\mathcal{B}^*$.

Along the curve γ, we now consider the set of points $C_j = (z_{1,j}, z_{2,j})$ where some horizontal branch bifurcates on the left. A property of such points is given below.

Lemma 3.4 *In the above setting, for every j, one has*

$$\varphi(s) < z_{2,j} \qquad \textit{for all } s < z_{1,j}\,. \tag{3.15}$$

Proof. If (3.15) fails, there exists another point $C_j^* = (z_{1,j}^*, z_{2,j})$ along the curve γ, with $z_{1,j}^* < z_{1,j}$. We can now replace the measure μ by another measure $\widetilde{\mu}$ obtained as follows. All the mass lying on the horizontal half-line $\{(z_{1,j}, s)\,;\ \ s \geq z_{2,j}\}$ is shifted downward on the half-line $\{(z_{1,j}^*, s)\,;\ \ s \geq z_{2,j}\}$. Since the functional $\mathcal{S}^{\mathbf{n}}$ is invariant under vertical shifts, we have $\mathcal{S}^{\mathbf{n}}(\widetilde{\mu}) = \mathcal{S}^{\mathbf{n}}(\mu)$. However, the transportation cost is strictly smaller: $\mathcal{I}^\alpha(\widetilde{\mu}) < \mathcal{I}^\alpha(\mu)$. This contradicts the optimality of μ. □

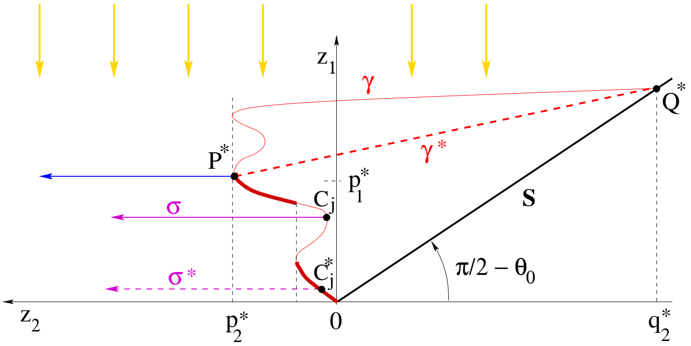


Figure 7: The thick portions of the curve γ are the only points where a left bifurcation can occur. If a horizontal branch σ bifurcates from C_j, all the mass on this branch can be shifted downward to another branch σ^* bifurcating from C_j^*. Furthermore, if some portion of the path γ between P^* and Q lies above the segment γ^* joining these two points, we can take a projection of γ on γ^*. In both cases, the transportation cost is strictly reduced.

Next, as shown in Fig. 7, we consider a point $P^* = (p_1^*, p_2^*) \in \gamma$ where the component z_2 achieves its maximum, namely

$$p_2^* \;=\; \max\{z_2\,;\;\; (z_1,z_2)\in\gamma\} \;\geq\; 0. \tag{3.16}$$

Notice that such a maximum exists because γ is a continuous curve, starting at the origin. If this maximum is attained at more than one point, we choose the one with smallest z_1-coordinate, so that

$$p_1^* \;=\; \min\{z_1\,;\;\; (z_1,p_2^*)\in\gamma\}. \tag{3.17}$$

Moreover, call

$$q_2^* \;\doteq\; \inf\{z_2\,;\;\; (z_1,z_2)\in \mathrm{Supp}(\mu)\},$$

and let $Q^* = (q_1^*, q_2^*) \in \mathbf{S}$ be the point on the ray $\mathbf{S}$ whose second coordinate is q_2^*. We observe that, by the optimality of the solution, all paths of the irrigation plan χ must lie within the convex set

$$\Sigma^* \;\doteq\; \{(z_1,z_2)\,;\;\; z_1\in[0,q_1^*], \quad z_2\geq q_2^*\}.$$

Otherwise, calling $\pi^* : \mathbb{R}^2 \mapsto \Sigma^*$ the perpendicular projection on the convex set Σ^*, the composed plan

$$\chi^*(\xi,t) \;\doteq\; \pi^*\big(\chi(\xi,t)\big)$$

would satisfy

$$\mathcal{S}^{\mathbf{n}}(\chi^*) \;=\; \mathcal{S}^{\mathbf{n}}(\chi), \qquad \mathcal{E}^\alpha(\chi^*) \;<\; \mathcal{E}^\alpha(\chi),$$

contradicting the optimality assumption.

By a projection argument we now show that, in an optimal solution, all the particle paths remain below the segment γ^* with endpoints P^* and Q^*.

Lemma 3.5 *In the above setting, let*

$$\gamma^* \;=\; \big\{(z_1,z_2)\,;\;\; z_1 = a + bz_2\,, \qquad z_2\in[q_2^*, p_2^*]\big\}$$

be the segment with endpoints P^*, Q^*. If

$$(\xi,t)\mapsto \chi(\xi,t) \;=\; \big(z_1(\xi,t), z_2(\xi,t)\big) \tag{3.18}$$

is an optimal irrigation plan for the problem (2.14), then we have the implication

$$z_2(\xi,t)\ \in\ [q_2^*,p_2^*] \qquad \Longrightarrow \qquad z_1(\xi,t)\ \leq\ a+b\,z_2(\xi,t). \tag{3.19}$$

Proof. 1. It suffices to show that the maximal curve γ lies below γ^*. If this is not the case, consider the set of particles which go through the point P^* and then move to the right of P^*, namely

$$\Omega^*\ =\ \Big\{\xi\in[0,M]\,;\ \ \chi(\xi,t^*)=P^*\ \text{ for some } t^*\geq 0,\quad z_2(\xi,t)<p_2^*\ \text{ for } t>t^*\Big\}. \tag{3.20}$$

2. Consider the convex region below γ^*, defined by

$$\Sigma\ \doteq\ \Big\{(z_1,z_2)\,;\ \ 0\leq z_1\leq a+bz_2\,,\qquad z_2\in[q_2^*,p_2^*]\Big\}.$$

Let $\pi:\mathbb{R}^2\mapsto\Sigma$ be the perpendicular projection. Then the irrigation plan

$$\chi^\dagger(\xi,t)\ \doteq\ \begin{cases}\pi\Big(\chi(\xi,t)\Big) & \text{if } \ \xi\in\Omega^*,\ t>t^*,\\ \quad\chi(\xi,t) & \text{otherwise,}\end{cases} \tag{3.21}$$

has total cost strictly smaller than χ. Indeed, for all x,ξ,t we have

$$\big|\pi(x)\big|_{\chi^\dagger}\ \geq\ |x|_\chi\,,\qquad \big|\dot\chi^\dagger(\xi,t)\big|\ \leq\ \big|\dot\chi(\xi,t)\big|. \tag{3.22}$$

Notice that, in (3.22), equality can hold for a.e. ξ,t only in the case where $\chi=\chi^\dagger$.

3. We now observe that the perpendicular projection on Σ can decrease the z_2-component. As a consequence, the measures μ and $\mu^\dagger$ irrigated by χ and $\chi^\dagger$ may have a different projections on the z_2 axis. If this happens, we may have $\mathcal{S}^{\mathbf{n}}(\mu)\neq\mathcal{S}^{\mathbf{n}}(\mu^\dagger)$.

To address this issue, we observe that all particles $\xi\in\Omega^*$ satisfy $\chi^\dagger(\xi,t^*)=\chi(\xi,t^*)=P^*$. In terms of the z_1,z_2 coordinates, this implies

$$z_2^\dagger(\xi,t^*)\ =\ z_2(\xi,t^*)\ =\ p_2^*,\qquad z_2^\dagger(\xi,T(\xi))\ \leq\ z_2(\xi,T(\xi))\ <\ p_2^*\,. \tag{3.23}$$

By continuity, for each $\xi\in\Omega^*$ we can find a stopping time $\tau(\xi)\in[t^*,T(\xi)]$ such that

$$z_2^\dagger(\xi,\tau(\xi))\ =\ z_2(\xi,T(\xi)).$$

Call $\widetilde\chi$ the truncated irrigation plan, such that

$$\widetilde\chi(\xi,t)\ \doteq\ \begin{cases}\chi^\dagger(\xi,t) & \text{if }\ \xi\in\Omega^*,\ t\leq\tau(\xi),\\ \chi(\xi,\tau(\xi)) & \text{if }\ \xi\in\Omega^*,\ t\geq\tau(\xi),\\ \chi(\xi,t) & \text{if }\ \xi\notin\Omega^*.\end{cases} \tag{3.24}$$

By construction, the measures μ and $\widetilde\mu$ irrigated by χ and $\widetilde\chi$ have exactly the same projections on the z_2 axis. Hence $\mathcal{S}^{\mathbf{n}}(\widetilde\mu)\ =\ \mathcal{S}^{\mathbf{n}}(\mu)$. On the other hand, the corresponding costs satisfy

$$\mathcal{E}^\alpha(\widetilde\chi)\ \leq\ \mathcal{E}^\alpha(\chi^\dagger)\ <\ \mathcal{E}^\alpha(\chi).$$

This contradicts optimality, thus proving the lemma. □

4 Proof of Theorem 2.7

In this section we give a proof of Theorem 2.7. As shown in Fig. 7, let $P^* = (p_1^*, p_2^*)$ be the point defined at (3.16). We consider two cases:

(i) $P^* = 0 \in \mathbb{R}^2$,

(ii) $P^* \neq 0$.

Assume that case (i) occurs. Then, by Lemma 3.4, the only branch that can bifurcate to the left of γ must lie on the z_2-axis. Moreover, by Lemma 3.5, the path γ cannot lie above the segment with endpoints P^*, Q^*. Therefore, the restriction of the measure μ to the half space $\{z_2 \leq 0\}$ is supported on the line $\mathbf{S}$. Combining these two facts we achieve the conclusion of the theorem.

The remainder of the proof will be devoted to showing that the case (ii) cannot occur, because it would contradict the optimality of the solution.

To illustrate the heart of the matter, we first consider the elementary configuration shown in Fig. 8, left, where all trajectories are straight lines. We call κ the flux along the segment P^*Q and σ the flux along the horizontal line bifurcating to the left of P^*. As in Fig. 8, right, we then replace the segments PP^* and P^*Q by a single segment with endpoints P, Q. To fix the ideas, the lengths of these two segments will be denoted by

$$\ell_a = |P - P^*|, \qquad \ell_b = |Q - P^*|. \tag{4.1}$$

The angles between these segments and a horizontal line will be denoted by θ_a, θ_b, respectively. Our main assumption is

$$0 \leq \theta_a \leq \frac{\pi}{2}, \qquad 0 \leq \theta_b < \frac{\pi}{2} - \theta_0 \,. \tag{4.2}$$

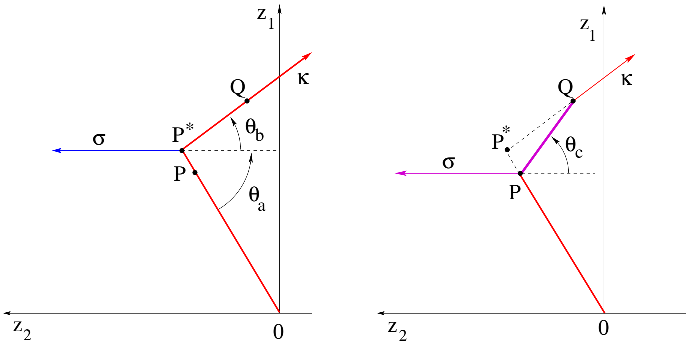


Figure 8: The basic case: in a neighborhood of P^* the trajectories are straight lines. To show that the configuration on the left is not optimal, we replace the portion of the trajectory between P and Q with a single segment.

Having performed this modification, the previous transportation cost along PP^* and P^*Q

$$(\kappa + \sigma)^\alpha \ell_a + \kappa^\alpha \ell_b$$

is replaced by

$$\kappa^\alpha \sqrt{\ell_a^2 + \ell_b^2 - 2\ell_a\ell_b\cos(\theta_a+\theta_b)} + \sigma^\alpha \ell_a \cos\theta_a\,. \tag{4.3}$$

Notice that the last term in (4.3) accounts for the fact that an amount σ of particles need to cover a longer horizontal distance, reaching P instead of P^*.

The difference in the cost is thus expressed by the function

$$f(\ell_a,\ell_b) \;=\; (\kappa+\sigma)^\alpha \ell_a - \sigma^\alpha \ell_a\cos\theta_a + \kappa^\alpha\left[\ell_b - \sqrt{\ell_a^2+\ell_b^2-2\ell_a\ell_b\cos(\theta_a+\theta_b)}\right].$$

Notice that this function is positively homogeneous of degree 1 w.r.t. the variables ℓ_a, ℓ_b. We observe that, by choosing the angle θ_c between the segment PQ and a horizontal line to be just slightly larger than θ_b, we can render the ratio ℓ_a/ℓ_b as small as we like. Taking advantage of this fact, we set

$$\ell_a = \varepsilon\ell, \qquad \ell_b \;=\; \ell$$

for some $\varepsilon > 0$ small. By the homogeneity of f it follows

$$f(\varepsilon\ell,\ell) \;=\; \ell\left[\varepsilon(\kappa+\sigma)^\alpha - \varepsilon\sigma^\alpha\cos\theta_a + \kappa^\alpha\Big(1-\sqrt{1+\varepsilon^2-2\varepsilon\cos(\theta_a+\theta_b)}\Big)\right].$$

This yields

$$\begin{aligned}\frac{d}{d\ell}f(\varepsilon\ell,\ell) \;&=\; \varepsilon(\kappa+\sigma)^\alpha - \varepsilon\sigma^\alpha\cos\theta_a + \kappa^\alpha\Big(1-\sqrt{1+\varepsilon^2-2\varepsilon\cos(\theta_a+\theta_b)}\Big)\\ &=\; \varepsilon\Big[(\kappa+\sigma)^\alpha - \sigma^\alpha\cos\theta_a + \kappa^\alpha\cos(\theta_a+\theta_b) + \mathcal{O}(1)\cdot\varepsilon\Big].\end{aligned} \tag{4.4}$$

Setting

$$\lambda \;=\; \frac{\sigma}{\kappa+\sigma}$$

we now study the function

$$F(\lambda,\theta_a,\theta_b) \;\doteq\; 1 - \lambda^\alpha\cos\theta_a + (1-\lambda)^\alpha\cos(\theta_a+\theta_b), \tag{4.5}$$

and find under which conditions on θ_b this function F remains positive for all $\lambda \in [0,1]$, $\theta_a \in [0,\pi/2]$.

Lemma 4.1 *(i) For $\alpha \geq 1/2$ and any $\theta_a,\theta_b \in [0,\pi/2]$, we always have $F(\lambda,\theta_a,\theta_b) \geq 0$.*

(ii) When $0 < \alpha < 1/2$ we have $F(\lambda,\theta_a,\theta_b) \geq 0$ for every $\theta_a,\theta_b \in [0,\pi/2]$ provided that θ_b satisfies the additional bound

$$\cos\theta_b \;\geq\; 1 - 2^{2\alpha-1}. \tag{4.6}$$

Proof. The function F in (4.5) can be written in terms of an inner product:

$$\begin{aligned}F(\lambda,\theta_a,\theta_b) \;&=\; 1 - \cos\theta_a\,[\lambda^\alpha - (1-\lambda)^\alpha\cos\theta_b] - \sin\theta_a(1-\lambda)^\alpha\sin\theta_b\\ &=\; 1 - \Big\langle (\cos\theta_a,\,\sin\theta_a)\;,\;\Big(\lambda^\alpha - (1-\lambda)^\alpha\cos\theta_b\;,\;(1-\lambda)^\alpha\sin\theta_b\Big)\Big\rangle.\end{aligned} \tag{4.7}$$

To prove that $F \geq 0$ it thus suffices to show that the second vector on the right hand side of (4.7) has length less than or equal to one, namely

$$\lambda^{2\alpha} + (1-\lambda)^{2\alpha} - 2\lambda^\alpha(1-\lambda)^\alpha\cos\theta_b \;\leq\; 1.$$

This inequality holds provided that

$$\cos\theta_b \;\geq\; \frac{\lambda^{2\alpha}+(1-\lambda)^{2\alpha}-1}{2\lambda^\alpha(1-\lambda)^\alpha}\,. \tag{4.8}$$

In the case where $\alpha \geq 1/2$ we have

$$\lambda^{2\alpha}+(1-\lambda)^{2\alpha} \;\leq\; 1 \qquad \text{for all } \lambda\in[0,1],$$

hence (4.8) holds.

To study the case where $\alpha < 1/2$, consider the function

$$g(\lambda) \;\doteq\; \frac{\lambda^{2\alpha}+(1-\lambda)^{2\alpha}-1}{2\lambda^\alpha(1-\lambda)^\alpha} \;=\; 1+\frac{\bigl(\lambda^\alpha-(1-\lambda)^\alpha\bigr)^2-1}{2\lambda^\alpha(1-\lambda)^\alpha}\,.$$

We observe that, for $0\leq\alpha\leq\frac{1}{2}$, one has

$$0 \;\leq\; g(\lambda) \;\leq\; g\Bigl(\frac{1}{2}\Bigr) \;=\; 1-2^{2\alpha-1}, \tag{4.9}$$

while

$$\lim_{\lambda\to0+} g(\lambda) \;=\; \lim_{\lambda\to1-} g(\lambda) \;=\; 0.$$

From (4.9) it now follows that the condition (4.6) guarantees that (4.8) holds, hence $F\geq 0$, as required. □

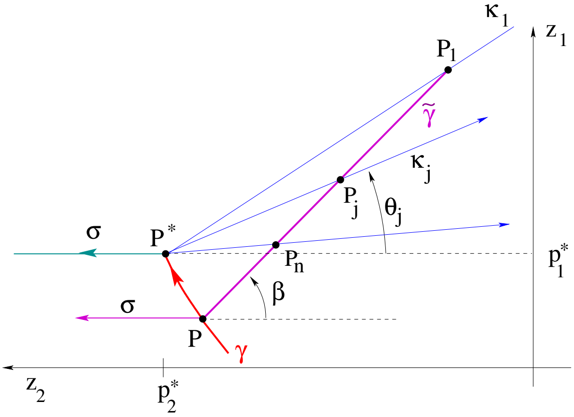


Figure 9: A more general configuration, compared with the one in Fig. 8.

We now consider the more general configuration shown in Fig. 9. Water is transported along the path γ up to the point P^*. Then the flux is split into a finite number of straight paths. One goes horizontally to the left, with flux $\sigma\geq 0$. The other pipes go to the right, with fluxes $\kappa_1,\ldots,\kappa_n>0$, at angles

$$0 \;\leq\; \theta_n \;<\; \cdots \;<\; \theta_2 \;<\; \theta_1 \;<\; \frac{\pi}{2}-\theta_0. \tag{4.10}$$

We compare this configuration with a modified irrigation plan, where a “bypass” is inserted along a segment $\widetilde\gamma$ with endpoints P, P_1, at an angle β satisfying

$$\theta_1 \;<\; \beta \;<\; \frac{\pi}{2}-\theta_0. \tag{4.11}$$

In this case, water particles travel along γ until they reach P. Then, an amount σ of particles bifurcates to the left. All the remaining particles are transported along the segment $\widetilde{\gamma}$, until they reach the points $P_n, \ldots, P_1$ along the old pipes. The next lemma estimates the saving in the irrigation cost achieved by inserting the "bypass" along the segment PP_1.

Lemma 4.2 *As in Theorem 2.7, assume that either* $1/2 \leq \alpha \leq 1$*, or else (2.16) holds. In the above setting, one has*

$$[\text{old cost}] - [\text{new cost}] \;\geq\; |P_1 - P^*| \cdot \delta(\theta_1, \kappa), \tag{4.12}$$

where $\delta(\theta_1, \kappa)$ *is a continuous function, strictly positive for* $0 \leq \theta_1 < \frac{\pi}{2} - \theta_0$ *and* $\kappa = \kappa_1 + \cdots + \kappa_n > 0$.

Proof. 1. As in the previous lemmas, we call θ_a the angle between the segment PP^* and a horizontal line. The difference between the old cost and the new cost can be expressed as

$$|P-P^*|\left(\sigma+\sum_{j=1}^n \kappa_j\right)^\alpha + \sum_{j=1}^n \kappa_j^\alpha |P^*-P_j| - \sigma^\alpha \cos\theta_a |P-P^*| - \sum_{j=1}^n \left(\sum_{i=1}^j \kappa_i\right)^\alpha |P_{j+1}-P_j|, \tag{4.13}$$

where, for notational convenience, we set $P_{n+1} \doteq P$. According to (4.13) we can write

$$[\text{old cost}] - [\text{new cost}] \;=\; A + S_n\,, \tag{4.14}$$

where

$$A \;\doteq\; |P-P^*|\left[\left(\sigma+\sum_{j=1}^n \kappa_j\right)^\alpha - \sigma^\alpha\cos\theta_a\right] + \left(\sum_{j=1}^n \kappa_j\right)^\alpha \Big(|P^*-P_1| - |P-P_1|\Big), \tag{4.15}$$

$$S_n \;=\; \sum_{j=1}^n \kappa_j^\alpha |P^*-P_j| - \Big(\sum_{j=1}^n \kappa_j\Big)^\alpha \Big(|P^*-P_1| - |P_{n+1}-P_1|\Big) - \sum_{j=1}^n \Big(\sum_{i=1}^j \kappa_i\Big)^\alpha |P_{j+1}-P_j|. \tag{4.16}$$

2. Notice that the quantity A in (4.15) would describe the difference in the costs if all the mass $\kappa = \kappa_1 + \cdots + \kappa_n$ were flowing through the point P_1. Using Lemma 4.1, we can thus choose $P = P_1$ close enough to P^* such that this difference is strictly positive. More precisely, for a fixed $\kappa > 0$, we claim that one can achieve the lower bound

$$\begin{aligned} A \;&\geq\; |P-P^*|\left[(\sigma+\kappa)^\alpha - \sigma^\alpha\cos\theta_a + \kappa^\alpha\cos(\theta_a+\theta_1) - \frac{\kappa^\alpha}{2}\frac{|P-P^*|}{|P_1-P^*|}\right] \\ &\geq\; |P_1-P^*|\cdot\delta(\theta_1,\kappa) \;>\; 0. \end{aligned} \tag{4.17}$$

Indeed, the last two terms within the square brackets in (4.17) are derived from

$$\begin{aligned} |P^*-P_1| - |P-P_1| \;&=\; |P^*-P_1|\left[1 - \sqrt{1 - 2\frac{|P-P^*|}{|P^*-P_1|}\cos(\theta_a+\theta_1) + \frac{|P-P^*|^2}{|P^*-P_1|^2}}\right] \\ &\geq\; |P^*-P_1|\left[1 - \left(1 - \frac{|P-P^*|}{|P^*-P_1|}\cos(\theta_a+\theta_1) + \frac{|P-P^*|^2}{2|P^*-P_1|^2}\right)\right]. \end{aligned}$$

Moreover, since we have the strict inequalities

$$\begin{cases} \theta_1 \;<\; \frac{\pi}{2} & \text{if} \quad \alpha \geq \frac{1}{2}\,, \\ \theta_1 < \frac{\pi}{2} - \theta_0 & \text{if} \quad \alpha < \frac{1}{2}\,, \end{cases} \tag{4.18}$$

the same argument used in the proof of (4.8) in Lemma 4.1 now yields the strict inequality

$$\cos\theta_1 \ > \ \frac{\lambda^{2\alpha}+(1-\lambda)^{2\alpha}-1}{2\lambda^\alpha(1-\lambda)^\alpha}\,. \tag{4.19}$$

Given $\kappa>0$ and P_1, we can then choose P close enough to P^* so that

- the term within the square brackets in (4.17) is strictly positive,
- the ratio $|P-P^*|/|P_1-P^*|$ is small but uniformly positive, as long as θ_1 remains bounded away from $\frac{\pi}{2}$ or from $\frac{\pi}{2}-\theta_0$ respectively, in the two cases considered in (4.18).

This proves our claim (4.17).

3. To complete the proof of the lemma, it remains to prove that $S_n\geq 0$. This will be proved by induction on n. Starting from (4.16) and using the inequalities

$$|P_n-P_1| \ \leq \ |P^*-P_1|, \qquad \Big(\sum_{i=1}^n \kappa_i\Big)^\alpha \ \leq \ \kappa_n^\alpha + \Big(\sum_{i=1}^{n-1}\kappa_i\Big)^\alpha,$$

we obtain

$$\begin{array}{rl} S_n & = \ \displaystyle\sum_{j=1}^n \kappa_j^\alpha|P^*-P_j| - \Big(\sum_{j=1}^n\kappa_j\Big)^\alpha \underbrace{\big(|P^*-P_1|-|P_n-P_1|\big)}_{\geq 0} - \sum_{j=1}^{n-1}\Big(\sum_{i=1}^j\kappa_i\Big)^\alpha|P_{j+1}-P_j| \\ & \geq \ \displaystyle\sum_{j=1}^{n-1} \kappa_j^\alpha|P^*-P_j| - \Big(\sum_{j=1}^{n-1}\kappa_j\Big)^\alpha \big(|P^*-P_1|-|P_{n-1}-P_1|\big) - \sum_{j=1}^{n-2}\Big(\sum_{i=1}^j\kappa_i\Big)^\alpha|P_{j+1}-P_j| \\ & \qquad +\kappa_n^\alpha|P^*-P_n| - \kappa_n^\alpha\Big(|P^*-P_1|-|P_n-P_1|\Big) \\ & = \ S_{n-1}+\kappa_n^\alpha\Big(|P^*-P_n|-|P^*-P_1|+|P_n-P_1|\Big) \ \geq \ S_{n-1}\,. \end{array} \tag{4.20}$$

Repeating this same argument, by induction we obtain

$$S_n \ \geq \ S_{n-1} \ \geq \ \cdots \ \geq S_1\,.$$

Observing that

$$S_1 \ = \ \kappa_1^\alpha|P^*-P_1| - \kappa_1^\alpha\Big(|P^*-P_1|-|P_2-P_1|\Big) - \kappa_1^\alpha|P_2-P_1| \ = \ 0,$$

we complete the proof of the lemma. □

We now consider the most general situation, shown in Fig. 10. Differently from the setting of Lemma 4.2, various scenarios must be considered.

- In addition to the horizontal path bifurcating to the left of P^* with flux σ, there can be countably many additional horizontal branches bifurcating to the left of γ, below P^*. We shall denote by σ_n, $n\geq 1$, the fluxes through these branches, at the bifurcation points.
- There can be countably many distinct branches bifurcating to the right of P^*, say with fluxes κ_j^*, $j\geq 1$.

- Furthermore, there can be countably many additional branches bifurcating to the right of γ, at points close to P^*. We shall denote by κ_i', $i \geq 1$, the fluxes through these branches, at the bifurcation points.
- Finally, the measure μ could concentrate a positive mass along the arc PP^*.

We observe that, by optimality, all particle trajectories to the right of γ move in the right-upward direction. Namely, setting $\chi(\xi,t) = (z_1(\xi,t),\, z_2(\xi,t))$, for these paths we have

$$\dot z_1(\xi,t) \,\geq\, 0, \qquad \dot z_2(\xi,t) \,\leq\, 0.$$

We now construct a "bypass", choosing a segment PQ with endpoints both lying on the curve γ, making an angle β with the horizontal direction such that

$$\beta^* \,<\, \beta \,<\, \frac{\pi}{2} - \theta_0\,. \tag{4.21}$$

Here β^* denotes the angle between the segment P^*Q^* and a horizontal line.

Given $\varepsilon > 0$, we can choose $N \geq 1$ large enough so that, among the branches bifurcating from P^*, one has

$$\sum_{j>N} \kappa_j^* \,<\, \varepsilon. \tag{4.22}$$

Moreover, by choosing Q sufficiently close to P^*, the following can be achieved:

(i) The total flux along the horizontal branches bifurcating to the left of γ below P^* satisfies

$$\sum_{n\geq 1} \sigma_n \,<\, \varepsilon. \tag{4.23}$$

(ii) The total flux along the branches bifurcating to the right of γ between P and P^*, and between P^* and Q satisfies

$$\sum_{i\geq 1} \kappa_i' \,<\, \varepsilon. \tag{4.24}$$

(iii) For each $j = 1,\ldots,N$, there exists a path γ_j connecting P^* with a point P_j on the segment PQ, along which the flux remains $\geq \kappa_j \,\geq\, \kappa_j^* - (\varepsilon/N)$. Here we denote by κ_j the flux reaching P_j.

In other words, even if the j-th branch through P^* further bifurcates, most of the particles along this branch cross the segment PQ at the same point P_j.

(iv) The total mass of μ along γ between P and P^* is $< \varepsilon$.

We estimate the difference in the new cost produced by these additional branches. Call $P = (p_1, p_2)$, $Q = (q_1, q_2)$.

- The additional mass on the left branches, together with the mass of μ present between P and P^* now travels along a horizontal line through P. By (i) and (iv) this mass is $< 2\varepsilon$. Hence:

$$[\text{additional cost}] \,\leq\, (2\varepsilon)^{1-\alpha}(z_2^* - z_2). \tag{4.25}$$

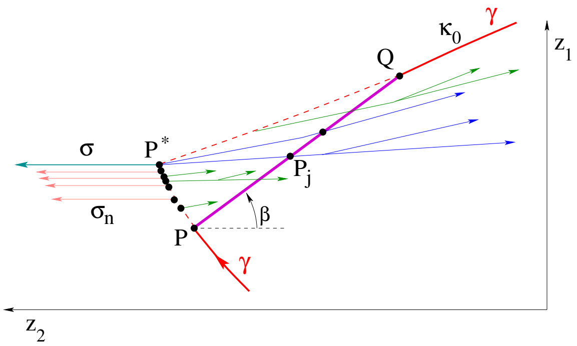


Figure 10: In the fully general situation, we have additional branches bifurcating to the left of γ between P and P^*, and to the right of γ at any point between P and Q. In addition, there can be an additional absolutely continuous source along the arc PP^*.

- The additional mass bifurcating to the right of γ, not crossing the segment PQ at one of the finitely many points $P_1, \ldots, P_N$ is $< 3\varepsilon$. The additional cost in transporting this mass from P to some point between P and Q satisfies

$$\text{[additional cost]} \;<\kappa_0^{\alpha-1}\cdot 3\varepsilon|P-Q|. \tag{4.26}$$

We now use Lemma 4.2. Combining (4.12) with (4.25)-(4.26) we obtain

$$\text{[old cost]} - \text{[new cost]} \;\geq\; |P_1-P^*|\cdot\delta(\theta_1,\kappa)-(2\varepsilon)^{1-\alpha}|P-P^*|-\kappa_0^{\alpha-1}\cdot 3\varepsilon|P-Q|. \tag{4.27}$$

By choosing $\varepsilon>0$ small enough, the right hand side of (4.27) is strictly positive. Hence the configuration with $P^*\neq 0$ is not optimal. This completes the proof of Theorem 2.7. □

5 The case $d=2$, $\alpha=0$

We give here a proof of Theorem 2.8.

1. Assume that there exists a unit vector $\mathbf{w}^*\in\mathbb{R}^2$ such that

$$K \;=\; \int_{\mathbf{n}\in S^1}\Big|\langle\mathbf{w}^*,\mathbf{n}\rangle\Big|\,\eta(\mathbf{n})\,d\mathbf{n}\;>\;c.$$

Let $\mathbf{v}=(\cos\beta,\sin\beta)$ be a unit vector perpendicular to $\mathbf{w}^*$, with $\beta\in[0,\pi]$. Let μ be the measure supported on the segment $\{r\mathbf{v}\,;\ r\in[0,\ell]\}$, with constant density λ w.r.t. 1-dimensional Lebesgue measure.

Then the payoff achieved by μ is estimated by

$$\begin{aligned}\mathcal{S}^\eta(\mu)-c\mathcal{I}^0(\mu) \;&=\; \ell\cdot\int_{S^1}\left(1-\exp\Big\{-\frac{\lambda}{\big|\langle\mathbf{w}^*,\mathbf{n}\rangle\big|}\Big\}\right)\Big|\langle\mathbf{w}^*,\mathbf{n}\rangle\Big|\,\eta(\mathbf{n})\,d\mathbf{n}-c\,\ell\\ &\geq\; \ell\cdot(1-e^{-\lambda})\int_{S^1}\Big|\langle\mathbf{w}^*,\mathbf{n}\rangle\Big|\,\eta(\mathbf{n})\,d\mathbf{n}-c\,\ell\\ &=\;\Big[(1-e^{-\lambda})\,K-c\Big]\,\ell.\end{aligned} \tag{5.1}$$

By choosing $\lambda > 0$ large enough, the first factor on the right hand side of (5.1) is strictly positive. Hence, by increasing the length ℓ, we can render the payoff arbitrarily large.

2. Next, assume that $K \leq c$. Consider any Lipschitz curve $s \mapsto \gamma(s)$, parameterized by arc-length $s \in [0, \ell]$. Then, for any measure μ supported on γ, the total amount of sunlight from the direction $\mathbf{n}$ captured by μ satisfies the estimate

$$\mathcal{S}^{\mathbf{n}}(\mu) \;\leq\; \int_0^\ell \Big|\langle \dot\gamma(s)^\perp,\, \mathbf{n}\rangle\Big|\, ds.$$

Indeed, it is bounded by the length of the projection of γ on the line $E_{\mathbf{n}}^\perp$ perpendicular to $\mathbf{n}$. Integrating over the various sunlight directions, one obtains

$$\mathcal{S}^\eta(\mu) \;\leq\; \int_0^\ell \int_{S^1} \Big|\langle \dot\gamma(s)^\perp,\, \mathbf{n}\rangle\Big|\, \eta(\mathbf{n})\, d\mathbf{n}\, ds \;\leq\; K\,\ell.$$

More generally, $\mu = \sum_i \mu_i$ can be the sum of countably many measures supported on Lipschitz curves γ_i. In this case, since the sunlight functional is sub-additive, one has

$$\mathcal{S}^\eta(\mu) \;\leq\; \sum_i \mathcal{S}^\eta(\mu_i) \;\leq\; \sum_i K\ell_i\,.$$

Hence

$$\mathcal{S}^\eta(\mu) - c\mathcal{I}^0(\mu) \;\leq\; \sum_i K\ell_i - c\sum_i \ell_i \;\leq\; 0.$$

This concludes the proof of case (ii) in Theorem 2.8. □

Acknowledgments. The research of the first author was partially supported by NSF with grant DMS-1714237, “Models of controlled biological growth”. The research of the second author was partially supported by a grant from the U.S.-Norway Fulbright Foundation.

References

[1] M. Bernot, V. Caselles, and J. M. Morel, *Optimal transportation networks. Models and theory.* Springer Lecture Notes in Mathematics **1955**, Berlin, 2009.

[2] M. Bernot, V. Caselles, and J. M. Morel, The structure of branched transportation networks. *Calculus of Variations* (2008), 279-317.

[3] A. Brancolini, and S. Solimini, Fractal regularity results on optimal irrigation patterns. *J. Math. Pures Appl.* **102** (2014), 854–890.

[4] A. Brancolini and B. Wirth, Optimal energy scaling for micropatterns in transport networks. *SIAM J. Math. Anal.* **49** (2017), 311-359.

[5] L. Brasco and F. Santambrogio, An equivalent path functional formulation of branched transportation problems. *Discrete Contin. Dyn. Syst.* **29** (2011), 845–871.

[6] A. Bressan, S. Galtung, A. Reigstad, and J. Ridder, Competition models for plant stems, *J. Differential Equations*, to appear.

[7] A. Bressan, M. Palladino, and Q. Sun, Variational problems for tree roots and branches, *Calc. Var. & Part. Diff. Equat.*, **57** (2020).

[8] A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*, AIMS Series in Applied Mathematics, Springfield Mo. 2007.

[9] A. Bressan and Q. Sun, On the optimal shape of tree roots and branches, *Math. Models & Methods Appl. Sci.* **28** (2018), 2763–2801.

[10] A. Bressan and Q. Sun, Weighted irrigation plans, submitted.

[11] L. Cesari, *Optimization - Theory and Applications*, Springer-Verlag, 1983.

[12] G. Devillanova and S. Solimini, Some remarks on the fractal structure of irrigation balls. *Adv. Nonlinear Stud.* **19** (2019), 55–68.

[13] E. N. Gilbert. Minimum cost communication networks. *Bell System Tech. J.* **46** (1967), 2209–2227.

[14] F. Maddalena, J. M. Morel, and S. Solimini, A variational model of irrigation patterns, *Interfaces Free Bound.* **5** (2003), 391–415.

[15] J. M. Morel and F. Santambrogio, The regularity of optimal irrigation patterns. *Arch. Ration. Mech. Anal.* **195** (2010), 499–531.

[16] P. Pegon, F. Santambrogio, and Q. Xia, A fractal shape optimization problem in branched transport. *J. Math. Pures Appl.* **123** (2019), 244–269.

[17] F. Santambrogio, Optimal channel networks, landscape function and branched transport. *Interfaces Free Bound.* **9** (2007), 149–169.

[18] Q. Xia, Optimal paths related to transport problems, *Comm. Contemp. Math.* **5** (2003), 251–279.

[19] Q. Xia, Motivations, ideas and applications of ramified optimal transportation. *ESAIM Math. Model. Numer. Anal.* **49** (2015), 1791–1832.