Springer Optimization and Its Applications 135

Themistocles M. Rassias *Editor*

Current Research in Nonlinear Analysis

In Honor of Haim Brezis and Louis Nirenberg

Springer

Springer Optimization and Its Applications

Volume 135

Managing Editor
Panos M. Pardalos (University of Florida)

Editor–Combinatorial Optimization
Ding-Zhu Du (University of Texas at Dallas)

Advisory Boards
J. Birge (University of Chicago)
C.A. Floudas (Texas A & M University)
F. Giannessi (University of Pisa)
H.D. Sherali (Virginia Polytechnic and State University)
T. Terlaky (Lehigh University)
Y. Ye (Stanford University)

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics and other sciences.

The series *Springer Optimization and Its Applications* aims to publish state-of-the-art expository works (monographs, contributed volumes, textbooks) that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Themistocles M. Rassias
Editor

Current Research in Nonlinear Analysis

In Honor of Haim Brezis and Louis Nirenberg

Springer

Editor
Themistocles M. Rassias
Department of Mathematics
National Technical University of Athens
Athens, Greece

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-89799-8 ISBN 978-3-319-89800-1 (eBook)
https://doi.org/10.1007/978-3-319-89800-1

Library of Congress Control Number: 2018945052

Mathematics Subject Classification: 26-XX, 28-XX, 30-XX, 32-XX, 34-XX, 35-XX, 37-XX, 39-XX, 41-XX, 43-XX, 45-XX, 46-XX, 47-XX, 49-XX, 52-XX, 53-XX, 54-XX, 57-XX, 58-XX, 65-XX

© Springer International Publishing AG, part of Springer Nature 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Dedicated to Professors Haim Brezis and Louis Nirenberg with deep admiration



Haim Brezis and Louis Nirenberg in a mathematical discussion, Jerusalem, January 2002. By the courtesy of Rachel S. Brezis

Preface

Nonlinear Analysis: In Honor of Haim Brezis and Louis Nirenberg focuses on various important areas of mathematical research that have been either introduced, inspired, or influenced by two pioneering researchers of contemporary nonlinear analysis, who greatly contributed to shaping this field. The contributed papers have been written by eminent experts from the international mathematical community. These works broaden our understanding of several theories, methods, and techniques which have been found very applicable toward the solution of several important problems, some of which had resisted solution for long stretches of time. Some of these papers have been contributed by former students of these two influential mathematicians. This volume serves as a tribute to Brezis and Nirenberg for their great service to mathematical analysis and its applications.

I feel very grateful to all the mathematicians who enthusiastically responded with their contributing chapters in this publication. I would also like to acknowledge the superb assistance of the staff of Springer for the production of this volume.

Athens, Greece — Themistocles M. Rassias

Contents

Anisotropic Surface Measures as Limits of Volume Fractions 1
Luigi Ambrosio and Giovanni E. Comi

Failure of the Chain Rule in the Non Steady Two-Dimensional Setting ... 33
Stefano Bianchini and Paolo Bonicatto

On the Maximality of the Sum of Two Maximal Monotone Operators 61
Yuqing Chen, Yeol Je Cho, and Themistocles M. Rassias

Analysis and Simulation of a Model of Phosphorus Uptake by Plant Roots 85
Myriam Comte

Elliptic Operators with Infinitely Many Variables 99
Giuseppe Da Prato

On Generalized Pseudo- and Quasiconvexities for Nonsmooth Functions 129
Ville-Pekka Eronen, Marko M. Mäkelä, and Napsu Karmitsa

On the Equivalence of Eulerian and Lagrangian Variables for the Two-Component Camassa–Holm System 157
Markus Grasmair, Katrin Grunert, and Helge Holden

Sum-Intersection Property of Sobolev Spaces 203
Petru Mironescu

Operator-Norm Convergence of the Trotter Product Formula on Hilbert and Banach Spaces: A Short Survey 229
Hagen Neidhardt, Artur Stephan, and Valentin A. Zagrebnov

Multiple Periodic Solutions of Lagrangian Systems of Relativistic Oscillators 249
Biagio Ricceri

Applications of a Theorem of Brezis-Nirenberg 259
Martin Schechter

Porous Medium Equation with Nonlocal Pressure 277
Diana Stan, Félix del Teso, and Juan Luis Vázquez

Nondiscrete Lassonde-Revalski Principle and Dependent Choice 309
Mihai Turinici

Anisotropic Surface Measures as Limits of Volume Fractions



Luigi Ambrosio and Giovanni E. Comi

1 Introduction

The literature on approximation of Sobolev and BV norms, and on the characterizations of the corresponding spaces in terms of these approximations, is by now very wide, see in particular [4] for the case of Sobolev spaces, [16] and the more recent papers [6, 7] which deal with non-local approximations, in the sense of Γ-convergence of (a multiple of) the total variation norm, with intriguing connection to problems considered in image processing. Still in connection with non-local functionals, it is worth to mention the paper [8] which gave origin to the theory of nonlocal minimal surfaces.

Somehow in the same vein, motivated by [5], the first author, Bourgain, Brezis and Figalli recently studied in [2] and [1] a new characterization of the perimeter of a set in $\mathbb{R}^n$ by considering the following functionals originating from a BMO-type seminorm

$$\mathrm{I}_\varepsilon(f) = \varepsilon^{n-1} \sup_{\mathscr{G}_\varepsilon} \sum_{Q' \in \mathscr{G}_\varepsilon} \fint_{Q'} |f(x) - \fint_{Q'} f| \, dx, \tag{1}$$

where $\mathscr{G}_\varepsilon$ is any disjoint collection of ε-cubes Q' with arbitrary orientation and cardinality not exceeding ε^{1-n}.

In particular, they studied the case $f = \mathbf{1}_A$; that is, the characteristic function of a measurable set A, and proved that

L. Ambrosio (✉) · G. E. Comi
Scuola Normale Superiore, Pisa, Italy
e-mail: l.ambrosio@sns.it; giovanni.comi@sns.it

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_1

$$\lim_{\varepsilon\to 0} \mathsf{I}_\varepsilon(\mathbf{1}_A) = \frac{1}{2}\min\{1, \mathsf{P}(A)\}. \tag{2}$$

This theme has been further investigated in [11], for BV functions, see also [12] for a variant of this construction leading to Sobolev norms and spaces.

In this paper we study more in detail the structure of the optimization problem in (1). We remove the upper bound on cardinality that seems to be very special of the case of cubes, at least if one is willing to get a precise formula as (2) and not only upper and lower bounds on I_ε. With this simplification, we prove that the existence of the limit and the emergence of a surface measure are general phenomena. In particular we prove that, for some dimensional constant $\xi = \xi(n)$, one has

$$\lim_{\varepsilon\to 0} H^B_\varepsilon(\mathbf{1}_A) = \xi \mathsf{P}(A), \tag{3}$$

where H^B_ε is defined as (1), without the bound on cardinality and using disjoint ε-balls.

More generally, if C is a bounded open set with Lipschitz boundary and if we define

$$H^C_\varepsilon(A) := \varepsilon^{n-1} \sup_{\mathscr{G}_\varepsilon} \sum_{C'\in\mathscr{G}_\varepsilon} \fint_{C'} |\mathbf{1}_A(x) - \fint_{C'} \mathbf{1}_A| \, dx, \tag{4}$$

where $\mathscr{G}_\varepsilon$ is any disjoint family of translations C' of the set εC with no bounds on cardinality, we are able to prove the following result.

Theorem 1 *There exists $\varphi^C : \mathbb{S}^{n-1} \to (0, +\infty)$, bounded and lower semicontinuous, such that, for any set of finite perimeter A, one has*

$$\lim_{\varepsilon\to 0} H^C_\varepsilon(A) = \int_{\mathscr{F}A} \varphi^C(\nu_A(x)) \, d\mathscr{H}^{n-1}(x), \tag{5}$$

where $\mathscr{F}A$ and ν_A are respectively the reduced boundary of A and the approximate unit normal to $\mathscr{F}A$. Moreover, if A is measurable and $\mathsf{P}(A) = \infty$, one has

$$\lim_{\varepsilon\to 0} H^C_\varepsilon(A) = +\infty. \tag{6}$$

The right hand side of (5) can be seen as an anisotropic version of the perimeter, $\mathsf{P}_\varphi(A)$. This result, while shows that the particular geometry of the covering sets is not essential, raises indeed some open questions. The most important is maybe the following one:

$$\text{Is the function } \tilde{\varphi}^C(p) := \begin{cases} |p|\varphi^C\left(\frac{p}{|p|}\right) & \text{if } p \neq 0 \\ \\ 0 & \text{if } p = 0 \end{cases} \quad \text{convex?}$$

This question is natural, in view of the fact that the anisotropic perimeter $\int_{\mathscr{F}A} \varphi(\nu_A)\, d\mathscr{H}^{n-1}$ is lower semicontinuous w.r.t. the convergence in measure if and only if φ is the restriction to the unit sphere of a positively 1-homogeneous and convex function. The problem is nontrivial since we were able to prove that, if C is the unit square $(0, 1)^2$ in $\mathbb{R}^2$, then $\tilde{\varphi}^C$ is not convex, as it is shown in Sect. 4. In particular, the convexity of C is not a sufficient condition to obtain $\tilde{\varphi}^C$ convex.

The paper is organized as follows: after a brief section containing preliminary results, in Sect. 3 we provide first the proof of (6), by a simple comparison argument based on the results of [1]. Then, we define suitable localized versions $H_\varepsilon(A, \Omega)$ and $H_\pm(A, \Omega)$ of our functionals; the latter arise by taking the lim sup and the lim inf w.r.t. the scale parameter ε. We can use symmetry and superadditivity arguments to show that $H_+ = H_-$ when both are evaluated in halfspaces A and in cubical domains Ω with faces parallel or orthogonal to the normal to the halfspace. Eventually, we use covering theorems as well as the fine properties of sets of finite perimeter to extend the result to general sets of finite perimeter and general domains.

In Sect. 4 we discuss examples and variants of our result. Finally, we provide in the Appendix, for the reader's convenience, a proof of the existence of the "best volume fraction" in Kepler's packing problem (related to our problem when we choose as C a ball), whose value is presently known only in dimensions 2 and 3 [14, 19].

2 Notation and Preliminary Results

In this paper we assume $n \geq 2$. We denote by A^c the complement of A, by $|A|$ the Lebesgue measure of a Lebesgue measurable set $A \subset \mathbb{R}^n$, by $\mathscr{H}^{n-1}$ the Hausdorff $(n-1)$-dimensional measure and by $A \Delta B := (A \setminus B) \cup (B \setminus A)$ the symmetric difference of the sets A, B. If E is a set of finite perimeter, we denote by $\partial^* E$ its essential boundary, namely the complement of the sets of density and rarefaction points. We refer to Section 3.5 of [3] for the basic definitions and the general properties of sets of finite perimeter needed in this paper.

In order to avoid confusion, in the rest of the paper we will mostly work with the reduced boundary $\mathscr{F}E \subset \partial^* E$, using the larger essential boundary $\partial^* E$ only when strictly necessary, see for instance (25). The reason for this choice is that at every point of the reduced boundary the measure theoretic exterior normal is well defined, and we need it in order to prove results on densities of set functions. On the other hand, it is well-known that $\mathscr{H}^{n-1}(\partial^* E \setminus \mathscr{F}E) = 0$, hence every property

which holds for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$ holds also for $\mathscr{H}^{n-1}$-a.e. $x \in \partial^* E$ and, when integrating w.r.t. $\mathscr{H}^{n-1}$, the difference is not seen.

An important result we need is the linear form of the relative isoperimetric inequality (see for instance Theorem 3.44 in [3]): for any open bounded set C with Lipschitz boundary there exists a constant $\gamma = \gamma(C)$ such that

$$\frac{|C \cap E||C \setminus E|}{|C|^2} \leq \gamma \mathsf{P}(E, C), \tag{7}$$

for any measurable set E. By scaling, it follows that if $C' = \varepsilon C$ we have

$$\frac{|C' \cap E||C' \setminus E|}{|C'|^2} \leq \varepsilon^{1-n} \gamma \mathsf{P}(E, C'),$$

for any measurable set E.

3 Proof of Theorem 1

In order to prove Theorem 1, we define a localized version of H_ε: for any measurable set A and any open set Ω we set

$$H_\varepsilon^C(A, \Omega) := \varepsilon^{n-1} \sup_{\mathscr{G}_\varepsilon} \sum_{C' \in \mathscr{G}_\varepsilon} \fint_{C'} |\mathbf{1}_A(x) - \fint_{C'} \mathbf{1}_A| \, dx, \tag{8}$$

where the supremum runs among all disjoint families $\mathscr{G}_\varepsilon$ made with translations of the set εC in Ω. Since

$$\fint_{C'} |\mathbf{1}_A(x) - \fint_{C'} \mathbf{1}_A| \, dx = \fint_{C'} \fint_{C'} |\mathbf{1}_A(x) - \mathbf{1}_A(y)| \, dxdy = 2\frac{|C' \cap A||C' \setminus A|}{|C'|^2}, \tag{9}$$

we have the following equivalent definition

$$H_\varepsilon^C(A, \Omega) := \varepsilon^{n-1} \sup_{\mathscr{G}_\varepsilon} \sum_{C' \in \mathscr{G}_\varepsilon} 2\frac{|C' \cap A||C' \setminus A|}{|C'|^2}, \tag{10}$$

which we are going to use mostly.

In addition, it is convenient to define the following set functions

$$H_+^C(A, \Omega) := \limsup_{\varepsilon \to 0} H_\varepsilon^C(A, \Omega), \tag{11}$$

$$H_-^C(A, \Omega) := \liminf_{\varepsilon \to 0} H_\varepsilon^C(A, \Omega). \tag{12}$$

Clearly, we have $H^C_-(A,\Omega) \le H^C_+(A,\Omega)$. In order to show the existence of the limit in the case of a set of finite perimeter A, we need to prove the converse inequality $H_-(A,\Omega) \ge H_+(A,\Omega)$.

The following scaling properties will be useful:

$$H^{\lambda C}_\varepsilon(A,\Omega) = \lambda^{1-n} H^C_{\varepsilon\lambda}(A,\Omega), \qquad H^{\lambda C}_\pm(A,\Omega) = \lambda^{1-n} H^C_\pm(A,\Omega). \tag{13}$$

In the sequel, we also often assume with no loss of generality that $\mathrm{diam}(C) = 1$. Indeed, if we set $\tilde{C} := C/\mathrm{diam}(C)$, then (13) with $\lambda = \mathrm{diam}(C)$, so that $C = \lambda\tilde{C}$, implies

$$H^C_\varepsilon(A,\Omega) = \mathrm{diam}(C)^{1-n} H^{\tilde{C}}_{\varepsilon\mathrm{diam}(C)}(A,\Omega),\ H^C_\pm(A,\Omega) = \mathrm{diam}(C)^{1-n} H^{\tilde{C}}_\pm(A,\Omega).$$

It is also not difficult to compare H^C_ε to H^D_ε when $D \subset C$ and D is an open set containing the origin.[1] Indeed, it is clear that that for any measurable set A one has

$$\frac{|D\cap A||D\setminus A|}{|D|^2} \le \frac{|C|^2}{|D|^2}\frac{|C\cap A||C\setminus A|}{|C|^2}, \tag{14}$$

and that the same holds for any translated and dilated copies of C and D. Now, for any disjoint family $\mathscr{G}_{\varepsilon,D}$ of translations of εD we can find a family $\mathscr{G}_{\varepsilon,C}$ of translations of εC such that for any $D_j \in \mathscr{G}_{\varepsilon,D}$ there exists $C_j \in \mathscr{G}_{\varepsilon,C}$ with $D_j \subset C_j$. Even though the family $\mathscr{G}_{\varepsilon,C}$ is not disjoint in general, it is easily seen, using the inclusions

$$B(x_j,\lambda\varepsilon) \subset D_j \subset C_j \subset B(x_j,\varepsilon) \qquad \text{for some } x_j \in \mathbb{R}^n$$

(where $\lambda > 0$ satisfies $B(0,\lambda) \subset D$), that it has bounded overlap. More precisely, there exists $\theta = \theta(n,\lambda) > 0$ such that for any fixed j we have $\#\{k :\ B(x_k,\varepsilon)\cap B(x_j,\varepsilon) \ne \emptyset\} \le \theta$ and so the same property holds if we replace the balls by the corresponding sets C_j. Therefore, the family $\mathscr{G}_{\varepsilon,C}$ can be seen as the union of at most θ disjoint subfamilies $\mathscr{G}_{\varepsilon,C,i}$. This argument yields

$$\begin{aligned}\varepsilon^{n-1}\sum_{D'\in\mathscr{G}_{\varepsilon,D}} 2\frac{|D'\cap A||D'\setminus A|}{|D'|^2} &\le \frac{|C|^2}{|D|^2}\sum_{i=1}^{\theta}\varepsilon^{n-1}\sum_{C'\in\mathscr{G}_{\varepsilon,C,i}} 2\frac{|C'\cap A||C'\setminus A|}{|C'|^2}\\ &\le \frac{|C|^2}{|D|^2}\theta H^C_\varepsilon(A,\Omega),\end{aligned}$$

[1] Without loss of generality, we can always assume $0 \in D \subset C$.

and, taking the supremum over the families $\mathscr{G}_{\varepsilon,D}$, we obtain

$$H^D_\varepsilon(A,\Omega) \le \frac{|C|^2}{|D|^2}\theta H^C_\varepsilon(A,\Omega), \qquad H^D_\pm(A,\Omega) \le \frac{|C|^2}{|D|^2}\theta H^C_\pm(A,\Omega). \tag{15}$$

In addition, we notice that, for any rotation R we have $H^{R(C)}_\varepsilon(R(A),R(\Omega)) = H^C_\varepsilon(A,\Omega)$ and $H^{R(C)}_\pm(R(A),R(\Omega)) = H^C_\pm(A,\Omega)$.

Since in the following the set C will be mostly fixed, we drop the superscript C from H^C_ε, $H^C_\pm$.

3.1 Proof of (6)

In this section we prove (6), which follows easily from the results of [1] and from comparison arguments.

Let

$$\mathsf{I}_\varepsilon(\mathbf{1}_A,\Omega) := \varepsilon^{n-1}\sup_{\mathscr{G}_\varepsilon}\sum_{Q'\in\mathscr{G}_\varepsilon} 2\frac{|Q'\cap A||Q'\setminus A|}{|Q'|^2},$$

where $\mathscr{G}_\varepsilon$ denotes a collection of disjoint open cubes $Q'\subset\Omega$ with side length ε and arbitrary orientation. In [1] it was shown that, for any Borel set A, one has

$$\lim_{\varepsilon\to 0}\mathsf{I}_\varepsilon(\mathbf{1}_A,\mathbb{R}^n) = \frac{1}{2}\mathsf{P}(A). \tag{16}$$

For later purposes, we recall also a local version of (16) which is proved in [1] in order to get the global version, namely

$$\liminf_{\varepsilon\to 0}\mathsf{I}_\varepsilon(\mathbf{1}_A,\Omega) \ge \frac{1}{2}\mathsf{P}(A,\Omega) \qquad \text{for any open set } \Omega\subset\mathbb{R}^n. \tag{17}$$

Arguing as in the proof of (15), we observe that for any cube Q' with arbitrary orientation and side length $2\varepsilon/\sqrt{n}$, we can find an open ε-ball $B'\supset Q'$. Hence, for any collection $\mathscr{G}_{2\varepsilon/\sqrt{n},Q}$ of disjoint cubes Q' with arbitrary orientation and side length $2\varepsilon/\sqrt{n}$, we find a family $\mathscr{G}_{\varepsilon,B}$ of ε-balls with bounded overlap; that is, there exists $\theta_n>0$ such that for any fixed $B'\in\mathscr{G}_{\varepsilon,B}$ we have

$$\#\{B''\in\mathscr{G}_{\varepsilon,B} : B''\cap B'\neq\emptyset\}\le\theta_n.$$

Then, if we denote by H^B_ε the functional where we take a covering with ε-balls, we get

$$\theta_n H^B_\varepsilon(A,\Omega) \ge \frac{4^n}{n^n\omega_n^2}\mathsf{I}_{2\varepsilon/\sqrt{n}}(\mathbf{1}_A,\Omega) \tag{18}$$

If $\mathsf{P}(A) = +\infty$, inequalities (16) and (18) clearly give

$$\liminf_{\varepsilon\to 0} H_\varepsilon^B(A) \geq \liminf_{\varepsilon\to 0} \frac{4^n}{n^n\omega_n^2\theta_n} \mathsf{I}_{\frac{2}{\sqrt{n}}\varepsilon}(\mathbf{1}_A) = +\infty.$$

The case of a general bounded open set C containing the origin follows immediately by (15), since $C \supset B(0,\lambda)$ for some $\lambda = \lambda(C) > 0$.

3.2 First Properties of H_ε and $H_\pm$

We show now some elementary properties of the functionals H_ε and $H_\pm$, omitting the proof of the simplest ones and assuming the normalization $\mathrm{diam}(C) = 1$.

1. Translation invariance: for any $\tau \in \mathbb{R}^n$, we have $H_\varepsilon(A+\tau, \Omega+\tau) = H_\varepsilon(A,\Omega)$; taking limits, one has also $H_\pm(A+\tau, \Omega+\tau) = H_\pm(A,\Omega)$;
2. Monotonicity: $H_\varepsilon(A,\cdot)$ and $H_\pm(A,\cdot)$ are increasing set functions on the class of open sets in $\mathbb{R}^n$;
3. Homogeneity: for any $t > 0$, $H_{t\varepsilon}(tA, t\Omega) = t^{n-1}H_\varepsilon(A,\Omega)$. Indeed, $tC' \subset t\Omega$ if and only if $C' \subset \Omega$, and

$$\frac{|tC' \cap tA||tC' \setminus tA|}{|tC'|^2} = \frac{|C' \cap A||C' \setminus A|}{|C'|^2}.$$

It follows immediately that

$$H_\pm(tA, t\Omega) = t^{n-1}H_\pm(A,\Omega). \tag{19}$$

4. Superadditivity of H_-: it is easy to see that

$$H_\varepsilon(A, \Omega_1 \cup \Omega_2) = H_\varepsilon(A,\Omega_1) + H_\varepsilon(A,\Omega_2) \tag{20}$$

whenever $\Omega_1 \cap \Omega_2 = \emptyset$. From (20) we get

$$H_-(A, \Omega_1 \cup \Omega_2) \geq H_-(A,\Omega_1) + H_-(A,\Omega_2). \tag{21}$$

5. Almost subadditivity of H_+:

$$H_\varepsilon(A, \Omega_1 \cup \Omega_2) \leq H_\varepsilon(A, I_\varepsilon(\Omega_1)) + H_\varepsilon(A, I_\varepsilon(\Omega_2)), \tag{22}$$

for any open set Ω_1, Ω_2, where $I_t(\Omega) := \{x \in \mathbb{R}^n : \mathrm{dist}(x,\Omega) < t\}$. Indeed, if $C' \subset \Omega_1 \cup \Omega_2$, then it must be contained in the ε-neighbourhood of one of the two open sets, since $\mathrm{diam}(C') = \varepsilon\mathrm{diam}(C) = \varepsilon$. From (22) we get

$$H_+(A, \Omega_1 \cup \Omega_2) \le H_+(A, W_1) + H_+(A, W_2), \tag{23}$$

for any open sets $W_i \supset I_\delta(\Omega_i)$, $i = 1, 2$, for some $\delta > 0$.

6. Upper bound for H_+: using (7), we see that

$$H_\varepsilon(A, \Omega) \le 2\gamma \mathsf{P}(A, \Omega)$$

and so

$$H_+(A, \Omega) \le 2\gamma \mathsf{P}(A, \Omega). \tag{24}$$

3.3 Lower and Upper Density of $H_\pm$

We set

$$\varphi_+(\nu) := H_+(S_\nu, Q_\nu),$$
$$\varphi_-(\nu) := H_-(S_\nu, Q_\nu),$$

where $\nu \in \mathbb{S}^{n-1}$, $S_\nu := \{x \in \mathbb{R}^n : x \cdot \nu \ge 0\}$ and Q_ν a unit cube centered in the origin having one face orthogonal to ν and bisected by the hyperplane ∂S_ν.

Due to the translation invariance, this definition does not actually depend on the choice of the origin, since we could take any hyperplane $\{(x - x_0) \cdot \nu \ge 0\}$ and cubes centered in x_0.

It is obvious that $\varphi_-(\nu) \le \varphi_+(\nu)$. We collect in the next proposition a few elementary properties of $\varphi_\pm$ (more refined estimates in some special cases will be given in Sect. 4) and then we prove that these two functions coincide.

Proposition 1 *We have the following upper and lower bounds for $\varphi_\pm$:*

1. $\varphi_+ \le 2\gamma$, *where γ is the same constant in (24);*
2. $\varphi_- \ge \lambda^{n+1} \dfrac{2^{2n-1}}{|C|^2 \theta n^n \theta_n}$, *where $\lambda = \lambda(C) := \sup\{r > 0 : B(0, r) \subset C\}$, $\theta = \theta(n, \lambda)$ and θ_n are defined in the proofs of (15) and (6), respectively.*

In addition, $\varphi_- = \varphi_+$ and φ_- is lower semicontinuous.

Proof The inequality $\varphi_+ \le 2\gamma$ is easy, since by (7) we have

$$H_+(S_\nu, Q_\nu) \le 2\gamma \mathsf{P}(S_\nu, Q_\nu)$$

and $\mathsf{P}(S_\nu, Q_\nu) = 1$, by the definition of S_ν and Q_ν.

As for the lower bound on $\varphi_-(\nu)$, it can be obtained as follows: first we take $r > 0$ such that $B(0, r) \subset C$, then we apply (15), (13) and eventually (18) to get

$$
\begin{aligned}
H_\varepsilon^C(S_\nu, Q_\nu) &\geq \frac{|B(0,r)|^2}{|C|^2\theta} H_\varepsilon^{B(0,r)}(S_\nu, Q_\nu) \\
&= r^{1-n}\frac{|B(0,r)|^2}{|C|^2\theta} H_{\varepsilon r}^{B}(S_\nu, Q_\nu) \\
&\geq r^{n+1}\frac{|B|^2}{|C|^2\theta}\frac{4^n}{n^n\omega_n^2\theta_n} \mathrm{I}_{2\varepsilon\lambda/\sqrt{n}}(S_\nu, Q_\nu) \\
&= r^{n+1}\frac{1}{|C|^2\theta}\frac{4^n}{n^n\theta_n} \mathrm{I}_{2\varepsilon\lambda/\sqrt{n}}(S_\nu, Q_\nu).
\end{aligned}
$$

Now we let $\varepsilon \to 0$, using (17) with $A = S_\nu$ and $\Omega = Q_\nu$, and finally we take the supremum over $r > 0$ such that $B(0,r) \subset C$.

Finally, homogeneity implies that

$$
\varphi_-(\nu) = H_-(S_\nu, Q_\nu) = \liminf_{\varepsilon\to 0} \varepsilon^{n-1} H_1(S_\nu, (1/\varepsilon)Q_\nu),
$$

since $(1/\varepsilon)S_\nu = S_\nu$ for any $\varepsilon > 0$.

We observe that $(1/\varepsilon)Q_\nu$ contains the union of at least $\lfloor (t/\varepsilon)\rfloor^{n-1}$ open disjoint cubes of side length $1/t$, for any $t > \varepsilon$, which are translations of $(1/t)Q_\nu$ centered in points of ∂S_ν. Clearly, $H_\varepsilon(S_\nu, x + Q_\nu) = H_\varepsilon(S_\nu, Q_\nu)$ for any $x \in \partial S_\nu$. Hence, the monotonicity in the second argument, the additivity of H_ε and the homogeneity imply

$$
\begin{aligned}
\varphi_-(\nu) \geq \liminf_{\varepsilon\to 0} \varepsilon^{n-1} \lfloor (t/\varepsilon)\rfloor^{n-1} H_1(S_\nu, (1/t)Q_\nu) &= t^{n-1} H_{t(1/t)}((1/t)S_\nu, (1/t)Q_\nu) \\
&= H_t(S_\nu, Q_\nu),
\end{aligned}
$$

which implies $\varphi_-(\nu) \geq \sup_{t>0} H_t(S_\nu, Q_\nu)$.

On the other hand, it is clear that

$$
\varphi_-(\nu) \leq \varphi_+(\nu) = \limsup_{\varepsilon\to 0} H_\varepsilon(S_\nu, Q_\nu) = \lim_{\varepsilon\to 0} \sup_{0<s<\varepsilon} H_s(S_\nu, Q_\nu) \leq \sup_{s>0} H_s(S_\nu, Q_\nu)
$$

from which we deduce that $\varphi_-(\nu) = \varphi_+(\nu) = \sup_{t>0} H_t(S_\nu, Q_\nu)$.

As a byproduct, we also obtain that φ_- is lower semicontinuous in ν, being the supremum with respect to the parameter t of the supremum over the families $\mathscr{G}_t$ of translations of tC of the quantities

$$
t^{n-1} \sum_{C' \in \mathscr{G}_t} 2\frac{|C' \cap S_\nu||C' \setminus S_\nu|}{|C'|^2},
$$

which are continuous functions of ν. □

We define $\varphi(\nu) := \lim_{\varepsilon\to 0} H_\varepsilon(S_\nu, Q_\nu)$, since Proposition 1 showed the existence of the limit.

To prove that the upper and lower densities of $H_\pm$ coincide with φ, we need a modulus of continuity for $E \to H_\varepsilon(E, \Omega)$ similar to the one shown in [1, Lemma 3.6]. We recall that for any E, F sets of finite perimeter in Ω we have

$$\mathscr{H}^{n-1}(\partial^*(E\Delta F)\cap\Omega) = \mathscr{H}^{n-1}((\partial^* E\Delta\partial^* F)\cap\Omega), \tag{25}$$

see for instance [1, Section 2].

Lemma 1 *For any E, F sets of finite perimeter in Ω and any $\varepsilon > 0$ we have*

$$H_\varepsilon(F,\Omega) \le H_\varepsilon(E,\Omega) + 4\gamma\,\mathscr{H}^{n-1}((\mathscr{F}E\Delta\mathscr{F}F)\cap\Omega). \tag{26}$$

In particular one has

$$H_\pm(F,\Omega) \le H_\pm(E,\Omega) + 4\gamma\,\mathscr{H}^{n-1}((\mathscr{F}E\Delta\mathscr{F}F)\cap\Omega). \tag{27}$$

Proof For any C' and any measurable set $L \subset C'$ we have the relative isoperimetric inequality (7) and, combining it with the inequality $\min\{t, 1-t\} \le 2t(1-t)$ for any $t \in [0,1]$, we obtain also

$$\min\{|L|, |C'\setminus L|\} \le 2\gamma|C|\varepsilon \mathsf{P}(L, C'). \tag{28}$$

Let now $\mathscr{G}_\varepsilon$ be a disjoint family of translations of εC in Ω.

For any $C' \in \mathscr{G}_\varepsilon$, we have

$$\begin{aligned}\fint_{C'}\fint_{C'} |\mathbf{1}_F(x) - \mathbf{1}_F(y)|\,dx\,dy \le{}& \fint_{C'}\fint_{C'} |\mathbf{1}_E(x) - \mathbf{1}_E(y)|\,dx\,dy \\ &+ \frac{2}{|C|}\varepsilon^{-n}|C'\cap(F\Delta E)|.\end{aligned} \tag{29}$$

Indeed,

$$\begin{aligned}&\fint_{C'}\fint_{C'} |\mathbf{1}_F(x) - \mathbf{1}_E(x) - \mathbf{1}_F(y) + \mathbf{1}_E(y)|\,dx\,dy \\ &= \frac{2}{|C'|^2}(2|C'\cap(F\setminus E)||C'\cap(E\setminus F)| + |C'\setminus(F\Delta E)||C'\cap(F\Delta E)|) \\ &= \frac{2}{|C'|^2}(2|C'\cap(F\setminus E)||C'\cap(E\setminus F)| + |C'||C'\cap(F\Delta E)| + \\ &\quad - (|C'\cap(E\setminus F)| + |C'\cap(F\setminus E)|)^2) \le \frac{2}{|C|}\varepsilon^{-n}|C'\cap(F\Delta E)|.\end{aligned}$$

Since $\mathbf{1}_{E^c}(x) - \mathbf{1}_{E^c}(y) = \mathbf{1}_E(y) - \mathbf{1}_E(x)$, then we have also

$$\fint_{C'}\fint_{C'} |\mathbf{1}_F(x) - \mathbf{1}_F(y)|\,dx\,dy \le \fint_{C'}\fint_{C'} |\mathbf{1}_E(x) - \mathbf{1}_E(y)|\,dx\,dy + \frac{2}{|C|}\varepsilon^{-n}|C' \cap (F\Delta E^c)|. \tag{30}$$

It is clear that $F\Delta E = \Omega \setminus (F\Delta E^c)$, hence we can apply (28) to $L = C' \cap (F\Delta E)$.

Therefore, by (9), we obtain

$$\varepsilon^{n-1}\sum_{C'\in\mathcal{G}_\varepsilon}\fint_{C'} |\mathbf{1}_F(x) - \fint_{C'}\mathbf{1}_F|\,dx \le H_\varepsilon(E,\Omega) + 4\gamma\sum_{C'\in\mathcal{G}_\varepsilon}\mathsf{P}(F\Delta E, C').$$

Since $\sum_{C'\in\mathcal{G}_\varepsilon}\mathsf{P}(F\Delta E, C') \le \mathsf{P}(F\Delta E,\Omega) = \mathscr{H}^{n-1}((\mathscr{F}E\Delta\mathscr{F}F)\cap\Omega)$ by (25), we can pass to the supremum at the left hand side and we get (26). □

Let now $x \in \partial S_\nu$. If we denote by $Q_\nu(x,r)$ the cube of side length r centered in $x \in \mathbb{R}^n$ and with one face orthogonal to ν, by homogeneity we have

$$\lim_{r\to 0}\frac{H_\pm(S_\nu, Q_\nu(x,r))}{r^{n-1}} = H_\pm(S_\nu, Q_\nu(x,1)) = \varphi(\nu). \tag{31}$$

Theorem 2 *Let E be a set of finite perimeter and ν_E be its measure theoretic interior normal. Then, for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$, we have*

$$\liminf_{r\to 0}\frac{H_-(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \ge \varphi(\nu_E(x)). \tag{32}$$

Proof By our previous remarks, the result holds if E is the half-space $\{y : (y-x)\cdot\nu_E(x) \ge 0\}$. Indeed, $\mathsf{P}(E, Q_{\nu_E(x)}(x,r)) = r^{n-1}$, so that (31) implies (32).

If E is a set of finite perimeter, for any $x \in \mathscr{F}E$ there exists the measure theoretic interior normal $\nu_E(x)$ and the approximate tangent space to the measure $|D\chi_E|$ is $\nu_E^\perp(x)$, namely

$$\mathrm{Tan}^{n-1}(|D\chi_E|, x) = \mathscr{H}^{n-1}\llcorner\nu_E^\perp(x).$$

This implies that

$$\lim_{r\to 0}\frac{\mathscr{H}^{n-1}(\mathscr{F}E\cap Q_{\nu_E(x)}(x,r))}{r^{n-1}} = \mathscr{H}^{n-1}((\nu_E^\perp(x))\cap Q_{\nu_E(x)}(x,1)) = 1.$$

Therefore, since $\mathsf{P}(E,\cdot) = \mathscr{H}^{n-1}\llcorner\mathscr{F}E$, we deduce that for all $x \in \mathscr{F}E$ one has

$$\mathsf{P}(E, Q_{\nu_E(x)}(x,r)) = r^{n-1} + o(r^{n-1}). \tag{33}$$

If F is the subgraph of a C^1 function in a neighbourhood of x, then $(F-x)/\rho$ is bi-Lipschitz equivalent to the half-space $S_{\nu_F(x)}$ in $Q_{\nu_F(x)}(0,1)$, with bi-Lipschitz constants converging to 1 as $\rho \to 0$. Hence, we can use a C^1 deformation map Φ with bi-Lipschitz constant close to 1 near to x to transform any disjoint family C_i' admissible for F into a disjoint family $D_i = \Phi(C_i)$; we can then find $C_i'' \subset D_i \subset C_i'''$ translated and scaled copies of C_i' whose diameters satisfy $\operatorname{diam}(C_i')/\operatorname{diam}(C_i) \sim 1$, $\operatorname{diam}(C_i''')/\operatorname{diam}(C_i) \sim 1$. Summing up, for $r > 0$ small enough there exists a nonnegative modulus of continuity $\omega(r)$ satisfying

$$(1-\omega(r))|C''\cap S_{\nu_F(x)}| \le |C'\cap F| \le (1+\omega(r))|C'''\cap S_{\nu_F(x)}|$$

for $0<\rho<r$ and any translated copy C' of ρC contained in $Q_{\nu_F(x)}(x,r)$. We can choose the modulus of continuity in such a way that similar inequalities hold with the roles of F and $S_{\nu_F(x)}$ reversed. Hence, we have

$$\frac{|C'\cap F||C'\setminus F|}{|C'|^2} \le (1+\omega(r))^2 \frac{|C'''\cap S_{\nu_F(x)}||C'''\setminus S_{\nu_F(x)}|}{|C'''|^2}, \tag{34}$$

and

$$\frac{|C'\cap F||C'\setminus F|}{|C'|^2} \ge (1-\omega(r))^2 \frac{|C''\cap S_{\nu_F(x)}||C''\setminus S_{\nu_F(x)}|}{|C''|^2}. \tag{35}$$

In particular, (35) and (31) imply

$$H_-(F, Q_{\nu_E(x)}(x,r)) \ge \varphi(\nu_F(x))r^{n-1} + o(r^{n-1}). \tag{36}$$

Now, in order to obtain (36) also for E, we are going to use the rectifiability of $\mathscr{F}E$ and apply Lemma 1 to E and to the subgraph of one of the C^1 hypersurfaces Γ_i whose union covers $\mathscr{H}^{n-1}$-almost all of $\mathscr{F}E$ and such that $\nu_E|_{\Gamma_i}$ is the interior normal of the subgraph of Γ_i. Indeed, we fix i and observe that for $\mathscr{H}^{n-1}$-a.e. $x \in \Gamma_i \cap \mathscr{F}E$ one has

$$\mathscr{H}^{n-1}((\Gamma_i \Delta \mathscr{F}E)\cap B(x,r)) = o(r^{n-1}),$$

arguing as in the proof of Lemma 3.7 of [1] and using the density properties of the Hausdorff measure (see [3, Theorem 2.56 and Eq. (2.41)]). It follows easily that we have also

$$\mathscr{H}^{n-1}((\Gamma_i \Delta \mathscr{F}E)\cap Q_{\nu_E(x)}(x,r)) = o(r^{n-1})$$

for $\mathscr{H}^{n-1}$-a.e. $x \in \Gamma_i \cap \mathscr{F}E$. Now we use (27) choosing $\Omega = Q_{\nu_E(x)}(x,r)$ and F to be the subgraph of Γ_i inside $Q_{\nu_E(x)}(x,r)$, obtaining

$$H_-(F, Q_{\nu_E(x)}(x,r)) \le H_-(E, Q_{\nu_E(x)}(x,r)) + 4\gamma\, \mathscr{H}^{n-1}((\Gamma_i \Delta \mathscr{F}E)\cap Q_{\nu_E(x)}(x,r)).$$

Since Γ_i is a C^1 hypersurface, we have (36) for F, with $\nu_F(x) = \nu_E(x)$. Since i is arbitrary this implies (36) for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$.

Combining (33) and (36), we get the desired result. □

Theorem 3 *For any Borel set $B \subset \mathscr{F}E$ and $t > 0$, we have that*

$$\liminf_{r\to 0} \frac{H_-(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \geq t \tag{37}$$

for all $x \in B$ implies $H_-(E,U) \geq t\mathscr{H}^{n-1}(B)$ for any open set $U \supset B$.

Proof Without loss of generality, let $U \supset B$ be a bounded open set, since $\mathscr{H}^{n-1} \llcorner B$ is inner regular. For a given $\delta \in (0,1)$, we consider the family $\mathscr{B}$ of all the closed cubes inside U centered in the points $x \in B$ with one face oriented as $\nu_E(x)$, such that, if we denote their interior by $Q_{\nu_E(x)}(x,r)$, we have

$$H_-(E, Q_{\nu_E(x)}(x,r)) \geq t(1-\delta)\mathsf{P}(E, Q_{\nu_E(x)}(x,r))$$

and $|D\chi_E|(\partial Q_{\nu_E(x)}(x,r)) = 0$.

In this way, we can apply the version of Vitali theorem for cubes (see Theorem 5.13 of [15]) and find a disjoint countable subfamily $\{\overline{Q_j}\}$ which covers $\mathscr{H}^{n-1}$-almost all of B. It is also clear that $\mathsf{P}(E, \overline{Q_j}) = \mathsf{P}(E, Q_j)$, hence we can use an open covering. Therefore, the superadditivity of $H_-(E,\cdot)$ implies

$$\begin{aligned} t\mathscr{H}^{n-1}(B) &\leq t\mathsf{P}(E, \bigcup_j Q_j) = t\sum_j \mathsf{P}(E, Q_j) \leq (1-\delta)^{-1}\sum_j H_-(E, Q_j) \\ &\leq (1-\delta)^{-1} H_-(E, \bigcup_j Q_j) \leq (1-\delta)^{-1} H_-(E,U). \end{aligned}$$

Letting $\delta \to 0$, we prove the theorem. □

We can now extend the result of Theorem 2 to $H_+(E,\cdot)$ using similar techniques.

Theorem 4 *Let E be a set of finite perimeter and ν_E be its measure theoretic interior normal. Then, for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$, we have*

$$\limsup_{r\to 0} \frac{H_+(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \leq \varphi(\nu_E(x)). \tag{38}$$

Proof In the beginning of the proof of Theorem 2 we showed that

$$\mathsf{P}(E, Q_{\nu_E(x)}(x,r)) = r^{n-1} + o(r^{n-1}) \text{ for all } x \in \mathscr{F}E.$$

By (31), (38) holds if E is a half space S_ν. Then we need to use estimate (34) in order to prove the inequality in the case that E is a subgraph of a C^1 function in a neighbourhood of x.

Finally, we switch the roles of F and E in (27) and we repeat the steps of the last part of the proof of Theorem 2 to obtain

$$H_+(E, Q_{\nu_E(x)}(x, r)) \leq \varphi(\nu_E(x))r^{n-1} + o(r^{n-1})$$

for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$.

Combining these results, we obtain (38). □

In order to prove the upper estimate for H_+, we need to consider the inner regularization of the nondecreasing set functions $H_+(E, \cdot)$ defined on the open sets of $\mathbb{R}^n$.

Definition 1 Let $\mathscr{A}$ be the family of open sets in $\mathbb{R}^n$ and let $\alpha : \mathscr{A} \to [0, +\infty]$ be a nondecreasing set function. The inner regular envelope of α is the function $\alpha^* : \mathscr{A} \to \overline{\mathbb{R}}$ defined by

$$\alpha^*(A) := \sup\{\alpha(A') : A' \Subset A\}.$$

It is not hard to show (see for instance [10]) that α^* is the largest inner regular function smaller than α (namely $\alpha^*(A) = \sup\{\alpha^*(A') : A' \Subset A\}$). Recall also that any inner regular and subadditive function α is σ-subadditive, namely

$$\alpha(A) \leq \sum_{i=0}^{\infty} \alpha(A_i) \qquad \text{whenever } A \subset \bigcup_{i=0}^{\infty} A_i.$$

The proof of this statement can be adapted also to the case when α is weakly subadditive as our set function H_+, this leads to the following result.

Proposition 2 *$H_+^*(E, \cdot)$ is σ-subadditive and*

$$H_+^*(E, \Omega) \leq 2\gamma \mathsf{P}(E, \Omega). \tag{39}$$

Proof Given open sets Ω_i, $i = 1, 2$, let $0 < t < H_+^*(E, \Omega_1 \cup \Omega_2)$. Then there exists $W \Subset \Omega_1 \cup \Omega_2$ such that $H_+(E, W) \geq t$. By Lemma 14.20 of [10], there exist open sets Ω_i', $i = 1, 2$, such that $W \Subset \Omega_1' \cup \Omega_2'$ and $\Omega_i' \Subset \Omega_i$, $i = 1, 2$. Hence, we can find open sets W_i such that $\Omega_i' \Subset W_i \Subset \Omega_i$, $i = 1, 2$, and, by (23), we obtain

$$t \leq H_+(E, W) \leq H_+(E, W_1) + H_+(E, W_2) \leq H_+^*(E, \Omega_1) + H_+^*(E, \Omega_2).$$

Since $t < H_+^*(E, \Omega_1 \cup \Omega_2)$ is arbitrary, this proves the subadditivity.

Since $H_+^*(E, \cdot)$ is inner regular and subadditive the σ-subadditivity follows. The last statement follows by (24) and the inner regularity of $\mathscr{H}^{n-1} \llcorner \mathscr{F}E$. □

We are now able to show the same result of Theorem 3 for H_+.

Theorem 5 *For any Borel set $B \subset \mathscr{F}E$ and $t > 0$, we have that*

$$\limsup_{r\to 0} \frac{H_+(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \le t \tag{40}$$

for all $x \in B$ implies $H_+^*(E,U) \le t\mathsf{P}(E,U) + 2\gamma\mathsf{P}(E, U\setminus B)$ for any open set $U \supset B$.

Proof Since H_+^* is inner regular, we may assume $U \supset B$ to be a bounded open set without loss of generality. We fix $\delta \in (0,1)$ and we consider the family $\mathscr{B}$ of all the closed cubes inside U centered in the points $x \in B$ with one face oriented as $\nu_E(x)$, such that, if we denote their interior by $Q_{\nu_E(x)}(x,r)$, we have

$$H_+^*(E, Q_{\nu_E(x)}(x,r)) \le H_+(E, Q_{\nu_E(x)}(x,r)) \le (1+\delta)t\mathsf{P}(E, Q_{\nu_E(x)}(x,r)),$$

and $|D\chi_E|(\partial Q_{\nu_E(x)}(x,r)) = 0$.

As in the proof of Theorem 3, we can apply the version of Vitali theorem for cubes (see Theorem 5.13 of [15]) and find a disjoint countable subfamily $\{\overline{Q_j}\}$ which covers $\mathscr{H}^{n-1}$-almost all B. It is also clear that, since $\mathsf{P}(E, \overline{Q_j}) = \mathsf{P}(E, Q_j)$, then we have

$$\mathscr{H}^{n-1}(B \setminus \bigcup_j Q_j) = 0. \tag{41}$$

Therefore the subadditivity of $H_+^*(E,\cdot)$ and (39) imply

$$\begin{aligned} H_+^*(E,U) &\le H_+^*(E, \bigcup_{j=1}^N Q_j) + H_+^*(E, U \setminus \bigcup_{j=1}^N (1-\delta)\overline{Q_j}) \\ &\le (1+\delta)t\sum_{j=1}^N \mathsf{P}(E, Q_j) + 2\gamma\mathsf{P}(E, U\setminus \bigcup_{j=1}^N (1-\delta)\overline{Q_j}) \\ &\le (1+\delta)t\mathsf{P}(E,U) + 2\gamma\mathsf{P}(E, U\setminus \bigcup_{j=1}^N (1-\delta)\overline{Q_j}). \end{aligned}$$

Letting first $\delta \to 0$ and then $N \to +\infty$, we obtain

$$\begin{aligned} H_+^*(E,U) &\le t\mathsf{P}(E,U) + 2\gamma\mathsf{P}(E, U\setminus \bigcup_j Q_j) \\ &= t\mathsf{P}(E,U) + 2\gamma\mathsf{P}(E, U\setminus (B\cup \bigcup_j Q_j)) + 2\gamma\mathsf{P}(E, B\setminus \bigcup_j Q_j) \\ &\le t\mathsf{P}(E,U) + 2\gamma\mathsf{P}(E, U\setminus B), \end{aligned}$$

because of (41). □

Remark 1 We notice that, by combining Theorems 2 and 4, for $\mathscr{H}^{n-1}$-a.e. $x \in \mathscr{F}E$ we obtain

$$\varphi(\nu_E(x)) \le \liminf_{r\to 0} \frac{H_-(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \le \limsup_{r\to 0} \frac{H_+(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} \le \varphi(\nu_E(x)),$$

which yields the following equalities:

$$\liminf_{r\to 0} \frac{H_-(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} = \limsup_{r\to 0} \frac{H_+(E, Q_{\nu_E(x)}(x,r))}{\mathsf{P}(E, Q_{\nu_E(x)}(x,r))} = \varphi(\nu_E(x)). \tag{42}$$

3.4 Final Estimates

Now we use the results of the previous section to adapt the classical results concerning differentiation of Radon measures to the nondecreasing set functions $H_\pm(E, \cdot)$.

Theorem 6 *For any set of finite perimeter E in $\mathbb{R}^n$ one has*

$$H_+(E) = H_-(E) = \int_{\mathscr{F}E} \varphi(\nu_E(x))\, d\mathscr{H}^{n-1}(x). \tag{43}$$

Proof We consider first the case of H_-. Then, fixed $t > 1$, we define the Borel sets

$$D_k := \{x \in \mathscr{F}E : \varphi(\nu_E(x)) \in (t^k, t^{k+1}]\}$$

for $k \in \mathbb{Z}$.

For any $\varepsilon_k > 0$ we can find compact sets $K_k \subset D_k$ such that

$$\mathscr{H}^{n-1}(D_k \setminus K_k) < \varepsilon_k. \tag{44}$$

Since this family of compact sets is disjoint, it is then clear that

$$\min_{-J\le k\ne k'\le J} \operatorname{dist}(K_k, K_{k'}) > 0. \qquad \forall J \in \mathbb{N}.$$

Hence, for any J, we can find a disjoint family of open sets $U_k \supset K_k$, for $-J \le k \le J$. By the superadditivity of H_-, Theorem 3 and (42), we get

$$H_-(E) \ge H_-(E, \bigcup_{-J\le k\le J} U_k) \ge \sum_{-J\le k\le J} H_-(E, U_k) \tag{45}$$

$$\begin{aligned}
&\geq \sum_{-J\leq k\leq J} t^k \mathscr{H}^{n-1}(K_k)\\
&\geq \sum_{-J\leq k\leq J} t^{-1}\int_{K_k} \varphi(\nu_E)\, d\mathscr{H}^{n-1}\\
&= t^{-1}\int_{\bigcup_{-J\leq k\leq J} K_k} \varphi(\nu_E)\, d\mathscr{H}^{n-1}
\end{aligned}$$

for any $J \in \mathbb{N}$. Since the measure $\mathscr{H}^{n-1}\llcorner \mathscr{F}E$ is regular and ε_k are arbitrary, we can pass to the supremum to get

$$H_-(E) \geq t^{-1}\int_{\bigcup_{-J\leq k\leq J} D_k} \varphi(\nu_E)\, d\mathscr{H}^{n-1}$$

Finally, we pass to the supremum over J and then send $t \to 1$ to get

$$H_-(E) \geq \int_{\mathscr{F}E} \varphi(\nu_E)\, d\mathscr{H}^{n-1}. \tag{46}$$

Now we deal with H_+. Fixed $t > 1$, we define the Borel sets D_k as above. For $\varepsilon > 0$, we can therefore find open sets $U_k \supset D_k$ with

$$\sum_k t^{k+1}\mathsf{P}(E, U_k \setminus D_k) < \varepsilon, \quad \sum_k 2\gamma\mathsf{P}(E, U_k \setminus D_k) < \varepsilon.$$

Since $\bigcup_k U_k$ covers $\mathscr{H}^{n-1}$-almost all of $\mathscr{F}E$ we can cover $\mathbb{R}^n \setminus \bigcup_k U_k$ with an open set U_0 with $\mathsf{P}(E, U_0)$ arbitrarily small and use the σ-subadditivity of H_+^* and (39) to get

$$H_+^*(E, \mathbb{R}^n) \leq \sum_k H_+^*(E, U_k).$$

Now, using Theorem 5, we estimate

$$\begin{aligned}
H_+^*(E, \mathbb{R}^n) &\leq \sum_k H_+^*(E, U_k)\\
&\leq \sum_k t^{k+1}\mathsf{P}(E, U_k) + 2\gamma\mathsf{P}(E, U_k \setminus D_k)\\
&\leq \sum_k t^{k+1}\mathsf{P}(E, D_k) + 2\varepsilon
\end{aligned}$$

$$\leq \sum_k t \int_{(\mathscr{F}E)\cap D_k} \varphi(\nu_E)\, d\mathscr{H}^{n-1} + 2\varepsilon$$

$$\leq t \int_{\mathscr{F}E} \varphi(\nu_E)\, d\mathscr{H}^{n-1} + 2\varepsilon. \tag{47}$$

Now we let $\varepsilon \downarrow 0$ and $t \downarrow 1$ to get

$$H^*_+(E, \mathbb{R}^n) \leq \int_{\mathscr{F}E} \varphi(\nu_E)\, d\mathscr{H}^{n-1}.$$

We show now that $H^*_+(E, \mathbb{R}^n) = H_+(E, \mathbb{R}^n)$. Indeed, we need only to show $H_+(E, \mathbb{R}^n) \leq H^*_+(E, \mathbb{R}^n)$. Fix $W \Subset \mathbb{R}^n$ open and let Ω such that $W \Subset \Omega$; by (23) we have

$$H_+(E, \mathbb{R}^n) \leq H_+(E, \Omega) + H_+(E, \mathbb{R}^n \setminus \overline{W}),$$

since we can take $\tilde{\Omega}$ and $\tilde{W}$ such that $W \Subset \tilde{W} \Subset \tilde{\Omega} \Subset \Omega$ and write $\mathbb{R}^n = \tilde{\Omega} \cup (\mathbb{R}^n \setminus \tilde{W})$. By (24), we have

$$H_+(E, \mathbb{R}^n) \leq H^*_+(E, \mathbb{R}^n) + 2\gamma \mathsf{P}(E, \mathbb{R}^n \setminus \overline{W}),$$

which implies $H_+(E, \mathbb{R}^n) \leq H^*_+(E, \mathbb{R}^n)$, since W is arbitrary. In this way we obtain the inequality

$$H_+(E) \leq \int_{\mathscr{F}E} \varphi(\nu_E)\, d\mathscr{H}^{n-1}. \tag{48}$$

Combining (46) and (48), we prove the theorem.

Remark 2 (A Local Version of Theorem 6) By similar arguments one can prove that $\mathsf{P}(E, \mathbb{R}^n) < \infty$ implies that the family

$$\mathscr{R} := \left\{ A \subset \mathbb{R}^n : \ A \text{ open}, H_\pm(E, A) = \int_{A\cap\mathscr{F}E} \varphi(\nu_E)\, d\mathscr{H}^{n-1} \right\}$$

is rich, namely the set $\{i \in [0,1] : \ A_i \notin \mathscr{R}\}$ is at most countable whenever the family $\{A_i\}_{i\in[0,1]}$ satisfies $A_i \Subset A_j$ for $i < j$.

Indeed, since the density arguments are local, one need just to start with $H_-(E, A)$ in (45) and with $H^*_+(E, A)$ in (47) and to estimate in a finer way. Then, we recall that $H^*_+(E, \cdot) = H_+(E, \cdot)$ on a rich family of open sets. More specifically, one can use (39) and an argument similar to the last part of the proof of Theorem 6 to prove that any open set $A \subset \mathbb{R}^n$ such that $|D\chi_E|(\partial A) = 0$ belongs to this family.

4 Examples and Variants

In this section we discuss a few examples and estimates of the function φ. We also introduce a variant of the functionals H_ε in which we allow for dilations ηC, for any $\eta \in (0, \varepsilon]$ (i.e. the size of the sets in the family need not be the same).

4.1 Covering with Balls

If we choose the set C to be the unit ball $B(0, 1)$, it is easy to see that the function φ is a constant ξ_n depending only on the space dimension. Indeed, in this case the functionals H_ε and $H_\pm$ are rotationally invariant.

We are also able to estimate ξ_n, see (52) below. A result due to Cianchi [9] shows that we have the following sharp form of the relative linear isoperimetric inequality in the unit ball B:

$$\frac{|E \cap B||B \setminus E|}{|B|^2} \leq \frac{1}{4\omega_{n-1}} \mathsf{P}(E, B) \qquad \text{for any measurable set } E.$$

This inequality clearly gives us the upper bound

$$\xi_n = H_+^B(S_\nu, Q_\nu) \leq \frac{1}{2\omega_{n-1}} \mathsf{P}(S_\nu, Q_\nu) = \frac{1}{2\omega_{n-1}}. \tag{49}$$

On the other hand, the derivation of a lower bound is related to the well-known Kepler's problem (see for instance [13, 17]). This problem, also called "packing problem", consists in looking for the best way to place finite unions of disjoint open balls with the same (small) radius inside a unit cube in $\mathbb{R}^n$ in order to cover as much volume as possible. As the radius tends to 0, this problem identifies the best fraction $\rho_n \in (0, 1]$ of volume covered. Kepler's problem is highly non trivial, since only in 1998 Hales [14] was able to prove that in three dimensions the best packing is the face centered cubic lattice (which is the one used to pack oranges and cannon balls), and that $\rho_3 = \frac{\pi}{3\sqrt{2}}$, as Kepler conjectured. In two dimensions the best packing is the hexagonal lattice and therefore $\rho_2 = \frac{\pi}{2\sqrt{3}}$, as it was proved by Thue in 1890 (see [18] and [19]), while in dimensions higher than 3 the problem is still essentially open. Nevertheless, it is not difficult to prove the existence of the constant ρ_n by standard subadditivity arguments; for the reader's convenience, we include a proof of this fact in the Appendix.

Our aim is to give a lower estimate of the number of disjoint ε-balls which can stay inside Q_ν and are bisected by ∂S_ν. Thus, it is clear that this problem is related to the one of looking for the optimal fraction $\rho_n \in (0, 1]$ of the volume of the n-dimensional unit cube covered by finite unions of disjoint balls with the same radii. We claim that we have

$$\xi_n \geq \frac{\rho_{n-1}}{2\omega_{n-1}}. \tag{50}$$

Indeed, we can cover $\partial S_\nu \cap Q_\nu$ with a number N_ε of $(n-1)$-dimensional ε-balls satisfying

$$N_\varepsilon \sim \rho_{n-1} \frac{1}{\omega_{n-1}\varepsilon^{n-1}}. \tag{51}$$

Such $(n-1)$-dimensional ε-balls can be seen as the sections $\partial S_\nu \cap B'$ for some disjoint n-dimensional ε-balls B' which are bisected by the hyperplane ∂S_ν and lie inside the cube Q_ν. Therefore, we get

$$\xi_n = H^B_-(S_\nu, Q_\nu) \geq \liminf_{\varepsilon \to 0} \varepsilon^{n-1} \frac{1}{2} N_\varepsilon = \frac{\rho_{n-1}}{2\omega_{n-1}}.$$

Combining (49) and (50), we obtain

$$\frac{\rho_{n-1}}{2\omega_{n-1}} \leq \xi_n \leq \frac{1}{2\omega_{n-1}}. \tag{52}$$

In particular, it is easy to see that $\rho_1 = 1$, since the ball centered in the origin of radius r coincides with the cube, being the interval $(-r, r)$. Therefore, we conclude that $\xi_2 = 1/(2\omega_1) = 1/4$.

We notice that we can use the above arguments to estimate φ also in the case when C is the spherical shell $B(0,1) \setminus \overline{B(0,r)}$, for some $r \in (0,1)$.

Indeed, it is clear that φ is a constant $\xi_{n,r}$ depending only on the interior radius and the space dimension, due to the rotational invariance. If we choose the arrangement of disjoint copies of εC which are bisected by ∂S_ν and cover the maximum fraction of surface area, then their number will be the same N_ε as in (51): in fact, $C \cap (\partial S_\nu)$ occupies the same surface area as $B(0,1) \cap (\partial S_\nu)$. Hence, we have

$$\xi_{n,r} \geq \frac{\rho_{n-1}}{2\omega_{n-1}}.$$

On the other hand, it is clear $C \subset B(0,1)$ and that any disjoint family of translations of εC generates a disjoint family of full ε-balls. Hence, the inequalities (14) and (49) imply

$$\xi_{n,r} \leq \frac{|B(0,1)|^2}{|B(0,1) \setminus \overline{B(0,r)}|^2} \frac{1}{2\omega_{n-1}} = \frac{1}{(1-r^n)^2 2\omega_{n-1}}.$$

4.2 Isotropic Coverings

If we redefine H_ε in an isotropic way; that is, allowing for any orientation of the sets C' in the covering, we clearly get the rotational invariance for the modified functionals $H_\varepsilon^{\mathrm{iso}}$ and so the associated function φ^{iso} is a constant $\xi(C)$. This was done in [1] with C equal to the unit cube Q and it is not difficult to show that $\xi(Q) = 1/2$, as Ambrosio, Brezis, Bourgain and Figalli proved. Indeed, by the relative isoperimetric inequality in the unit cube

$$|E|(1-|E|) \le \frac{1}{4}\mathsf{P}(E, Q) \tag{53}$$

for any measurable set $E \subset Q$ (see [1, (2.2)]), we have that

$$\frac{|Q' \cap S_\nu||Q' \setminus S_\nu|}{|Q'|^2} \le \frac{1}{4}\varepsilon^{1-n}\mathsf{P}(S_\nu, Q')$$

for any ε-cube Q'. This gives $H_+^{\mathrm{iso},Q}(S_\nu, Q_\nu) \le \frac{1}{2}\mathsf{P}(S_\nu, Q_\nu) = \frac{1}{2}$.

On the other hand, we can take the ε-cubes with one face oriented as ν, bisected by ∂S_ν and whose intersection with it gives the canonical partition of $\partial S_\nu \cap Q_\nu$ in order to obtain $H_\varepsilon^{\mathrm{iso},Q}(S_\nu, Q_\nu) \ge \frac{1}{2}\varepsilon^{n-1}\left\lfloor \varepsilon^{1-n}\right\rfloor$. This gives the result, coherently with [1].

Actually, using (17) and (53) we have immediately

$$\lim_{\varepsilon\to 0} H_\varepsilon^{\mathrm{iso},Q}(A, \Omega) = \frac{1}{2}\mathsf{P}(A, \Omega) \tag{54}$$

for any measurable set A and any open sets Ω.

It is also possible to show that we obtain a similar result if C is the pluri-rectangle $R = \prod_{j=1}^n (-a_j/2, a_j/2)$, for $a_j > 0$.

Indeed, we can take the copies of εR having one face oriented as ν, bisected by ∂S_ν and whose intersection with it gives a partition of $\partial S_\nu \cap Q_\nu$ with the largest cardinality; that is, at least $\left\lfloor \varepsilon^{1-n}/m\right\rfloor$, where

$$m := \min_{i=1,\dots,n} \prod_{j\ne i} a_j.$$

Thus, we obtain the lower bound $H_\varepsilon^{\mathrm{iso},R}(S_\nu, Q_\nu) \ge \frac{1}{2}\varepsilon^{n-1}\left\lfloor \varepsilon^{1-n}/m\right\rfloor$ and so $\xi(R) \ge \frac{1}{2m}$.

As for the upper bound, by (65) in the following subsection, we have $\xi(R) = 1/(2m) = 1/(2\min\{a_1, a_2\})$ if $n = 2$.

We notice that in these isotropic cases the result of Theorem 1 for sets of finite perimeter follows directly from Theorems 3 and 5 with $B = \mathscr{F}E$. Indeed, these theorems still hold true since $H_\varepsilon^{\mathrm{iso}}$ has the same properties of H_ε.

Then, if we take $t = \xi(C)$, Theorem 3 implies $H^{\mathrm{iso}}_{-}(E, U) \geq \xi(C)\mathscr{H}^{n-1}(\mathscr{F}E)$ for any open set $U \supset \mathscr{F}E$: it follows immediately that $H^{\mathrm{iso}}_{-}(E) \geq \xi(C)\mathsf{P}(E)$.

On the other hand, if we take an open set U containing $\mathscr{F}E$ and an open set $W \Subset U$, the subadditivity of $H^{\mathrm{iso},*}_{+}$ gives

$$H^{\mathrm{iso},*}_{+}(E) \leq H^{\mathrm{iso},*}_{+}(E, U) + H^{\mathrm{iso},*}_{+}(E, \mathbb{R}^n \setminus \overline{W}) \leq H^{\mathrm{iso},*}_{+}(E, U) + 2\gamma\mathsf{P}(E, \mathbb{R}^n \setminus \overline{W})$$

and clearly $\mathsf{P}(E, \mathbb{R}^n \setminus \overline{W}) \downarrow 0$ as $\overline{W} \uparrow U$.

Now, Theorem 5 yields $H^{\mathrm{iso},*}_{+}(E) \leq \xi(C)\mathsf{P}(E, U) = \xi(C)\mathsf{P}(E)$. It suffices now to repeat the same argument at the end of the proof of Theorem 6 in order to obtain $H^{\mathrm{iso}}_{+}(E) \leq \xi(C)\mathsf{P}(E)$, which concludes the proof.

4.3 Examples of Anisotropic Coverings

We present now three examples of estimates of φ in some special cases in $\mathbb{R}^2$.

Let at first C be the unit cube $Q = (0, 1)^n$ in $\mathbb{R}^n$.

In order to evaluate $\varphi^Q(\nu)$, we want to maximize in $x \in \mathbb{R}^n$ for any fixed unit vector ν the function

$$f(x, \nu) := \begin{cases} \frac{|(x+Q)\cap S_\nu||(x+Q)\setminus S_\nu|}{\mathsf{P}(S_\nu,(x+Q))} & \text{if } \mathsf{P}(S_\nu, (x + Q)) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{55}$$

We define then

$$g(\nu) := \sup_{x\in\mathbb{R}^n} f(x, \nu). \tag{56}$$

We observe that g is well defined and that the supremum is a maximum.

Indeed, for any fixed $\nu \in \mathbb{S}^{n-1}$, $f(x, \nu)$ is continuous in x. Clearly, $f(x, \nu) = 0$ if $x \notin \{y : -\sqrt{n} \leq y \cdot \nu \leq \sqrt{n}\}$, and, if $v \cdot \nu = 0$, then $f(x + v, \nu) = f(x, \nu)$. Thus, by symmetry, we can restrict ourselves to a compact set K_ν containing the origin inside the stripe $\{y : -\sqrt{n} \leq y \cdot \nu \leq \sqrt{n}\}$ such that $f(K_\nu, \nu) = f(\mathbb{R}^n, \nu)$. Next, we notice that if we have a sequence $y_k \to x$ with $\mathsf{P}(S_\nu, (y_k + Q)) > 0$ for any k and $\mathsf{P}(S_\nu, (y_k + Q)) \to 0$ as $y_k \to x$, then $\min\{|(y_k + Q) \cap S_\nu|, |(y_k + Q) \setminus S_\nu|\} = o(\mathsf{P}(S_\nu, (y_k + Q)))$, because one of the two parts of $(y_k + Q)$ reduces to a simplex, for k large enough, and so its volume is proportional to the product of the basis area, $\mathsf{P}(S_\nu, (y_k + Q))$, and the relative height, which is going to zero.

Hence, $\sup_{x\in\mathbb{R}^n} f(x, \nu) = \max_{x\in K_\nu} f(x, \nu)$.

By the definition of φ^Q, it follows that $\varphi^Q(\nu) \leq 2g(\nu)$, since

$$\varphi^Q(\nu) = \lim_{\varepsilon\to 0} \varepsilon^{n-1} \sup_{\mathscr{H}_\varepsilon} \sum_{Q'\in\mathscr{H}_\varepsilon} 2\frac{|Q' \cap S_\nu||Q' \setminus S_\nu|}{|Q'|^2}$$

$$\leq \lim_{\varepsilon\to 0}\sup_{\mathscr{H}_\varepsilon} 2g(\nu)\mathsf{P}\left(S_\nu,\ \bigcup_{Q'\in\mathscr{H}_\varepsilon} Q'\right) \leq 2g(\nu)\mathsf{P}(S_\nu, Q_\nu) = 2g(\nu).$$

On the other hand, by symmetry, there exists $\tau \geq 0$ such that $g(\nu) = f(\pm\tau\nu + tv, \nu)$, for any $v \in \mathbb{S}^{n-1}$ orthogonal to ν and $t > 0$. Then, for any ε, we can choose the disjoint family $\mathscr{G}_\varepsilon$ of translations of εQ inside Q_ν which corresponds to a subset of $\{\pm\tau\nu + tv : v \in \mathbb{S}^{n-1}, v\cdot\nu = 0, t > 0\}$ and which covers ∂S_ν up to a set of $\mathscr{H}^{n-1}$-measure going to zero as $\varepsilon \to 0$. The existence of such a family of translation for any fixed $\varepsilon > 0$ follows easily from the fact that one can cover $\mathbb{R}^n$ with a tessellation of open disjoint cubes, up to a Lebesgue negligible set. For such a sequence of families we obtain

$$\varphi^Q(\nu) \geq \lim_{\varepsilon\to 0} 2g(\nu)\mathsf{P}\left(S_\nu,\ \bigcup_{Q'\in\mathscr{G}_\varepsilon} Q'\right) = 2g(\nu).$$

Thus, we conclude that $\varphi^Q(\nu) = 2g(\nu)$.

We consider now the case $n = 2$. By the symmetries of the problem, we can redefine the function f as

$$f(q, m) := \frac{|Q \cap S_m^q||Q \setminus S_m^q|}{\mathsf{P}(S_m^q, Q)}, \tag{57}$$

where $Q = (0, 1)\times(0, 1)$, $S_m^q := \{(x, y) \in \mathbb{R}^2 : y \geq mx + q\}$, $q \in [-m, 1]$, $m = -(\nu_1/\nu_2) \in [0, +\infty)$. It is enough now to distinguish between the cases $0 \leq m \leq 1$ and $m \geq 1$.

If $0 \leq m \leq 1$, then we need only to consider $q \in [0, 1]$. The line $\{y = mx + q\}$ intersects the edges of Q in the points $A = (0, q)$ and

$$B = \begin{cases} (1, m + q) & \text{if } 0 \leq q \leq 1 - m \\ \left(\frac{1-q}{m}, 1\right) & \text{if } 1 - m \leq q \leq 1 \end{cases}.$$

Hence, we have

$$f(q, m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \left(q + \frac{m}{2}\right) - \left(q + \frac{m}{2}\right)^2 & \text{if } 0 \leq q \leq 1 - m \\ \frac{1-q}{2} - \frac{(1-q)^3}{4m} & \text{if } 1 - m \leq q \leq 1 \end{cases}. \tag{58}$$

It is easy to see that, for any fixed m, the partial derivative in q is

$$\frac{\partial f}{\partial q}(q, m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} 1 - 2q - m & \text{if } 0 < q < 1 - m \\ -\frac{1}{2} + \frac{3}{4m}(1-q)^2 & \text{if } 1 - m < q < 1 \end{cases}.$$

Hence, $\frac{\partial f}{\partial q} \geq 0$ if and only if

$$\begin{cases} q \leq \frac{1-m}{2} & \text{if } 0 < q < 1-m \\ q \leq 1 - \sqrt{\frac{2m}{3}} & \text{if } 1-m < q < 1 \end{cases},$$

which means that

$$\max_{q \in [0,1-m]} f(q,m) = \frac{1}{4\sqrt{1+m^2}}$$

and

$$\max_{q \in [1-m,1]} f(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \frac{1}{3}\sqrt{\frac{2m}{3}} & \text{if } \frac{2}{3} < m \leq 1 \\ \frac{m}{2}\left(1 - \frac{m}{2}\right) & \text{if } 0 \leq m \leq \frac{2}{3} \end{cases}.$$

Since

$$\frac{m}{2}\left(1 - \frac{m}{2}\right) \leq \frac{1}{4}$$

for any $m \in [0, 1]$ and

$$\frac{1}{3}\sqrt{\frac{2m}{3}} \leq \frac{1}{4}$$

only for $0 \leq m \leq (27/32)$, it follows that

$$\max_{q \in [0,1]} f(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \frac{1}{3}\sqrt{\frac{2m}{3}} & \text{if } \frac{27}{32} \leq m \leq 1 \\ \frac{1}{4} & \text{if } 0 \leq m \leq \frac{27}{32} \end{cases}. \tag{59}$$

If $m > 1$, we need only to consider $q \in [1-m, 1]$ and the intersections are

$$A = \begin{cases} (0, q) & \text{if } 0 \leq q \leq 1 \\ \left(-\frac{q}{m}, 0\right) & \text{if } 1-m \leq q \leq 0 \end{cases}$$

and $B = \left(\frac{1-q}{m}, 1\right)$. Hence, we have

$$f(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \frac{1-q}{2} - \frac{(1-q)^3}{4m} & \text{if } 0 \leq q \leq 1 \\ \left(\frac{1}{2} - q\right) - \frac{1}{m}\left(\frac{1}{2} - q\right)^2 & \text{if } 1-m \leq q \leq 0 \end{cases}. \tag{60}$$

We have that, for any fixed m, the partial derivative in q is

$$\frac{\partial f}{\partial q}(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} -\frac{1}{2} + \frac{3}{4m}(1-q)^2 & \text{if } 0 < q < 1 \\ -1 + \frac{2}{m}\left(\frac{1}{2} - q\right) & \text{if } 1-m < q < 0 \end{cases}.$$

Hence, $\frac{\partial f}{\partial q} \geq 0$ if and only if

$$\begin{cases} q \leq 1 - \sqrt{\frac{2m}{3}} & \text{if } 0 < q < 1 \\ q \leq \frac{1-m}{2} & \text{if } 1-m < q < 0 \end{cases},$$

which means that

$$\max_{q \in [1-m,0]} f(q,m) = \frac{m}{4\sqrt{1+m^2}}$$

and

$$\max_{q \in [0,1]} f(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \frac{1}{3}\sqrt{\frac{2m}{3}} & \text{if } 1 < m < \frac{3}{2} \\ \frac{2m-1}{4m} & \text{if } m \geq \frac{3}{2} \end{cases}.$$

Since

$$\frac{2m-1}{m} \leq m$$

for any $m > 1$ and

$$\frac{1}{3}\sqrt{\frac{2m}{3}} \leq \frac{m}{4}$$

only for $1 < m \leq (32/27)$, it follows that

$$\max_{q \in [1-m,1]} f(q,m) = \frac{1}{\sqrt{1+m^2}} \begin{cases} \frac{1}{3}\sqrt{\frac{2m}{3}} & \text{if } 1 < m \leq \frac{32}{27} \\ \frac{m}{4} & \text{if } m \geq \frac{32}{27} \end{cases}. \tag{61}$$

Because of the symmetry of the cube, we can conclude that

$$g(\nu) = \begin{cases} \frac{|\nu_2|}{4} & \text{if } |\nu_1| \le \frac{27}{32}|\nu_2| \\ \frac{1}{3}\sqrt{\frac{2}{3}|\nu_1||\nu_2|} & \text{if } \frac{27}{32}|\nu_2| \le |\nu_1| \le \frac{32}{27}|\nu_2| \\ \frac{|\nu_1|}{4} & \text{if } |\nu_1| \ge \frac{32}{27}|\nu_2| \end{cases},$$

which means

$$g(\nu) = \begin{cases} \frac{1}{3}\sqrt{\frac{2}{3}|\nu_1||\nu_2|} & \text{if } \frac{27}{32}|\nu_2| \le |\nu_1| \le \frac{32}{27}|\nu_2| \\ \frac{\|\nu\|_\infty}{4} & \text{if } |\nu_1| \le \frac{27}{32}|\nu_2| \text{ or } |\nu_1| \ge \frac{32}{27}|\nu_2| \end{cases} \tag{62}$$

and so

$$\varphi^Q(\nu) = \begin{cases} \frac{2}{3}\sqrt{\frac{2}{3}|\nu_1||\nu_2|} & \text{if } \frac{27}{32}|\nu_2| \le |\nu_1| \le \frac{32}{27}|\nu_2| \\ \frac{\|\nu\|_\infty}{2} & \text{if } |\nu_1| \le \frac{27}{32}|\nu_2| \text{ or } |\nu_1| \ge \frac{32}{27}|\nu_2| \end{cases}.$$

It is clear that its 1-homogeneous extension $\Phi^Q(x, y) := \sqrt{x^2+y^2}\varphi^Q\left(\frac{(x,y)}{\sqrt{x^2+y^2}}\right)$ is indeed $\varphi^Q(x)$ and that it is not convex in the region $\{\frac{27}{32}|y| \le |x| \le \frac{32}{27}|y|\}$.

We also notice that $\max_{\nu\in\mathbb{S}^1} \varphi^Q(\nu) = (1/2)$, coherently with the results of [1] in the isotropic case.

Let us now consider anisotropic coverings with rectangles. Let $R = \prod_{i=1}^n (0, a_i)$, $a_i > 0$, then we can argue as before to show that

$$\varphi^R(\nu) = 2 \sup_{x\in\mathbb{R}^n} \frac{|(x+R)\cap S_\nu||(x+R)\setminus S_\nu|}{|R|^2 \mathsf{P}(R_\nu, (x+R))}.$$

If $n = 2$, we can work with $R_\lambda = (0, 1)\times(0, \lambda)$, $\lambda > 0$, since for a generic rectangle $R = (0, a)\times(0, b)$, we have $R = aR_\lambda$ if $\lambda = (b/a)$, and so $\varphi^R(\nu) = (1/a)\varphi^{R_\lambda}(\nu)$.

In order to deal with the explicit calculation, we can proceed in a similar way as before, by considering the function

$$f_\lambda(q, m) := \frac{|R_\lambda \cap S_m^q||R_\lambda \setminus S_m^q|}{|R_\lambda|^2 \mathsf{P}(S_m^q, R_\lambda)}, \tag{63}$$

where $S_m^q := \{(x, y) \in \mathbb{R}^2 : y \ge mx+q\}$, $q \in [-m, \lambda]$, $m = -(\nu_1/\nu_2) \in [0, +\infty)$, and dividing in the two cases $0 \le m \le \lambda$ and $m \ge \lambda$. Then, it is not difficult to show that we have

$$g^{R_\lambda}(\nu) = \begin{cases} \frac{|\nu_2|}{4} & \text{if } |\nu_1| \le \frac{27}{32}\lambda|\nu_2| \\ \frac{1}{3}\sqrt{\frac{2}{3\lambda}|\nu_1||\nu_2|} & \text{if } \frac{27}{32}\lambda|\nu_2| \le |\nu_1| \le \frac{32}{27}\lambda|\nu_2| \\ \frac{|\nu_1|}{4\lambda} & \text{if } |\nu_1| \ge \frac{32}{27}\lambda|\nu_2| \end{cases}. \tag{64}$$

Since $\varphi^{R_\lambda}(\nu) = 2g^{R_\lambda}(\nu)$, then neither the 1-homogeneous extension of this function is convex.

In conclusion, for the rectangle $R = (0, a) \times (0, b)$ we have

$$\varphi^{R}(\nu) = \begin{cases} \frac{|\nu_2|}{2a} & \text{if } a|\nu_1| \le \frac{27}{32}b|\nu_2| \\ \frac{2}{3}\sqrt{\frac{2}{3ab}|\nu_1||\nu_2|} & \text{if } \frac{27}{32}b|\nu_2| \le a|\nu_1| \le \frac{32}{27}b|\nu_2| \\ \frac{|\nu_1|}{2b} & \text{if } a|\nu_1| \ge \frac{32}{27}b|\nu_2| \end{cases}.$$

It is also easy to see that

$$\max_{\nu\in\mathbb{S}^1} \varphi^{R}(\nu) = \frac{1}{2\min\{a, b\}}, \tag{65}$$

which gives the value of the constant $\xi(R)$ in the case of isotropic coverings with rectangles, when arbitrary rotations are allowed.

As a last example, let now C be the ellipse $E = \{(x, y) : (x/a)^2 + (y/b)^2 < 1\}$, for some $a,\ b > 0$.

In order to estimate φ from below, we choose the arrangement of copies of εE such that each one is bisected by the line ∂S_ν and the contiguous copies are tangent in the intersection between their boundaries and ∂S_ν. Hence, we need to evaluate the length of the segment intersected by a copy of εE on the line $\partial S_\nu = \{(x, y)\cdot\nu = 0\}$.

If $m = -\nu_1/\nu_2$, then $\partial S_\nu = \{y = mx\}$ and the intersections with ∂E are the points $\pm\frac{ab}{\sqrt{b^2 + m^2a^2}}(1, m)$. Therefore,[2] the length of the segment intersected is $2\varepsilon ab\sqrt{(1 + m^2)/(b^2 + m^2a^2)}$. Since the copies of εE need to cover the unitary segment $(\partial S_\nu) \cap Q_\nu$, we obtain

$$\varphi^{E}(\nu) \ge \lim_{\varepsilon\to 0} \frac{1}{2}\varepsilon \left\lfloor \frac{\sqrt{b^2 + m^2a^2}}{\varepsilon 2ab\sqrt{1 + m^2}} \right\rfloor = \frac{1}{4ab}\sqrt{b^2\nu_2^2 + a^2\nu_1^2}.$$

In particular, if $a = b = 1$, we get $\varphi^{B}(\nu) \ge (1/4)$, coherently with (52).

[2] If $\nu_2 = 0$, the length is $2\varepsilon b$.

4.4 A Variant

One may define a family of functionals similar to H_ε allowing for different dilations of the set C under a fixed level $\varepsilon > 0$. More specifically, we set

$$\tilde{H}_\varepsilon(A, \Omega) := \sup_{\mathscr{G}_\varepsilon} \sum_{C' \in \mathscr{G}_\varepsilon} 2(\varepsilon(C'))^{n-1} \frac{|C' \cap A||C' \setminus A|}{|C'|^2}, \tag{66}$$

where $C' = \varepsilon(C')(C + a)$, for some translation vector a, and $\mathscr{G}_\varepsilon$ is a disjoint family inside Ω of translations of the set ηC, for any $\eta \in (0, \varepsilon]$. It is clear that $H_\varepsilon(A, \Omega) \leq \tilde{H}_\varepsilon(A, \Omega)$, and so (6) follows for $\tilde{H}_\varepsilon$ in the case $\mathsf{P}(A) = +\infty$.

We can define the functionals $\tilde{H}_\pm$ as liminf and limsup of $\tilde{H}_\varepsilon$. It is also not difficult to see that $\tilde{H}_\varepsilon$ and $\tilde{H}_\pm$ satisfy the same elementary properties of H_ε shown in Sect. 3.2.

For instance, the homogeneity $\tilde{H}_{t\varepsilon}(tA, t\Omega) = t^{n-1}\tilde{H}_\varepsilon(A, \Omega)$ follows from the fact that each set $C'' \in \mathscr{G}_{t\varepsilon}$ can be seen as $C'' = tC'$, with $C' = \varepsilon(C')C$, for $\varepsilon(C') \leq \varepsilon$, and so $C' \in \mathscr{G}_\varepsilon$.

Since these functionals satisfy the same properties of H_ε and $H_\pm$, we can define the functions $\tilde{\varphi}_\pm(\nu) := \tilde{H}_\pm(S_\nu, Q_\nu)$ and show an analogous version of Proposition 1 for them.

Then, one may follow the same steps in order to prove Theorem 1 for $\tilde{H}_\varepsilon$ in the rectifiable case. Thus, we obtain that for any set of finite perimeter A

$$\lim_{\varepsilon \to 0} \tilde{H}_\varepsilon(A) = \int_{\mathscr{F}A} \tilde{\varphi}(\nu_A)\, d\mathscr{H}^{n-1}.$$

Let us now consider the case in which the set C is the unit ball. Then $\tilde{\varphi}$ is a constant, since $\tilde{H}_\varepsilon$ is rotation invariant, and $\tilde{\varphi} \equiv 1/(2\omega_{n-1})$, since arbitrarily small radii are allowed.

Indeed, the upper estimate is given by (49). On the other hand, we notice that we can find a lower bound by considering only the family of balls which are bisected by the hyperplane ∂S_ν.

For any fixed $\varepsilon > 0$, we can apply Vitali-Besicovitch Theorem (Theorem 2.19 in [3]) to the measure $\mu = \mathscr{H}^{n-1} \llcorner Q''$, where Q'' is a unit cube in $\mathbb{R}^{n-1}$ and to a fine cover of balls with radii smaller than ε, in order to find a disjoint family $\mathscr{G}_{\varepsilon,(n-1)}$ of $(n-1)$-dimensional balls with radii smaller than ε such that

$$\mathscr{H}^{n-1}\left(Q'' \setminus \bigcup_{B'' \in \mathscr{G}_{\varepsilon,(n-1)}} B''\right) = 0. \tag{67}$$

Hence, we can take the family $\mathscr{G}_\varepsilon$ of n-dimensional balls bisected by $(\partial S_\nu) \cap Q_\nu$ and whose intersections with it generate the family $\mathscr{G}_{\varepsilon,(n-1)}$. Then, we use (67) to obtain

$$\begin{aligned}\tilde{\varphi}(\nu) &\geq \lim_{\varepsilon\to 0}\frac{1}{2}\sum_{B'\in\mathcal{G}_\varepsilon}\varepsilon^{n-1}\\ &= \lim_{\varepsilon\to 0}\frac{1}{2\omega_{n-1}}\sum_{B''\in\mathcal{G}_{\varepsilon,(n-1)}}\omega_{n-1}\varepsilon^{n-1}\\ &= \frac{1}{2\omega_{n-1}}\mathcal{H}^{n-1}((\partial S_\nu)\cap Q_\nu) = \frac{1}{2\omega_{n-1}}.\end{aligned}$$

Finally, we observe that if we redefine $\tilde{H}_\varepsilon$ allowing for the possibility to rotate the sets C' in the covering, we obtain that $\tilde{\varphi}$ is a constant, as it happens for H_ε.

In particular, if we take C to be the unit cube Q as in [1], then, by (54) and (53), we have

$$\frac{1}{2}\mathsf{P}(A,\Omega) = \lim_{\varepsilon\to 0} H_\varepsilon^{\mathrm{iso},Q}(A,\Omega) \leq \lim_{\varepsilon\to 0}\tilde{H}_\varepsilon^{\mathrm{iso},Q}(A,\Omega) \leq \frac{1}{2}\mathsf{P}(A,\Omega),$$

which gives $\lim_{\varepsilon\to 0}\tilde{H}_\varepsilon^{\mathrm{iso},Q}(A,\Omega) = \frac{1}{2}\mathsf{P}(A,\Omega)$ for any measurable set A and open set Ω.

Appendix: Proof of the Existence of ρ_n

For any $\varepsilon > 0$ and for any open set $U \subset \mathbb{R}^n$ we denote by $\Lambda_\varepsilon(U)$ the supremum of the Lebesgue measure of the finite unions of disjoint ε-balls inside U. In a similar way as we did in the beginning of Sect. 3, we list now some elementary properties of the set functions Λ_ε:

1. translation invariance and monotonicity;
2. homogeneity:

$$\Lambda_{t\varepsilon}(tU) = t^n\Lambda_\varepsilon(U) \qquad t > 0; \tag{68}$$

3. superadditivity: for any open sets $U,\ V$ such that $U\cap V = \emptyset$, one has

$$\Lambda_\varepsilon(U\cup V) \geq \Lambda_\varepsilon(U) + \Lambda_\varepsilon(V);$$

4. almost subadditivity: for any open sets $U,\ V$,

$$\Lambda_\varepsilon(U\cup V) \leq \Lambda_\varepsilon(I_{2\varepsilon}(U)) + \Lambda_\varepsilon(V); \tag{69}$$

5. upper and lower bounds:

$$0 \le \Lambda_\varepsilon(U) \le |U|. \tag{70}$$

For $U \subset \mathbb{R}^n$ open, we then set

$$\Lambda_-(U) := \liminf_{\varepsilon \downarrow 0} \Lambda_\varepsilon(U), \qquad \Lambda_+(U) := \limsup_{\varepsilon \downarrow 0} \Lambda_\varepsilon(U).$$

These set functions inherit from Λ_ε the translation invariance, the monotonicity, the homogeneity. In addition Λ_- is superadditive and subadditivity for Λ_+ holds in this form: for any bounded open sets $U,\ V$,

$$\Lambda_+(U \cup V) \le \Lambda_+(W) + \Lambda_+(V), \text{ whenever } U \Subset W. \tag{71}$$

We have also the upper and lower bounds

$$0 \le \Lambda_-(U) \le \Lambda_+(U) \le |U|.$$

Theorem 7 *There exists a constant $\rho_n \in [\omega_n/2^n, 1]$ such that*

$$\Lambda_+(U) = \Lambda_-(U) = \rho_n |U|.$$

Proof We set $\rho_\pm := \Lambda_\pm(Q)$, where $Q = [0,1]^n$. Because of the translation invariance, we see that $\rho_\pm = \Lambda_\pm(Q)$ for any unit cube Q in $\mathbb{R}^n$ and, using the homogeneity, we have $\Lambda_\pm(Q(x,r)) = \rho_\pm r^n$ for any $x \in \mathbb{R}^n$ and any $r > 0$. If we fix an open set U with finite measure and $\delta \in (0,1)$, there exists a finite disjoint family of open cubes $\{Q_i\}_{1 \le i \le N}$ inside U which covers a fraction larger or equal to $(1-\delta)$ of the volume of U; that is, $|U \setminus \bigcup_{i=1}^N Q_i| \le \delta |U|$. Then, monotonicity and superadditivity yield

$$\Lambda_-(U) \ge \Lambda_-\Big(\bigcup_{i=1}^N Q_i\Big) \ge \sum_{i=1}^N \Lambda_-(Q_i) = \rho_- \sum_{i=1}^N |Q_i| \ge \rho_-(1-\delta)|U|.$$

Therefore, since U and δ are arbitrary, by monotonicity and inner approximation we see that $\Lambda_-(U) \ge \rho_- |U|$.

We take now U open with finite measure, $\delta > 0$ and $\{Q_i\}_{i=1,\dots,N}$ as above. In addition, let Q_i' be cubes concentric with Q_i, satisfying $Q_i \Subset Q_i'$. Since we have $I_r(\bigcup_{i=1}^N Q_i) \subset \bigcup_{i=1}^N I_r(Q_i)$, applying (69), for ε small enough we obtain

$$\Lambda_\varepsilon(U) \le \Lambda_\varepsilon\Big(U \setminus \bigcup_{i=1}^N (1-\varepsilon)\overline{Q_i}\Big) + \Lambda_\varepsilon\Big(I_{2\varepsilon}\Big(\bigcup_{i=1}^N Q_i\Big)\Big)$$

$$\leq \beta(\varepsilon)\delta|U| + \sum_{i=1}^{N} \Lambda_\varepsilon(I_{2\varepsilon}(Q_i))$$

$$\leq \beta(\varepsilon)\delta|U| + \sum_{i=1}^{N} \Lambda_\varepsilon(Q_i'),$$

where $\beta(\varepsilon) \downarrow 1$ as $\varepsilon \downarrow 0$. Taking the limsup yields

$$\Lambda_+(U) \leq \delta|U| + \rho_+ \sum_{i=1}^{N} |Q_i'|$$

and if we let first Q_i' tend to Q_i and then $\delta \downarrow 0$, we deduce $\Lambda_+(U) \leq \rho_+|U|$.

Therefore, to complete the proof it is enough to show that $\rho_+ \leq \rho_-$. Let r_i and ε_i be infinitesimal sequences for which

$$\lim_{i\to\infty} \Lambda_{\varepsilon_i}(Q) = \rho_+, \qquad \lim_{i\to\infty} \Lambda_{r_i}(Q) = \rho_-,$$

where $Q = [0,1]^n$. Up to subsequences, we may assume that $\varepsilon_i = o(r_i)$ and we set $\lambda_i = \varepsilon_i/r_i \to 0$. We can cover Q with $k_i = [\lambda_i^{-n}]$ disjoint subcubes Q_j with side length λ_i, up to a set of measure less or equal to $\eta_i := 1 - k_i\lambda_i^n$. Since $I_s(tQ) = tI_{s/t}(Q)$, properties (69), (70) and homogeneity imply

$$\Lambda_{\varepsilon_i}(Q) \leq \sum_{j=1}^{k_i} \Lambda_{\varepsilon_i}(I_{2\varepsilon_i}(Q_j)) + \eta_i \leq \lambda_i^{-n}\Lambda_{r_i\lambda_i}(\lambda_i I_{2r_i}(Q)) + \eta_i = \Lambda_{r_i}(I_{2r_i}(Q)) + \eta_i.$$

We take now a cube Q' such that $Q \Subset Q'$. Then, for i large enough we have

$$\Lambda_{\varepsilon_i}(Q) \leq \Lambda_{r_i}(Q') + \eta_i$$

and passing to the limit we obtain $\rho_+ \leq \rho_-|Q'|$ for any such cube Q'. Hence, it follows that $\rho_+ \leq \rho_-$; that is, $\rho_+ = \rho_- =: \rho_n \in [0,1]$. This means that $\Lambda_\varepsilon(U) \to \rho_n|U|$ for any open set $U \subset \mathbb{R}^n$.

If $\rho_\pm$ were defined using cubes instead of balls, we would obviously get $\rho_n = 1$. Comparing cubes with balls we get $\rho_n \geq \omega_n/2^n$. □

Acknowledgements The authors would like to thank Emanuele Paolini for fruitful discussions concerning the examples of anisotropic coverings (Sect. 4.3).

References

1. L. Ambrosio, H. Brezis, J. Bourgain, A. Figalli, BMO-type norms related to the perimeter of sets. Comm. Pure Appl. Math. **69**, 1062–1086 (2016)
2. L. Ambrosio, H. Brezis, J. Bourgain, A. Figalli, Perimeter of sets and BMO-type norms. C.R. Math. CRAS **352**, 697–698 (2014)
3. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems* (Oxford University Press, Oxford, 2000)
4. J. Bourgain, H. Brezis, P. Mironescu, Another look at Sobolev spaces, in *Optimal Control and Partial Differential Equations* ed. by J.L. Menaldi, E. Rofman, A. Sulem. A Volume in Honour of A. Bensoussan's 60th Birthday (IOS Press, Amsterdam, 2001), pp. 439–455
5. J. Bourgain, H. Brezis, P. Mironescu, A new function space and applications. J. Eur. Math. Soc. **17**, 2083–2101 (2015)
6. H. Brezis, H.M. Nguyen, Two subtle convex nonlocal approximation of the BV norm. Nonlinear Anal. **137**, 222–245 (2016)
7. H. Brezis, H.M. Nguyen, Non-local functionals related to the total variation and connections with image processing. Annals of PDE 4.1. **9** (2016)
8. L. Caffarelli, J.M. Roquejoffre, O. Savin, Nonlocal minimal surfaces. Commun. Pure Appl. Math. **63**, 1111–1144 (2010)
9. A. Cianchi, A sharp form of Poincaré type inequalities on balls and spheres. Z. Angew. Math. Phys. **40**(4), 558–569 (1989)
10. G. Dal Maso, *An Introduction to Γ-Convergence* (Birkhauser, Boston, 1993)
11. N. Fusco, C. Sbordone, G. Moscariello, A formula for the total variation of SBV functions. J. Funct. Anal. **270**, 419–446 (2016)
12. N. Fusco, C. Sbordone, G. Moscariello, BMO-type seminorms and Sobolev functions. ESAIM: Control, Optimisation and Calculus of Variations (2017)
13. T.C. Hales, Cannonballs and honeycombs. Not. Am. Math. Soc. **47**, 440–449 (2000)
14. T.C. Hales, A proof of the Kepler conjecture. Ann. Math. **162**, 1065–1185 (2005)
15. A.P. Morse, Perfect blankets. Trans. Am. Math. Soc. **61**, 418–442 (1947)
16. H.M. Nguyen, Γ-convergence, Sobolev norms, and BV functions. Duke Math. J. **157**, 495–533 (2011)
17. G.C. Szpiro, *Kepler's Conjecture* (Wiley, Hoboken, 2003)
18. A. Thue, *Über die dichteste Zusammenstellung von kongruenten Kreisen in einer Ebene*. Videnskabs-Selskabets Skrifter. I. Math.-Naturv. Klasse, No. 1 (Stanford University, Stanford, 1910), pp. 1–9
19. L.F. Toth, Über die dichteste Kugellagerung. Math. Z. **48**, 676–684 (1943)

Failure of the Chain Rule in the Non Steady Two-Dimensional Setting



Stefano Bianchini and Paolo Bonicatto

1 Introduction

In this paper we consider a variant of the classical problem of the chain rule for the divergence of a bounded vector field. Specifically, the classical problem of the chain rule reads as follows:

Problem 1 (Chain Rule) Let $d \geq 2$ and assume that it is given a bounded, Borel vector field $\mathbf{b}\colon \mathbb{R}^d \to \mathbb{R}^d$, a bounded, scalar function $u\colon \mathbb{R}^d \to \mathbb{R}$ and Radon measures $\lambda, \mu \in \mathscr{M}(\mathbb{R}^d)$ such that

$$\operatorname{div} \mathbf{b} = \lambda, \tag{1a}$$

$$\operatorname{div}(u\mathbf{b}) = \mu, \tag{1b}$$

in the sense of distributions on $\mathbb{R}^d$. Characterize (compute) the distribution

$$\nu := \operatorname{div}(\beta(u)\mathbf{b}),$$

where $\beta\colon \mathbb{R} \to \mathbb{R}$ is a fixed C^1 function.
In the smooth setting one can use the standard chain rule formula to get

$$\begin{aligned} \nu = \operatorname{div}(\beta(u)\mathbf{b}) &= \beta'(u) \operatorname{div}(u\mathbf{b}) + (\beta(u) - u\beta'(u)) \operatorname{div} \mathbf{b} \\ &= \beta'(u)\mu + (\beta(u) - u\beta'(u))\lambda. \end{aligned} \tag{2}$$

S. Bianchini (✉) · P. Bonicatto
S.I.S.S.A., Trieste, Italy
e-mail: bianchin@sissa.it; paolo.bonicatto@sissa.it

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_2

The extension of (2) to a non-smooth setting is far from being trivial and this is exactly the aim of the chain rule problem.

As noted in [4], if one replaces "divergence" by "derivative", the problem boils down to the one of writing a chain rule for weakly differentiable functions (a theme that has been investigated in several papers, see e.g. [5, 16] for the BV setting). However, the "divergence" problem seems to be harder than the "derivative" one, due to stronger cancellation effects.
Problem 1 arises naturally in the study of partial differential equations, like the transport equation, the continuity equation or, more generally, hyperbolic conservations laws: indeed, they all can be written in the form $\operatorname{div}(uB) = c$, where $B : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}^d$ is vector field which has a space-time structure and $c \in \mathscr{D}'(\mathbb{R}^d)$ is some distribution. For instance, considering Problem 1 for a particular choice of B and β, one can establish uniqueness and comparison principles for weak solutions of scalar conservation laws (in the spirit of Kružkov's theory, see [15]).

1.1 Positive Results

If we assume Sobolev regularity on the vector field, i.e. $\mathbf{b} \in W^{1,p}_{\mathrm{loc}}(\mathbb{R}^d)$ and $u \in L^q_{\mathrm{loc}}(\mathbb{R}^d)$ with p, q dual exponents, the chain rule has been established in [13]. In this case, it turns out that ν can be computed in terms of λ and μ just as in the classical (smooth) setting: it holds

$$\nu = (\beta(u) - u\beta'(u))\lambda + \beta'(u)\mu,$$

provided $\mu = \operatorname{div}(u\mathbf{b})$ is absolutely continuous with respect to the Lebesgue measure $\mathscr{L}^d$ on $\mathbb{R}^d$. This result has been extended in [4] to the case when $\mathbf{b} \in \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^d)$, $u \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$ and both $\operatorname{div}\mathbf{b}$, $\operatorname{div}(u\mathbf{b})$ are absolutely continuous with respect to $\mathscr{L}^d$. More precisely, using the commutator estimate due to Ambrosio [7], in [4] it is first proved that the distribution $\nu = \operatorname{div}(\beta(u)\mathbf{b})$ is a Radon measure which satisfies $\nu \ll |\lambda| + |\mu|$. Furthermore, the authors decompose λ, μ, ν into three parts (the *absolutely continuous* part λ^a, the *jump part* λ^j and the *Cantor part* λ^c, as in the standard BV setting) and treat them separately. They obtained that:

- the **absolutely continuous** part behaves as in the Sobolev case:

$$\nu^a = (\beta(u) - u\beta'(u))\lambda^a + \beta'(u)\mu^a, \qquad \text{as measures on } \mathbb{R}^d.$$

- For the **jump part**, they use the results obtained in [3] to prove that ν^j can be computed in terms of the *traces* u^+ and u^- of u on the (countably) rectifiable set Σ where λ^j and μ^j are concentrated on.
- The **Cantor part** is harder and it is not characterized completely in [4], but only up to an error term. More precisely, they proved

$$v^c = (\beta(\widetilde{u}) - \widetilde{u}\beta'(\widetilde{u}))\lambda^c \llcorner_{\Omega \setminus S_u} + \beta'(\widetilde{u})\mu^c \llcorner_{\Omega \setminus S_u} + \sigma$$

where $\widetilde{u}$ is the L^1 approximately continuous representative of u, S_u is the set of points where the L^1 approximate limit does not exist and σ is an error term (which is a measure concentrated on S_u, with $\sigma \ll \lambda^c + \mu^c$).

Further results in this directions have been obtained in [11], where the problem is completely solved in the case $d = 2$ with $\mathbf{b}$ of bounded variation, and in the recent preprint [9] where the analysis is completed (in the BV setting) for every $d \geq 2$.

1.2 Negative Results

If we assume no regularity on $\mathbf{b}$ and u apart from measurability and boundedness, it can happen that λ and μ give no information about v. This is related to the so-called problem of (non) *locality of the divergence operator*: indeed, in [1] the authors constructed an example of a bounded vector field $\mathbf{v}$ (defined in $\mathbb{R}^2$) such that $\operatorname{div} \mathbf{v} \neq 0$, $\operatorname{div} \mathbf{v} \in L^\infty$ and $\operatorname{div} \mathbf{v}$ is supported on the set where $\mathbf{v}$ vanishes. Notice that this phaenomenon cannot occur for distributional derivatives, as they enjoy locality properties [6, Prop. 3.73].
In the same spirit, in the recent work [12], using the abstract machinery of convex integration, the authors construct several examples of vector fields $\mathbf{b}\colon \mathbb{R}^d \to \mathbb{R}^d$ and densities $u\colon \mathbb{R}^d \to \mathbb{R}$ such that $\lambda = 0$, $\mu = 0$ but $\operatorname{div}(u^2\mathbf{b}) \neq 0$ in the sense of distributions in $\mathbb{R}^d$ for $d \geq 3$. More precisely, they show the following

Theorem 1 ([12]) *Let $d \geq 3$ and $\Omega \subset \mathbb{R}^d$ a smooth domain. Let f be a distribution such that the equation $\operatorname{div} \mathbf{w} = f$ admits a bounded, continuous solution $\mathbf{w}\colon \Omega \to \mathbb{R}^d$ on Ω. Then there exists a bounded vector field $\mathbf{b} \in L^\infty(\Omega; \mathbb{R}^d)$ and a density $u\colon \mathbb{R}^d \to \mathbb{R}$, with $0 < C^{-1} \leq u \leq C$ a.e. for some constant $C > 0$, such that*

$$\operatorname{div} \mathbf{b} = 0$$
$$\operatorname{div}(u\mathbf{b}) = 0$$
$$\operatorname{div}\left(u^2\mathbf{b}\right) = f$$

in the sense of distributions in Ω.

1.3 The Two-Dimensional Case

The aim of this note is to address the two-dimensional case, i.e. $d = 2$. Notice that the assumption $d \geq 3$ is essential in [12], in view of the result of [11]. More precisely, in [11], the authors proved that if $d = 2$, $\mathbf{b}$ is bounded and of class BV and $u\colon \mathbb{R}^d \to \mathbb{R}$, with $0 < C^{-1} \leq u \leq C$ a.e. for some constant $C > 0$, are such that

$$\operatorname{div} \mathbf{b} = 0$$
$$\operatorname{div}(u\mathbf{b}) = 0$$

then the Chain rule property holds, i.e. we have necessarily $\operatorname{div}\left(u^2\mathbf{b}\right) = 0$. Actually, the same conclusion is true if the assumption $\mathbf{b} \in \mathrm{BV}$ is replaced by $\mathbf{b} \neq 0$ a.e. in Ω.
However, still remaining in the planar setting, in view of the results obtained in [10], it seems reasonable to consider the chain rule also in the non steady setting, i.e. assuming that the vector field has a (special) space-time structure (and letting the divergence operator acting also on the time variable). More precisely, we are led to consider the following variant of Problem 1:

Problem 2 (Non Steady Chain Rule) Let $T > 0$ be fixed and assume that it is given a bounded, Borel vector field $\mathbf{b}\colon \mathbb{R}^d \to \mathbb{R}^d$, a bounded, scalar function $u\colon (0,T) \times \mathbb{R}^d \to \mathbb{R}$ and Radon measures λ and μ such that

$$\operatorname{div} \mathbf{b} = \lambda,$$
$$\partial_t u + \operatorname{div}(u\mathbf{b}) = \mu,$$

in the sense of distributions on $(0,T) \times \mathbb{R}^d$. Characterize (compute) the distribution

$$\nu := \partial_t \beta(u) + \operatorname{div}(\beta(u)\mathbf{b}),$$

where $\beta\colon \mathbb{R} \to \mathbb{R}$ is a fixed C^1 function.

In this note we want to show the following

Theorem 2 *There exists an autonomous, compactly supported vector field* $\mathbf{b}\colon \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{b} \in L^\infty(\mathbb{R}^2)$, *and a bounded, scalar function* $u\colon (0,T) \times \mathbb{R}^2 \to \mathbb{R}$, *such that*

$$\operatorname{div} \mathbf{b} = 0,$$
$$\partial_t u + \operatorname{div}(u\mathbf{b}) = 0,$$

in $\mathscr{D}'((0,T) \times \mathbb{R}^2)$ *but the distribution*

$$\partial_t \left(u^2\right) + \operatorname{div}\left(u^2\mathbf{b}\right) \notin \mathscr{M}((0,T) \times \mathbb{R}^2)$$

i.e. it is not *(representable by) a Radon measure.*

2 Preliminaries

In order to fix the notation, we collect in this section some preliminary results we will be using in the rest of the paper.

2.1 A Particular Change of Variables

As in [2, § 2.16], we will denote by I the interval $[0, L]$, by $\mathscr{L}^1$ the Lebesgue measure on I and, in general, λ will be an arbitrary measure on I, which is singular with respect to $\mathscr{L}^1$ and has $\mathscr{A}$ as the set of its atoms (points with positive measure). We set $\hat{L} := (\mathscr{L}^1 + \lambda)(I)$ and $\hat{I} := [0, \hat{L}]$. We denote by $\hat{\mathscr{L}}^1$ the Lebesgue measure restricted to $\hat{I}$. We denote by $\hat{\sigma}$ the multifunction from I to $\hat{I}$ that to every $s \in I$ associates the interval

$$\hat{\sigma}(s) := [\hat{\sigma}_-(s), \hat{\sigma}_+(s)]$$

where

$$\hat{\sigma}_-(s) := (\mathscr{L} + \lambda)\big([0, s)\big), \qquad \hat{\sigma}_+(s) := (\mathscr{L} + \lambda)\big([0, s]\big).$$

It is immediate to see that $\hat{\sigma}$ is surjective on I, strictly increasing, and uni-valued for every $s \notin \mathscr{A}$, because σ_- and σ_+ are strictly increasing, and $\sigma_-(s) = \sigma_+(s)$ whenever $s \notin \mathscr{A}$. Moreover it is obvious that the map is expanding, i.e.

$$s_2 - s_1 \leq \hat{s}_2 - \hat{s}_1 \tag{3}$$

for every $s_1, s_2 \in I$ with $s_1 < s_2$, and every $\hat{s}_1 \in \hat{\sigma}(s_1)$, $\hat{s}_2 \in \hat{\sigma}(s_2)$. Accordingly σ is surjective from $\hat{I}$ onto I, uni-valued and 1-Lipschitz (because of (3)); furthermore, it is constant on the interval $\sigma(s)$ for every $s \in \mathscr{A}$ and strictly increasing at every point outside $\sigma(\mathscr{A})$.

We recall the following

Lemma 1 ([2, Lemma 2.17]) *Let F a $\mathscr{L}^1$-null set in I which supports the measure λ and $\hat{F} := \hat{\sigma}(F)$. Then*

1. *it holds $\sigma_{\#}\hat{\mathscr{L}}^1 = \mathscr{L}^1 + \lambda$;*
2. *the derivative of σ agrees with $\mathbf{1}_{\hat{I}\setminus\hat{F}}$ a.e. in $\hat{I}$.*

2.2 Solutions to Singular, One-Dimensional Transport Equations

In the following we will be dealing with 1d-transport equations involving singular terms, i.e. equations of the form

$$\partial_t\big(v(1+\mathscr{L}^1\times\lambda)\big)+\partial_s v=0, \tag{4}$$

where $v\colon[0,T]\times I\to\mathbb{R}$ is a function of t,s and λ is a singular measure on I. Clearly, Eq. (4) has to be understood in the sense of distributions on $(0,T)\times I$: we say that v is a solution to (4) if for every $\phi\in C_c^\infty\big((0,T)\times I\big)$ it holds

$$\int_0^T\int_I v(t,s)\big(\phi_t(t,s)+\phi_s(t,s)\big)\,ds\,dt=-\int_0^T\int_I \phi_t(t,s)\,d\lambda(s)\,dt.$$

It is very well known that such equations present a severe phenomenon of non-uniqueness (for the associated initial value problem). In order to clarify what we mean, we begin by discussing an example.

Assume for simplicity that $I=\mathbb{R}$ and λ is the Dirac mass at 0, so that we are considering the equation

$$\partial_t\big(v(\mathscr{L}^1\times\delta_0)\big)+\partial_s v=0\qquad\text{in }\mathscr{D}'((0,T)\times\mathbb{R}). \tag{5}$$

If v represents the density of a distribution of particles, then Eq. (5) is saying that each particle moves at constant speed 1 from left to right, except when it reaches the point 0, where it may stop for any given amount of time. Therefore, if v_0 is an arbitrary, bounded initial datum (for simplicity, suppose its support is contained in $(-\infty,0)$), then a solution of (5) with initial condition $v(0,s)=v_0(s)$ is the function $v\colon[0,+\infty)\times\mathbb{R}\to\mathbb{R}$ defined by

$$v(t,s)=\begin{cases}v_0(s-t) & s\neq 0\\ 0 & s=0\end{cases},$$

which physically means that no particle stops at 0. Another solution can be constructed by stopping all particles at 0, i.e.

$$\widetilde{v}(t,s)=\begin{cases}v_0(s-t) & s<0\\ 0 & s=0\\ \displaystyle\int_{-t}^0 u_0(\tau)\,d\tau & s>0\end{cases}.$$

More in general, for every $\alpha>0$ one can construct a solution for which the particles arrive at 0, stay there exactly for time α and then leave (see Fig. 1):

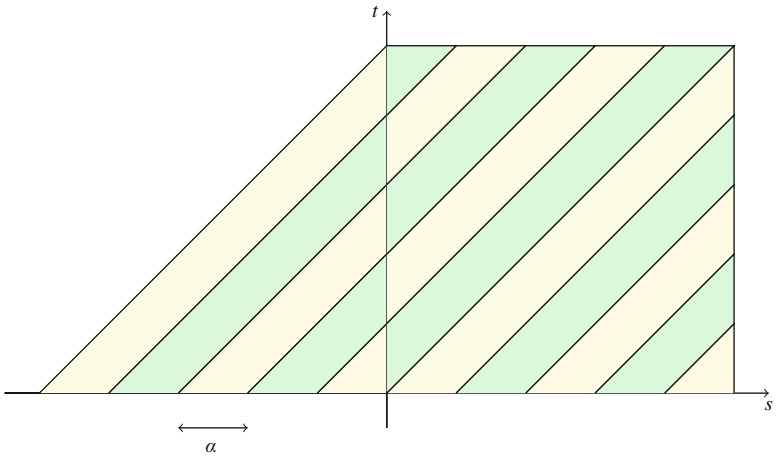


Fig. 1 A particular solution to Eq. (5): the particles at the initial time are of two different colors (yellow and green): they start moving following characteristic lines, arrive at 0 and stay there for a prescribed time α before leaving

$$u^\alpha(t,s) := \begin{cases} u_0(s-t) & s<0 \\ \displaystyle\int_{-t}^{-t+\alpha} u_0(\tau)\,d\tau & s=0 \\ u_0(s-t+\alpha) & s>0. \end{cases}$$

More precisely, we recall the following result, which is used in the proof of [2, Lemma 4.5].

Lemma 2 *Let λ be a non trivial measure on $[0,L]$, singular w.r.t. to $\mathscr{L}^1{}_{\llcorner[0,L]}$. Let furthermore $K \subset (0,L)$ be a closed, $\mathscr{L}^1$-negligible set, with $\lambda(K)>0$. Then the problem*

$$\begin{cases} \partial_t\big(v(1+\mathscr{L}^1\times\lambda)\big)+\partial_s v=0 \\ v(0,\cdot)=\mathbf{1}_K(\cdot) \end{cases} \tag{6}$$

admits a non trivial bounded solution.

We recall here the main steps of the proof, as it will be useful in the following.

Proof Clearly, the function $v(t,s) := \mathbf{1}_K(s)$ is a stationary solution of (6). Following [2], we construct a second solution by exploiting the change of variable $s=\sigma(\hat{s})$ defined in Sect. 2.1. We thus define

$$v(t,s) := \begin{cases} w(t,\hat{\sigma}(s)) & \text{for } s \notin \mathscr{A} \\ \fint_{\hat{\sigma}(s)} w(t,\hat{s})\, d\hat{s}, & \text{for } s \in \mathscr{A}, \end{cases} \tag{7}$$

where we recall $\mathscr{A}$ is the set of atoms of λ and $w : [0,T) \times [0,L] \to \mathbb{R}$ is the (unique) bounded, distributional solution of

$$\begin{cases} \partial_t w + \partial_s w = 0 \\ w(0,\cdot) = \mathbf{1}_{\hat{\sigma}(K)}(\cdot). \end{cases} \tag{8}$$

To see that (7) actually solves Problem 6 we proceed as follows: first observe that (6) can be explicitly written as

$$\int_0^T \int_0^L \big(\partial_t \phi + \mathbf{1}_{I\setminus F} \partial_s \phi\big) v \, d(\mathscr{L}^1 + \lambda)\, dt = \int_0^L \phi(0,\cdot)\mathbf{1}_K d(\mathscr{L}^1 + \lambda). \tag{9}$$

By changing variable $s = \sigma(\hat{s})$, i.e. setting $\hat{v}(t,\hat{s}) := v(t,\sigma(\hat{s}))$ and $\hat{\phi}(t,\hat{s}) := \phi(t,\sigma(\hat{s}))$ and using Lemma 1, we can rewrite (9) as

$$\int_0^T \int_0^{\hat{L}} \big(\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}\big) \hat{v}\, d\hat{s}\, dt = \int_0^{\hat{L}} \hat{\phi}(0,\cdot)\mathbf{1}_{\hat{\sigma}(K)} d\hat{s}.$$

Since on the complement of $\hat{\sigma}(\mathscr{A})$ it holds $\hat{v} = w$, to conclude we only need to show that

$$\int_{\hat{\sigma}(\mathscr{A})} \big(\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}\big) \hat{v} d\hat{s} = \int_{\hat{\sigma}(J)} \big(\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}\big) w d\hat{s}.$$

Indeed,

$$\begin{aligned} \int_{\hat{\sigma}(\mathscr{A})} (\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}) \hat{v} d\hat{s} &= \sum_{a \in \mathscr{A}} \int_{\hat{\sigma}(a)} (\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}) \hat{v} d\hat{s} \\ &= \sum_{a \in \mathscr{A}} \partial_t \phi(t,s) \int_{\hat{\sigma}(a)} \hat{v} d\hat{s} \\ &= \sum_{a \in \mathscr{A}} \partial_t \phi(t,s) \int_{\hat{\sigma}(a)} w = \int_{\hat{\sigma}(\mathscr{A})} (\partial_t \hat{\phi} + \partial_{\hat{s}} \hat{\phi}) w d\hat{s}, \end{aligned}$$

since $\partial_{\hat{s}} \hat{\phi}(t,\hat{s}) = 0$ and $\partial_t \hat{\phi}(t,\hat{s}) = \partial_t \phi(t,s)$ for all $\hat{s} \in \hat{\sigma}(s)$ and by direct definition of $\hat{v}$. To conclude the proof it is enough to show that the solution $\hat{v}$ does not coincide with the stationary one, and for this a possible strategy is to show that the maximum $M(t)$ of the support of $v(t,\cdot)$ is strictly increasing at $t = 0$ (see [2, Lemma 4.5])).

2.3 Structure of Level Sets of Lipschitz Functions and Weak Sard Property

Since we will need some results on the structure of level sets of Lipschitz functions defined in the plane, we recall them here. Suppose that $\Omega \subset \mathbb{R}^2$ is an open, simply connected domain and $H\colon \Omega \to \mathbb{R}$ is a compactly supported Lipschitz function. For any $h \in \mathbb{R}$, let $E_h := H^{-1}(h)$. We recall the following deep

Theorem 3 ([1, Theorem 2.5]) *Then the following statements hold for $\mathscr{L}^1$-a.e. $h \in H(\Omega)$:*

(1) *$\mathscr{H}^1(E_h) < \infty$ and E_h is countable $\mathscr{H}^1$-rectifiable (in what follows, we will say E_h is* regular*);*
(2) *for $\mathscr{H}^1$-a.e. $x \in E_h$ the function H is differentiable at x with $\nabla H(x) \neq 0$;*
(3) *$\mathrm{Conn}^\star(E_h)$ is countable and every $C \in \mathrm{Conn}^\star(E_h)$ is a closed simple curve;*
(4) *$\mathscr{H}^1(E_h \setminus E_h^\star) = 0$.*

Let us now define the critical set S of H as the set of all $x \in \Omega$ where H is not differentiable or $\nabla H(x) = 0$. We will be interested in the following *Weak Sard Property*, introduced in [2, Section 2.13]:

$$H_\# \left(\mathscr{L}^2 \llcorner_{S \cap E^\star} \right) \perp \mathscr{L}^1,$$

where the set $E^\star$ is the union of all connected components with positive length of all level sets of H. The relevance of the Weak Sard Property in the framework of transport and continuity equation has been completely understood in [2], to whom we refer the reader for further details. Here we simply point out that it is possible to prove that in some sense the Weak Sard Property is satisfied by a generic Lipschitz function (in Baire's category sense), as the class of all Lipschitz functions $H\colon \Omega \to \mathbb{R}$ satisfying the Weak Sard Property is residual in the Banach space of Lipschitz functions $\mathrm{Lip}(\Omega)$ (see [1, Thm. 4]). However, an explicit construction of a Lipschitz function f without the Weak Sard property was proposed in [1] and we will recall it in the Sect. 3, as it will be the building block of our counterexample.

2.4 Local Disintegration of Lebesgue Measure and of the Equation $\mathbf{div}(u\mathbf{b}) = \mu$

From now onwards, let $\mathbf{b}\colon \Omega \to \mathbb{R}^2$ a bounded, Borel, divergence-free vector field. From $\mathrm{div}(\mathbf{b}) = 0$ in Ω we deduce that there exists a Lipschitz potential $H\colon \Omega \to \mathbb{R}$ such that

$$\nabla^\perp H(x) = \mathbf{b}(x), \qquad \text{for } \mathscr{L}^2\text{-a.e. } x \in \Omega.$$

Using Theorem 3 on the Lipschitz function H, we can define the negligible set N_1 such that E_h is regular in Ω whenever $h \notin N_1$; moreover, let N_2 denote the negligible set on which the measure $(H_\# \mathscr{L}^2)^{\text{sing}}$ is concentrated, where $(H_\# \mathscr{L}^2)^{\text{sing}}$ is the singular part of $(H_\# \mathscr{L}^2)$ with respect to $\mathscr{L}^1$. Then we set

$$N := N_1 \cup N_2 \qquad \text{and} \qquad E^* := \cup_{h \notin N} E_h^* \tag{10}$$

For any $x \in E$ let C_x denote the connected component of E such that $x \in C_x$. By definition of E for any $x \in E$ the corresponding connected component C_x has strictly positive length. We recall the following Lemma which studies the disintegration of the measure $\mathscr{L}^2 \llcorner_\Omega$ w.r.t. the map H:

Lemma 3 ([2, Lemma 2.8]) *There exist Borel families of measures σ_h, κ_h, $h \in \mathbb{R}$, such that*

$$\mathscr{L}^2 \llcorner_\Omega = \int \left(c_h \mathscr{H}^1 \llcorner_{E_h} + \sigma_h \right) dh + \int \kappa_h \, d\zeta(h), \tag{11}$$

where

1. *$c_h \in L^1(\mathscr{H}^1 \llcorner_{E_h^\star})$, $c_h > 0$ a.e.; moreover, by Coarea formula, we have $c_h = 1/|\nabla H|$ a.e. (w.r.t. $\mathscr{H}^1 \llcorner_{E_h^\star}$);*
2. *σ_h is concentrated on $E_h^\star \cap \{\nabla H = 0\}$ and $\sigma_h \perp \mathscr{H}^1$ for $\mathscr{L}^1$-a.e. $h \notin N$;*
3. *κ_h is concentrated on $E_h^\star \cap \{\nabla H = 0\}$;*
4. *$\zeta := H_\# \mathscr{L}^2 \llcorner_{B \setminus E^\star}$ is concentrated on N (hence $\zeta \perp \mathscr{L}^1$).*

2.4.1 Reduction of the Equation on the Level Sets

We now show how it is possible to reduce an equation of the form $\operatorname{div}(u\mathbf{b}) = \mu$, where u is a bounded Borel function on $\mathbb{R}^2$ and μ is a Radon measure on $\mathbb{R}^2$, into a family of 1d problems on the level sets of H. For all the following Lemmas we refer the reader to [10].

The first step is the disintegration of the equation:

Lemma 4 ([10, Lemma 3.5]) *Suppose that μ is a Radon measure on $\mathbb{R}^2$ and $u \in L^\infty(\mathbb{R}^2)$. Then the equation*

$$\operatorname{div}(u\mathbf{b}) = \mu \tag{12}$$

holds in $\mathscr{D}'(\Omega)$ if and only if:

- *the disintegration of μ with respect to H has the form*

$$\mu = \int \mu_h \, dh + \int \nu_h \, d\zeta(h), \tag{13}$$

where ζ is defined in Point (4) of Lemma 3;

- *for $\mathscr{L}^1$-a.e. $h \in \mathbb{R}$ it holds*

$$\operatorname{div}\left(uc_h\mathbf{b}\mathscr{H}^1\llcorner_{E_h}\right) + \operatorname{div}(u\mathbf{b}\sigma_h) = \mu_h; \tag{14}$$

- *for ζ-a.e. $h \in \mathbb{R}$*

$$\operatorname{div}(u\mathbf{b}\kappa_h) = \nu_h. \tag{15}$$

2.4.2 Reduction on the Connected Components

The next step is to reduce further the analysis of the Eq. (14) on the nontrivial connected components of the level sets. In view of Lemma 4 in what follows we always assume that $h \notin N$ (see (10)).

Lemma 5 ([10, Lemma 3.7]) *The Eq. (14) holds iff*

- *for any nontrivial connected component C of E_h it holds*

$$\operatorname{div}\left(uc_h\mathbf{b}\mathscr{H}^1\llcorner_C\right) + \operatorname{div}(u\mathbf{b}\sigma_h\llcorner_C) = \mu_h\llcorner_C; \tag{16}$$

- *it holds*

$$\operatorname{div}(u\mathbf{b}\sigma_h\llcorner_{E_h\setminus E_h^\star}) = \mu_h\llcorner_{E_h\setminus E_h^\star}. \tag{17}$$

Now we can split further and obtain the following

Lemma 6 ([10, Lemma 3.8]) *Equation (16) holds iff*

$$\operatorname{div}\left(uc_h\mathbf{b}\mathscr{H}^1\llcorner_C\right) = \mu_h\llcorner_C, \tag{18a}$$

$$\operatorname{div}(u\mathbf{b}\sigma_h\llcorner_C) = 0. \tag{18b}$$

2.4.3 Reduction of the Equation on Connected Components in Parametric Form

Finally, we would like to discuss the parametric version of the Eq. (18a). Let $\gamma\colon I \to \mathbb{R}^2$ be an injective Lipschitz parametrization of C, where $I = \mathbb{R}/\ell\mathbb{Z}$ or $I = (0, \ell)$ (for some $\ell > 0$) is the domain of γ. The existence of such a parameterization is granted by [1, Thm. 2.5 (iv)].

Lemma 7 ([10, Lemma 3.9]) *Equation* (18a) *holds iff for any admissible parametrization* γ *of* C

$$\partial_s \hat{u}_h = \hat{\mu}_h \tag{19}$$

where $\gamma_\# \hat{\mu}_h = \mu_h \llcorner_C$, $\hat{u}_h = u \circ \gamma$.

2.5 Local Disintegration of a Balance Law

We now pass to consider a general balance law associated to the Hamiltonian vector field $\mathbf{b}$, i.e. $\partial_t u + \operatorname{div}(u\mathbf{b}) = \nu$, being ν a Radon measure on $(0,T) \times \Omega$ and $u \in L^\infty((0,T) \times \Omega)$. A reduction on the connected components of the Hamiltonian H can be performed, similarly to what we have done for equation $\operatorname{div}(u\mathbf{b}) = \mu$ to above. In some sense, we are presenting now the time-dependent version of Lemmas 4–7.

Lemma 8 *A function* $u \in L^\infty([0,T] \times \Omega)$ *is a solution to the problem*

$$\begin{cases} u_t + \operatorname{div}(u\mathbf{b}) = \nu, \\ u(0,\cdot) = u_0(\cdot), \end{cases} \qquad \text{in } \mathscr{D}'((0,T) \times \Omega) \tag{20}$$

if and only if

- $\hat{u}_h(t,s) := u(t, \gamma_h(s))$ *solves*

$$\begin{cases} \partial_t \hat{u}_h + \partial_s \hat{u}_h = \hat{\nu}_h \\ \hat{u}_h(0,\cdot) = \hat{u_0}_h(\cdot), \end{cases} \qquad \text{in } \mathscr{D}'((0,T) \times I)$$

- *it holds*

$$\operatorname{div}(u\mathbf{b}\sigma_h) = 0$$

for $\mathscr{L}^1$*-a.e.* h, *where* $\gamma_h \colon I \to \mathbb{R}^2$ *is an admissible parametrization of a connected component* C *of the level set* E_h *of the Hamiltonian* H *and* $\hat{\nu}_h$ *is a measure such that* $\hat{\nu}_h = (\gamma_h^{-1})_\# \nu$.

Proof Multiplying equation in (20) by a function $\psi \in C_c^\infty([0,T))$ and formally integrating by parts we get

$$u_t\psi + \operatorname{div}(u\psi\mathbf{b}) = \psi\nu \;\Rightarrow\; \operatorname{div}\left(\int_0^T u\psi \, dt\, \mathbf{b}\right) = \int_0^T u\psi_t \, dt - \psi(0)u_0 + \left(\int_0^T \psi \, dt\right)\nu,$$

which can be written in the form

$$\operatorname{div}(w\mathbf{b}) = \mu, \tag{21}$$

where $w := \int_0^T u\psi \, dt$ and

$$\mu := \left(\int_0^T u\psi_t \, dt - \psi(0)v_0 \right) \mathscr{L}^2 + \left(\int_0^T \psi \, dt \right) \nu.$$

Applying Lemmas 4 and 6 to (21), we obtain that continuity equation is equivalent to

$$\operatorname{div}\left(w c_h \mathbf{b} \mathscr{H}^1 \llcorner E_h \right) = \mu_h \tag{22}$$

and

$$\operatorname{div}(u\mathbf{b}\sigma_h) = 0 \tag{23}$$

for $\mathscr{L}^1$-a.e. h, where the measure μ_h can be computed explicitly, using Coarea Formula and disintegration Theorem

$$\mu_h = \left(\int_0^T u\psi_t \, dt - \psi(0)v_0 \right) \mathscr{H}^1 \llcorner E_h + \left(\int_0^T \psi \, dt \right) \nu_h.$$

Thanks to Lemma 7, Eq. (22) is *equivalent* to

$$\partial_s \hat{u} = \hat{\mu}_h,$$

in $\mathscr{D}'(I)$. Now being γ_h Lipschitz and injective, we have

$$(\gamma_h^{-1})_{\#}\left(\mathscr{H}^1 \llcorner E_h \right) = |\gamma_h'| \mathscr{L}^1,$$

and this allows us to compute explicitly

$$\begin{aligned} \hat{\mu}_h &= (\gamma_h^{-1})_{\#} \mu_h \\ &= (\gamma_h^{-1})_{\#} \left(\int_0^T u\psi_t \, dt \, c_h \mathscr{H}^1 \llcorner E_h - \int_{\mathbb{R}^2} \psi(0) v_0 c_h d\mathscr{H}^1 \llcorner E_h + \int_0^T \psi \, dt \, \nu_h \right) \\ &= \int_0^T v(\tau, \gamma(s)) \psi_\tau(\tau) \, d\tau - \psi(0) u_0(\gamma_h(s)) c_h(\gamma(s)) + \left(\int_0^T \psi(\tau) d\tau \right) \hat{\nu}_h, \end{aligned} \tag{24}$$

where

$$\hat{\nu}_h = (\gamma_h^{-1})_{\#} \nu.$$

Formally, (24) means

$$\hat{\mu}_h = -\int_0^T \partial_t \hat{u} + \hat{v}_h.$$

To sum up, we have obtained that Problem (20) is equivalent to

$$\begin{cases} \partial_t \hat{u}_h + \partial_s \hat{u}_h = \hat{v}_h, \\ \hat{u}_h(0, \cdot) = \hat{u_0}_h(\cdot), \end{cases}$$

and

$$\operatorname{div}(u\mathbf{b}\sigma_h) = 0$$

in $\mathscr{D}'((0, T) \times I)$ for $\mathscr{L}^1$-a.e. $h \in \mathbb{R}$. We explicitly notice that the last relation is always satisfied, as $\mathbf{b} = 0$ on the critical set (where σ_h is concentrated).

2.6 Anzellotti Traces for Measure Divergence L^∞ Vector Fields

Let now $\rho \in L^\infty(\mathbb{R}^d)$ and $\mathbf{v} \in L^\infty(\mathbb{R}^d; \mathbb{R}^d)$. We consider the vector field $\mathbf{V} := \rho\mathbf{v}$ and we assume that its distributional divergence $\operatorname{div} \mathbf{V}$ is represented by some Radon measure, so that $\mathbf{V}$ is a so called *measure divergence vector field*. There are well known results that allows to give a meaning and to characterize the *trace* of such vector fields over rectifiable sets. We list here the main ones and we refer for more details to [3, 8, 14].

Definition 1 Given a bounded, open domain with C^1 boundary $U \subset \mathbb{R}^d$, the *(Anzellotti) normal trace* of $\mathbf{V}$ over ∂U is the distribution defined by

$$\langle \mathsf{Tr}(\mathbf{V}, U) \cdot \mathbf{n}, \psi \rangle := \int_U \psi(x)\, d(\operatorname{div} \mathbf{V})(x) + \int_U \mathbf{V} \cdot \nabla\psi(x)\, d\mathscr{L}^d(x)$$

for every compactly supported smooth test function $\psi \in C_c^\infty(\mathbb{R}^d)$.

We have the following Proposition, which says that the trace of a measure divergence vector field is not an arbitrary distribution, but is induced by integration of a bounded function defined on ∂U.

Proposition 1 *There exists a unique* $g \in L^\infty_{\mathrm{loc}}(\partial U; \mathscr{H}^{d-1}\llcorner_{\partial U})$ *such that*

$$\langle \mathsf{Tr}(\mathbf{V}, U) \cdot \mathbf{n}, \phi \rangle = \int_{\partial U} g\phi\, \mathscr{H}^{d-1}, \qquad \forall \phi \in C_c^\infty(\mathbb{R}^d).$$

One can also define the traces of $\mathbf{V}$ on a *oriented* hypersurface of class C^1, say Σ. Indeed, choosing an open C^1 domain $U \Subset U$ such that $\Sigma \subset \partial U$ and the unit outer normals agree $\nu_U = \nu_\Sigma$ we can define

$$\mathsf{Tr}^-(\mathbf{V}, \Sigma) \cdot \mathbf{n} := \mathsf{Tr}(\mathbf{V}, U).$$

Analogously, choosing an open C^1 domain U' such that $\Sigma \subset \partial U'$ and $\nu_{U'} = -\nu_\Sigma$ we define

$$\mathsf{Tr}^+(\mathbf{V}, \Sigma) \cdot \mathbf{n} := -\mathsf{Tr}(\mathbf{V}, U') \cdot \mathbf{n}.$$

We remark that one can replace C^1 regularity with Lipschitz, so that it is possible to give the definition of normal trace of a measure divergence vector field on countable H^{d-1}-rectifiable sets.

We collect here other important results on Anzellotti's weak traces:

Proposition 2 *If $\mathbf{V}$ is a bounded, measure divergence vector field, then:*

- $\operatorname{div} \mathbf{V} \ll \mathscr{H}^{d-1}$ *as measures in* $\mathbb{R}^d$*;*
- *for any oriented, C^1 hypersurface Σ it holds*

$$\operatorname{div} \mathbf{V} \llcorner_\Sigma = \big(\mathsf{Tr}^+(\mathbf{V}, \Sigma) \cdot \mathbf{n} - \mathsf{Tr}^-((\mathbf{V}, \Sigma) \cdot \mathbf{n}\big) \mathscr{H}^{d-1} \llcorner_\Sigma.$$

Finally, an interesting case is when we assume more regularity on the vector field, for instance $\mathbf{v} \in \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^d; \mathbb{R}^d)$. In this situation, one has the usual definition of the trace of $\mathbf{v}$ over ∂U as BV function. We recall that the trace of BV functions $\mathbf{v}$ for open sets $U \subset \mathbb{R}^d$ of class C^1 is a measure which is absolutely continuous w.r.t. $\mathscr{H}^{d-1} \llcorner_{\partial U}$. We conclude this section by recalling the following chain rule for traces, proved when $\mathbf{v} \in \mathrm{BV}$ in [4] (see also [3] for the case of vector fields of *bounded deformation*).

Theorem 4 (Change of Variables for Traces) *Let $U \subset \mathbb{R}^d$ be an open domain of class C^1 and let $\mathbf{v} \in \mathrm{BV}_{\mathrm{loc}}(\mathbb{R}^d; \mathbb{R}^d)$ and $\beta \in \mathrm{Lip}(\mathbb{R})$. Then if $\mathbf{V} = \rho \mathbf{v}$ is a measure divergence vector field, then also $\beta(\rho)\mathbf{v}$ is a measure divergence vector field and, moreover, it holds*

$$\mathsf{Tr}^\pm(\beta(\rho)\mathbf{v}, U) \cdot \mathbf{n} = \beta\left(\frac{\mathsf{Tr}^\pm(\rho \mathbf{v}, U) \cdot \mathbf{n}}{\mathsf{Tr}^\pm(\mathbf{v}, U) \cdot \mathbf{n}}\right) \mathsf{Tr}^\pm(\mathbf{v}, U) \cdot \mathbf{n}, \qquad \mathscr{H}^{d-1} \text{ -a.e. on } \partial U,$$

where the ratio is arbitrarily defined at points where the trace $\mathsf{Tr}(\mathbf{v}, U)$ vanishes.

3 The Construction of the Hamiltonian

In this section, we construct a suitable Lipschitz function $H\colon \mathbb{R}^2 \to \mathbb{R}$ that will be the building block of our counterexample. The construction presented in this paragraph goes back to [1].

3.1 A Function That Does Not Have Weak Sard Property

Let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ be decreasing sequences of positive numbers with asymptotic behaviour given by

$$a_n \sim b_n \sim \frac{1}{n^2 2^n}.$$

Hence, the following quantities

$$\hat{a} := \sum_{n=0}^{\infty} 2^{n+2} a_n, \qquad \hat{b} := \sum_{n=0}^{\infty} 2^{n+1} b_n$$

are finite. Chosen a real number $\delta > 0$, we set

$$c_0 := \delta + \hat{a}, \qquad d_0 := \delta + \hat{b}.$$

3.1.1 The Construction of the Set

We consider the set C_0, which is the closed rectangle with width c_0 and height d_0. Then we define C_1 to be the union of four closed rectangles with sizes

$$c_1 := \frac{c_0}{2} - 2a_0, \qquad d_1 := \frac{d_0}{2} - b_0$$

like in Fig. 2.

If we iterate the above construction, we obtain a sequence of nested sets: more precisely, if C_n is the union of 4^n pairwise disjoint, closed rectangles with width c_n and height d_n, then C_{n+1} is the union of 4^{n+1} pairwise disjoint closed rectangles with width

$$c_{n+1} := \frac{c_n}{2} - 2a_n, \qquad d_{n+1} := \frac{d_n}{2} - b_n.$$

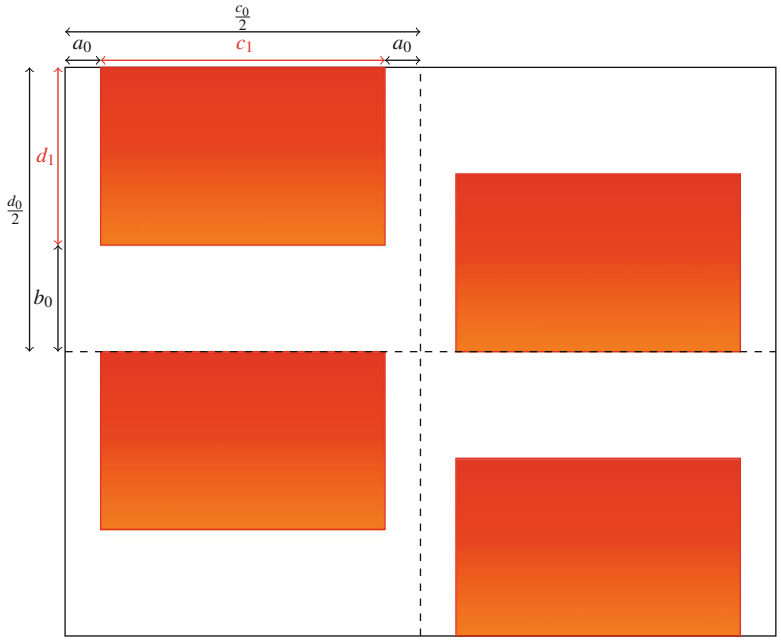


Fig. 2 The sets C_0 (black) and C_1 (red)

It is easy to see that from this recursion we have

$$2^n c_n = c_0 - \sum_{m=0}^{n-1} 2^{m+2} a_m \searrow \delta \qquad \text{and} \qquad 2^n d_n = d_0 - \sum_{m=0}^{n-1} 2^{m+2} b_m \searrow \delta$$

which implies that c_n, d_n are always strictly positive and satisfy

$$c_n \sim d_n \sim \frac{\delta}{2^n}.$$

If C denotes the intersection of the closed sets C_n we have

$$\mathscr{L}^2(C) = \lim_n \mathscr{L}^2(C_n) = \lim_n 4^n d_n c_n = \delta^2.$$

3.1.2 Construction of the Function

We now turn to the construction of a suitable sequence of Lipschitz and piecewise smooth functions $f_n : \mathbb{R}^2 \to \mathbb{R}$. The function f_0 is defined by its level sets, drawn in Fig. 3a.

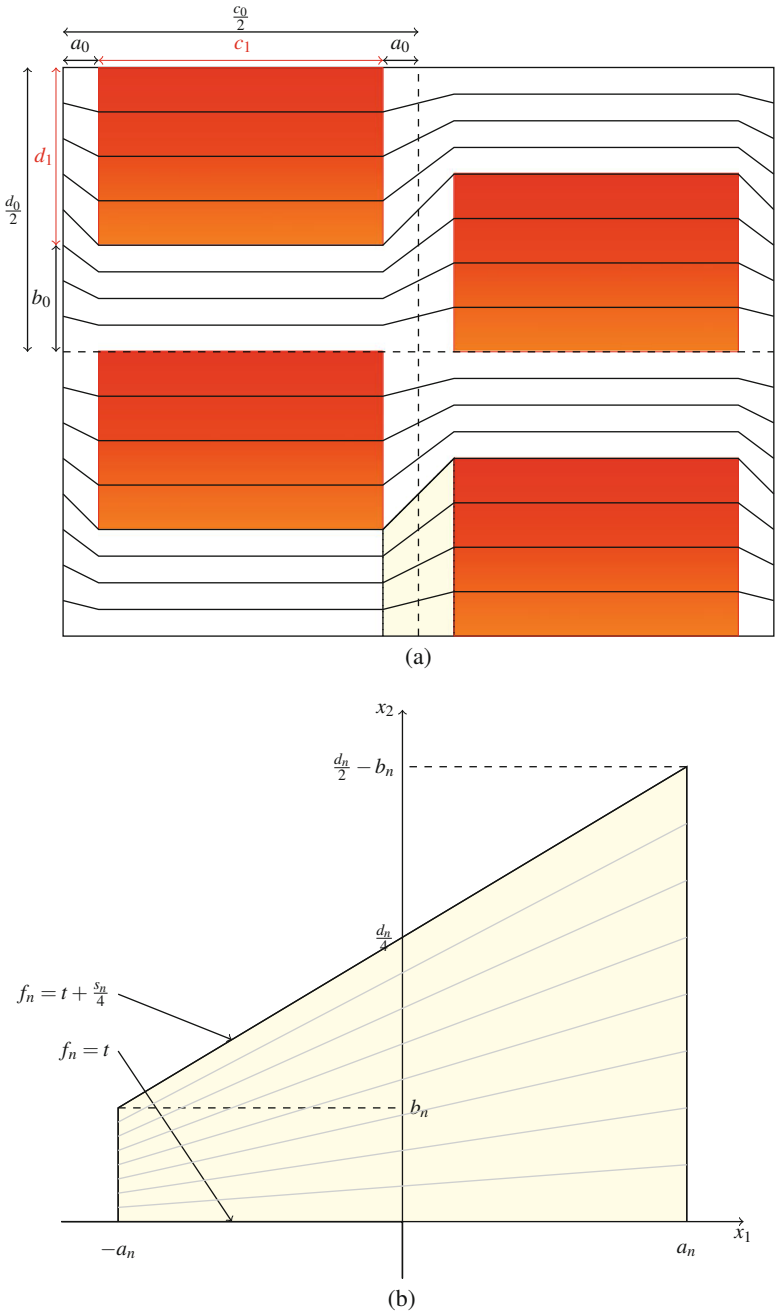


Fig. 3 Level sets of the function f_0 and estimates for $|\nabla f_n|$. (**a**) Level sets of the function f_0. (**b**) Estimate of $|\nabla f_n|$: the level sets of f_n in the set E

Let s_n be the oscillation of the function f_n on the component of C_n; it is clear from the picture that

$$s_{n+1} = \frac{s_n}{4}, \tag{25}$$

hence $s_n = 4^{-n} s_0 = 4^{-n} d_0$.

3.1.3 L^∞ Gradient Estimates

We can now estimate the gradient of the functions f_n. It is easy to see that the supremum of $|\nabla f_n|$ in the set C_n is attained in the set E defined in Fig. 3a. Choosing the axes as in Fig. 3b we can write an explicit formula for f_n; in particular, the line that passes through the points $(-a_n, b_n)$ and $(a_n, \frac{d_n}{2} - b_n)$ has equation

$$x_2 = b_n + \frac{1}{4a_n}(x_1 + a_n)(d_n - 4b_n).$$

Then if we pick a $\tau \in (0, b_n)$ we impose the similarity of the triangles, hence

$$\frac{\tau b_n}{b_n} = \frac{x_2}{b_n + \frac{1}{4a_n}(x_1 + a_n)(d_n - 4b_n)}$$

hence we get

$$\tau = \frac{4a_n x_2}{(d_n - 4b_n)x_1 + a_n d_n}.$$

Therefore, the function f_n has the following explicit formula in E:

$$f_n(x_1, x_2) = (1 - \tau)t + \tau\left(t + \frac{s_n}{4}\right) = t + \frac{s_n}{4}\tau = t + \frac{a_n s_n x_2}{(d_n - 4b_n)x_1 + a_n d_n}.$$

A direct computation shows that

$$\nabla f_n(x) = \frac{1}{a_n d_n + (d_n - 4b_n)x_1}\left(-(d_n - 4b_n)(f_n(x) - t), a_n s_n\right).$$

Taking into account that $x_1 \geq -a_n$ and that $d_n - 4b_n > 0$ (due to the asymptotic behaviour) we can estimate from below the denominator:

$$a_n d_n + (d_n - 4b_n)x_1 \geq 4a_n b_n.$$

On the other hand, we clearly have $|f_n - t| \leq s_n$ and thus we obtain the following estimate:

$$\|\nabla f_n\|_{L^\infty(C_n)} \leq \frac{(d_n - 4b_n)s_n + a_n s_n}{4a_n b_n} = O(n^4 2^{-n}). \tag{26}$$

Now let us define the function $h_n := f_n - f_{n-1}$. Clearly, by definition of f_n, the support of h_n lies in C_n; moreover,

$$\|\nabla h_n\|_\infty \leq \|\nabla f_n\|_{L^\infty(C_n)} + \|\nabla f_{n-1}\|_{L^\infty(C_{n-1})} \sim n^4 2^{-n}. \tag{27}$$

Since the distance of a point in C_n from $\mathbb{R}^2 \setminus C_n$ is of order $c_n \sim 2^{-n}$, by the Mean Value Theorem

$$\|h_n\|_{L^\infty} \sim n^4 4^{-n}.$$

For every $x \in \mathbb{R}^2$ set

$$f(x) := \lim_{n \to +\infty} f_n(x) = f_0(x) + \sum_{n=1}^{\infty} h_n(x).$$

We sum up the properties of the function f in the following

Theorem 5 ([1, Prop. 4.7]) *If C is the set above and f is the function built in the previous sections, then:*

(i) *f is differentiable at every $x \in C$ with $\nabla f(x) = 0$;*
(ii) *$\mathscr{L}^1(f(C)) = d_0$;*
(iii) *$f_\sharp(\mathscr{L}^2 \llcorner C) = m\mathscr{L}^1 \llcorner f(C)$, where $m = \delta^2/d_0$; in particular, f does not satisfy the Weak Sard Property.*

Proof

(i) The estimates (26) and (27) yield

$$\|\nabla f\|_{L^\infty(C_n)} = \mathscr{O}(n^4 2^{-n})$$

which means that the Lipschitz constant of f on each component of C_n is of order $\mathscr{O}(n^4 2^{-n})$; being C contained in the interior of C_n, it follows that for every $x \in C$

$$\limsup_{y \to x} \frac{|f(x) - f(y)|}{|y - x|} = \mathscr{O}(n^4 2^{-n}).$$

Letting $n \to +\infty$ we obtain the assertion, i.e. f is differentiable at x with $\nabla f(x) = 0$.

(ii) The range $f(C)$ is the intersection of all $f(C_n)$, and $f(C_n)$ agrees with $f_n(C_n)$, and therefore it is the union of 4^n pairwise disjoint, closed intervals each of length s_n (defined as the oscillation of f_n in C_n). Thus from (25),

$$\mathscr{L}^1(f(C)) = \lim_n \mathscr{L}^1(f_n(C_n)) = \lim_n 4^n s_n = d_0.$$

(iii) We must show that the measures $\mu := f_\#(\mathscr{L}^2 \llcorner S)$ and $\lambda := m\mathscr{L}^1 \llcorner f(C)$ are the same. Since both μ and λ are supported on the compact set $f(C)$, we apply Lemma 4.6 in [1] to the partitions F_n given by the sets $R' := f(R \cap C)$ where R is a component of C_n, and deduce that it suffices to prove $\mu(R') = \lambda(R')$ for every such R'. Since C can be written as a disjoint union of 4^n translated copies of $R \cap C$, we have

$$\mu(R') = \mathscr{L}^2(R \cap C) = 4^{-n}\mathscr{L}^2(C) = 4^{-n}\delta^2.$$

On the other hand, as already observed, $f(C)$ can be written as a disjoint union of 4^n translated copies of R', and then

$$\lambda(R') = m\mathscr{L}^1(R') = 4^{-n}m\mathscr{L}^1(f(C)) = 4^{-n}md_0 = 4^{-n}\delta^2.$$

3.1.4 Further Remarks on the Hamiltonian Without Weak Sard Property

The Lipschitz function constructed in (3.1) will be denoted as $f_{c_0,d_0,\delta}$, since c_0, d_0, δ are free parameters in the construction. Recall also that $\operatorname{osc} f_{c_0,d_0,\delta} = d_0$ so that, up to a translation, we can suppose directly that

$$f_{c_0,d_0,\delta}(\mathbb{R}^2) = (0, d_0).$$

The critical set S of $f_{c_0,d_0,\delta}$ has area $\mathscr{L}^2(S) = \delta^2$ and, as shown in Theorem 5

$$(f_{c_0,d_0,\delta})_\#(\mathscr{L}^2 \llcorner C) = \frac{\delta^2}{d_0}\mathscr{L}^1 \llcorner f(C).$$

Therefore, we can apply Disintegration Theorem to the probability measure $\frac{1}{\delta^2}\mathscr{L}^2 \llcorner C$ w.r.t. the map $f_{c_0,d_0,\delta}$. We thus write

$$\frac{1}{\delta^2}\mathscr{L}^2 \llcorner C = \frac{1}{d_0}\int \nu_h \, dh$$

where $h \mapsto \nu_h$ is a measurable measure-valued map, ν_h being a probability measure concentrated on $f^{-1}_{c_0,d_0,\delta}(h) \cap C$ for $\mathscr{L}^1$-a.e. $h \in \mathbb{R}$. We can actually say more, characterizing completely the measure ν_h. In particular, we want to show that for a.e. h the intersection

$$f^{-1}_{c_0,d_0,\delta}(h) \cap C$$

is a single point. We have indeed

$$f^{-1}_{c_0,d_0,\delta}(h) \cap C = \bigcap_n \left(f^{-1}_{c_0,d_0,\delta}(h) \cap C_n \right)$$

and for every h it is possible to prove that $f^{-1}_{c_0,d_0,\delta}(h) \cap C_n$ is a sequence of nested intervals whose measure goes to 0 as $n \to +\infty$. For instance, if $h \in (d_0/2^n, d_0)$, we have that

$$f^{-1}(h) \cap C_n = \left(a_{n-1}, \frac{c_{n-1}}{2} - a_{n-1} \right) \times \{d_0\}.$$

The lenght of the interval is clearly $\frac{c_{n-1}}{2} - 2a_{n-1} = c_n \simeq \delta \cdot 2^{-n} \to 0$ as $n \to +\infty$. This shows that $f^{-1}(h) \cap C = \{x_h\}$ for every $h \in (0, d_0) = f([0, c_0] \times [0, d_0])$. So ν_h has to be δ_{x_h}. Finally notice that we can write

$$\mathscr{L}^2 \llcorner C = \int \delta_{x_h} m \, dh.$$

3.1.5 Scaling the Hamiltonian $f_{1,1,\delta}$

Set now $H_1 := f_{1,1,\delta}$ whose range is $(0, 1)$. The disintegration now looks like

$$\mathscr{L}^2 \llcorner C = \delta^2 \int \delta_{x_h} \, dh = |S| \int \delta_{x_h} \, dh$$

which will be written from now onwards as

$$\mathscr{L}^2 \llcorner C = \int c_h \delta_{x_h} \, dh,$$

where we have set for $\mathscr{L}^1$-a.e. h the coefficient $c_h := |S|$. The map $h \mapsto c_h$ is thus constant and it simply represents the density of $f_\sharp(\mathscr{L}^2 \llcorner C)$ along the level sets. We will see that this map plays a significant role in the construction: we will suitably modify it, in order to obtain a piecewise constant map which is integrable but not square-integrable. To do this, we perform some scaling transformations: for fixed $n \in \mathbb{N}$ and $\alpha \in \mathbb{R}$ we first scale the *domain* of H_1 with the following linear map:

$$\mathscr{D}_n : (x, y) \mapsto \left(x, \frac{y}{2^n} \right)$$

The area of the critical set was $|S| = \delta^2 = \int_0^1 c_h \, dh$, while after the operation the area becomes

$$\det \mathscr{Q}_n \cdot |S| = \frac{|S|}{2^n}$$

hence we set

$$c'_h := \frac{|S|}{2^n}.$$

Now we rescale the range $(0, 1) \mapsto (0, 2^{-n\alpha})$ via a map $\mathscr{R}_{n,\alpha} : \mathbb{R} \to \mathbb{R}$ so that if we impose

$$\frac{|S|}{2^n} = \int_0^{2^{-n\alpha}} c''_h \, dh$$

we have to set accordingly

$$c''_h := \frac{|S|}{2^n} 2^{n\alpha} = \frac{|S|}{2^{n(1-\alpha)}}$$

Finally, we define the translation operator $\mathscr{T}_{n,\alpha}$ which acts both in the domain and in the target in the following way: if a function is defined in the square $[0, 1] \times \left[0, \frac{1}{2^n}\right] \subset \mathbb{R}^2$ with range $\left[0, \frac{1}{2^{n\alpha}}\right]$ then under the action of $\mathscr{T}_{n,\alpha}$ the domain becomes the rectangle $[0, 1] \times \left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right]$ while the range turns to the interval $\left[\frac{1}{2^{n\alpha}}, \frac{1}{2^{n\alpha-1}}\right]$. We call the resulting function $\mathscr{T}_{n,\alpha} \circ \mathscr{R}_{n,\alpha} \circ \mathscr{Q}_{n,\alpha} \circ H_1 := H_{n,\alpha}$ and we define now

$$H_\alpha(x, y) := \sum_{n\in\mathbb{N}} H_{n,\alpha}(x, y), \ (x, y) \in D := \bigcup_{n\in\mathbb{N}} \left([0, 1] \times \left[\frac{1}{2^n}, \frac{1}{2^{n-1}}\right] \right)$$
$$= [0, 1] \times [0, 1].$$

In other words, we have "patched together" the rescaled Hamiltonians, one above the other, with ranges that are adjacent intervals. Notice that the function is well defined, as the domains of the different $H_{n,\alpha}$ are disjoint, so that for any $(x, y) \in D$ the sum is locally finite (actually it reduces to a single term).

3.1.6 Properties of H_α

Some remarks about the properties of H_α are now in order.

- For $\alpha > 0$, the function H_α is bounded. Indeed, its range is

$$H_\alpha(D) = \bigcup_{n\in\mathbb{N}} \left[2^{-n\alpha}, 2^{-n\alpha+1}\right]$$

whose measure is

$$\mathscr{L}^2(H_\alpha(D)) = \sum_n \frac{1}{2^{n\alpha}} =: \ell_\alpha < +\infty,$$

for $\alpha > 0$.

- For any $\alpha \in \mathbb{R}$, the area of the critical set of H_α is always finite:

$$\int_0^{\ell_\alpha} c_h'' \, dh = \sum_n \frac{|S|}{2^{n(1-\alpha)}} \times \frac{1}{2^{n\alpha}} = \sum_n \frac{|S|}{2^n} = |S| < +\infty.$$

- On the contrary, we have that

$$\int_0^{\ell_\alpha} (c_h'')^2 \, dh = \sum_n \frac{|S|^2}{2^{2n(1-\alpha)}} \times \frac{1}{2^{n\alpha}} = \sum_n \frac{|S|^2}{2^{n(2-\alpha)}}.$$

In particular, if we take $\alpha \geq 2$ we have that

$$\int_0^{\ell_\alpha} (c_h'')^2 \, dh = +\infty.$$

In other words, for $\alpha \geq 2$, the function $h \mapsto c_h''$ belongs to $L^1([0, \ell_\alpha]) \setminus L^2([0, \ell_\alpha])$ (it behaves essentially like $n\mathbf{1}_{[0,n^{-2}]}$ in $[0, 1]$).

4 The Counterexample

We now fix $\alpha > 2$ and we consider the corresponding Hamiltonian H_α constructed in paragraph above and we set $\mathbf{b} := \nabla H_\alpha$. By construction, setting $\sigma_h := c_h'' \delta_{x_h}$, we have that H_α satisfies the following

$$\mathscr{L}^2 = \int_{\mathbb{R}} \left(\frac{1}{|\nabla H|} \mathscr{H}^1 \llcorner E_h + \sigma_h \right) dh. \tag{28}$$

For typographical reasons, we will write from now onward simply c_h instead of c_h''.

By applying Lemma 8 to H_α we get at once the following

Proposition 3 *The problem*

$$\begin{cases} \partial_t u + \operatorname{div}(u\mathbf{b}) = 0 \\ u(0, \cdot) = u_0(\cdot) \end{cases} \tag{29}$$

is equivalent to

$$\begin{cases} \partial_t \hat{u}_h + \partial_s \hat{u}_h + c_h \partial_t \hat{u} \mathscr{L}^1 \otimes \delta_{s_h} = 0 \\ \hat{u}_0(s) = u_{0h} \\ \partial_s \left(\hat{u}_h c_h \mathscr{L}^1 \otimes \delta_{s_h} \right) = 0 \end{cases} \tag{30}$$

for $\mathscr{L}^1$-a.e. h.

Remark 1 Notice that, by splitting

$$u = m\mathbf{1}_S + u\mathbf{1}_{S^c}$$

the equation can be written as

$$\partial_t(u\mathbf{1}_{S^c}) + \partial_t(m\mathbf{1}_S) + \underbrace{\operatorname{div}(m\mathbf{1}_S\mathbf{b})}_{=0} + \operatorname{div}(u\mathbf{1}_{S^c}\mathbf{b}) = 0$$

because $\mathbf{b} = 0$ on S by construction. Hence, taking into account that $\mathbf{b} = 0$ on the critical set, Proposition 3 is actually establishing that

$$\begin{cases} \partial_t(u\mathbf{1}_{S^c}) + \operatorname{div}(u\mathbf{1}_{S^c}\mathbf{b}) = -\partial_t(m\mathbf{1}_S) \\ u(0,\cdot) = u_0(\cdot) \end{cases}$$

is *equivalent* to

$$\begin{cases} \partial_t \hat{u}_h + \partial_s \hat{u}_h + c_h \partial_t \hat{m}_h \mathscr{L}^1 \otimes \delta_{s_h} = 0 \\ \hat{u}_0(s) = u_{0h} \end{cases} \quad \text{for } \mathscr{L}^1\text{-a.e. } h.$$

Now we consider the Cauchy problem for the transport equation associated to $\mathbf{b}$ with initial condition $u_0 := \mathbf{1}_S$:

$$\begin{cases} \partial_t u + \operatorname{div}(u\mathbf{b}) = 0 \\ u(0,\cdot) = \mathbf{1}_S(\cdot) \end{cases}.$$

We disintegrate the equation on the level sets and we obtain, denoting for typographic simplicity by $v_h(t,s) := \hat{u}_h(t,s)$, we have

$$\begin{cases} \partial_t v_h + \partial_s v_h = -c_h \partial_t (v_h \mathscr{L}^1 \times \delta_{s_h}) \\ v_h(0,\cdot) = c_h \mathbf{1}_{\{s_h\}}(\cdot) \end{cases}.$$

i.e.

$$\begin{cases} \partial_t \big(v_h(1 + \mathscr{L}^1 \times c_h \delta_{s_h})\big) + \partial_s v_h = 0 \\ v_h(0, \cdot) = c_h \mathbf{1}_{\{s_h\}}(\cdot) \end{cases}, \tag{31}$$

which is exactly of the form (6). Applying Lemma 2, we have that the function

$$v_h(t, s) := \begin{cases} c_h \mathbf{1}_{\hat{\sigma}(s_h)}(\hat{\sigma}(s) - t) & s \neq s_h \\ \fint_{\hat{\sigma}(s_h)} c_h \mathbf{1}_{\hat{\sigma}(s_h)}(\hat{s} - t) d\hat{s} & s = s_h \end{cases}$$

is a non-stationary solution to (31). Some easy computations show that

$$\begin{aligned} v_h(t, s_h) = \fint_{\hat{\sigma}(s_h)} \mathbf{1}_{\hat{\sigma}(s_h)}(\hat{s} - t) d\hat{s} &= \frac{1}{c_h} \int_{s_h}^{s_h + c_h} c_h \mathbf{1}_{[s_h, s_h + c_h]}(\hat{s} - t) d\hat{s} \\ &= \int_{s_h - t}^{s_h + c_h - t} \mathbf{1}_{[s_h, s_h + c_h]}(\tau) d\tau \\ &= \begin{cases} c_h - t & t < c_h \\ 0 & t > c_h \end{cases} \end{aligned}$$

In particular, we have that for a.e. $h \in \mathbb{R}$ and for every $t \in (0, T)$ it holds

$$\partial_t v_h(t, s_h) = -\mathbf{1}_{[0, c_h]}(t).$$

Hence, for this particular solution, the 1D equation on the level set E_h is explicit:

$$\partial_t v_h + \partial_s v_h = c_h \mathbf{1}_{[0, c_h]},$$

which can be written also in the divergence form

$$\mathrm{div}_{t,s}(v_h(1, 1)) = c_h \mathbf{1}_{[0, c_h]}. \tag{32}$$

From (32), we deduce immediately that, for a.e. $h \in \mathbb{R}$, the vector field $v_h(1, 1)$ is a bounded, divergence-measure vector field in $(0, T) \times \mathbb{R}_s$. Applying Point 2 of Proposition 2 we can write for a.e. $t \in (0, T)$

$$v_h^-(t) - v_h^+(t) = +c_h \mathbf{1}_{[0, c_h]}(t) \tag{33}$$

where $v_h^\pm$ are the (L^∞ functions representing) Anzellotti traces on the surface $\Sigma_h := \{s = s_h\}$, defined as

$$v_h^\pm := \frac{\mathsf{Tr}^\pm\big(v_h(1, 1), \Sigma_h\big)}{\mathsf{Tr}^\pm\big((1, 1), \Sigma_h\big)} = \mathsf{Tr}^\pm\big(v_h(1, 1), \Sigma_h\big).$$

We observe that by construction $v_h^- = 0$ a.e., hence (33) reduces to

$$-v_h^+ = c_h \mathbf{1}_{[0,c_h]}.$$

Taking now $\beta(\tau) = \tau^2$ and applying the Chain rule for Anzellotti traces (4) (being the vector field $v = (1, 1)$ clearly of bounded variation) we obtain that for a.e. $h \in \mathbb{R}$ the vector field $w_h(1, 1) := v_h^2(1, 1)$ is still a divergence-measure vector field and it holds

$$w_h^- = 0, \qquad w_h^+ = +c_h^2 \mathbf{1}_{[0,c_h]},$$

i.e.

$$w_h^- - w_h^+ = -c_h^2 \mathbf{1}_{[0,c_h]}$$

so that, applying again Point 2 of Proposition 2, we can write

$$\mathrm{div}_{t,s}(w_h(1, 1))\llcorner_{\Sigma_h} = (\partial_t \hat{w} + \partial_s \hat{w})\llcorner_{\Sigma_h} = -c_h^2 \mathbf{1}_{[0,c_h]}.$$

which in turn can be written as (recall $\hat{m}_h = c_h v_h \mathbf{1}_{s_h}$)

$$\partial_t \hat{w}_h + \partial_s \hat{w}_h = -c_h \partial_t (\hat{m}_h \mathscr{L}^1 \times \delta_{s_h}), \qquad \text{for a.e. } h \in \mathbb{R}.$$

Integrating and using Remark (1), we obtain the equation satisfied by u^2:

$$\partial_t(u^2) + \mathrm{div}(u^2 \mathbf{b}) = T,$$

being T the distribution defined by

$$T := -\partial_t(u^2 \mathbf{1}_S). \tag{34}$$

We conclude with the following

Proposition 4 *The distribution T defined in (34) is not representable by a Radon measure.*

Proof By contradiction, assume that T is induced by some measure ξ: being the divergence of the bounded, measure-divergence vector field $w(1, \mathbf{b})$, we would necessarily have $\xi \ll \mathscr{H}^d$. On the other hand, it is immediate to see, directly from the construction of the Hamiltonian, that for any $\phi \in C_c^\infty$, $\|\phi\|_\infty \leq 1$ we have

$$\int_{[0,T]\times\mathbb{R}^2} \phi d\xi = \int_0^T \int_S u^2 \phi_t dt dx = \int_{\mathbb{R}} \int_0^T u(t, x_h) c_h^2 \phi_t(t, x_h)\, dt\, dh$$

which diverges being $c_h \notin L^2(\mathbb{R})$. Thus

$$\sup\left\{\langle T, \phi\rangle : \phi \in C_c^\infty((0,T)\times\mathbb{R}^2), \|\phi\|_\infty \le 1\right\} = +\infty$$

which shows that T cannot be a distribution of order 0, hence it is not representable by a measure.

Acknowledgements The second author would like to thank Nikolay A. Gusev for bringing to his attention the reference [12] and for useful discussions on the topic.

References

1. G. Alberti, S. Bianchini, G. Crippa, Structure of level sets and Sard-type properties of Lipschitz maps. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **12**(4), 863–902 (2013)
2. G. Alberti, S. Bianchini, G. Crippa, A uniqueness result for the continuity equation in two dimensions. J. Eur. Math. Soc. **16**(2), 201–234 (2014)
3. L. Ambrosio, G. Crippa, S. Maniglia, Traces and fine properties of a BD class of vector fields and applications. Ann. Fac. Sci. Toulouse Math. (6) **14**(4), 527–561 (2005)
4. L. Ambrosio, C. De Lellis, J. Malý, On the chain rule for the divergence of BV-like vector fields: applications, partial results, open problems, in *Perspectives in Nonlinear Partial Differential Equations*. Contemporary Mathematics, vol. 446 (American Mathematical Society, Providence, 2007), pp. 31–67
5. L. Ambrosio, G. Dal Maso, A general chain rule for distributional derivatives. Proc. Amer. Math. Soc. **108**(3), 691–702 (1990)
6. L. Ambrosio, N. Fusco, D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems* (Oxford Science Publications/Clarendon Press, Oxford, 2000)
7. L. Ambrosio, Transport equation and Cauchy problem for BV vector fields. Invent. Math. **158**(2), 227–260 (2004)
8. G. Anzellotti, *Traces of Bounded Vectorfields and the Divergence Theorem* (Dipartimento di matematica, Università degli studi di Trento, Trento, 1983)
9. S. Bianchini, P. Bonicatto, A uniqueness result for the decomposition of vector fields in $\mathbb{R}^d$ (2017, Preprint)
10. S. Bianchini, P. Bonicatto, N.A. Gusev, Renormalization for autonomous nearly incompressible BV vector fields in two dimensions. SIAM J. Math. Anal. **48**(1), 1–33 (2016)
11. S. Bianchini, N.A. Gusev, Steady nearly incompressible vector fields in two-dimension: chain rule and renormalization. Arch. Ration. Mech. Anal. **222**(2), 451–505 (2016)
12. G. Crippa, N.A. Gusev, S. Spirito, E. Wiedemann, Failure of the chain rule for the divergence of bounded vector fields. Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **XVII**, 1–18 (2017)
13. R.J. DiPerna, P.-L. Lions, Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**(3), 511–547 (1989)
14. C. De Lellis, Notes on hyperbolic systems of conservation laws and transport equations, in *Handbook of Differential Equations: Evolutionary Equations*, Vol. III (Elsevier/North-Holland, Amsterdam, 2007), pp. 277–382
15. S.N. Kružkov, First order quasilinear equations with several independent variables. Mat. Sb. (N.S.) **81**(123), 228–255 (1970)
16. A.I. Volpert, Spaces BV and quasilinear equations. Mat. Sb. (N.S.) **73**(115), 255–302 (1967)

On the Maximality of the Sum of Two Maximal Monotone Operators



Yuqing Chen, Yeol Je Cho, and Themistocles M. Rassias

Dedicated to Haim Brezis and Louis Nirenberg in deep admiration

1 Introduction

Maximal monotone operators play very important role in proving various existence results for nonlinear partial differential equations, variational inequalities and nonlinear evolution equations (see [3–9, 14]).

In this paper, let E be a real reflexive Banach space, E^* be the dual space of E and $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators. It is well known that the sum $T + S$ of two maximal monotone operators need not to be maximal monotone (see [14]) and, also, we can find some well known results on the maximality of the sum of maximal monotone operators in Brézis et al. [5] and Rockafellar [15, 16]. For more results on the maximality of sums of two maximal monotone mappings, see [1, 2, 10, 14, 18].

In this paper, we pose the following conditions on two maximal monotone operators T and S:

(A) There exists $x_0 \in E$ such that $\inf_{f \in Tx}(f, x - x_0)$ is lower bounded on each bounded subset of $D(T)$, i.e., for each bounded subset $K \subseteq D(T)$, there exists a constant $\alpha_K \in \mathbb{R}$ such that

Y. Chen
Guangdong University of Technology, Guangzhou, Guangdong, People's Republic of China

Y. J. Cho (✉)
Department of Mathematics Education and RINS, Gyeongsang National University, Jinju, Korea
Center for General Education, China Medical University, Taichung, Taiwan
e-mail: yjcho@gnu.ac.kr

T. M. Rassias
Department of Mathematics, National Technical University of Athens, Athens, Greece
e-mail: trassias@math.ntua.gr

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_3

$$(f, x - x_0) \geq \alpha_K$$

for all $x \in K$ and $f \in Tx$;

(B) There exist $r > 0$ and $\lambda_0 > 0$ such that, if, for each $y \in B(x_0, r)$, $g \in E^*$, $x_n \in D(T)$ and $\lambda_n \in (0, \lambda_0)$ with $g \in Tx_n + S_{\lambda_n}x_n + Jx_n$ for each $n = 1, 2, \cdots$, $\{R^S_{\lambda_n}x_n\}_{n=1}^{\infty}$ is bounded, then we have $\inf_{n\geq 1}(S_{\lambda_n}x_n, R^S_{\lambda_n}x_n - y) > -\infty$, i.e., there exists $\beta(g, P, y) \in \mathbb{R}$ such that

$$\inf_{n\geq 1}(S_{\lambda_n}x_n, R^S_{\lambda_n}x_n - y) \geq \beta(g, P, y),$$

where $P = \{x_n : n \geq 1\}$.

Under these conditions (A) and (B), we prove that $S + T$ is maximal monotone. We note that a special case of the assumption (B) was used in a recent paper by Afsaw [1] to prove the maximality of the sum of two maximal monotone operators, where he assume that, for each $y \in B(x_0, r)$, $\inf_{g\in Sx}(g, x - y)$ is lower bounded on each bounded subset of $D(S)$, where r is a constant. In Sect. 4, we show that the assumption (B) is more flexible in applications. Further, if $\Omega \subset E$ is an open bounded subset such that

$$\Omega \cap D(T) \cap D(S) \neq \emptyset, \quad 0 \notin (T + S)(D(T) \cap D(S) \cap \partial\Omega)$$

and the assumptions (A) and (B) are only satisfied on $\overline{\Omega} \cap D(T)$, then we construct a degree theory for the sum $T+S$ of two maximal monotone operators, where this sum may not be maximal monotone. Also, we use this degree theory to study the operator equation $0 \in (T + S)x$. Finally, we give some examples to show applications of our results.

2 Maximality of the Sum of Two Maximal Monotone Operators

In the following, let E be a real reflexive Banach spaces and E^* be the dual space of E. We may assume that both E and E^* are locally uniform convex (otherwise, there exists equivalent norms on E and E^* such that both E and E^* are locally uniform convex, see [19], since renorming doesn't change the monotonicity (maximal monotonicity) of an operator and boundedness of subsets), so the duality operator $J : E \to E^*$ defined by $(Jx, x) = \|x\|^2 = \|Jx\|^2$ is a single-valued continuous maximal monotone operator and is also an operator of class (S_+) (see [7, 13, 20]).

Let $S : D(S) \subseteq E \to 2^{E^*}$ be a maximal monotone operator and let

$$S_\lambda x = (S^{-1} + \lambda J^{-1})^{-1}x$$

for any $x \in E$ and $\lambda > 0$ be the Yosida approximation of S and $R_\lambda^S x = x - \lambda J^{-1} S_\lambda x$ be the resolvent with respect to S_λ. Then it is well known that S_λ is a continuous maximal monotone operator.

Lemma 2.1 *If $\{\lambda_n\}_{n=1}^{\infty} \subset (0, +\infty)$ and $\{x_n\}_{n=1}^{\infty}$ are bounded, $x_0 \in D(S)$ and*

$$(S_{\lambda_n} x_n, x_n - x_0) \leq L$$

for some $L \in \mathbb{R}$, then $\{R_{\lambda_n}^S x_n\}_{n=1}^{\infty}$ is bounded.

Proof Since $S_{\lambda_n} x_n \in S R_{\lambda_n}^S x_n$, we have

$$(S_{\lambda_n} x_n, R_{\lambda_n}^S x_n - x_0) \geq (f_0, R_{\lambda_n}^S x_n - x_0),$$

where $f_0 \in S x_0$ is a fixed element. Note that

$$(S_{\lambda_n} x_n, x_n - x_0) = (S_{\lambda_n} x_n, x_n - R_{\lambda_n}^S x_n) + (S_{\lambda_n} x_n, R_{\lambda_n}^S x_n - x_0)).$$

So, we have

$$(S_{\lambda_n} x_n, x_n - R_{\lambda_n}^S x_n) \leq L + \|f_0\| \|R_{\lambda_n}^S x_n - x_0\|. \tag{2.1}$$

Since $S_{\lambda_n} x_n = \lambda_n^{-1} J(x_n - R_{\lambda_n}^S x_n)$, it follows from (2.1) that

$$\|x_n - R_{\lambda_n}^S x_n\|^2 \leq \lambda_n (L + \|f_0\| \|x_n - R_{\lambda_n}^S x_n\|). \tag{2.2}$$

By the boundedness of $\{\lambda_n\}_{n=1}^{+\infty}$, $\{x_n\}_{n=1}^{\infty}$ and (2.2), it follows that $\{R_{\lambda_n}^S x_n\}_{n=1}^{\infty}$ is bounded. This completes the proof.

Remark 2.1 A special case of Lemma 2.1 was proved in [1] under the assumption that $\lambda_n \to 0^+$, where we only require that $\{\lambda_n\}_{n=1}^{\infty}$ is bounded.

Lemma 2.2 *Let $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators and let $x_0 \in D(S) \cap D(T)$, $x_n \in D(T)$, $f_n \in T x_n$ and $\lambda_n \in (0, \lambda)$ be such that $f_n + S_{\lambda_n} x_n + J x_n = f$ for each $n = 1, 2, \cdots$. Then $\{x_n\}_{n=1}^{\infty}$ is bounded and*

$$(S_{\lambda_n} x_n, x_n - x_0) \leq L$$

for some $L \in \mathbb{R}$.

Proof Take $f_0 \in Tx_0$. Then, by the assumption $f_n + S_{\lambda_n}x_n + Jx_n = f$ for each $n = 1, 2, \cdots$, we have

$$\begin{aligned}&(f - f_0 - S_{\lambda_n}x_0, x_n - x_0)\\ &= (f_n - f_0, x_n - x_0) + (S_{\lambda_n}x_n - S_{\lambda_n}x_0, x_n - x_0) + (Jx_n, x_n - x_0).\end{aligned}$$

The monotonicity of T and S_λ yields

$$(f - f_0 - S_{\lambda_n}y_0, x_n - y_0) \geq (Jx_n, x_n - y_0). \tag{2.3}$$

By (2.3) and $\lim_{n\to\infty} S_{\lambda_n}y_0 = g_0 \in Sy_0$ (see [5, 14]), we deduce that $\{x_n\}_{n=1}^{\infty}$ is bounded. Again, by the assumption $f_n + S_{\lambda_n}x_n + Jx_n = f$ for each $n = 1, 2, \cdots$, we have

$$(f, x_n - x_0) = (f_n + S_{\lambda_n}x_n + Jx_n, x_n - x_0)$$

and

$$(f_n, x_n - x_0) \geq (f_0, x_n - x_0).$$

Therefore, we have

$$(S_{\lambda_n}x_n, x_n - x_0) \leq (f - f_0 - Jx_n, x_n - x_0)$$

for each $n = 1, 2\cdots$. Thus there exists $L > 0$ such that $(S_{\lambda_n}x_n, x_n - x_0) \leq L$ for each $n = 1, 2, \cdots$. This completes the proof.

Lemma 2.3 *Let $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators such that $D(S) \cap D(T) \neq \emptyset$ and let $x_n \in D(T)$, $f_n \in Tx_n$ and $\lambda_n \in (0, \lambda)$ be such that $f_n + S_{\lambda_n}x_n + Jx_n = f$ for each $n = 1, 2, \cdots$. Suppose that there exist $x_0 \in E$, $\alpha \in \mathbb{R}$ and $r > 0$ such that*

(1) $(f_n, x_n - x_0) \geq \alpha$;
(2) *If $\{R^S_{\lambda_n}x_n\}_{n=1}^{\infty}$ is bounded and, for each $y \in B(x_0, r)$, there exists $\beta(y) \in \mathbb{R}$ such that*

$$(S_{\lambda_n}x_n, R^S_{\lambda_n}x_n - y) \geq \beta(y)$$

for each $n = 1, 2, \cdots$.

Then both $\{x_n\}_{n=1}^{\infty}$ and $\{S_{\lambda_n}x_n\}_{n=1}^{\infty}$ are bounded.

Proof Take $y_0 \in D(T) \cap D(S)$. Then, by Lemma 2.2, $\{x_n\}_{n=1}^{+\infty}$ is bounded and there exists $L \in \mathbb{R}$ such that $(S_{\lambda_n}x_n, x_n - y_0) \leq L$ for each $n = 1, 2, \cdots$. By Lemma 2.1, we know that $\{R^S_{\lambda_n}x_n\}_{n=1}^{\infty}$ is bounded. Once again, by the assumption $f_n + S_{\lambda_n}x_n + Jx_n = f$ for each $n = 1, 2, \cdots$, we have

$$(f_n, x_n - x_0) + (S_{\lambda_n} x_n, x_n - x_0) + (Jx_n, x_n - x_0) = (f, x_n - x_0). \tag{2.4}$$

By (2.4) and the assumption (1), it follows that there exists a constant $k \in \mathbb{R}$ such that

$$(S_{\lambda_n} x_n, x_n - x_0) \leq k. \tag{2.5}$$

Since

$$\begin{aligned}&(S_{\lambda_n} x_n, z)\\ &= (S_{\lambda_n} x_n, R^S_{\lambda_n} x_n - x_0 + z) + (S_{\lambda_n} x_n, x_n - R^S_{\lambda_n} x_n) - (S_{\lambda_n} x_n, x_n - x_0)\end{aligned}$$

for all $z \in B(0, r)$, by the assumption (2) and (2.5), we have

$$(S_{\lambda_n} x_n, z) \geq \beta(x_0 - z) - k \tag{2.6}$$

for all $z \in B(0, r)$. Note that $-B(0, r) = B(0, r)$ and so, by (2.6), we know that $\{(S_{\lambda_n} x_n, z)\}_{n=1}^{\infty}$ is bounded for each $z \in B(0, r)$. By the uniform bounded principle (see [17]), it follows that $\{S_{\lambda_n} x_n\}_{n=1}^{\infty}$ is bounded. This completes the proof.

The following lemma is a strong version of Theorem 2.1 in [5] and we also need this lemma in the next section to construct a degree theory:

Lemma 2.4 *If $\lambda_n \to 0^+$ as $n \to \infty$, $f \in E^*$, $\{x_n\}_{n=1}^{\infty} \subset D(T)$ is bounded, $f_n \in Tx_n$ and $\{S_{\lambda_n} x_n\}_{n=1}^{\infty}$ is bounded such that*

$$f_n + S_{\lambda_n} x_n + Jx_n = f \tag{2.7}$$

for each $n = 1, 2, \cdots$, then $x_n \to y_0 \in D(T) \cap D(S)$ in E and $f \in Ty_0 + Sy_0 + Jy_0$.

Proof Without loss of generality, we may assume that $x_n \rightharpoonup y_0 \in E$, $f_n \rightharpoonup g_1 \in E^*$ and $Jx_n \rightharpoonup g_2 \in E^*$ as $n \to \infty$ and thus $S_{\lambda_n} x_n \rightharpoonup f - g_1 - g_2$. By the monotonicity of the operators T, S_λ and (2.7), we have

$$\limsup_{n,m \to \infty} (Jx_n - Jx_m, x_n - x_m) \leq 0. \tag{2.8}$$

Since

$$\begin{aligned}(Jx_n - Jx_m, x_n - x_m) &= \|x_n\|^2 - Jx_n(x_m) - Jx_m(x_n) + \|x_m\|^2\\ &\geq (\|x_n\| - \|x_m\|)^2,\end{aligned}$$

we have $\lim_{n \to \infty} \|x_n\|$ exists, which is denoted by r. Therefore, we must have

$$\lim_{n,m \to \infty} Jx_n(x_m) = r^2, \qquad \lim_{n,m \to \infty} Jx_m(x_n) = r^2,$$

i.e., for any $\varepsilon > 0$, there exists $N > 0$ such that

$$|Jx_n(x_m) - r^2| < \varepsilon \tag{2.9}$$

for each $n, m > N$. First, letting $m \to \infty$ and then, letting $n \to \infty$ in (2.9), we have $g_2(y_0) = r^2$. Since

$$\|g_2\| \le \liminf_{n\to\infty} \|Jx_n\|, \quad \|y_0\| \le \liminf_{n\to\infty} \|x_n\|,$$

we must have $\|g_2\| = \|y_0\| = r$. The locally uniform convexity of E implies that $x_n \to y_0$ and so $Jx_n \to Jy_0 = g_2$. By the monotonicity of S and T, we have

$$(S_{\lambda_n}x_n - g, R^S_{\lambda_n}x_n - x) \ge 0$$

for all $x \in D(S)$, $g \in Sx$ and $(f_n - h, x_n - y) \ge 0$ for all $y \in D(T)$ and $h \in Ty$. By letting $n \to \infty$ and noting that $R^S_{\lambda_n}x_n \to y_0$, we have

$$(g_1 - h, y_0 - y) \ge 0$$

for all $y \in D(T)$, $h \in Ty$ and $(f - g_1 - Jy_0 - g, y_0 - x) \ge 0$ for all $x \in D(S)$ and $g \in Sx$. The maximality of T and S implies that

$$g_1 \in Ty_0, \quad f - g_1 - Jy_0 \in Sy_0,$$

i.e., $f \in Sy_0 + Ty_0 + Jy_0$. This completes the proof.

Now, we give the main results in this section.

Theorem 2.1 *Let $S : D(S) \subseteq E \to 2^{E^*}$ and $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators with $D(S) \cap D(T) \neq \emptyset$. Suppose that there exist $x_0 \in E$, $r > 0$ and $\lambda_0 > 0$ such that*

(1) *For each bounded subset $K \subseteq D(T)$, there exists a constant $\alpha_K \in \mathbb{R}$ such that*

$$(f, x - x_0) \ge \alpha_K$$

for all $x \in K$ and $f \in Tx$;

(2) *There exist $r > 0$ and $\lambda_0 > 0$ such that, if, for each $y \in B(x_0, r)$, $g \in E^*$, $x_n \in D(T)$ and $\lambda_n \in (0, \lambda_0)$ with $g \in Tx_n + S_{\lambda_n}x_n + Jx_n$ for each $n = 1, 2, \cdots$, $\{R^S_{\lambda_n}x_n\}_{n=1}^{\infty}$ is bounded, then we have $\inf_{n\ge1}(S_{\lambda_n}x_n, R^S_{\lambda_n}x_n - y) > -\infty$, i.e., there exists $\beta(g, P, y) \in \mathbb{R}$ such that*

$$\inf_{n\ge1}(S_{\lambda_n}x_n, R^S_{\lambda_n}x_n - y) \ge \beta(g, P, y),$$

where $P = \{x_n : n \ge 1\}$.

Then $S + T$ is maximal monotone.

Proof We only need to prove that $S + T + J$ is surjective. For any $f \in E^*$ and $\lambda > 0$, it is well known that $T + S_\lambda + J$ is surjective. Take $\lambda_n \to 0^+$, then there exist $x_n \in D(T)$ and $f_n \in Tx_n$ such that

$$f_n + S_{\lambda_n} x_n + J x_n = f \tag{2.10}$$

for each $n = 1, 2, \cdots$. By (2.10), Lemmas 2.2 and 2.1, $\{R^S_{\lambda_n} x_n\}_{n=1}^\infty$ is bounded and so, by Lemma 2.3, we know that both $\{x_n\}_{n=1}^\infty$ and $\{S_{\lambda_n} x_n\}_{n=1}^\infty$ are bounded. By Lemma 2.4, $\{x_n\}_{n=1}^\infty$ converges to a point $y_0 \in E^*$ as $n \to \infty$ and $f \in Sy_0 + Ty_0 + Jy_0$. Thus $S + T + J$ is surjective. Therefore, $S + T$ is maximal monotone. This completes the proof.

Corollary 2.1 *Let $S : D(S) \subseteq E \to 2^{E^*}$ and $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators with $D(S) \cap D(T) \neq \emptyset$. Suppose that there exist $x_0 \in E$, $r > 0$ and $\lambda_0 > 0$ such that*

(1) *For each bounded subset $K \subseteq D(T)$, there exists a constant $\alpha_K \in \mathbb{R}$ such that*

$$(f, x - x_0) \geq \alpha_K$$

for all $x \in K$ and $f \in Tx$;

(2) *For each $y \in B(y_0, r)$, each bounded subset Q of $D(T)$ and each bounded subset P of $\cup_{\lambda \in (0,\lambda_0)} R^S_\lambda(Q)$, there exists a constant $\beta(P, Q, y) \in \mathbb{R}$ such that*

$$(g, x - y) \geq \beta(P, Q, y)$$

for all $x \in P$ and $g \in Sx$.

Then $T + S$ is maximal monotone.

Proof For any $g \in E^*$ and $\lambda_n \in (0, \lambda_0)$ with $g \in Tx_n + S_{\lambda_n} x_n + Jx_n$ for each $n = 1, 2, \cdots$, Lemmas 2.2 and 2.1 imply that $Q = \{x_n\}_{n=1}^\infty \subset D(T)$ and $P = \{R^S_{\lambda_n} x_n\}_{n=1}^\infty \subset \cup_{\lambda \in (0,\lambda_0)} R^S_\lambda(Q)$ are bounded. By the assumption (2), we have

$$\inf_{n \geq 1} (S_{\lambda_n} x_n, R^S_{\lambda_n} x_n - y) \geq \beta(P, Q, y)$$

and so the conclusion follows from Theorem 2.1. This completes the proof.

Corollary 2.2 *Let $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators, and $D(S) \cap D(T) \neq \emptyset$. Suppose that there exist $x_0 \in E$, $r > 0$, such that*

(1) *For each bounded subset $K \subseteq D(T)$, there exists a constant $\alpha_K \in \mathbb{R}$ such that*

$$(f, x - x_0) \geq \alpha_K$$

for all $x \in K$ and $f \in Tx$;

(2) *For each* $y \in B(x_0, r)$ *and each bounded subset* $P \subseteq D(S)$*, there exists a constant* $\beta(P, y) \in \mathbb{R}$ *such that*

$$(g, x - y) \geq \beta(P, y)$$

for all $x \in P$ *and* $g \in Sx$.

Then $S + T$ *is maximal monotone.*

Proof Since $\cup_{\lambda>0} R_\lambda^S(D(T)) \subseteq D(S)$, the conclusion of Corollary 2.2 follows from Corollary 2.1.

Remark 2.2 Corollary 2.2 was proved in [1].

Corollary 2.3 *Let* $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ *be two maximal monotone operators with* $x_0 \in D(S) \cap D(T)$, $L : E \to \mathbb{R}$ *be a bounded function and* $\gamma : E \to \mathbb{R}$ *be a function. Suppose that* $(g, x - y) \geq L(x) + \gamma(y)$ *for all* $x \in D(T)$, $g \in Tx$ *and* $y \in B(x_0, r_0)$*, where* $r_0 > 0$ *is a constant. Then* $S + T$ *is maximal monotone.*

Proof Take $f_0 \in Tx_0$, then we have $(f, x - x_0) \geq (f_0, x - x_0)$ for all $x \in D(T)$ and $f \in Tx$. For any bounded subset $P \subseteq D(S)$, set $\beta(P, y) = \inf_{x \in P} L(x) + \gamma(y)$. It is easy to see that the conditions of Corollary 2.2 are satisfied and so the conclusion of Corollary 2.3 follows from Corollary 2.2.

Remark 2.3 Corollary 2.3 was proved in [11].

Corollary 2.4 *Let* $T : D \subseteq E \to 2^{E^*}$, $S : D \subseteq E \to 2^{E^*}$ *be two maximal monotone mappings and* $\Omega \subset E$ *be an open bounded subset. Suppose that there exist* $x_0 \in D \cap \Omega$ *and* $r > 0$ *such that, for each* $y \in B(x_0, r)$*, each bounded subset* $Q \subseteq D$ *and each* $P \subseteq \cup_{0<\lambda<\lambda_0} R_\lambda^S(Q)$*, there exists a constant* $\beta(P, Q, y) \in \mathbb{R}$ *such that* $(g, x - y) \geq \beta(P, Q, y)$ *for all* $x \in P$ *and* $g \in Sx$ *and*

$$0 \notin \overline{\cup_{t \in [0,1]} [tT + (1 - t)S](\partial\Omega \cap D)}.$$

Then $deg(T, \Omega \cap D, 0) = deg(S, \Omega \cap D, 0)$.

Proof For each $t \in (0, 1)$, we have $(tf, x - x_0) \geq t(f_0, x - x_0)$ for all $x \in D$, $f \in Tx$, where $f_0 \in Tx_0$ is a fixed element. For each $y \in B(x_0, r)$, each bounded subset $Q \subseteq D(T)$ and each $P \subseteq \cup_{0<\lambda<\lambda_0} R_\lambda^S(Q)$, we have

$$((1 - t)g, x - y) \geq (1 - t)\beta(P, Q, y)$$

for all $x \in P$, $g \in Sx$ and $t \in (0, 1)$. By Theorem 2.1, $(1 - t)S + tT$ is maximal monotone for all $t \in (0, 1)$. By Theorem 2.9 in [10], we have

$$deg(T, \Omega \cap D, 0) = deg(T, \Omega \cap D, 0).$$

This completes the proof.

For homotopy properties of degree theory for monotone mappings, see [7, 10, 13, 20].

3 Degree for the Sum of Two Maximal Monotone Operators

Let $T : D(T) \subseteq E \to 2^{E^*}$, $(S : D(S) \subseteq E \to 2^{E^*}$ be two maximal monotone operators, $\Omega \subset E$ be an open bounded subset with $\Omega \cap D(T) \cap D(S) \neq \emptyset$ and $0 \notin (T + S)(D(T) \cap D(S) \cap \partial\Omega$.

In fact, we do not have maximal monotonicity condition on $T+S$. In this section, we construct a degree theory $deg(T + S, \Omega \cap D(T) \cap D(S), 0)$. For this, we always assume that there exist $y_0 \in E$, $r > 0$ and $\lambda_0 > 0$ such that

(C) There exists a constant $\alpha \in \mathbb{R}$ such that $(f, x - y_0) \geq \alpha$ for all $x \in \overline{\Omega} \cap D(T)$ and $f \in Tx$;

(D) For each $y \in B(y_0, r)$, there exists a constant $\beta(y) \in \mathbb{R}$ such that

$$(g, x - y) \geq \beta(y)$$

for all $x \in \{\overline{\Omega} \cap D(T) \cap D(S)\} \cup \{\cup_{\lambda \in (0,\lambda_0)} R_\lambda^S(D(T) \cap \overline{\Omega})\}$ and $g \in Sx$.

First, we give some lemmas for the main results.

Lemma 3.1 *Under the above assumptions (C) and (D), there exists $\varepsilon_0 > 0$ such that*

$$0 \notin (T + S + \varepsilon J)(D(T) \cap D(S) \cap \partial\Omega)$$

for all $\varepsilon \in (0, \varepsilon_0)$.

Proof Suppose that there exist a sequence $\{\varepsilon_n\}$ with $\varepsilon_n \to 0^+$, $x_n \in D(T) \cap D(S) \cap \partial\Omega$, $f_n \in Tx_n$ and $g_n \in Sx_n$ such that

$$f_n + g_n + \varepsilon_n J x_n = 0 \tag{3.1}$$

for each $n = 1, 2, \cdots$. We may also assume that $x_n \rightharpoonup x_0$ by taking a subsequence $\{x_n\}$. By the monotonicity of T, S and (3.1), we have

$$(\varepsilon_n J x_n - \varepsilon_m J x_m, x_n - x_m) \leq 0 \tag{3.2}$$

for each $n = 1, 2, \cdots$. By letting $m \to \infty$ in (3.2), we have $(\varepsilon_n J x_n, x_n - x_0) \leq 0$ for each $n = 1, 2, \cdots$. Since J is a mapping of class (S_+), we have $x_n \to x_0 \in \partial\Omega$. Again, by (3.1), we have

$$(f_n, x_n - y) + (g_n, x_n - y) + \varepsilon_n (J x_n, x_n - y) = 0 \tag{3.3}$$

for each $n = 1, 2, \cdots$. Set $y = y_0 - z$ for any $z \in B(0, r_0)$ and using the assumptions (C) and (D), it follows from (3.3) that

$$(f_n, z) \leq -\alpha - \beta(y_0 - z) + \varepsilon_n \|x_n\| \|x_n - y_0 + z\|$$

for each $n = 1, 2, \cdots$. Note that $-B(0, r_0) = B(0, r_)$ and so we know that $\{(f_n, z)\}_{n=1}^{\infty}$ is bounded for all $z \in B(0, r_0)$. Therefore, it follows from the uniform bounded principle that $\{f_n\}_{n=1}^{\infty}$ is bounded and so $\{g_n\}_{n=1}^{\infty}$ is also bounded. We may assume that $f_n \rightharpoonup f_0$ and so $g_n \rightharpoonup -f_0$. By using the maximality of T and S, it is easy to check that $x_0 \in D(T) \cap D(S)$, $f_0 \in D(T)$ and $-f_0 \in Sx_0$ and so $0 \in Tx_0 + Sx_0$, which is a contradiction. This completes the proof.

Lemma 3.2 *For each $\varepsilon \in (0, \varepsilon_0)$, where ε_0 is as in Lemma 3.1, there exists $\lambda_0 > 0$ such that*

$$0 \notin (T + S_\lambda + \varepsilon J)(D(T) \cap \partial\Omega)$$

for all $\lambda \in (0, \lambda_0)$.

Proof Suppose that there exist $\lambda_n \to 0^+$, $x_n \in D(T) \cap \partial\Omega$ and $f_n \in Tx_n$ such that

$$f_n + S_{\lambda_n} x_n + \varepsilon J x_n = 0 \tag{3.4}$$

for each $n = 1, 2, \cdots$. By using the same proof as in Lemma 2.3, we know that $\{S_{\lambda_n} x_n\}_{n=1}^{\infty}$ is bounded. By Lemma 2.4, we know that $\{x_n\}_{n=1}^{\infty}$ converges to a point $x_0 \in \partial\Omega$ as $n \to \infty$ and so $0 \in Tx_0 + Sx_0$, which is a contradiction. This completes the proof.

Lemma 3.3 *For any $\lambda_1, \lambda_2 > 0$, $\{T_{t\lambda_1 + (1-t)\lambda_2} + \varepsilon J\}_{t \in [0,1]}$ is a homotopy of mappings of class (S_+).*

Proof Suppose that $x_n \rightharpoonup x_0 \in E$, $t_n \to t_0$ and

$$\limsup_{n \to \infty} (S_{t_n\lambda_1 + (1-t_n)\lambda_2} x_n + \varepsilon J x_n, x_n - x_0) \leq 0.$$

By the monotonicity of S_λ, we have

$$(S_{t_n\lambda_1 + (1-t_n)\lambda_2} x_n - (S_{t_n\lambda_1 + (1-t_n)\lambda_2} x_0, x_n - x_0) \geq 0$$

and, by Lemma 2.7 in [10], it follows that

$$\lim_{n \to \infty} S_{t_n\lambda_1 + (1-t_n)\lambda_2} x_0 = S_{t_0\lambda_1 + (1-t_0)\lambda_2} x_0$$

and so

$$\limsup_{n \to \infty} (\varepsilon J x_n, x_n - x_0) \leq 0.$$

Thus $x_n \to x_0$ and $S_{t_n\lambda_1+(1-t_n)\lambda_2}x_n + \varepsilon Jx_n \to S_{t_0\lambda_1+(1-t_0)\lambda_2}x_0 + \varepsilon Jx_0$. This completes the proof.

Since $S_\lambda + \varepsilon J$ is a mapping of class (S_+) for each $\lambda > 0$ and $\varepsilon > 0$, under the above assumptions, by Lemma 3.2 and [7] (see also [14]), we know that

$$deg(T + S_\lambda + \varepsilon J, \Omega \cap D(T), 0)$$

is well defined for any $\lambda \in (0, \lambda_0(\varepsilon))$ and $\varepsilon \in (0, \varepsilon_0)$ and we define

$$deg(T + S + \varepsilon J, D(T) \cap D(S) \cap \Omega, 0) = \lim_{\lambda \to 0^+} deg(T + S_\lambda + \varepsilon J, D(T) \cap \Omega, 0). \tag{3.5}$$

By Lemma 3.3, we know that $deg(T + S_\lambda + \varepsilon J, \Omega \cap D(T), 0)$ does not depend on $\lambda \in (0, \lambda_0(\varepsilon))$ and so $deg(T + S + \varepsilon J, D(T) \cap D(S) \cap \Omega, 0)$ is well defined.

Finally, we define

$$deg(T + S, D(T) \cap D(S) \cap \Omega, 0) = \lim_{\varepsilon \to 0^+} deg(T + S + \varepsilon J, D(T) \cap D(S) \cap \Omega, 0). \tag{3.6}$$

Now, we need to check that (3.6) is well defined. To do this, we need the following:

Lemma 3.4 *For any $\varepsilon_i \in (0, \varepsilon_0)$, $i = 1, 2$, there exists $\eta > 0$ such that*

$$0 \notin \cup_{t\in[0,1]}[T + S_\lambda + t\varepsilon_1 J + (1 - t)\varepsilon_2 J](D(T) \cap \partial\Omega)$$

for all $\lambda \in (0, \eta)$.

Proof Suppose that the conclusion is not true. Then there exist a sequence $\{\lambda_n\}_{n=1}^\infty$ with $\lambda_n \to 0^+$, $x_n \in D(T) \cap \partial\Omega$, $f_n \in Tx_n$ and $t_n \in [0, 1]$ such that

$$f_n + S_{\lambda_n}x_n + [\varepsilon_1 t_n + (1 - t_n)\varepsilon_2]Jx_n = 0$$

for each $n = 1, 2, \cdots$. For simplicity, we denote by $\varepsilon_n = t_n\varepsilon_1 + (1 - t_n)\varepsilon_2$. From the same reasoning as in Lemma 2.3, it follows that $\{S_{\lambda_n}x_n\}_{n=1}^\infty$ is bounded and so $\{f_n\}_{n=1}^\infty$ is also bounded. We may assume that $t_n \to t_0$, $x_n \rightharpoonup x_0$, $f_n \rightharpoonup g_0 \in E^*$ and $Jx_n \rightharpoonup g_1 \in E^*$ and so

$$S_{\lambda_n}x_n \rightharpoonup -g_0 - [t_0\varepsilon_1 + (1 - t_0)\varepsilon_2]g_1.$$

Since

$$(f_n - f_m, x_n - x_m) + (S_{\lambda_n}x_n - S_{\lambda_m}x_m, x_n - x_m) + (\varepsilon_n Jx_n - \varepsilon_m Jx_m, x_n - x_m) = 0,$$

for each $n = 1, 2\cdots$, by using the monotonicity of T and S, we have

$$\limsup_{n,m\to\infty}(\varepsilon_n Jx_n - \varepsilon_m Jx_m, x_n - x_m) \le 0. \tag{3.7}$$

Note that $\varepsilon_n \to t_0\varepsilon_1 + (1-t_2)\varepsilon_2 \in (0, \varepsilon_0)$ as $n \to \infty$, it follows from (3.7) that

$$\limsup_{n,m\to\infty}(Jx_n - Jx_m, x_n - x_m) = 0.$$

The same proof as in Lemma 2.4, we have $x_n \to x_0 \in \partial\Omega$ and so $Jx_n \to Jx_0$.

Finally, by using the monotonicity of T and S, we have $(f_n - f, x_n - x) \geq 0$ for all $x \in D(T)$, $f \in Tx$ and

$$(S_{\lambda_n}x_n - g, R_{\lambda_n}x_n - y) \geq 0$$

for all $y \in D(S)$ and $g \in Sy$. By letting $n \to \infty$ and using the maximality of T and S, we have

$$x_0 \in D(T) \cap D(S), \quad g_0 \in Tx_0, \quad -g_0 - [t_0\varepsilon_1 + (1-t_0)\varepsilon_2]Jx_0 \in Sx_0,$$

i.e., $0 \in Tx_0 + Sx_0 + [t_0\varepsilon_1 + (1-t_0)\varepsilon_2]Jx_0$, which is a contradiction to Lemma 3.1. This completes the proof.

Lemma 3.5 *$\{S_\lambda + t\varepsilon_1 J + (1-t)\varepsilon_2 J\}_{t\in[0,1]}$ is a homotopy of mappings of class (S_+).*

Proof The proof is straightforward, so that we omit the details.

By Lemmas 3.4, 3.5 and [10] (see also Theorem 6.3.2 in [13]), we know

$$deg(T + S_\lambda + \varepsilon_1 J, D(T) \cap \Omega, 0) = deg(T + S_\lambda + \varepsilon_2 J, D(T) \cap \Omega, 0)$$

for all $\lambda \in (0, \eta)$ and so

$$deg(T + S + \varepsilon_1 J, D(T) \cap D(S) \cap \Omega, 0) = deg(T + S + \varepsilon_2 J, D(T) \cap D(S) \cap \Omega)$$

for and $\varepsilon_i \in (0, \varepsilon_0)$, $i = 1, 2$. Thus (3.6) is well defined.

Theorem 3.1 *If $deg(T + S, D(T) \cap D(S) \cap \Omega, 0) \neq 0$, then $0 \in (T + S)x$ has a solution in $D(T) \cap D(S) \cap \Omega$.*

Proof Since $deg(T + S, D(T) \cap D(S) \cap \Omega, 0) \neq 0$, we have

$$deg(T + S + \varepsilon J, D(T) \cap D(S) \cap \Omega, 0) \neq 0$$

for all $\varepsilon \in (0, \varepsilon_0)$. By (3.5), we have $deg(T + S_\lambda + \varepsilon J, D(T) \cap \Omega, 0) \neq 0$ for all $\lambda \in (0, \lambda_0(\varepsilon))$ and so $0 \in (T + S_\lambda + \varepsilon J)x$ has a solution in $D(T) \cap \Omega$. Take $\lambda_n \to 0^+$, then there exist $x_n \in \Omega \cap D(T)$ and $f_n \in Tx_n$ such that $f_n + S_{\lambda_n}x_n + \varepsilon Jx_n = 0$ and so

$$(f_n, x_n - y) + (S_{\lambda_n}x_n, x_n - y) + \varepsilon(Jx_n, x_n - y) = 0$$

for all $y \in B(y_0, r_0)$ and $n = 1, 2, \cdots$. By using the assumptions (C) and (D), we know that $\{S_{\lambda_n} x_n\}_{n=1}^{\infty}$ is bounded. By Lemma 2.4, we may assume that $x_n \to z_\varepsilon$ and $0 \in (T + S + \varepsilon J)x_\varepsilon$.

Now, we take $\varepsilon_j \to 0^+$, then there exist $x_j \in D(T) \cap D(S) \cap \overline{\Omega}$, $x_j \rightharpoonup z_0$, $f_j \in Tx_j$ and $g_j \in Sx_j$ such that

$$f_j + g_j + \varepsilon_j J x_j = 0 \tag{3.8}$$

for each $j = 1, 2, \cdots$. By (3.8) and the monotonicity of T and S, we have

$$(\varepsilon_i J x_i - \varepsilon_j J x_i - x_j) \leq 0 \tag{3.9}$$

for each $i, j = 1, 2, \cdots$. The same reason as in (3.2), we have $x_i \to x_0 \in \overline{\Omega}$ as $i \to \infty$.

Again, by (3.8), we have

$$(f_j, x_j - y) + (g_j, x_j - y) + \varepsilon_j (J x_j, x_j - y) = 0$$

for all $y \in B(y_0, r_0)$ and $j = 1, 2, \cdots$. By using the assumptions (C) and (D), we have

$$(f_j, z) \geq -\alpha - \beta(y_0 - z) - \varepsilon_j \|x_j\| \|x_j - y_0 - z\|$$

for all $z \in B(0, r)$ and the symmetry of $B(0, r)$ implies that $\{(f_j, z)\}_{j=1}^{\infty}$ is bounded for each $z \in B(0, r)$. By using the uniform bounded principle, $\{f_j\}_{j=1}^{\infty}$ is bounded and so $\{g_j\}_{j=1}^{\infty}$ is bounded. We may assume that $f_j \rightharpoonup f_0$ by taking a subsequence $\{j_j\}$ and so $g_j \rightharpoonup -f_0$. The maximal monotonicity of T and S imply that $f_0 \in Tx_0$ and $-f_0 \in Sx_0$, which yields $0 \in (T + S)x_0$. Obviously, $x_0 \notin \partial\Omega$ and so $x_0 \in D(T) \cap D(S) \cap \Omega$. This completes the proof.

Theorem 3.2 *Let $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators and $\Omega \subset E$ be an open bounded subset with $0 \in \Omega \cap D(S) \cap D(T)$. Suppose that the assumptions (C) and (D) hold, $0 \notin (S + T)(\partial\Omega)$ and*

$$(f, x) \geq 0$$

for all $x \in D(S) \cap D(T) \cap \partial\Omega$ and $f \in Sx + Tx$. Then

$$deg(S + T, D(S) \cap D(T) \cap \Omega, 0) = 1.$$

Proof For any $\varepsilon > 0$, it is easy to see that $(tf + \varepsilon J x, x) > 0$ for all $x \in D(S) \cap D(T) \cap \partial\Omega$, $f \in Sx + Tx$ and $t \in [0, 1]$ and so

$$0 \notin [t(S + T) + \varepsilon J](D(S) \cap D(T) \cap \partial\Omega$$

for all $t \in [0, 1]$.

Next, we prove that there exists $\lambda_0 > 0$ such that

$$0 \notin [tS_\lambda + tT + \varepsilon J](\partial\Omega \cap D(T)) \tag{3.10}$$

for all $t \in [0.1]$ and $\lambda \in (0, \lambda_0)$. If this is not true, then there exist a sequence $\{t_n\}_{n=1}^{\infty}$ with $t_n \to t_0$, $x_n \in \partial\Omega$ with $x_n \rightharpoonup x_0$, $\lambda_n \to 0^+$ and $g_n \in Tx_n$ such that

$$t_n S_{\lambda_n} + t_n g_n + \varepsilon J x_n = 0 \tag{3.11}$$

for each $n = 1, 2 \cdots$.

Case (1) $t_0 = 0$: Since $\{S_{\lambda_n} x_n\}_{n=1}^{\infty}$ converges as $n \to \infty$ (see [5] or [14]) and $t_n g \to 0$ for all $g \in T0$, we have

$$\liminf_{n\to\infty} t_n(S_{\lambda_n} x_n, x_n) \geq 0, \quad \liminf_{n\to\infty} t_n(g_n, x_n) \geq 0$$

for all $g_n \in Tx_n$. Therefore, we have

$$\limsup_{n\to\infty}(Jx_n, x_n) \leq 0$$

and so $x_n \to 0 \in \partial\Omega$, which is a contradiction to $0 \in \Omega$.

Case (2) $t_0 \neq 0$: We have

$$(t_n(S_{\lambda_n} x_n, x_n - y) + (t_n g_n, x_n - y) + \varepsilon(Jx_n, x_n - y) = 0$$

for all $y \in B(y_0, r_0)$ and $n = 1, 2, \cdots$. By using the assumptions (C) and (D) and the same reasoning as in Lemma 2.3, it follows that $\{S_{\lambda_n} x_n\}_{n=1}^{+\infty}$ is bounded and so $\{g_n\}_{n=1}^{\infty}$ is bounded. Without loss of generality, we may assume that $g_n \rightharpoonup g_0 \in E^*$ and $S_{\lambda_n} x_n \rightharpoonup f_0 \in E^*$ as $n \to \infty$.

Now, we claim that

$$\liminf_{n\to\infty}(S_{\lambda_n} x_n, x_n - x_0) \geq 0, \quad \liminf_{n\to\infty}(g_n, x_n - x_0) \geq 0. \tag{3.12}$$

Otherwise, we have

$$\lim_{k\to\infty}(S_{\lambda_{n_k}} x_{n_k}, x_{n_k} - x_0) < 0 \tag{3.13}$$

and

$$\lim_{k\to\infty}(g_{n_k} x_{n_k}, x_{n_k} - x_0) < 0. \tag{3.14}$$

By the monotonicity of S and T, we have

$$(S_{\lambda_{n_k}} x_{n_k} - f, R^S_{\lambda_{n_k}} x_{n_k} - x) \geq 0 \tag{3.15}$$

for all $x \in D(S)$ and $f \in Sx$ and

$$(g_{n_k} - g, x_{n_k} - y) \geq 0 \tag{3.16}$$

for all $y \in D(T)$ and $g \in Ty$. From (3.13)–(3.16) and $x_n - R^S_{\lambda_n} x_n \to 0$, it follows that

$$(f_0 - f, x_0 - x) \geq 0$$

for all $x \in D(S)$, $f \in Sx$ and $(g_0 - g, x_0 - y) \geq 0$ for all $y \in D(S)$ and $g \in Ty$. The maximality of S and T implies that $x_0 \in D(S) \cap D(T)$, $f_0 \in Sx_0$ and $g_0 \in Tx_0$ and so

$$\liminf_{k\to\infty}(S_{\lambda_{n_k}} x_{n_k}, x_{n_k} - x_0) \geq 0, \quad \liminf_{k\to\infty}(g_{n_k}, x_{n_k} - x_0) \geq 0,$$

which are contradictions to (3.13) and (3.14), respectively, and so our claim is true. By (3.11) and (3.12), we have $\lim_{n\to\infty}(\varepsilon Jx_n, x_n - x_0) \leq 0$. Since J is a mapping of class (S_+), $x_n \to x_0 \in \partial\Omega$ and $Jx_n \to Jx_0$. Consequently, $x_0 \in D(S) \cap D(T)$, $f_0 \in Sx_0$, $g_0 \in Tx_0$ and $0 \in t_0(S + T) + \varepsilon Jx_0$, which is a contradiction. So, (3.10) is true.

Finally, for each $\lambda \in (0, \lambda_0)$, we claim that there exists $\eta_0 > 0$ such that

$$0 \neq t(S_\lambda + T_\eta)x + \varepsilon Jx \tag{3.17}$$

for all $t \in [0, 1]$, $x \in \partial\Omega$ and $\eta \in (0, \eta_0)$. If this is not true, then there exist $\eta_j \to 0$, $t_j \to t_0$, $x_j \in \partial\Omega$ and $x_j \rightharpoonup x_0$ such that $t_j T_{\eta_j} x_j + t_j S_\lambda x_j + \varepsilon Jx_j = 0$ for each $j = 1, 2, \cdots$. If $t_0 = 0$, then $t_j T_{\eta_j} 0 \to 0$ and $t_j S_\lambda x_j \to 0$ and so we have $x_j \to 0 \in \partial\Omega$, which is a contradiction. Therefore, since $t_0 \neq 0$ and $\{T_{\eta_j} x_j\}_{j=1}^\infty$ is bounded, we may assume that $T_{\eta_j} x_j \rightharpoonup g_0$ by taking a subsequence.

Now, we prove that $\liminf_{j\to\infty}(T_{\eta_j} x_j, x_j - x_0) \geq 0$. Otherwise, we have

$$\limsup_{k\to\infty}(T_{\eta_{j_k}} x_{j_k}, x_{j_k} - x_0) < 0, \tag{3.18}$$

i.e., $\limsup_{k\to\infty}(T_{\eta_{j_k}} x_{j_k}, x_{j_k}) < (g_0, x_0)$. By the monotonicity of T, we have

$$(T_{\eta_{j_k}} x_{j_k} - g, R^T_{\eta_{j_k}} x_{j_k} - x) \geq 0$$

for all $x \in D(T)$ and $g \in Tx$ and $R^T_{\eta_{j_k}} x_{j_k} - x_{j_k} \to 0$ and so we have $(g_0 - g, x_0 - x) \geq 0$ for all $x \in D(T)$ and $g \in Tx$. The maximal monotonicity of T implies that $x_0 \in D(T)$ and $g_0 \in Tx_0$ and so

$$\liminf_{k\to\infty}(T_{\eta_{j_k}} x_{j_k}, x_{j_k} - x_0) \geq 0,$$

which contradicts (3.18). Thus we have $\liminf_{j\to\infty}(T_{\eta_j}x_j, x_j - x_0) \geq 0$ and so the monotonicity of S_λ implies that $\lim_{j\to\infty}(\varepsilon J x_j, x_j - x_0) \leq 0$ and so $x_j \to x_0 \in \partial\Omega$. Consequently, $Jx_j \to Jx_0$, $S_\lambda x_j \to S_\lambda x_0$ and $T_{\eta_j}x_j \to -S_\lambda x_0 - t_0^{-1}Jx_0 \in Tx_0$, which is a contradiction to (3.10). Hence (3.17) is true. It is easy to see that $\{t(S_\lambda + T_\eta) + \varepsilon J\}_{t\in[0,1]}$ is a homotopy mappings of class (S_+) and so we have

$$deg(S_\lambda + T_\eta + \varepsilon J, \Omega, 0) = deg(\varepsilon J, \Omega, 0) = 1$$

for all $\eta \in (0, \eta_0)$. Therefore, $deg(S_\lambda + T + \varepsilon J, D(T)\cap\Omega, 0) = 1$ for all $\lambda \in (0, \lambda_0)$. By (3.5), we have

$$deg(S + T + \varepsilon J, D(S) \cap D(T) \cap \Omega, 0) = 1$$

for all $\varepsilon > 0$ and so (3.6) yields the conclusion of Theorem 3.2. This completes the proof.

Corollary 3.2 *Let $S : D(S) \subseteq E \to 2^{E^*}$, $T : D(T) \subseteq E \to 2^{E^*}$ be two maximal monotone operators and $\Omega \subset E$ be an open bounded subset with $0 \in \Omega \cap D(S) \cap D(T)$. Suppose that the assumptions (C) and (D) hold and*

$$(f, x) \geq 0, \text{ for all } x \in D(S) \cap D(T) \cap \partial\Omega, f \in (T + S)x.$$

Then $0 \in (T + S)x$ has a solution in $D(T) \cap D(S) \cap \overline{\Omega}$.

Proof Set $T_\varepsilon = T + \varepsilon J$, then it is easy to see that T_ε satisfies the assumption (3) and we have

$$(f, x) > 0$$

for all $x \in D(T) \cap D(S) \cap \partial\Omega$ and $f \in (T_\varepsilon + S)x$. By Theorem 3.2, we have

$$deg(T_\varepsilon + S, D(T) \cap D(S) \cap \Omega, 0) = 1$$

and, by Theorem 3.1, $0 \in (T + S + \varepsilon J)x$ has a solution in $D(T) \cap D(S) \cap \Omega$. Take $\varepsilon_n \to 0^+$, then there exist $x_n \in D(T) \cap D(S) \cap \Omega$, $f_n \in Tx_n$ and $g_n \in Sx_n$ such that

$$f_n + g_n + \varepsilon_n J x_n = 0$$

for each $n = 1, 2, \cdots$. By using the same proof as in Lemma 3.1, we know that $\{x_n\}$ converges to a point $x_0 \in D(T) \cap D(S) \cap \overline{\Omega}$ as $n \to \infty$ and $0 \in (T + S)x_0$. This completes the proof.

When $S = 0$ in Corollary 3.2, we have the following:

Corollary 3.3 *Let $T : D(T) \subseteq E \to 2^{E^*}$ be a maximal monotone operator and $\Omega \subset E$ be an open bounded subset with $0 \in \Omega \cap \cap D(T)$. Suppose that*

$$(f, x) \geq 0$$

for all $x \in D(T) \cap \partial\Omega$ and $f \in Tx$. Then $0 \in Tx$ has a solution in $D(T) \cap \overline{\Omega}$.

Remark 3.1 The assumption (D) in this section can be weaken to the following assumptions:

(4A) There exists $e_0 > 0$ such that, if, for $x_n \in D(T) \cap D(S) \cap \overline{\Omega}$ and $\varepsilon_n \in (0, e_0)$ with $f_n + g_n + \varepsilon_n J x_n = 0$, where $f_n \in Tx_n$ and $g_n \in Sx_n$ for each $n = 1, 2, \cdots$, then we have

$$\inf_{n \geq 1} (g_n, x_n - y) > -\infty$$

for all $y \in B(y_0, r)$;

(4B) There exists $e_0 > 0$ such that, if, for $x_n \in D(T) \cap \overline{\Omega}$, $\varepsilon \in (0, e_0)$, $\lambda_n \in (0, \lambda_0)$ with $f_n + S_{\lambda_n} x_n + \varepsilon J x_n = 0$, where $f_n \in Tx_n$ for each $n = 1, 2, \cdots$ and $\{R^s_{\lambda_n} x_n\}_{n=1}^{\infty}$ is bounded, then we have

$$\inf_{n \geq 1} (S_{\lambda_n} x_n, R_{\lambda_n} x_n - y) > -\infty$$

for all $y \in B(y_0, r)$.

4 Examples

In this section, we give some examples to illustrate the main results in this paper.

Example 4.1 Let $\Omega \subset \mathbb{R}^N$ be an open bounded subset with smooth boundary and $b_i : \mathbb{R} \to \mathbb{R}$ be differentiable functions for each $i = 1, 2, \cdots, N$. Suppose the following conditions are satisfied:

(1) $\Sigma_{i=1}^N [b_i(x_i) - b_i(y_i)](x_i - y_i) \geq 0$, where $x = (x_i)$, $y = (y_i) \in \mathbb{R}^N$;
(2) $|b_i(t)| \leq \beta |t| + \delta$ for all $t \in \mathbb{R}$ and $i = 1, 2, \cdots, N$, where $\delta, \beta > 0$ are constants;
(3) $\Sigma_{i=1}^N b_i(x_i) x_i \geq \alpha \|x\|^2 - \gamma$ for all $x = (x_i) \in \mathbb{R}^N$, where $\alpha > 0$, $\gamma > 0$.

Assume that $V : \Omega \to \mathbb{R}$ satisfying $V(x) \geq 0$ for a.e. $x \in \Omega$ and $V(\cdot) \in L^p(\Omega)$ with $p \geq 2$ and $p > N$ and $f(\cdot) \in L^2(\Omega)$. Consider the following equation:

$$\begin{cases} -\Sigma_{i=1}^N D_i [b_i((D_i u)] + V(x) u(x) = f(x), \quad a.e.\ x \in \Omega, \\ u(x) = 0, \quad a.e.\ x \in \partial\Omega. \end{cases} \tag{E4.1}$$

Let $A : H_0^1(\Omega) \to (H_0^1(\Omega))^*$ be an operator defined as follows:

$$(Au, v) = \int_\Omega [\Sigma_{i=1}^N b_i(D_i u) D_i v] dx$$

for all $u, v \in H_0^1(\Omega)$. Then we have the following:

(a) $\|Au\| \le \beta \sqrt{\int_\Omega (\Sigma_{i=1}^N |D_i u|^2) dx} + \delta \sqrt{mes(\Omega)}$ for all $u \in H_0^1(\Omega)$;
(b) $(Au - Av, u - v) \ge 0$ for $t \in \mathbb{R}$ and $u, v \in H_0^1(\Omega)$ and A is hemi-continuous and monotone and so it is maximal monotone;
(c) $(Au, u) \ge \alpha \|u\|^2_{H_0^1(\Omega)} - \gamma mes(\Omega)$.

Let $T : D(T) \subset H_0^1 \to L^2(\Omega)$ be an operator defined by

$$D(T) = \{u \in H_0^1(\Omega) : Au \in L^2(\Omega)\}$$

and $Tu = Au$ for $u \in D(T)$. Then $T : D(T) \subset L^2(\Omega) \to L^2(\Omega)$ is obviously maximal monotone. Set $Su = Vu$ for $u \in D(S) = \{u \in L^2(\Omega) : Vu \in L^2(\Omega)\}$, then it is well known that S is maximal monotone (see [5]).

For each $u \in H^1(\Omega)$, since $p > N$ and $\frac{2}{p} + \frac{1}{q} = 1$, by Sobolev's theorem, $H^1(\Omega) \subset L^{2q}(\Omega)$ and we have

$$\|Su\| \le \Big(\int_\Omega V^2 u^2 dx \Big)^{\frac{1}{2}} \le \|V\|_{L^p(\Omega)} \|u\|_{L^{2q}(\Omega)} \le k \|V\|_{L^p(\Omega)} \|u\|_{H^1(\Omega)}$$

for some $k > 0$. Therefore, we have

$$(Su, u - y) \ge -(Su, y) \ge -k \|V\|_{L^p(\Omega)} \|u\|_{H^1(\Omega)} \|y\|_{L^2(\Omega)} \tag{4.1}$$

for all $u \in H^1(\Omega)$, $y \in L^2(\Omega)$. For each $u \in D(T)$, we have $R_\lambda^S(u) \in H_0^1(\Omega)$ since R_λ^S is Lipschitz for all $\lambda > 0$.

If $u_n \in D(T)$, $\lambda_n \in (0, \lambda_0)$ and $Tu_n + S_{\lambda_n} u_n + u_n = v \in L^2(\Omega)$ for each $n = 1, 2, \cdots$, then, by (c), $\{u_n\}_{n=1}^\infty$ is bounded in $H_0^1(\Omega)$ and so (4.1) imply that the assumptions (1) and (2) in Theorem 2.1 are satisfied by taking $x_0 = 0$ and any ball $B(0, r)$ in $L^2(\Omega)$ with $r > 0$ and so $T + S$ is maximal monotone.

Note that it is impossible to check that the assumption (2) in Corollary 2.2 is satisfied. By [3], we know that

$$\begin{cases} \frac{\partial u}{\partial t} = \Sigma_{i=1}^N D_i [b_i(D_i u(t, x))] - v(x) u(t, x), \quad (t, x) \in (0, +\infty) \times \Omega, \\ u(t, x) = 0, \quad (t, x) \in (0, +\infty) \times \partial\Omega, \\ u(0, x) = u_0(x) \in H_0^1 \cap D(T) \end{cases} \tag{E4.2}$$

has a unique solution $u(t,x) \in C((0,+\infty); L^2(\Omega))$ such that $u(t,x) \in H_0^1(\Omega) \cap D(T)$ for each fixed $t \geq 0$ and $\frac{\partial u}{\partial t} \in L^\infty((0,+\infty); L^2(\Omega))$. By (c) and $(Su,u) \geq 0$ for all $u \in D(S)$, $T+S$ is coercive and so the Eq. (E4.1) has a generalized solution $u \in H_0^1(\Omega) \cap D(T)$ for all $f \in L^2(\Omega)$.

One can easily see that the solution of the Eq. (E4.1) is unique if $V(x) > 0$ for a.e. $x \in \Omega$.

Example 4.2 Let H be a real separable Hilbert space and $\phi : D(\phi) \subseteq H \to (0,+\infty]$ be a proper lower semi-continuous even function, where $D(\phi) = -D(\phi)$, $B : D(B) \subseteq H \to 2^H$ be an odd maximal monotone operator with $D(B) = -D(B)$ and $f(\cdot) \in L^2([0,T]; H)$. Consider the following equation:

$$\begin{cases} x'(t) \in -\partial\phi(x(t)) - Bx(t) + f(t), \quad a.e.\ t \in (0,T), \\ x(0) = -x(T). \end{cases} \tag{E4.3}$$

Assume that $0 \in D(\partial\phi) \cap D(B)$ and

$$\|y\| \leq L\|x\| + \alpha$$

for all $x \in D(B)$ and $y \in Bx$, where $L, \alpha > 0$ are constants and $\frac{LT}{2} < 1$.

Let $A : D(A) \subset L^2([0,T]; H) \to L^2([0,T]; H)$ be an operator defined as follows:

$$Ax(t) = x'(t) \tag{4.2}$$

for all $x(\cdot) \in D(A)$, where

$$L^2([0,T]; H) = \left\{ h(t) : [0,T] \to H \text{ is measurable}, \int_0^T \|h(t)\|^2 dt < \infty \right\}$$

and

$$D(A) = \big\{ x(t) : [0,T] \to H \text{ is continuous with } x(0) = -x(T), \\ x'(\cdot) \in L^2([0,T]; H) \big\}.$$

The norms in $L^2([0,T]; H)$ and $D(A)$ are denoted by

$$\|h(\cdot)\|_{L^2} = \sqrt{\int_0^T \|h(t)\|^2 dt}$$

and

$$\|x(\cdot)\|_\infty = \max_{t \in [0,T]} \|x(t)\|,$$

respectively. Then it is easy to see that

$$\|x(\cdot)\|_\infty \le \frac{1}{2}\sqrt{T}\|x'(\cdot)\|_{L^2} \tag{4.3}$$

for all $x(\cdot) \in D(A)$. We have

$$(Ax(\cdot), x(\cdot)) = \int_0^T (x'(t), x(t))dt = 0$$

for all $x(\cdot) \in D(A)$ and so A is monotone and, for all $g(\cdot) \in L^2([0, T]; H)$, the equation

$$\begin{cases} x'(t) + x(t) = g(t), \quad a.e.\ t \in (0, T), \\ x(0) = -x(T) \end{cases} \tag{E4.4}$$

has a unique solution

$$x(t) = -\frac{e^{-(t+T)}}{1 + e^{-T}} \int_0^T e^s g(s)ds + e^{-t} \int_0^t e^s g(s)ds$$

and so $A : D(A) \subset L^2([0, T]; H) \to L^2([0, T]; H)$ is maximal monotone.

Next, we define an operator $K : D(K) \subset L^2([0, T]; H) \to L^2([0, T]; H)$ by

$$Kx(t) = \{k(\cdot) \in L^2([0; T]; H) : k(t) \in \partial\phi(x(t)),\ \text{a.e. } t \in [0, T]\},$$

where $D(K) = \{x(\cdot) \in L^2([0, T]; H)$ such that $x(t) \in D(\partial\phi)$ a.e. $t \in [0, T]\}$. It is known that K is maximal monotone and $A + K_\lambda$ is maximal monotone for $\lambda > 0$. We also have

$$(Ax(\cdot), K_\lambda x(\cdot)) = \int_0^T x'(t)(\partial\phi)_\lambda x(t)dt = 0$$

for all $x(\cdot) \in D(A)$ and $\lambda > 0$.

Also, we define an operator $S : D(S) \to L^2([0, T]; H)$ by

$$Sx(t) = \{b(t) : b(t) \in L^2([0, T]; H) \text{ and } b(t) \in Bx(t),\ \text{a.e. } t \in [0, T]\},$$

where

$$D(S) = \{x(\cdot) \in L^2([0, T]; H) \text{ such that } x(t) \in D(B),\ \text{a.e. } t \in [0, T]\}.$$

Then S is maximal monotone and, by (4.2), we have

$$\begin{aligned}(b(\cdot), x(\cdot) - y(\cdot)) \ \ge & -\frac{3L}{2}\|x(\cdot)\|_{L^2}^2 - \alpha\sqrt{T}\|x(\cdot)\|_{L^2} - \frac{L}{2}\|y(\cdot)\|_{L^2}^2 \\ & -\alpha\sqrt{T}\|y(\cdot)\|_{L^2}\end{aligned}$$

for all $x(\cdot) \in D(S)$, $b(\cdot) \in Sx(\cdot)$ and $y(\cdot) \in L^2([0,T];H)$ and so, by Corollary 2.3, $A + K_\lambda + S$ is maximal monotone for all $\lambda > 0$. Thus $A + K_\lambda + S + \varepsilon I$ is surjective for all $\lambda > 0$ and $\varepsilon > 0$, i.e., the equation

$$\begin{cases} x'(t) + (\partial\phi)_\lambda x(t) + Sx(t) + \varepsilon x(t) \ni f(t), \quad a.e.\ t \in (0,T), \\ x(0) = -x(T) \end{cases} \tag{E4.5}$$

has a solution for all $\varepsilon > 0$, $\lambda > 0$.

For any fixed $\lambda > 0$, take $\varepsilon_n \to 0^+$, then there exist $x_n(\cdot) \in D(A) \cap D(S)$ and $f_n(\cdot) \in Sx_n(\cdot)$ such that

$$x_n'(t) + (\partial\phi)_\lambda x_n(t) + f_n(t) + \varepsilon_n x_n(t) = f(t) \tag{4.4}$$

for a.e. $t \in (0,T)$ and $n = 1, 2, \cdots$. Both sides of (4.4) take inner products with $x_n'(t)$ and integrate over $[0,T]$, we have

$$\int_0^T \|x_n'(s)\|^2 ds + \int_0^T (f_n(s), x_n'(s)) ds = \int_0^T (f(s), x_n'(s)) ds,$$

which, together with (4.2) and (4.3), yields

$$\|x_n'(\cdot)\| \le \left(1 - \frac{LT}{2}\right)^{-1} \|f(\cdot)\|_{L^2} + \alpha\sqrt{T}$$

and

$$\|x_n(\cdot)\|_\infty \le \frac{\sqrt{T}}{2}\left(1 - \frac{LT}{2}\right)^{-1} \|f(\cdot)\|_{L^2} + \frac{\alpha T}{2}.$$

Therefore, $\|f_n(t)\| \le L\|x_n(\cdot)\|_\infty + \alpha$ for a.e. $t \in [0,T]$ and so $\{\partial\phi_\lambda x_n(\cdot)\}_{n=1}^\infty$ is bounded in $L^2([0,T];H)$.

We may assume that $x_n'(\cdot) \rightharpoonup y(\cdot)$, $(\partial\phi)_\lambda x_n(\cdot) \rightharpoonup g_1(\cdot)$, $f_n(\cdot) \rightharpoonup g(\cdot)$ and $x_n(\cdot) \rightharpoonup x(\cdot)$ by taking subsequences in $L^2([0,T];H)$. By (4.4), we have

$$\begin{aligned} &\int_0^T (x_n'(t) - x_m'(t) + (\partial\phi)_\lambda x_n(t) - (\partial\phi)_\lambda x_m(t), x_n(t) - x_m(t)) dt \\ &\quad + \int_0^T (f_n(t) - f_m(t) + \varepsilon_n x_n(t) - \varepsilon_m x_m(t), x_n(t) - x_m(t)) dt \\ &= 0. \end{aligned}$$

The monotonicity of A, K_λ and S imply that

$$\int_0^T (\varepsilon_n x_n(t) - \varepsilon_m x_m(t), x_n(t) - x_m(t)) dt \le 0.$$

The same reason as in (3.2), we have $x_n(\cdot) \to x(\cdot)$ in $L^2([0,T];H)$ and so, from the monotonicity of A, K_λ and S,

$$\lim_{n,m\to\infty} \int_0^T (x_n'(t) - x_m'(t), x_n(t) - x_m(t))dt = 0,$$

$$\lim_{n,m\to\infty} \int_0^T ((\partial\phi)_\lambda x_n(t) - (\partial\phi)_\lambda x_m(t), x_n(t) - x_m(t))dt = 0$$

and

$$\lim_{n,m\to\infty} \int_0^T (f_n(t) - f_m(t), x_n(t) - x_m(t))dt = 0.$$

By the maximality of A, K_λ, S and Lemma 1.2 in [5], we have

$$y(\cdot) = Ax(\cdot), \quad g_1(\cdot) = (\partial\phi)_\lambda x(\cdot), \quad g(\cdot) \in Sx(\cdot), \quad y(\cdot) + g_1(\cdot) + g(\cdot) = f(\cdot),$$

i.e.,

$$\begin{cases} x'(t) + (\partial\phi)_\lambda x(t) + g(t) = f(t), \quad a.e.\ t \in (0,T), \\ x(0) = -x(T). \end{cases} \tag{E4.6}$$

Now, take $\lambda_n \to 0^+$ and let $y_n(\cdot)$ be a solution of the Eq. (E4.6), then, from the same reason as before, we know that $y_n'(\cdot)$, $y_n(\cdot)$, $(\partial\phi)_{\lambda_n} y_n(\cdot)$ and $g_n(\cdot) \in Sy_n(\cdot)$ satisfying (E4.6) are bounded in $L^2([0,T];H)$. Assume that $y_n(\cdot) \rightharpoonup y(\cdot)$, $y_n'(\cdot) \rightharpoonup z(\cdot)$, $(\partial\phi)_{\lambda_n} y_n(\cdot) \rightharpoonup h_1(\cdot)$ and $g_n(\cdot) \rightharpoonup h_2(\cdot)$ in $L^2([0,T];H)$. Again, by using the maximal monotonicity of A, K, S and Lemma 1.2 in [5], we have

$$z(\cdot) = Ay(\cdot), \quad h_1(\cdot) \in Ky(\cdot), \quad h_2(\cdot) \in Sy(\cdot), \quad z(\cdot) + h_1(\cdot) + h_2(\cdot) = f(\cdot),$$

i.e., $y(\cdot)$ is a solution of the Eq. (E4.3).

Remark 4.1 When $B = 0$, we have the result of Okochi [12].

Acknowledgements The authors thank the referee for helpful comments and suggestions on this manuscript. Also, we wish to express our thanks to Professor Mihai Turinici for reading the paper and providing very helpful comments.

References

1. T.M. Afsaw, Maximality theorems on the sum of two maximal monotone operators and application to variational inequality problems. Abstr. Appl. Anal. **2016**, 10 (2016)
2. H. Attouch, On the maximality of the sum of two maximal monotone operators. Nonlinear. Anal. **5**, 143–147 (1981)

3. V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces* (Noordhoff, Leyden, 1976)
4. H. Brézis, *Operateurs maxinaux monotones* (North-Holland, Amsterdam, 1973)
5. H. Brézis, M.G. Crandall, A. Pazy, Perturbations of nonlinear maximal monotone sets. Commun. Pure Appl. Math. **23**, 123–144 (1970)
6. F.E. Browder, Nonlinear operators and nonlinear equations of evolution in Banach spaces, in *Proceedings of Symposia in Pure Mathematics*, vol. 18, Part 2 (American Mathematical Society, Providence, 1976)
7. F.E. Browder, Fixed point theory and nonlinear problems. Bull. Amer. Math. Soc. **1**, 1–39 (1983)
8. F.E. Browder, Nonlinear maximal monotone operators in Banach space. Math. Ann. **175**, 89–113 (1968)
9. F.E. Browder, Nonlinear variational inequalities and maximal monotone mappings in Banach spaces. Math. Ann. **185**, 81–90 (1970)
10. Y.Q. Chen, D. O'Regan, On the homotopy property of topological degree for maximal monotone mappings. Appl. Math. Comput. **208**, 373–377 (2009)
11. Y.Q. Chen, Y.J. Cho, P. Kunum, On the maximality of sums of two maximal monotone operators. J. Math. Anal. **7**, 24–30 (2016)
12. H. Okochi, On the existence of anti-periodic solutions to a nonlinear evolution equation associated with odd sub-differential operators. J. Funct. Anal. **91**, 246–258 (1990)
13. D. O'Regan, Y.J. Cho, Y.Q. Chen, *Topological Degree Theory and Applications* (Chapman and Hall/CRC Press, Boca Raton, 2006)
14. D. Pascali, S. Sburlan, *Nonlinear Mappings of Monotone Type* (Noordhoff, Leyden, 1978)
15. R.T. Rockafellar, On the maximal monotonicity of subdifferential mappings. Pac. J. Math. **33**, 209–216 (1970)
16. R.T. Rockafellar, On the maximality of sums of two nonlinear monotone operators. Trans. Am. Math. Soc. **149**, 75–88 (1970)
17. W. Rudin, *Functional Analysis* (MacGraw-Hill, New York, 1973)
18. S. Simons, Sum theorems for monotone operators and convex functions. Trans. Am. Math. Soc. **350**, 2953–2972 (1998)
19. S.L. Trojansky, On locally uniformly convex and differentiable norms in certain nonseparable Banach spaces. Stud. Math. **37**, 173–180 (1971)
20. S. Zhang, Y. Chen, Degree theory for multivalued(S) type mappings and fixed point theorems. Appl. Math. Mech. **11**, 441–454 (1990)

Analysis and Simulation of a Model of Phosphorus Uptake by Plant Roots



Myriam Comte

1 Introduction

Among mineral nutrients, nitrogen (N) and phosphorus (P) are key elements for plant growth. However, in heavily fertilized agriculture, these nutrients can be potentially pollutants for water resources and, in the case of P, its availability might be limited in the future, as a non-renewable resource. As an answer to these problems, low-input agriculture shall be devised by optimizing the use of fertilizers (or soil natural resources) and water by plant roots in the soil. The aim of this paper is to better understand the interactive effects between root system architecture (geometry, growth) and P/water uptake, in relation with the soil spatial variability of resources. This will be undertaken by numerical modeling and mathematical optimization with different levels of details in the description of the root system. This should result in a modeling coupling soil water/nutrient transfer and uptake by a root system architecture, including retro-actions between aerial and subterranean parts of the plant.

In this paper we:

- simulate water movement in soil and water uptake by plant roots, together with the transport and uptake of nutrients,
- explicitly take into account the geometry of a root system,
- study how water and nutrient uptake is affected by the type and shape of root systems.

M. Comte (✉)
Sorbonne Université, Laboratoire Jacques-Louis Lions, LJLL, F75005 Paris, France
e-mail: myriam.comte@sorbonne-universite.fr

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_4

In Sect. 2 we describe the mathematical analysis of a model of phosphorus uptake by plant roots. The evolution of the concentration of P in the soil solution is governed by a convection-diffusion equation with a nonlinear boundary condition at the root surface, which is included as a boundary of the soil domain. A shape optimization problem is formulated in order to find root shapes maximizing P uptake. In Sect. 3, we use unstructured mesh adaptation and parallel computing to develop numerical models of soil water and solute movement with root water and nutrient uptake at the plant scale.

2 Nutrient Uptake: Problem Statement

2.1 Soil Solute Transport

The Richards equation represents commonly the movement of water in unsaturated soils. Combining Darcy's law

$$\mathbf{q} = -K(h)\nabla(h+z),$$

with the continuity equation [19] gives the Richards equation:

$$\frac{\partial\theta(h)}{\partial t} = \nabla.(K(h))\nabla(h+z) + S$$

where

- h is the matric head,
- $\mathbf{q}$ is the Darcy flux,
- $\theta(h)$ is the volumetric water content,
- $K(h)$ is the hydraulic conductivity,
- z is the elevation
- S represents sources or sinks.

The nonlinear relationships $\theta(h)$ and $K(h)$ are given by empirical models whose parameters depend on the soil physical properties.

Several models can be used, such as the Brooks-Corey model or the van Genuchten model.

The evolution of the concentration c of a nutrient in the soil solution is governed by the following mechanisms:

- diffusion of nutrient ions in the soil solution, *(dominant for phosphate)*,
- transport of nutrients by mass flow, *(dominant for nitrate)*
- absorption of nutrient ions in the soil solid phase, *(strong for phosphate, negligible for nitrate)*.

Active nutrient uptake and its ion-selectivity can be regarded as a kinetic process equivalent to that described by Michaelis-Menten type kinetics, used for the description of enzyme-catalyzed reactions [1]. Here, the Michaelis-Menten model relates the uptake rate to the concentration in the soil solution. The uptake rate h is related to the concentration in the soil solution and is given by

$$h(c) = \frac{I_m c}{K_m + c}\ ,\ I_m > 0\ ,\ K_m > 0$$

where I_m is the maximum uptake rate. The Michaelis constant K_m is the concentration at which the uptake rate is half of I_m. The Michaelis-Menten parameters I_m and K_m are ion-specific and depend on several factors such as plant species and plant age.

2.2 Root Nutrient Uptake

Let us consider a shape modeling the root surface. The exterior domain around the root is the studied section of the soil. Let us denote by $\Omega \subset \mathbb{R}^d (d = 2, 3)$ the soil domain, delimited by the root surface and the domain boundaries. Let Γ_1 be the boundary representing the root surface and $\Gamma_2 = \partial\Omega \setminus \Gamma_1$ (Fig. 1).

Let $T > 0$ be given and $I = [0, T]$. The concentration of P in the soil solution c is given by the following convection-diffusion equation:

$$\begin{cases} \partial_t(\theta c + \varphi(c)) = \nabla\big(A\nabla c - \mathbf{q}c\big) - R & \text{in} \quad I \times \Omega, \\ \alpha h(c) = -\big(A\nabla c - \mathbf{q}c\big)\cdot \mathbf{n} & \text{on} \quad I \times \Gamma_1 \\ 0 = \big(A\nabla c - \mathbf{q}c\big)\cdot \mathbf{n} & \text{on} \quad I \times \Gamma_2 = I \times (\partial\Omega \setminus \Gamma_1), \\ c(0, x) = c^0(x) & \text{in} \quad \Omega, \end{cases} \tag{1}$$

where

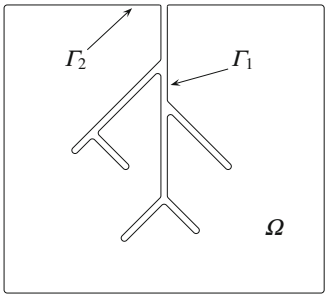


Fig. 1 Configuration of the domain

- **n** is the unit outward normal to the boundary of the domain,
- c^0 is the initial P concentration,
- A is the diffusion coefficient of P in soil,
- θ is the volumetric water content,
- **q** is the Darcy flux,
- R represent additional optional source or sink terms to the system; we will only consider source terms, such as fertilizer application,
- α is a parameter we introduce in order to obtain sufficient regularity of the boundary condition in the case $\Gamma_1 \cap \Gamma_2 \neq 0$: $\alpha \in C^2(\partial\Omega)$ such that for $x \in \partial\Omega$

$$\begin{cases} 0 < \alpha(x) \leq 1 & \text{on } \Gamma_1 \\ \alpha(x) = 0 & \text{on } \Gamma_2, \end{cases}$$

- φ is an adsorption isotherm relating the amount of adsorbed P to the equilibrium concentration of P in solution ; an example is the Freundlich adsorption isotherm, defined by [17]:

$$\varphi(c) = \kappa \times c^b \text{ for } c \in [0, +\infty[\ ,\ \kappa > 0\ ,\ b \in]0, 1[,$$

- h is a model of enzyme kinetics, relating in this case the root uptake rate of P to its concentration at the root surface ; an example has been given previously: the Michaelis-Menten model.

We don't restrict ourselves to considering explicit forms for φ or h: we only use general properties of these functions.

With J.-M. Coron, S. Guerrero, and P.-H. Tournier, we prove in [2] the following result:

Theorem 1 *Let $\beta \in (0, 1)$.*

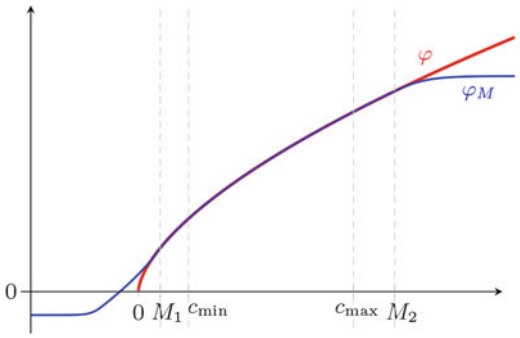
Under suitable regularity conditions on the data and assuming $c^0 > 0$, problem (1) has a unique strictly positive solution in the space $C^{1+\beta/2,2+\beta}(\overline{Q})$.

Proof (Some Steps)

- we derive a priori estimates for positive solutions in $C^{1,2}(\overline{Q})$ using a comparison principle:

$$0 < c_{min} \leq c(t, x) \leq c_{max}, \quad (t, x) \in Q,$$

- we prove uniqueness of solution in $C^{1,2}(\overline{Q})$,
- we define φ_M by extending φ beyond arbitrary bounds M_1, M_2 with $0 < M_1 < M_2$ such that $\varphi_M \in C^3(\mathbb{R})$, $\varphi'_M \geq 0$ and $\varphi'_M(c) = \varphi''_M(c) = \varphi'''_M(c) = 0$ for $|c| > M_2 + \epsilon_M$,



- the new problem (Pb_M) has a unique solution c_M in the space

$$C^{1+\beta/2,2+\beta}(\overline{Q}),$$

- c_M can be bounded independently of M_1, M_2:

$$0 < c_{min} \leq c_M(t,x) \leq c_{max}, \quad (t,x) \in Q,$$

- we choose M_1, M_2 such that $M_1 < c_{min}$ et $c_{max} < M_2$. It follows that the unique solution c_M of problem (Pb_M) is also a solution of the original problem (1). Thus, using the fact that problem (1) has at most one solution in the space $C^{1,2}(\overline{Q})$, we can deduce that problem (1) has a unique solution in the space $C^{1+\beta/2,2+\beta}(\overline{Q})$.

2.3 Shape Optimization

Experiments have shown that the uptake of phosphorus especially for young plants whose growth depends strongly on the uptake capacity of phosphorus, is influenced by the structure and distribution of the root system: phosphorus uptake creates a depletion profile in the vicinity of the surface of the root, and due to the slow diffusion of phosphorus, a root system that explores a greater volume of soil is more effective. In order to study this phenomenon, we consider two caricature forms of root systems in dimension 2 of equal surface (Figs. 2, 3, 4).

We note that for short times, here $t < 6$ h, the differences in local and global structures between the two forms do not influence the P uptake: the absorbed quantities are identical with equal surface. However, for longer times, the shape that explores a larger useful volume is becoming more and more efficient, with a total gain of more than 8% after 6 days.

We will prove mathematically the results of this experiments. For this we use the tools of shape optimization presented in [9, 21] and [11] in order to find root shapes

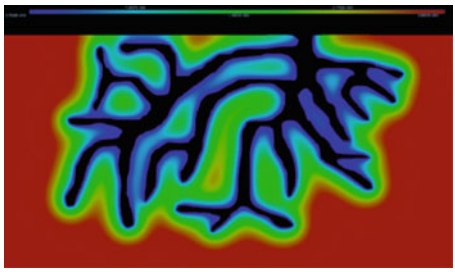


Fig. 2 Dense root, concentration profile after 6 days

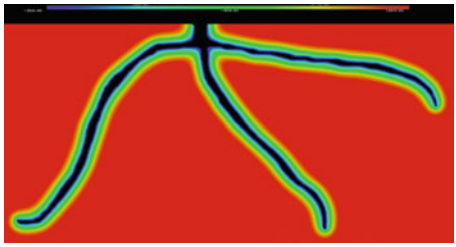


Fig. 3 Root explorer, concentration profile after 6 days

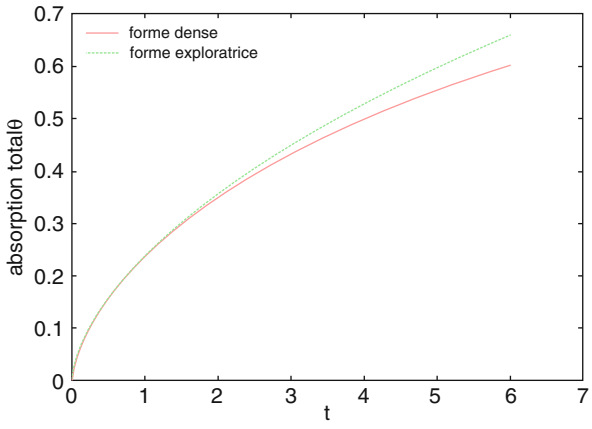


Fig. 4 Root explorer

that increase the amount of absorbed P. More specifically, we want to deform Ω in order to maximize the shape functional

$$J(c) = \int_0^T \int_{\Gamma_1} \alpha h(c),$$

where $|\Omega| =$ is a given constant.

We also assume that A and $\mathbf{q}$ are constants and we take $R = 0$.

Introducing a deformation field $\mathbf{V}$ and considering $\Omega_s = (Id + s\mathbf{V})(\Omega)$, $s \ll 1$, we prove in [2] that the shape derivative of J at Ω in the direction $\mathbf{V}$ is

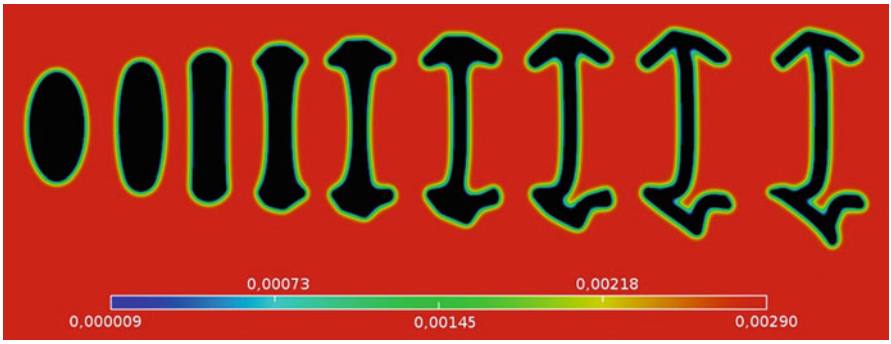


Fig. 5 Snapshots of the domain and P concentration at different steps of the shape optimization process

$$dJ(c, \mathbf{V}) = \int_{\Gamma_1 \cup \Gamma_2} j(\mathbf{V} \cdot \mathbf{n}),$$

where j does not depend on $\mathbf{V}$.

The shape gradient

$$dJ(c, \mathbf{V}) = \int_{\Gamma_1 \cup \Gamma_2} j(\mathbf{V} \cdot \mathbf{n})$$

provides a natural ascent direction $\mathbf{V} = j\mathbf{n}$. This yields

$$dJ(c, \mathbf{V}) = \int_{\Gamma_1 \cup \Gamma_2} j^2 > 0,$$

which ensures that J increases as the domain is iteratively deformed.

Numerical resolution of the state and adjoint equations in two spatial dimensions is carried out using the free finite element software `FreeFEM++`[10], and we obtain in [2] (Fig. 5).

3 Coupling Soil and Root Water Flow

In this section we will describe the model used in [23]. With F. Hecht and P.-H. Tournier, we describe an adaptive finite element model of soil water and nutrient transport with root uptake. The model aims at including local processes in the rhizosphere at the whole root system scale by taking advantage of unstructured mesh adaptation. The model is comparable to [5], where the root system is represented as a tree-like network composed of cylindrical root segments and radial and axial water flows are defined for each segment. Soil and root water flows are coupled via

a sink term in Richards equation and the two problems are solved iteratively until convergence at each time step. The sink terms representing root water and nutrient uptake are built from a characteristic function of the root system, representative of its geometry. This characteristic function is used to drive the mesh adaptation procedure, so that the adapted mesh can accurately resolve the complex geometry of the root system as well as small-scale phenomena in the vicinity of the roots. Since such an approach can be computationally intensive, a parallelization technique based on a scalable two-level Schwarz domain decomposition method is used to solve linear systems arising from the discretization of the soil problems. Some numerical experiments are conducted to illustrate the capabilities of the model.



3.1 Water Flow Within the Root System

We consider that the root system is composed of cylindrical root segments. The geometry of the root system can then be represented as a series of interconnected nodes forming a network of root segments Σ, each segment with its own parameters (radius, conductivity, ...).

Such a representation can be generated by RootBox [14] which implements a root growth model using L-Systems.

For a cylindrical root segment of radius r and length l and according to the Ohm's law analogy, the volumetric radial water flow between the soil-root interface and the root xylem can be written as

$$J_r = L_r s_r (h_s - h_r), \tag{2}$$

where

- L_r is the radial conductivity of the root,
- $s_r = 2\pi r l$ is the root-soil interface area,

- h_s is the soil water potential at the root surface,
- h_r is the water potential in the xylem.

Water moves through the root system due to the transpiration-cohesion-tension mechanism. The longitudinal water flow in the xylem is defined as

$$J_x = -K_r \frac{dh_r}{dl}, \tag{3}$$

where K_r is the xylem conductance.

The radial and longitudinal flow equations

$$J_r = L_r s_r (h_s - h_r), \quad J_x = -K_x \frac{d(h_r + z)}{dl}$$

can be used to define the following water mass balance for a given root node i of parent node p in the tree-like structure:

$$-K_{r,i} \frac{h_{r,p} - h_{r,i}}{l_i} = -\sum_{j \in \text{childs}(i)} \left(K_{r,j} \frac{h_{r,i} - h_{r,j}}{l_j} \right)$$
$$+ L_{r,i} 2\pi r_i l_i \frac{(h_{s,i} - h_{r,i}) + (h_{s,p} - h_{r,p})}{2}.$$

Here $K_{r,i}$, $L_{r,i}$, r_i and l_i refer to the root segment (p, i) while $K_{r,j}$ and l_j relate to the root segment (i, j). $h_{s,i}$ and $h_{r,i}$ are the soil water potential at root node i and the xylem water potential at root node i respectively. We approximate the potentials h_s and h_r for segment (i, p) by averaging their value at the two nodes i and p. Parameters L_r and K_r are given for each segment and can depend on various data such as root type and age (Fig. 6).

The xylem water potential vector $(h_{r,i})_i$ is then solution of a linear system, with the right-hand side containing the soil factors $h_{s,i}$.

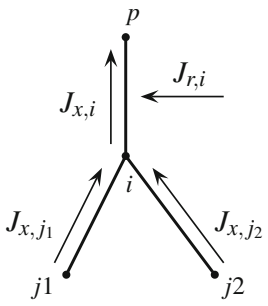


Fig. 6 Water mass balance for root node i

3.2 Coupling Soil and Root Water Flow

Root water uptake is taken into account in the soil model by defining a sink term S in the Richards equation.

The usual approach [6, 12] is to compute the sink term by summing contributions of root segments to water uptake in each soil voxel.

Our approach aims at

- defining an accurate sink term whose shape matches the geometry of the root system,
- resolving small-scale phenomena at the individual root level.

For this we build a characteristic function of the root system f_c representative of its geometry and use it to define the sink term as well as to guide the mesh adaptation procedure. The function f_c is constructed as follows:

- For a point x of the domain Ω the distance d from x to the root is computed:

$$d(x) = \min_{s \in \Sigma} d_s(x), \tag{4}$$

  with Σ the set of root segments in the tree-like network. For each root segment s, the distance $d_s(x)$ from the point x to the segment s is easily computed using distance from line and point routines.
- The distance function d is then used to compute the characteristic function. There is a variety of admissible transformations that we can use, and we choose the following:

$$f_c(x) = f_d\left(d(x)\right) = 1 - \tanh\left(\frac{6d(x)}{\varepsilon}\right). \tag{5}$$

  We can take ε equal to the radius of the root.
- Consider the case of a single cylindrical root segment (i,j):

$$J_r = L_r s_r \frac{(h_{s,i} - h_{r,i}) + (h_{s,j} - h_{r,j})}{2}.$$

  We build the corresponding sink term $S = -\lambda f_c h_l$, where h_l linearly interpolates $h_s - h_r$ along the segment and with $\lambda > 0$ such that

$$\int_{\Omega} S = -J_r.$$

3.3 Numerical Experiment

The algorithm we use in [23] is the following. The coupling between the root and soil models consists in iteratively solving the two problems until convergence. Let $h_s^{t_i}$ be the soil water potential distribution at time t_i, h_s^k and h_r^k the soil and xylem water potentials at inner iteration k and time t_{i+1}.

1. $h_s^0 = h_s^{t_i}$.
2. Solve the linear system arising from the problem defined on the tree-like root network with soil factors h_s^k, obtain h_r^k.
3. Compute the sink term S using h_s^k and h_r^k.
4. Solve the linearized problem corresponding to one Newton step of Richards equation, obtain h_s.
5. $h_s^{k+1} = h_s^k + \alpha_k(h_s - h_s^k)$, where $0 < \alpha_k \leq 1$ is a damping parameter that ensures convergence of the system.
6. If $||h_s - h_s^k|| > \varepsilon$, go to 2 with $k := k + 1$.

We then use an unstructured volume mesh adaptation (see[3]). Unstructured volume mesh adaptation is a flexible and powerful tool in the case of complex geometries. In our case, the tetrahedral mesh is adapted to the variations of the characteristic function f_c so as to resolve the geometry accurately and capture high gradients and small scale phenomena expected near the roots (local conductivity drop). To summarize the mesh adaptation procedure is an iterative algorithm which consists in

- computing the characteristic function f_c on the current mesh.
- defining a nodal-based anisotropic metric tensor field based on the interpolation error using the reconstructed Hessian of f_c (*mshmet*, P. Frey).
- building a unit mesh for which all edges are of unit length in the prescribed metric, using local mesh modifications and anisotropic Delaunay kernel (*mmg3d*, [3, 4]).

We obtain the following result (Fig. 7).

4 Conclusion

In this work, we only considered homogeneous soils, and the model should be extended to heterogeneous soils. Root water uptake and soil heterogeneity have not often been considered simultaneously, although the variability of soil hydraulic properties significantly affects flow processes and root uptake patterns as stressed out in [13]. In addition, a number of soil processes affecting solute movement and availability are simplified or neglected. In most plant nutrition models, chemical interactions between dissolved elements and the soil are reduced to a buffer power or simplified by using the Freundlich adsorption isotherm, as is the case in this work.

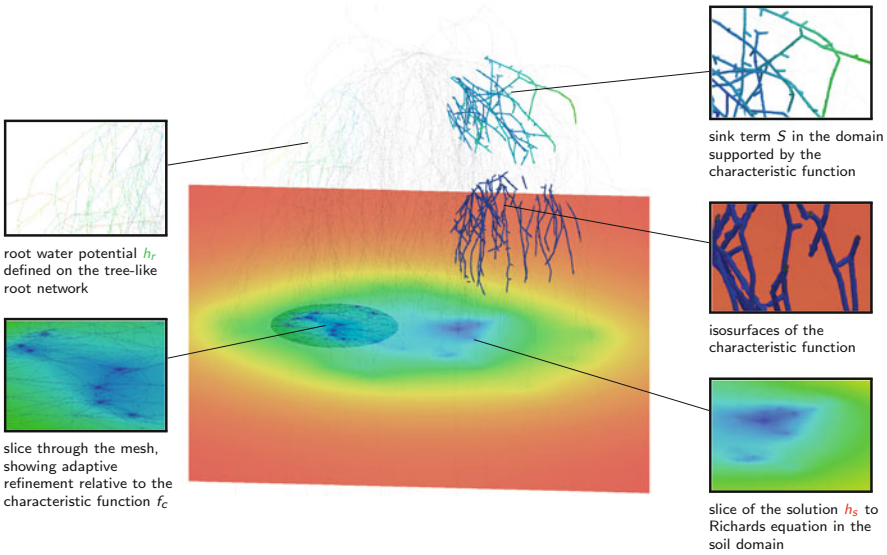


Fig. 7 Overview of the water model

The Freundlich equation is an empirical relation between the solute concentration in the liquid phase and the amount adsorbed to soil particles and fails to account for changes of nutrient availability arising from the range of biogeochemical and biochemical processes that occur in the rhizosphere [8, 20]. The model also makes a number of simplifications regarding root structure and function. Osmotic gradients were neglected, although they can significantly affect root water up- take, especially when considering salt accumulation at the root-soil interface caused by salt transport towards the roots by mass flow through the soil. This salinity buildup in the rhizosphere can lead to large osmotic gradients across the roots, effectively reducing root water uptake. In the simulations, we also used a simplified representation of the hydraulic architecture of the root system by considering constant radial and axial conductivities. In reality, root conductivity is a function of root type and age and varies along root axes. The radial conductivity is affected by the development of apoplastic barriers [22] and by the activity of aquaporins [16], while the axial conductivity depends on the development stage of the xylem. On a similar note, we used a simple Michaelis-Menten model assuming that the nutrient uptake capacity is evenly distributed over the whole root system, while there is experimental evidence that nutrient uptake is not uniform along root axes [7].

Future root uptake models should be coupled with mechanistic surface complexation models using thermodynamic and kinetic relationships in order to include soil processes affecting sorbing surfaces [15, 18]. In particular, biogeochemical interactions and microbial activity in the rhizosphere can have major effects on soil properties and ultimately impact the acquisition of nutrients by plants.

References

1. S.A. Barber, *Soil Nutrient Bioavailability: A Mechanistic Approach* (Wiley, New York, 1984)
2. M. Comte, J.-M. Coron, S. Guerrero, P.-H. Tournier, Analysis of a model of phosphorus uptake by plant roots. J. Evol. Equ. **13**(3), 595–615 (2013)
3. C. Dapogny, C. Dobrzynski, P. Frey, Three-dimensional adaptive domain remeshing, implicit domain meshing, and applications to free and moving boundary problems. J. Comput. Phys. **262**, 358–378 (2014)
4. C. Dobrzynski, MMG3D: User Guide. Rapport Technique RT-0422, INRIA, March 2012
5. C. Doussan, L. Pagès, G. Vercambre, Modelling of the hydraulic architecture of root systems: an integrated approach to water absorption-model description. Ann. Bot. **81**, 213–223 (1998)
6. C. Doussan, Al. Pierret, E. Garrigues, L. Pagès, Water up-take by plant roots: II – modelling of water transfer in the soil root-system with explicit account of flow within the root system – comparison with experiments. Plant Soil **283**(1–2), 99–117 (2006)
7. I.B. Ferguson, D.T. Clarkson, Ion transport and endodermal suberization in the roots of Zea mays. New Phytol. **75**(1), 69–79 (1975)
8. J.S. Geelhoed, W.H. Van Riemsdijk, G.R. Findenegg, Simulation of the effect of citrate exudation from roots on the plant availability of phosphate adsorbed on goethite. Eur. J. Soil Sci. **50**(3), 379–390 (1999)
9. J. Haslinger, R.A.E. Makinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation* (Society for Industrial and Applied Mathematics, Philadelphia, 2003)
10. F. Hecht, A. Le Hyaric, K. Ohtsuka, O. Pironneau, *Freefem++*, finite element software, http://www.freefem.org/ff++/
11. A. Henrot, M. Pierre, *Variation et optimisation de formes: une analyse géométrique* (Springer, Berlin, 2005)
12. M. Javaux, T. Schroder, J. Vanderborght, H. Vereecken, Use of a three-dimensional detailed modeling approach for predicting root water uptake. Vadose Zone J. **7**, 1079–1088 (2008)
13. A. Kuhlmann, I. Neuweiler, S. Zee, R. Helmig, Influence of soil structure and root water uptake strategy on unsaturated flow in heterogeneous media. Water Resour. Res. **48**(2), W02534 (2012). doi/10.1029/2011WR010651
14. D. Leitner, S. Klepsch, G. Bodner, A. Schnepf, A dynamic root system growth model based on L-systems. Plant Soil **332**(1–2) 177–192 (2010)
15. J. Luster, A. Gottlein, B. Nowack, G. Sarret, Sampling, defining, characterising and modeling the rhizosphere – the soil science tool box. Plant Soil **321**(1–2), 457–482 (2009)
16. C. Maurel, L. Verdoucq, D.-T. Luu, V. Santoni, Plant aquaporins: membrane channels with multiple integrated functions. Annu. Rev. Plant Biol. **59**, 595–624 (2008)
17. M.B. McGechan, D.R. Lewis, Sorption of phosphorus by soil, part 1: principles, equations and models. Biosyst. Eng. **82**, 1–24 (2002)
18. B. Nowack, K.U. Mayer, S.E Oswald, W. Van Beinum, C.A.J. Appelo, D. Jacques, P. Seuntjens, F. Gerard, B. Jaillard, A. Schnepf, et al., Verification and intercomparison of reactive transport codes to describe root-uptake. Plant Soil **285**(1–2), 305–321 (2006)
19. L.A. Richards, Capillary conduction of liquids through porous mediums. Physics **1**, 318–333 (1931)
20. M. Silberbush, S.A. Barber, Sensitivity of simulated phosphorus uptake to parameters used by a mechanistic-mathematical model. Plant Soil **74**(1), 93–100 (1983)
21. J. Sokolowski, J.P. Zolesio, *Introduction to Shape Optimization: Shape Sensitivity Analysis*, (Springer, Berlin, 1992)
22. E. Steudle, C.A. Peterson, How does water get through roots? J. Exp. Bot. **49**(322), 775–788 (1998)
23. P.-H. Tournier, F. Hecht, M. Comte, Finite element model of soil water and nutrient transport with root uptake: explicit geometry and unstructured adaptive meshing. Transp. Porous Media **106**(2), 487–504 (2015)

Elliptic Operators with Infinitely Many Variables



Giuseppe Da Prato

1 Introduction

The extension of the classical theory of linear elliptic PDE's in finite dimensions to infinite dimensional Hilbert spaces H is a widely open field. A typical equation is the following

$$\lambda\varphi - \frac{1}{2}\,\mathrm{Tr}\,[C D_x^2\varphi] - \langle Ax - D_x U, D_x\varphi\rangle = f, \tag{1}$$

where $\lambda > 0$, $C : H \to H$ is a positive bounded operator, $A : D(A) \subset H \to H$ is a self-adjoint negative operator, $U : H \to \mathbb{R}$ is a potential, whereas f (given) and φ (unknown) are mappings from H into $\mathbb{R}$. Moreover D_x (or shortly D) will denote the x derivative and $\mathrm{Tr}\,[C D_x^2\varphi]$ the trace of the operator $C D_x^2\varphi$.

Assume for instance that there is an orthonormal basis (e_h) in H and that C and A are diagonal with respect to this basis

$$Ae_h = -\alpha_h e_h, \quad Ce_h = c_h e_h, \quad h \in \mathbb{N},$$

where (α_h) and (c_h) are positive numbers. Setting $x_h = \langle x, e_h\rangle$ and denoting by D_h the partial derivative with respect to x_h, Eq. (1) can be written as

$$\lambda\varphi - \frac{1}{2}\sum_{h=1}^{\infty} c_h D_h^2\varphi + \sum_{h=1}^{\infty}(\alpha_h x_h + D_h U(x)\, D_h\varphi) = f. \tag{2}$$

G. Da Prato (✉)
Scuola Normale Superiore, Pisa, Italy
e-mail: giuseppe.daprato@sns.it

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_5

We shall call the elliptic operator

$$\mathscr{K}\varphi = \frac{1}{2}\,\mathrm{Tr}\,[C\,D_x^2\varphi] + \langle Ax - D_xU(x),\, D_x\varphi\rangle \tag{3}$$

a *Kolmogorov* operator, recalling the seminal paper [10].

Motivations for studying the elliptic equation (1) or the corresponding parabolic equation

$$D_t u = \mathscr{K} u, \quad u(0,\cdot) = u_0, \tag{4}$$

come from partial stochastic differential equations (SPDEs) arising in different domains as: quantum fields theory, statistical mechanics, biology, chemistry and mathematical finance. The SPDE corresponding to Eq. (1) is the following

$$dX = (AX - DU(X))dt + \sqrt{C}\,dW(t), \quad X(0) = x \tag{5}$$

where W is a cylindrical Wiener process on a given probability space $(\Omega, \mathscr{F}, \mathbb{P})$, see e.g. [7]. The connection between (1) and (5) is formally given by

$$u(t,x) = \mathbb{E}[u_0(X(t,x))]. \tag{6}$$

(The expectation $\mathbb{E}$ means integration in Ω with respect to $\mathbb{P}$).

In this note we shall present a review of some recent results on existence, uniqueness and regularity for the elliptic equation (1) in spaces $L^2(H,\mu)$ where μ is an invariant measure for $\mathscr{K}$ (see Sect. 3.1 below). All proofs will be only sketched with the exception of a maximal regularity result (Theorem 25 below) whose proof seems somewhat non satisfactory in the literature, see [7, 15].

To finish, let us note that in infinite dimensions several well established finite dimensional techniques fail, due for instance to the lack of translation invariant and doubling Borel measures and lack of compactness of bounded closed sets. These difficulties prevent to study equations by using approximation by convolution with mollifiers and localization methods based on comparison of integrals of functions over balls with integrals over larger balls. Therefore proofs of maximal regularity results are completely different.

2 Setting of the Problem and Preliminaries

Let us start with the *heat* equation in $\mathbb{R}^n$

$$\begin{cases} D_t u(t,x) = \frac{1}{2}\,\Delta u(t,x), \quad (t,x) \in (0,+\infty)\times\mathbb{R}^n \\ \\ u(0,x) = f(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{7}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and bounded. It is well known that the unique continuous solution of (7) is given by

$$u(t,x) = \frac{1}{(2\pi t)^{n/2}} \int_{\mathbb{R}^n} \varphi(y) e^{-\frac{|x-y|^2}{2t}}\, dy = \int_{\mathbb{R}^n} \varphi(y) N_{x,tI}(dy), \tag{8}$$

where $N_{x,tI}$ (I is the identity operator in $\mathbb{R}^n$) is the Gaussian measure in $\mathbb{R}^n$,

$$N_{x,tI}(dy) = \frac{1}{(2\pi t)^{n/2}}\, e^{-\frac{|y|^2}{2t}}(dy). \tag{9}$$

Let us consider a slightly more general problem

$$\begin{cases} D_t u(t,x) = \frac{1}{2}\operatorname{Tr}[C D_x^2 u(t,x)] + \langle Ax, D_x u(t,x)\rangle, \quad (t,x) \in (0,+\infty)\times\mathbb{R}^n \\ u(0,x) = f(x), \quad x \in \mathbb{R}^n, \end{cases} \tag{10}$$

where A, C are $n \times n$ matrices, with C symmetric positive definite, and $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and bounded. In this case the differential operator $\frac{1}{2}\operatorname{Tr}[C D_x^2 \varphi]$ is elliptic possibly degenerate (when $\det C = 0$.)

The solution to (10) is still explicit. As one can see by an elementary computation, it is given by

$$u(t,x) = \int_{\mathbb{R}^n} \varphi(y) N_{e^{tA}x, Q_t}(dy), \tag{11}$$

where $N_{e^{tA}x, Q_t}$ represents the Gaussian measure in $\mathbb{R}^n$ of mean $e^{tA}x$ and covariance operator

$$Q_t = \int_0^t e^{sA} C e^{sA^*} ds, \quad t \geq 0. \tag{12}$$

In particular, if $\det Q_t > 0$ we have the familiar formula

$$N_{e^{tA}x, Q_t}(dy) = \frac{1}{\sqrt{(2\pi t)^n \det Q_t}}\, e^{-\frac{1}{2}\langle Q_t^{-1} y, y\rangle}\, dy. \tag{13}$$

Now we aim to replace $\mathbb{R}^n$ with an infinite dimensional separable Hilbert space H (norm $|\cdot|$, inner product $\langle\cdot,\cdot\rangle$). To this purpose we have first to recall some results on probability measures in Hilbert spaces.

2.1 Probability Measures on Separable Hilbert Spaces

Let μ be a Borel probability measure in the Hilbert space H that we choose possessing a finite second moment for simplicity $\int_H |x|^2 \, d\mu < \infty$. Then by the Riesz representation theorem there exist $m \in H$ such that

$$\langle m, x \rangle = \int_H \langle x, y \rangle \, \mu(dy), \quad \forall \, x \in H$$

and a symmetric bounded operator $Q \in L(H)$ such that

$$\langle Qx, y \rangle = \int_H \langle x, z \rangle \langle y, z \rangle \, \mu(dz), \quad \forall \, x, y \in H. \tag{14}$$

m is called the *mean* and Q the *covariance* of μ. Note that Q is a positive operator because

$$\langle Qx, x \rangle = \int_H |\langle x, z \rangle|^2 \, \mu(dz) \geq 0, \quad \forall \, x \in H.$$

Moreover its *trace* is finite, in fact, for any orthonormal basis (e_h) in H we have, thanks to the monotone convergence theorem and Parseval's identity,

$$\operatorname{Tr} Q = \sum_{h=1}^{\infty} \langle Qe_h, e_h \rangle = \sum_{h=1}^{\infty} \int_H |\langle z, e_h \rangle|^2 \, \mu(dz)$$

$$= \int_H \sum_{h=1}^{\infty} |\langle z, e_h \rangle|^2 \, \mu(dz) = \int_H |z|^2 \, \mu(dz) < \infty.$$

We say that Q is of *trace class*. We shall denote by $L_2^+(H)$ the set of all positive bounded operators in H which are of trace class. It is not difficult to show that all such operators are compact.

Finally, we say that a Borel probability measure μ on H is Gaussian $N_{m,Q}$ (where $m \in H$ and $Q \in L_2^+(H)$), when its *Fourier transform* $\widehat{\mu}$ is given by

$$\widehat{\mu}(h) = \int_H e^{i\langle x, h \rangle} \, N_{m,Q}(x) = e^{i\langle m, h \rangle} \, e^{-\frac{1}{2}\langle Qh, h \rangle}, \quad \forall \, h \in H. \tag{15}$$

If $m = 0$ we set $N_{0,Q} = N_Q$ for short. The measure N_Q is said to be *non degenerate* if Ker $Q = \{0\}$.

Given a non degenerate Gaussian measure N_Q, an important role is played by the *Cameron–Martin* space $Q^{1/2}(H)$. One shows easily that $Q^{1/2}(H)$ is dense in H and that $N_Q(Q^{1/2}(H)) = 0$.

Also of great importance is the *white noise* function

$$Q^{1/2}(H) \to L^2(H, N_Q), \quad f \mapsto W_f(x) = \langle Q^{-1/2} f, x \rangle, \quad x \in H. \tag{16}$$

In view of (14), we have

$$\begin{aligned} \int_H W_f \, W_g \, dN_Q &= \int_H \langle Q^{-1/2} f, x \rangle \, \langle Q^{-1/2} g, x \rangle \, N_Q(dx) \\ &= \langle f, g \rangle^2, \quad \forall \, f, g \in Q^{1/2}(H). \end{aligned}$$

Therefore W is an isometry between $Q^{1/2}(H)$ and $L^2(H, N_Q)$ that we shall extend to the whole space H by density (recall that $Q^{1/2}(H)$ is dense in H). In some cases we shall write, "par abus de language",[1]

$$W_f(x) = \langle Q^{-1/2} f, x \rangle, \quad x, f \in H.$$

2.2 Back to Parabolic Equations

It is not possible to extend the validity of formula (8) to the infinite dimensional space H because for $t > 0$ the identity operator I is not of trace class. Instead, formula (11) makes sense under the following assumptions.

Hypothesis 1

(i) $A : D(A) \subset H \to H$ *is the infinitesimal generator of a strongly continuous semigroup denoted* e^{tA}, $t \geq 0$.

(ii) The trace of Q_t *is finite for all* $t > 0$, *where*[2]

$$Q_t = \int_0^t e^{sA} Q e^{sA^*} ds, \quad t > 0. \tag{17}$$

If Hypothesis 1 is fulfilled then for any $\varphi \in C_b(H)$[3] Eq. (10) has a unique continuous and bounded solution u which is given by

$$u(t, x) = P_t \varphi(x) = \int_H \varphi(y) N_{e^{tA}x, Q_t}(dy), \tag{18}$$

see [7, Theorem 6.1.2]. P_t is called an *Ornstein–Uhlenbeck* semigroup.

[1] Which is not meaningful in $L^2(H, N_Q)$ because $N_Q(Q^{1/2}(H)) = 0$.

[2] The integral is intended in the strong sense that is for any fixed $x \in H$.

[3] $C_b(H)$ represents the space of all real, bounded and continuous functions in H.

Remark 2 Hypothesis 1 is in a sense necessary for requiring that (10) has solutions for "sufficiently many" initial datas φ, see [7, page 48].

Now we assume Hypothesis 1 and ask the question whether the semigroup P_t is improving as it happens for the heat equation in finite dimensions. We note, looking at identity (18), that this fact will occur provided $N_{e^{tA}x,Q_t}$ is absolutely continuous with respect to N_{Q_t} with a regular density. To this purpose we recall the following basic *Cameron–Martin* formula, see e.g. [7, Theorem 6.2.2].

Theorem 3 *Let $\mu = N_Q$ be a nondegenerate Gaussian measure and let $\nu = N_{a,Q}$, where $a \in H$. Then μ and ν are equivalent*[4] *if and only if $a \in Q^{1/2}(H)$. In this case we have*

$$\frac{d\nu}{d\mu}(y) = e^{-\frac{1}{2}|Q^{-1/2}a|^2 + W_{Q^{-1/2}a}(y)}, \quad y \in H, \tag{19}$$

where W is the white noise function defined in (16). Moreover, if $a \notin Q^{1/2}(H)$ then $N_{a,Q}$ and $N_{0,Q}$ are singular.

Now we can prove the regularity of the solution $u(t,x)$ to (10) under the following *rank condition*, see [16].

Hypothesis 4 *For any $t > 0$ we have $e^{tA}(H) \subset Q^{1/2}(H)$. Consequently, by the closed graph theorem, the operator $\Lambda_t := Q^{-1/2}e^{tA}$ is bounded for all $t > 0$.*

Remark 5

(i) The rank condition is necessary in order that $u(t,\cdot) \in C_b^1(H)$ for any $\varphi \in C_b(H)$, see [7].
(ii) If $Q = I$ Hypothesis 4 is fulfilled under Hypothesis 1, see [7, Example B 3.2].
(iii) If H is finite dimensional the rank condition is equivalent to the hypoellipticity Hörmandar condition for the (possible degenerated) elliptic operator

$$\mathscr{L}\varphi = \frac{1}{2}\,\mathrm{Tr}\,[C D_x^2\varphi(x)] + \langle Ax, D_x\varphi(x)\rangle, \quad x \in \mathbb{R}^n.$$

see [16].

Theorem 6 *Assume Hypotheses 1 and 4. Let $\varphi \in B_b(H)$*[5] *and let $u(t,x)$ be the solution to (10). Then for any $t > 0$, $u(t,\cdot)$ belongs to $C_b^\infty(H)$.*[6]

Proof The proof is straightforward, because under the assumptions of the theorem we can write, thanks to Theorem 3,

[4] That is $\mu << \nu$ and $\nu << \mu$.

[5] $B_b(H)$ is the space of all $\varphi : H \to \mathbb{R}$ bounded and Borel.

[6] $C_b^\infty(H)$ is the space of all real mappings in $\mathbb{R}$ which are infinitely many differentiable with bounded derivatives of all order.

$$u(t,x) = P_t\varphi(x) = \int_H \varphi(y) N_{e^{tA}x,\,Q_t}(dy) = \int_H \varphi(y)\rho(t,x,y) N_{Q_t}(dy), \qquad (20)$$

where

$$\rho(t,x,y) = \frac{dN_{e^{tA}x,\,Q_t}}{dN_{Q_t}}(y) = e^{-\frac{1}{2}|\Lambda_t x|^2 + W_{\Lambda_t x}(y)}, \quad t > 0,\ x, y \in H. \qquad (21)$$

Since $\rho(t,x,y)$ is C^∞ on x the conclusion of the theorem follows by differentiating (17) how many times we wish.

Remark 7 Under the assumptions of Theorem 6 the semigroup P_t is improving. In particular, $P_t\varphi$ is differentiable for all $\varphi \in C_b(H)$ and we have

$$\langle DP_t\varphi(x), h\rangle = \int_H \langle \Lambda_t x,\, Q_t^{-1/2} y\rangle\, \varphi(e^{tA}x + y) N_{Q_t}(dy), \quad x, h \in H. \qquad (22)$$

Example 8

(i) Assume that $C = I$, that A is self-adjoint negative and $(-A)^{-1} \in L_2^+(H)$. Then we have

$$Q_t = \int_0^t e^{sA} C e^{sA^*} ds = -\frac{1}{2} A^{-1}(1 - e^{2tA}).$$

Consequently Q_t is of trace class and thus Hypothesis 1 is fulfilled. Hypothesis 4 is fulfilled as well because the operator Λ_t,

$$\Lambda_t = Q^{-1/2} e^{tA} = \sqrt{2}\, A e^{tA}(1 - e^{2tA})^{-1}$$

is clearly bounded for all $t > 0$.

(ii) We cannot take, as in the finite-dimensional case, $C = I$ and $A = 0$, because in this case we have

$$Q_t = tI,$$

which is not of trace class for $t > 0$.

(iii) Take $C \in L_2^+(H)$ and $A = -\frac{1}{2}$. Then P_t is given by

$$P_t\varphi(x) = \int_H \varphi(y) N_{e^{-t/2}x,\,Q(1-e^{-t})}(dy), \quad \varphi \in C_b(H),\ t > 0 \qquad (23)$$

and Hypothesis 1 is fulfilled whereas Hypothesis 4 is not as easily checked. So, in this case P_t is not improving except on the directions of the Cameron–Martin space $Q^{1/2}(H)$, see e.g. [7]. This semigroup is basic in the construction of the *Malliavin calculus*, see [12].

3 The Ornstein–Uhlenbeck Semigroup P_t and Its Generator

We assume here Hypotheses 1 and 4 and consider the Ornstein–Uhlenbeck semigroup:

$$P_t\varphi(x) = \int_H \varphi(y) N_{e^{tA}x, Q_t}(dy), \quad \varphi \in C_b(H). \tag{24}$$

We notice that P_t has an explicit nice expression on the set of all exponential functions $\mathscr{E}(H)$ defined as the linear span of the real parts of functions

$$\varphi_h(x) = e^{i\langle h,x\rangle}, \quad x \in H, \tag{25}$$

where $h \in H$. Later we shall need to consider also the subset $\mathscr{E}_A(H)$ of $\mathscr{E}(H)$ spanned by all φ_h such that $h \in D(A^*)$ (the adjoint of A). Any function from $C_b(H)$ can be approximated point-wise by elements of $\mathscr{E}_A(H)$ (and consequently by elements of $\mathscr{E}_A(H)$) as the following propositions shows [3, Proposition 1.2].

Proposition 9 *For any $\varphi \in C_b(H)$ there exists a sequence $(\varphi_n) \subset \mathscr{E}_A(H)$ such that*

(i) $\lim_{t\to\infty} \varphi_n(x) = \varphi(x)$ *for all* $x \in H$.
(ii) $\|\varphi_n\|_\infty \le \|\varphi\|_\infty$ *for all* $n \in \mathbb{N}$.

Recalling the expression (15) for the Fourier transform of a Gaussian measure we see that

$$P_t\varphi_h(x) = \int_H e^{i\langle h,x\rangle} N_{e^{tA}x, Q_t}(dy) = e^{i\langle h, e^{tA^*}x\rangle}\, e^{-\frac{1}{2}\langle Q_t h, h\rangle}. \tag{26}$$

Therefore the subspace $\mathscr{E}_A(H)$ (as well as $\mathscr{E}(H)$) is stable for P_t. P_t is not a strongly continuous semigroup on $C_b(H)$ in general (as it follows from (26)). We shall see, however, in the next subsections that P_t is strongly continuous in $L^2(H,\mu)$, where μ is a suitable probability measure, the *invariant measure* of P_t.

3.1 Invariant Measure of P_t

We say that a Borel probability measure in H is *invariant* for P_t whenever

$$\int_H P_t\varphi\, d\mu = \int_H \varphi\, d\mu, \quad \forall\, t \ge 0,\ \varphi \in C_b(H). \tag{27}$$

We are going to show that P_t has an invariant measure assuming, besides Hypothesis 1 the following

Hypothesis 10 *The semigroup e^{tA} from Hypothesis 1 is of negative type, that is there exist $M,\ \omega > 0$ such that*

$$\|e^{tA}\| \le Me^{-\omega t}, \quad \forall\, t \ge 0. \tag{28}$$

Obviously, under this assumption the linear operator

$$Q_\infty x := \int_0^\infty e^{tA} C e^{tA^*} x\, dt, \quad x \in H,$$

is well defined. Moreover, it is of trace class, because

$$Q_\infty = \sum_{k=1}^\infty \int_{k-1}^k e^{sA} C e^{sA^*} ds = \sum_{k=1}^\infty \int_0^1 e^{(s+k-1)A} C e^{(s+k-1)A^*} ds$$

$$= \sum_{k=1}^\infty e^{(k-1)A} Q_1 e^{(k-1)A^*}$$

and consequently,

$$\text{Tr } Q_\infty = \sum_{k=1}^\infty \text{Tr } [e^{(k-1)A} Q_1 e^{(k-1)A^*}] \le \sum_{k=1}^\infty \|e^{(k-1)A}\|^2 \text{ Tr } Q_1$$

$$\le M^2 \sum_{k=1}^\infty e^{-2\omega(k-1)} \text{ Tr } Q_1 < +\infty,$$

as claimed.

Proposition 11 *Under Hypotheses 1 and 10, $\mu = N_{Q_\infty}$ is an invariant measure of P_t.*[7]

Proof It is enough to show that identity (27) holds when $\varphi = \varphi_h,\ h \in D(A^*)$ (φ_h is defined in (25)), because $\mathscr{E}_A(H)$ is dense in $L^2(H, \mu)$ thanks to Proposition 9. In fact if $\varphi = \varphi_h$, (27) becomes

$$\widehat{\mu}(h) = e^{-\frac{1}{2}\langle Q_\infty h, h\rangle} = e^{-\frac{1}{2}\langle Q_t h, h\rangle} e^{-\frac{1}{2}\langle Q_\infty e^{tA^*} h, e^{tA^*} h\rangle}, \quad \forall\, t > 0,\ h \in H. \tag{29}$$

Equivalently,

$$\langle Q_\infty e^{tA^*} h, e^{tA^*} h\rangle + \langle Q_t h, h\rangle = \langle Q_\infty h, h\rangle, \quad h \in H,\ t \ge 0,$$

[7] One can show that μ is unique, see e.g. [3, Theorem 2.34].

which can be written as

$$e^{tA}Q_\infty e^{tA^*} + Q_t = Q_\infty, \quad t \geq 0 \tag{30}$$

(because all operators are symmetric.) So, it remains to check (30). Since for each $h \in H$

$$\begin{aligned} e^{tA}Q_\infty e^{tA^*}h &= \int_0^\infty e^{(t+s)A}Ce^{(t+s)A^*}h\,ds \\ &= \int_t^\infty e^{sA}Ce^{sA^*}h\,ds = Q_\infty h - Q_t h, \quad \forall\, h \in H, \end{aligned}$$

(30) follows.

Proposition 12 *Assume Hypotheses 1 and 10. Then P_t is uniquely extendable to a strongly continuous semigroup of contractions to $L^2(H,\mu)$, denoted still P_t.*

Proof We notice that for all $\varphi \in C_b(H)$ we have by Hölder's inequality

$$(P_t\varphi(x))^2 \leq P_t(\varphi(x)^2), \quad \forall\, x \in H,\ t \geq 0.$$

Therefore

$$\int_H (P_t\varphi)^2\,d\mu \leq \int_H P_t(\varphi^2)\,d\mu = \int_H \varphi^2\,d\mu, \quad t \geq 0,$$

by the invariance of μ. Since $C_b(H)$ is dense in $L^2(H,\mu)$, P_t is extendable to $L^2(H,\mu)$ and

$$\|P_t\varphi\|_{L^2(H,\mu)} \leq \|\varphi\|_{L^2(H,\mu)}, \quad \forall\, \varphi \in L^2(H,\mu).$$

It remains to show the strong continuity of P_t,

$$\lim_{t\to 0} P_t\varphi = \varphi \quad \text{in } L^2(H,\mu). \tag{31}$$

It is enough to check (31) for $\varphi = \varphi_h$, $h \in D(A^*)$. In this case we have

$$P_t\varphi(x) = e^{-\frac{1}{2}\langle Q_t h,h\rangle} e^{i\langle e^{tA}x,h\rangle} \to \varphi_h(x),$$

as $h \to 0$, as required.

3.2 The Infinitesimal Generator of P_t

We denote by $\mathscr{L}$ the infinitesimal generator of P_t on $L^2(H,\mu)$ defined as

$$\mathscr{L}\varphi = \lim_{t\to 0} \frac{1}{t}\,(P_t\varphi - \varphi) \ \text{ in } L^2(H,\mu), \quad \forall\,\varphi \in D(\mathscr{L}), \tag{32}$$

where $D(\mathscr{L})$ is the space of all $\varphi \in L^2(H,\mu)$ for which the limit above exists.

Proposition 13

(i) $\mathscr{E}_A(H) \subset D(\mathscr{L})$ *and*[8]

$$\mathscr{L}\varphi = \frac{1}{2}\,\mathrm{Tr}\,[C D_x^2\varphi] + \langle x, A^* D_x\varphi\rangle, \quad \forall\,\varphi \in D(\mathscr{L}). \tag{33}$$

(ii) $\mathscr{E}_A(H)$ *is a core for* $\mathscr{L}$*, that is it is dense in* $\mathscr{L}$ *(endowed with the graph norm).*

Proof (i) It is enough to show (33) when $\varphi(x) = \varphi_h(x) = e^{i\langle h,x\rangle}$. In this case by (26) we have

$$\frac{1}{t}\,(P_t\varphi_h(x) - \varphi_h(x)) = \frac{1}{t}\,(e^{i\langle h, e^{tA^*}x\rangle}\, e^{-\frac{1}{2}\langle Q_t h,h\rangle} - e^{i\langle h,x\rangle}) \tag{34}$$

and (33) follows letting $t \to 0$. Finally, (ii) follows from a classical result of semigroup theory, see [9], since $\mathscr{E}_A(H) \subset D(\mathscr{L})$ and it is stable for P_t.

4 Sobolev Spaces

We are given a non degenerate Gaussian measure $\mu = N_Q$.

4.1 The Sobolev Space $W^{1,2}(H,\mu)$

We start with a classical integration by parts formula, see [12].

Lemma 14 *For all* $\varphi, \psi \in \mathscr{E}(H)$ *and all* $z \in Q^{1/2}(H)$ *we have*

$$\int_H \langle D\varphi, z\rangle\,\psi\,d\mu = -\int_H \langle D\psi, z\rangle\,\varphi\,d\mu + \int_H \varphi\,\psi\,\langle Q^{-1/2}z, x\rangle\,\mu(dx). \tag{35}$$

[8] Note that the function φ_h belongs to $D(\mathscr{L})$ if and only if $h \in D(A^*)$.

Proof It suffices to check (35) when $\varphi(x) = e^{i\langle h,x\rangle}$, $\psi(x) = e^{i\langle k,x\rangle}$ and $h, k \in H$. This is a straightforward computation. In fact, taking into account (15), we have

$$\int_H \langle D\varphi, z\rangle\, \psi\, d\mu = i\langle h, z\rangle \int_H e^{i\langle h+k,x\rangle}\, \mu(dx) = i\langle h, z\rangle e^{-\frac{1}{2}\langle Q(h+k),h+k\rangle}, \tag{36}$$

$$\int_H \langle D\psi, z\rangle\, \varphi\, d\mu = i\langle k, z\rangle \int_H e^{i\langle h+k,x\rangle}\, \mu(dx) = i\langle k, z\rangle e^{-\frac{1}{2}\langle Q(h+k),h+k\rangle}, \tag{37}$$

and

$$\begin{aligned}
&\int_H \varphi\, \psi\, \langle Q^{-1/2}z, x\rangle\, \mu(dx) = \int_H \langle Q^{-1/2}z, x\rangle\, e^{i\langle h+k,x\rangle}\mu(dx)\\
&= -i\frac{d}{dt}\int_H e^{i\langle h+k+tQ^{-1/2}z,x\rangle}\mu(dx)\Big|_{t=0} = -i\frac{d}{dt}\int_H e^{i\langle h+k+tQ^{-1/2}z,x\rangle}\mu(dx)\Big|_{t=0}\\
&= -i\frac{d}{dt}\, e^{-\frac{1}{2}\langle Q(h+k+tQ^{-1/2}z),h+k+tQ^{-1/2}z\rangle}\Big|_{t=0} = i\langle h+k, z\rangle e^{-\frac{1}{2}\langle Q(h+k),h+k\rangle}.
\end{aligned} \tag{38}$$

The conclusion follows comparing (36), (37) with (38).

Let us show an important consequence of identity (35).

Proposition 15 *The gradient operator*

$$D : \mathscr{E}(H) \to L^2(H, \mu; H), \quad \varphi \to D\varphi,$$

is closable in $L^2(H, \mu)$. *The domain of its closure will be denoted by* $W^{1,2}(H, \mu)$.

Proof We have to show that if $(\varphi_n) \subset \mathscr{E}(H)$ is such that

$$\varphi_n \to 0 \text{ in } L^2(H, \mu), \quad D\varphi_n \to F \text{ in } L^2(H, \mu; H),$$

it results $F = 0$. To prove this fact let just apply (35) with φ_n replacing φ and then let $n \to \infty$. We obtain

$$\int_H \langle F, z\rangle \psi\, d\mu = 0,$$

which yields $F = 0$ by the arbitrariness of ψ and z (recall that $Q^{1/2}(H)$ is dense in H).

Remark 16 In a similar way one can prove that the operator $Q^{1/2}D$ is closable on $L^2(H, \mu)$. Its closure is called the *Malliavin derivative* operator.

4.2 The Sobolev Space $W^{2,2}(H, \mu)$

We need a generalization of identity (35) to second derivatives. Its proof is similar and straightforward.

Proposition 17 *For all $\varphi,\ \psi \in \mathscr{E}(H)$, $u, v \in Q^{1/2}(H)$ we have (D^2 denotes second derivative)*

$$\int_H \langle D^2\varphi\, u, v\rangle\, \psi\, d\mu = \int_H \langle D^2\psi\, u, v\rangle\, \varphi\, d\mu - \int_H W_{Q^{-1/2}v}\, \langle D\psi, u\rangle\, \varphi\, d\mu$$
$$- \int_H W_{Q^{-1/2}u}\, \langle D\psi, v\rangle\, \varphi\, d\mu + \int_H W_{Q^{-1/2}u}\, W_{Q^{-1/2}v}\, \varphi\, \psi\, d\mu. \tag{39}$$

Consider now the application γ,

$$\mathscr{E}(H) \to L^2(H, \mu; H) \oplus L^2(H, \mu; L_2(H)), \quad \varphi \mapsto \gamma(\varphi) := \{D\varphi, D^2\varphi\},$$

where $L_2(H)$ is the space of all *Hilbert–Schmidt* operators in H endowed with the norm

$$\|T\|^2_{L_2(H)} = \mathrm{Tr}\,[TT^*] = \sum_{h,k=1}^{\infty} T^2_{h,k}, \quad T_{h,k} = \langle Te_k, e_h\rangle,$$

and (e_h) is any orthonormal basis on H.

Proposition 18 *γ is closable.*

Proof Let $(\varphi_n) \subset \mathscr{E}(H)$, $F \in L^2(H, \mu; H)$ and $T \in L^2(H, \mu; L_2(H))$ be such that

$$\lim_{n\to\infty} \varphi_n = 0 \text{ in } L^2(H, \mu), \quad \lim_{n\to\infty} D\varphi_n = F \text{ in } L^2(H, \mu; H),$$

$$\lim_{n\to\infty} D^2\varphi_n = T \text{ in } L^2(H, \mu; L_2(H)).$$

We have to show that $F = 0$ and $T = 0$. It is clear that $F = 0$. To show that $T = 0$ notice that by (39) we have for any $u, v \in Q^{1/2}(H)$, $\psi \in \mathscr{E}(H)$, $n \in \mathbb{N}$,

$$\int_H \langle D^2\varphi_n\, u, v\rangle\, \psi\, d\mu = \int_H \langle D^2\psi\, u, v\rangle\, \varphi_n\, d\mu - \int_H W_{Q^{-1/2}v}\, \langle D\psi, u\rangle\, \varphi_n\, d\mu$$
$$- \int_H W_{Q^{-1/2}u}\, \langle D\psi, v\rangle\, \varphi_n\, d\mu + \int_H W_{Q^{-1/2}u}\, W_{Q^{-1/2}v}\, \varphi_n\, \psi\, d\mu.$$

Letting $n \to \infty$ yields

$$\int_H \langle T\, u, v\rangle\, \psi\, d\mu = 0.$$

By the arbitrariness of u, v and ψ we deduce that $T = 0$.

We shall denote by $W^{2,2}(H, \mu)$ the closure of the domain of γ. Let (e_h) be an orthonormal basis in H and let D_h and $D_h D_k$ denote the partial derivatives in the direction e_h and the second partial derivative in the directions e_h, e_k respectively. Then the norms of $W^{2,2}(H, \mu)$ and $W^{2,2}(H, \mu)$ can be written as

$$\|\varphi\|^2_{W^{1,2}(H,\mu)} = \int_H \varphi^2\, d\mu + \sum_{h=1}^{\infty} \int_H |D_h\varphi|^2\, d\mu \tag{40}$$

and

$$\|\varphi\|^2_{W^{2,2}(H,\mu)} = \|\varphi\|^2_{W^{1,2}(H,\mu)} + \sum_{h,k=1}^{\infty} \int_H |D_h D_k\varphi|^2\, d\mu, \tag{41}$$

respectively.

5 The Elliptic Equation

We are here concerned with the elliptic equation

$$\lambda\varphi - \mathscr{L}\varphi = f, \tag{42}$$

where $\mathscr{L}$ is the infinitesimal generator of P_t in $L^2(H, \mu)$, λ is a positive number and $f \in L^2(H, \mu)$. By the Hille–Yosida Theorem, see e.g. [14], there is a unique $\varphi \in D(\mathscr{L})$ solving (42) given by

$$\varphi = R(\lambda, \mathscr{L})f, \tag{43}$$

where $R(\lambda, \mathscr{L})$ is the resolvent of $\mathscr{L}$ at λ.

This section is devoted to prove regularity properties of φ (that a priory only belongs to $L^2(H, \mu)$). First, we show that any solution φ to (42) is the limit of regular solutions.

Lemma 19 *Let $\varphi = R(\lambda, \mathscr{L})f$ be a solution of (42). Then there is a sequence $(\varphi_n) \subset \mathscr{E}_A(H)$ such that*

$$\varphi_n \to \varphi, \quad \mathscr{L}\varphi_n \to \mathscr{L}\varphi \quad \text{in } L^2(H, \mu). \tag{44}$$

Proof It is enough to recall that $\mathscr{E}_A(H)$ is a core for $\mathscr{L}$.

We are first going to show that $\varphi \in W^{1,2}(H,\mu)$, then we shall prove that $\varphi \in W^{2,2}(H,\mu)$. Finally we shall give an explicit characterization of the domain $D(\mathscr{L})$ of $\mathscr{L}$ following [15] and [7]. As will shall see, contrary to what happens when H is finite dimensional, $D(\mathscr{L})$ is strictly narrow than $W^{2,2}(H,\mu)$. We shall perform this program assuming, besides Hypotheses 1 and 4 the following one (recall Example 8)

Hypothesis 20 *A is self-adjoint negative, A^{-1} is of trace class and $C = I$.*

Note that Hypothesis 20 implies Hypothesis 4, see Remark 5(ii).

5.1 Regularity of the First Derivative of φ

We first prove an integration by parts formula (a kind of Green formula).

Lemma 21 *Assume, Hypotheses 1, 10 and 20. Then for all $\varphi, \psi \in \mathscr{E}_A(H)$ we have*

$$\int_H \mathscr{L}\varphi\, \psi\, d\mu = -\frac{1}{2}\int_H \langle D\varphi, D\psi\rangle\, d\mu \tag{45}$$

Proof It is enough to check (45) when $\varphi(x) = e^{i\langle h,x\rangle}$, $\psi(x) = e^{i\langle k,x\rangle}$ with $h, k \in D(A)$. This follows by a direct computation, as we did in the proof of Lemma 14.

Proposition 22 *Assume Hypotheses 1, 10 and 20 and let $\lambda > 0$ and $f \in L^2(H,\mu)$. Then the solution φ to Eq. (42) belongs to $W^{1,2}(H,\mu)$.*

Proof By Lemma 19 there exists a sequence $(\varphi_n) \subset \mathscr{E}_A(H)$ such that

$$\varphi_n \to \varphi, \quad \mathscr{L}\varphi_n \to \mathscr{L}\varphi \quad \text{in } L^2(H,\mu).$$

On the other hand, by (45) we have for each $n, m \in \mathbb{N}$,

$$\int_H |D\varphi_n - D\varphi_m|^2\, d\mu = -2\int_H \mathscr{L}(\varphi_n - \varphi_m)\,(\varphi_n - \varphi_m)\, d\mu, \tag{46}$$

so, $(D\varphi_n)$ is Cauchy on $L^2(H,\mu;H)$ and the conclusion follows by the very definition of Sobolev space $W^{1,2}(H,\mu)$.

5.2 Characterization of the Domain of $\mathscr{L}$

We still consider the elliptic equation (42) assuming, Hypotheses 1, 10 and 20. In this case we have $Q = -\frac{1}{2}A^{-1}$ by (17) and, since Q is compact and Ker $Q = \{0\}$, there exist an orthonormal basis (e_k) in H and a sequence of positive

numbers (λ_k) such that $Qe_k = \lambda_k e_k$, $k \in \mathbb{N}$. We shall set $\alpha_k = -\frac{1}{2}\lambda_k^{-1}$ so that $Ae_k = -\alpha_k e_k$, $k \in \mathbb{N}$, and $x_k = \langle x, e_k\rangle$, $x \in H$, $k \in \mathbb{N}$. Moreover, we shall denote by D_k the derivative in the direction e_k.

Now let $\lambda > 0$, $\varphi \in \mathscr{E}_A(H)$ and set $f = \lambda\varphi - \mathscr{L}\varphi$, which we write as

$$\lambda\varphi - \sum_{k=1}^{\infty} D_k^2\varphi + \sum_{k=1}^{\infty} \alpha_k x_k D_k\varphi = f. \tag{47}$$

Differentiating with respect to x_h, yields

$$\lambda D_h\varphi - \frac{1}{2}\sum_{k=1}^{\infty} D_k^2 D_h\varphi + \sum_{k=1}^{\infty} \alpha_k x_k D_k D_h\varphi + \alpha_h D_h\varphi = D_h f, \tag{48}$$

which can be written as

$$\lambda D_h\varphi - \mathscr{L}(D_h\varphi) + \alpha_h D_h\varphi = D_h f. \tag{49}$$

Multiplying both sides by $D_h\varphi$, integrating with respect to μ over H and taking into account (45), yields

$$\lambda\int_H |D_h\varphi|^2 d\mu + \frac{1}{2}\int_H |DD_h\varphi|^2 d\mu + \alpha_h\int_H |D_h\varphi|^2 d\mu = \int_H D_h f\, D_h\varphi\, d\mu. \tag{50}$$

Summing up on h yields

$$\lambda\int_H |D\varphi|^2 d\mu + \frac{1}{2}\int_H \mathrm{Tr}\,[(D^2\varphi)^2] d\mu + \int_H |(-A)^{1/2} D\varphi|^2 d\mu = \int_H \langle Df, D\varphi\rangle\, d\mu. \tag{51}$$

We want now to find an identity involving f but not Df. So, we use again (45) and write

$$\int_H \langle Df, D\varphi\rangle\, d\mu = -2\int_H \mathscr{L}\varphi\, f\, d\mu.$$

Then by (51) we obtain the following estimate

$$\lambda\int_H |D\varphi|^2 d\mu + \frac{1}{2}\int_H \mathrm{Tr}\,[(D^2\varphi)^2] d\mu + \int_H |(-A)^{1/2} D\varphi|^2 d\mu = -2\int_H f\,\mathscr{L}\varphi\, d\mu$$

$$= 2\int_H f^2\, d\mu - \lambda\int_H \varphi\, f\, d\mu \le 2\int_H f^2\, d\mu + \frac{1}{2}\int_H f^2\, d\mu + \frac{1}{2}\lambda^2\int_H \varphi^2\, d\mu$$

$$\le 3\int_H f^2\, d\mu. \tag{52}$$

We are now in position to prove further regularity results for the solution φ to (42).

Proposition 23 *Assume Hypotheses 1, 10 and 20. Let $\lambda > 0$, $f \in L^2(H,\mu)$ and let φ be the solution to (42). Then $\varphi \in W^{2,2}(H,\mu)$, $(-A)^{1/2}D\varphi \in L^2(H,\mu;H)$ and the following estimate holds*

$$\lambda \int_H |D\varphi|^2\, d\mu + \frac{1}{2}\int_H \mathrm{Tr}\,[(D^2\varphi)^2]\, d\mu + \int_H |(-A)^{1/2}D\varphi|^2\, d\mu \le 3\int_H f^2 d\mu. \tag{53}$$

Proof By Lemma 19 there exists a sequence $(\varphi_n) \subset \mathscr{E}_A(H)$ such that

$$\varphi_n \to \varphi, \quad \mathscr{L}\varphi_n \to \mathscr{L}\varphi \quad \text{in } L^2(H,\mu).$$

Then $f_n := \lambda\varphi_n - \mathscr{L}\varphi_n \to f$. Now let $m,\ n \in \mathbb{N}$. Then replacing $\varphi,\ f$ with $(\varphi_m - \varphi_n),\ (f_m - f_n)$ respectively in (52) we see that the sequence (φ_n) is Cauchy in $W^{2,2}(H,\mu)$ and that the sequence $((-A)^{1/2}D\varphi_n)$ is Cauchy in $L^2(H,\mu;H)$. The conclusion follows.

5.2.1 The Special Case When H Is Finite Dimensional

We assume here that $\dim H = d < \infty$, A is a symmetric and strictly negative matrix and that $C = I$.

Proposition 24 *The domain of $\mathscr{L}$ is precisely $W^{2,2}(H,\mu)$ and we have*

$$\mathscr{L}\varphi = \frac{1}{2}\,\Delta\varphi + \langle Ax, D\varphi\rangle, \quad \forall \varphi \in W^{2,2}(H,\mu). \tag{54}$$

Proof We note first that if $\varphi \in W^{2,2}(H,\mu)$ then $|x||D\varphi| \in L^2(H,\mu)$ as easily seen, so that

$$\|\langle Ax, D\varphi\rangle\|_{L^2(H,\mu)} \le \|A\|\,\||x|\,|D\varphi|\|_{L^2(H,\mu)} \le K\|\varphi\|_{W^{2,2}(H,\mu)}. \tag{55}$$

We already know by Proposition 23 that $\varphi \in W^{2,2}(H,\mu)$ so that the following inclusion holds

$$D(\mathscr{L}) \subset W^{2,2}(H,\mu). \tag{56}$$

Let us show now the converse inclusion. Let $\varphi \in W^{2,2}(H,\mu)$. Then there exists $(\varphi_n) \subset \mathscr{E}_A(H)$ such that

$$\varphi_n \to \varphi, \quad \Delta\varphi_n \to \Delta\varphi \quad \text{in } L^2(H,\mu),$$

and, by (55) we have

$$\langle Ax, D\varphi_n\rangle \to \langle Ax, D\varphi\rangle \quad \text{in } L^2(H,\mu).$$

Then

$$\mathscr{L}\varphi_n = \frac{1}{2}\,\Delta\varphi_n + \langle Ax, D\varphi_n\rangle \to \mathscr{L}\varphi \quad \text{in } L^2(H,\mu), \tag{57}$$

and letting $n \to \infty$ we obtain $\mathscr{L}\varphi = \frac{1}{2}\,\Delta\varphi + \langle Ax, D\varphi\rangle$, as claimed.

5.2.2 The Infinite Dimensional Case

By Proposition 23 the following inclusion holds

$$D(\mathscr{L}) \subset \left\{\varphi \in W^{2,2}(H,\mu) :\ (-A)^{1/2}D\varphi \in L^2(H,\mu;H)\right\} \tag{58}$$

From this result we cannot conclude (as we did in the finite dimensional case) that

$$\mathscr{L}\varphi = \frac{1}{2}\mathrm{Tr}\,[D^2\varphi] + \langle x, AD\varphi\rangle, \quad \forall\,\varphi \in D(\mathscr{L}).$$

The reason is that by (58) it follows that $D^2\varphi$ is Hilbert–Schmidt which does not imply that it is of trace class. However, we can prove

Theorem 25 *Assume Hypotheses 1, 10 and 20. Then we have*

$$D(\mathscr{L}) = \left\{\varphi \in W^{2,2}(H,\mu) :\ (-A)^{1/2}D\varphi \in L^2(H,\mu;H)\right\} \tag{59}$$

We shall prove this result using the spectral decomposition of P_t with respect to the orthonormal basis $(H_\gamma)_{\gamma\in\Gamma}$ of Hermite polynomials in H whose definition we recall in the Appendix below. Let us first to prove a lemma.

Lemma 26 *For any $\gamma = (\gamma_n) \in \Gamma$[9] we have (see (95) for the definition of $\langle\gamma, \lambda^{-1}\rangle$)*

$$P_t H_\gamma(x) = e^{-\frac{t}{2}\langle\gamma,\lambda^{-1}\rangle} H_\gamma(x), \quad x \in H. \tag{60}$$

Therefore P_t is diagonal with respect to the basis $(H_\gamma)_{\gamma\in\Gamma}$.

Proof Write

$$P_t H_\gamma(x) = \prod_{n=1}^{\infty} p_t^{(n)} H_{\gamma_n}(\lambda_n^{-1/2}x_n), \tag{61}$$

[9]For the definition of Γ and H_γ, $\gamma \in \Gamma$, see the Appendix.

where

$$p_t^{(n)} f(\xi) = \int_{\mathbb{R}} f(y) N_{e^{-t/(2\lambda_n)}\xi, \lambda_n(1-e^{-t/\lambda_n})}(dy), \quad x \in \mathbb{R}.$$

So,

$$p_t^{(n)} H_{\gamma_n}(\lambda_n^{-1/2} x_n) = \int_{\mathbb{R}} H_{\gamma_n}(y) N_{e^{-t/(2\lambda_n)}\lambda_n^{-1/2} x_n, \lambda_n(1-e^{-t/\lambda_n})}(dy).$$

Then, to conclude the proof, it is enough to show that for any $n \in \mathbb{N}$ we have

$$p_t^{(n)} H_{\gamma_n}(\lambda_n^{-1/2} x_n) = e^{-\frac{t}{2}\gamma_n \lambda_n^{-1}} H_{\gamma_n}(\lambda_n^{-1/2} x_n). \tag{62}$$

Let us show (62), forgetting the index n for short and so writing

$$p_t H_\gamma(\lambda^{-1/2} x) = e^{-\frac{t}{2}\gamma\lambda^{-1}} H_\gamma(\lambda^{-1/2} x), \tag{63}$$

where

$$p_t H_\gamma(\lambda^{-1/2} x) = \int_{\mathbb{R}} H_\gamma(y) N_{e^{-t/(2\lambda)}\lambda^{-1/2} x, \lambda(1-e^{-t/\lambda})}(dy).$$

We first compute p_t on the generatrix function $F(s, \lambda^{-1/2} x)$ (see (91)). We have

$$p_t F(s, \lambda^{-1/2} x) = e^{-\frac{1}{2}s^2} r_t(e^{s\lambda^{-1/2}} x) = e^{-\frac{1}{2}s^2} \int_{\mathbb{R}} e^{s\lambda^{-1/2} y} N_{e^{-t/(2\lambda)} x, \lambda(1-e^{-t/\lambda})}(dy).$$

Therefore

$$\begin{aligned} p_t F(s, \lambda^{-1/2} x) &= e^{-\frac{1}{2}s^2} e^{s e^{-t/(2\lambda)} \lambda^{-1/2} x + \frac{1}{2}s^2(1-e^{-t/\lambda})} \\ &= e^{-\frac{1}{2}e^{-t/\lambda}s^2} e^{s e^{-t/(2\lambda)} \lambda^{-1/2} x} = F(s e^{-t/(2\lambda)}, \lambda^{-1/2} x). \end{aligned} \tag{64}$$

Now by the identity

$$p_t F(s, \lambda^{-1/2} x) = F(s e^{-t/(2\lambda)}, \lambda^{-1/2} x),$$

we deduce

$$\sum_{n=0}^{\infty} \frac{s^n}{\sqrt{n!}} p_t H_n(\lambda^{-1/2} x) = \sum_{n=0}^{\infty} \frac{s^n e^{-(tn)/(2\lambda)}}{\sqrt{n!}} H_n(\lambda^{-1/2} x)$$

which yields

$$p_t H_n(\lambda^{-1/2}x) = e^{-(tn)/(2\lambda)} H_n(\lambda^{-1/2}x).$$

By Lemma 26 we obtain the following spectral decomposition of P_t. For any $\varphi \in L^2(H,\mu)$, $\varphi = \sum_{\gamma\in\Gamma} \varphi_\gamma H_\gamma$, we have (see (95) for the definition of $\langle \gamma, \lambda^{-1}\rangle$)

$$P_t\varphi = \sum_{\gamma\in\Gamma} e^{-\frac{t}{2}\langle \gamma,\lambda^{-1}\rangle}\varphi_\gamma H_\gamma, \quad t \geq 0. \tag{65}$$

We are now ready to prove Theorem 25.

Proof We shall show that if

$$\varphi \in W^{2,2}(H,\mu), \quad (-A)^{1/2}D\varphi \in L^2(H,\mu;H), \tag{66}$$

then we have $\varphi \in D(\mathscr{L})$.

Claim If (66) holds we have

$$\sum_{\gamma\in\Gamma} \varphi_\gamma^2 \langle \gamma, \lambda^{-1}\rangle^2 < \infty. \tag{67}$$

Assume for the moment that the claim is proved. Then the series

$$\psi := \sum_{\gamma\in\Gamma} \varphi_\gamma \langle \gamma, \lambda^{-1}\rangle\, H_\gamma,$$

is convergent in $L^2(H,\mu)$. Now

$$\tfrac{1}{t}\,(P_t\varphi - \varphi) = \sum_{\gamma\in\Gamma} \tfrac{1}{t}\,\varphi_\gamma (e^{-\frac{t}{2}\langle \gamma,\lambda^{-1}\rangle} - 1).$$

Consequently

$$\lim_{t\to 0} \tfrac{1}{t}\,(P_t\varphi - \varphi) = \sum_{\gamma\in\Gamma} \varphi_\gamma \langle \gamma, \lambda^{-1}\rangle\, H_\gamma = \psi, \quad \text{in } L^2(H,\mu)$$

and so $\varphi \in D(\mathscr{L})$ and $\mathscr{L}\varphi = \psi$. Therefore the proposition is proved under the validity of the claim.

Proof of the Claim We already know that $\varphi \in W^{2,2}(H,\mu)$ and $(-A)^{1/2}D\varphi \in L^2(H,\mu;H)$, which imply

$$\int_H \mathrm{Tr}[(D^2\varphi)^2]d\mu = \sum_{\gamma\in\Gamma} |\varphi_\gamma|^2\langle\gamma, \lambda^{-1}\rangle^2 - \sum_{\gamma\in\Gamma} |\varphi_\gamma|^2\langle\gamma, \lambda^{-2}\rangle < \infty \tag{68}$$

and

$$\begin{aligned}\int_H |(-A)^{1/2}D\varphi|^2 d\mu &= \sum_{k=1}^{\infty}\int_H \alpha_k |D_k\varphi|^2 d\mu \\ &= \sum_{k=1}^{\infty}\sum_{\gamma\in\Gamma} |\varphi_\gamma|^2\alpha_k\frac{\gamma_k}{\lambda_k} = \frac{1}{2}\sum_{\gamma\in\Gamma} |\varphi_\gamma|^2\langle\gamma, \lambda^{-2}\rangle < \infty.\end{aligned} \tag{69}$$

So (67) is fulfilled, because, taking into account (96) and (97), we have

$$\frac{1}{2}\int_H \mathrm{Tr}[(D^2\varphi)^2]d\mu + \int_H |(-A)^{1/2}D\varphi|^2 d\mu = \frac{1}{2}\sum_{\gamma\in\Gamma} |\varphi_\gamma|^2\langle\gamma, \lambda^{-1}\rangle^2 < \infty. \tag{70}$$

The proof is complete.

Remark 27 Let $\mu = N_Q$ be a Gaussian measure. Denote by L_M be the Malliavin operator

$$\mathscr{L}_M\varphi = \frac{1}{2}\,\mathrm{Tr}\,[QD_x^2\varphi] - \frac{1}{2}\langle x, D_x\varphi\rangle, \quad \forall\,\varphi \in \mathscr{E}(H), \tag{71}$$

discussed in Example 8(iii). Consider the elliptic equation

$$\lambda\varphi - \mathscr{L}_M\varphi = f. \tag{72}$$

In this case the corresponding semigroup P_t^M is given by (23); one can easily show that the Gaussian measure N_Q is invariant for P_t. As we noticed in Example 8, P_t^M is only improving on the directions of the Cameron–Martin space $Q^{1/2}(H)$. For this reason, to study the regularity of the solution φ to (72), one has to introduce the Malliavin spaces $D^{1,2}(H, \mu)$ (the domain of the closure of $Q^{1/2}D$) and $D^{2,2}(H, \mu)$, whose norms as defined by

$$\|\varphi\|^2_{D^{1,2}(H,\mu)} = \int_H \varphi^2\, d\mu + \sum_{h=1}^{\infty}\lambda_h\int_H |D_h\varphi|^2\, d\mu \tag{73}$$

and

$$\|\varphi\|^2_{D^{2,2}(H,\mu)} = \|\varphi\|^2_{D^{1,2}(H,\mu)} + \sum_{h,k=1}^{\infty}\lambda_h\lambda_k\int_H |D_hD_k\varphi|^2\, d\mu, \tag{74}$$

respectively. Then a maximal regularity results has been proved by Meyer see [13], saying that the domain of $\mathscr{L}_M$ is precisely $D^{2,2}(H,\mu)$.

6 Generalizations

6.1 Elliptic Equations with an Additional Potential Term

We consider here, following [5], the elliptic equation

$$\lambda\varphi - \mathscr{K}\varphi = f, \tag{75}$$

where

$$\mathscr{K}\varphi = \mathscr{L}\varphi - \langle DU, D\varphi\rangle = f, \tag{76}$$

$\mathscr{L}$ is defined in (33) and $U : D(U) \subset H \to \mathbb{R}$ is a convex potential lower semicontinuous. Moreover, λ is a positive number and $f \in L^2(H,\nu)$, where ν is the invariant measure of $\mathscr{K}$.

It is not difficult to show that ν is given by

$$\nu(dx) = \frac{e^{-2U(x)}}{\int_H e^{-2U(y)}\mu(dy)}\,\mu(dx), \tag{77}$$

where $\mu = N_Q$ is the Gaussian measure defined earlier.

In statistical mechanics ν is called a *Gibbs* measure.

A weak solution to Eq. (75) is provided by the Lax–Milgram lemma applied to the coercive Dirichlet form

$$a(u,v) = \frac{1}{2}\int_H \langle Du, Dv\rangle\, d\nu, \quad u, v \in W^{1,2}(H,\nu). \tag{78}$$

In this case Theorem 25 generalizes (see [5])

Theorem 28 *Assume Hypotheses 1, 10 and 20 and that $U : H \to [0,+\infty]$ is convex and lower semicontinuous. Let $\lambda > 0$, $f \in L^2(H,\nu)$ and let φ be the weak solution to (75). Then $\varphi \in W^{2,2}(H,\nu)$, $(-A)^{1/2}D\varphi \in L^2(H,\nu;H)$ and the following estimate holds*

$$\lambda\int_H |D\varphi|^2\, d\nu + \frac{1}{2}\int_H \mathrm{Tr}\,[(D^2\varphi)^2]\, d\nu + \int_H |(-A)^{1/2}D\varphi|^2\, d\nu \le 3\int_H f^2 d\nu. \tag{79}$$

Remark 29 By Theorem 28 it follows that

$$D(\mathscr{K}) \subset \{\varphi \in W^{2,2}(H, \nu) : (-A)^{1/2} D\varphi \in L^2(H, \nu; H)\}.$$

A characterization of the domain of $\mathscr{K}$, however, is not known.

6.2 Neumann Problem

We still assume Hypotheses 1, 10 and 20. Then we consider the following *Neumann problem* in an open *convex* subset $\mathscr{O}$ of H

$$\begin{cases} \lambda\varphi - \mathscr{L}\varphi = f, & \text{in } \mathscr{O}, \\ \dfrac{\partial \varphi}{\partial n} = 0, & \text{on } \partial O, \end{cases} \tag{80}$$

where n is the exterior normal to ∂O. We choose $\mathscr{O} = \{x \in H : g(x) < r\}$, where $g : H \to \mathbb{R}$ is a convex mapping for which a notion of surface integral can be defined following Malliavin, see [1, 8, 12] and references therein. In particular, we can take $\mathscr{O} = B(0, R)$ the ball in H of center 0 and radius R. Other possible choices are domains of the form $\{g < r\}$ where $g : H \to \mathbb{R}$ is sufficiently regular. In all these cases the Sobolev spaces $W^{k,2}(\mathscr{O}, \mu)$, $k = 1, 2$, can be defined in a natural way.

Existence and uniqueness of a weak solution $\varphi \in W^{1.2}(\mathscr{O}, \mu)$ of problem (80) is easily obtained thanks to the Lax–Milgram lemma applied to the symmetric coercive Dirichlet form

$$a(u, v) = \frac{1}{2} \int_{\mathscr{O}} \langle Du, Dv \rangle \, d\mu, \quad u, v \in W^{1.2}(\mathscr{O}, \mu). \tag{81}$$

We aim to show that φ belongs to $W^{2.2}(\mathscr{O}, \mu)$ and that its normal derivative (defined thanks to a trace lemma) vanishes on $\partial\mathscr{O}$. We shall proceed by penalization as in [2] and [6], by studying the problem

$$\lambda\varphi_\varepsilon - \mathscr{L}\varphi_\varepsilon + \frac{1}{\varepsilon} \langle x - \Pi_K x, D\varphi_\varepsilon \rangle = f, \tag{82}$$

where K is the closure $\overline{\mathscr{O}}$ of $\mathscr{O}$ and Π_K denotes the projection on the convex set K.

Arguing as in the proof of (53) we arrive at the estimate

$$\lambda \int_H |D\varphi_\varepsilon|^2 \, d\mu_\varepsilon + \frac{1}{2} \int_H \mathrm{Tr}\,[(D^2\varphi_\varepsilon)^2] \, d\mu_\varepsilon + \int_H |(-A)^{1/2} D\varphi_\varepsilon|^2 \, d\mu_\varepsilon \le 3 \int_H f^2 d\mu_\varepsilon, \tag{83}$$

where here

$$\mu_\varepsilon(dx) = \frac{e^{-\frac{d_K(x)}{\varepsilon}}}{\int_H e^{-2d(y)} \mu(dy)} \, \mu(dx) \tag{84}$$

and $d_K(x)$ represents the distance of x from K,

$$d_K(x) = \inf\{|x - y|, \ y \in K\}, \quad x \in H.$$

One can show, letting $\varepsilon \to 0$ that $\varphi \in W^{2,2}(\mathscr{O}, \mu)$ and that $\frac{\partial \varphi}{\partial n} = 0$ on ∂O, see [2] and [6].

6.3 Dirichlet Problem

Few results seem to be devoted to Dirichlet problems in infinite dimensions. We limit ourselves, following [4], to present a $W^{2,2}$ regularity result for the *Dirichlet problem*

$$\begin{cases} \lambda\varphi - \mathscr{L}_M \varphi = f, & \text{in } \mathscr{O}, \\ \varphi = 0, & \text{on } \partial\mathscr{O}, \end{cases} \tag{85}$$

where $\mathscr{L}_M$ is the Malliavin operator introduced in Example 8(iii),

$$\mathscr{L}_M \varphi = \frac{1}{2} \mathrm{Tr}\,[Q D^2 \varphi] - \frac{1}{2} \langle x, D\varphi \rangle, \tag{86}$$

and $Q \in L_2^+(H)$, $\mu = N_Q$. Moreover, $\mathscr{O} = \{g < r\}$ and g is regular as in the previous subsection.

We shall denote by (e_h) an orthonormal basis in H and by (λ_h) a sequence of positive numbers such that $Qe_h = \lambda_h e_h$, $h \in \mathbb{N}$. Moreover we set $x_k = \langle x, e_k \rangle$, $k \in \mathbb{N}$, $P_n x = \sum_{k=1}^n x_k e_k$, $Q_n = Q P_n$ and $\mu_n = N_{Q_n}$, $n \in \mathbb{N}$.

We shall consider the Sobolev spaces $D^{1,2}(\mathscr{O}, \mu)$ and $D^{2,2}(\mathscr{O}, \mu)$ whose norms are natural generalizations of (73) and (74) respectively. Moreover we shall define

the space $\mathring{D}^{1,2}(\mathscr{O},\mu)$ as the subspace of those functions from $D^{1,2}(H,\mu)$ that vanish in the exterior of $\mathscr{O}$.

Existence and uniqueness of a *weak solution* φ of (85) is provided again by the Lax–Milgram theorem applied to the coercive Dirichlet form

$$a(u,v)=\frac{1}{2}\int_{\mathscr{O}}\langle Q^{1/2}Du,\,Q^{1/2}Dv\rangle\,d\mu,\quad u,v\in\mathring{D}^{1,2}(\mathscr{O},\mu). \tag{87}$$

To show that $\varphi\in D^{2,2}(\mathscr{O},\mu)$ we follow [4] and approximate (85) with a sequence of finite dimensional problems

$$\begin{cases}\lambda\varphi_n-\dfrac{1}{2}\displaystyle\sum_{h=1}^{n}\lambda_h D_h^2\varphi+\frac{1}{2}\sum_{h=1}^{n}x_h D_h\varphi=f,\quad \text{in } \mathscr{O}_n,\\[2ex] \varphi=0,\quad \text{on } \partial\mathscr{O}_n,\end{cases} \tag{88}$$

where $\mathscr{O}_n=\{g(P_nx)<r\}$. Then we look for an estimate

$$\|u_n\|_{D^{2,2}(O_n,\mu_n)}\le K\|f_n\|_{L^2(O_n,\mu_n)} \tag{89}$$

with a constant K independent of n. Procedures relying on maximal regularity for elliptic operators in $\mathbb{R}^n$ with respect to the Lebesgue measure do not work, because the final constant K depends on n in an uncontrollable way. Instead, we follow a more direct approach, following [11]. Let us consider for simplicity the case of the unit ball $\mathscr{O}_n=B(0,1)$ and $f\in C_0^\infty(B(0,1))$.

Dimension free bounds for $\|u\|_{D^{1,2}(O_n,\mu_n)}$ are easily found. To estimate the second order derivatives we proceed as in Sect. 5.2. We differentiate both members of (88) with respect to x_h, we multiply by $D_hu_n\lambda_h$, we sum up and we integrate by parts obtaining

$$\int_{B(0,1)}\mathrm{Tr}\,[(Q_nD^2u_n)^2]\,d\mu_n=\frac{1}{2}\int_{\partial B(0,1)}\langle D^2u_nQx,\,QDu_n\rangle\,d\sigma_n$$
$$+\text{terms under control}$$

The boundary integral still contains second order derivatives of u_n, however using the identities $u_n=f_n=0$ on $|x|=1$ we can express $\langle D^2u_nQ_nx,\,Q_nDu_n\rangle$ in terms of first order derivatives of u_n, and precisely, after some manipulations

$$\langle D^2u_nQ_nx,\,Q_nDu_n\rangle=\frac{\langle Q_n^{1/2}Du_n,\,Q_n^{1/2}x\rangle^2}{|Q_n^{1/2}x|^2}\left(1-\mathrm{Tr}\,[Q_n]+\frac{|Q_nx|^2}{|Q_n^{1/2}x|^2}\right)$$
$$:=\mathscr{H}_n(x)\langle Q_n^{1/2}Du_n,\,Q_n^{1/2}x\rangle^2.$$

On the other hand, by a suitable trace lemma it is possible to estimate the boundary integral

$$\int_{\partial B(0,1)} \langle Q_n^{1/2} Du_n, Q_n^{1/2} x\rangle^2 d\sigma_n,$$

in terms of the norm $W^{2,2}(O_n, \mu_n)$ of u as follows

$$\int_{\partial B(0,1)} \langle Q_n^{1/2} Du_n, Q_n^{1/2} x\rangle^2 d\sigma_n$$

$$\leq C_1 \int_{B(0,1)} \mathrm{Tr}\,[(Q_n D^2 u_n)^2]\, d\mu_n \| f_n\|_{L^2(B(0,1),\mu_n)} + C_2 \| f_n\|^2_{L^2(B(0,1),\mu_n)},$$

with $C_1, C_2 > 0$ independent of n.

Therefore, if

$$\sup_{x\in\partial B(0,1)} \mathscr{H}_n(x) \leq C, \quad \text{with } C \text{ independent of } n, \tag{90}$$

we are done: we get a dimension free estimate for $\int_{B(0,1)} \mathrm{Tr}\,[(Q_n D^2 u_n)^2]\, d\mu_n$ and the estimate $\|u_n\|_{W^{2,2}(O_n,\mu_n)} \leq K \|f_n\|_{L^2(O_n,\mu_n)}$ follows. Finally, letting $n \to \infty$, we arrive at the desired result $\varphi \in W^{2,2}(\mathscr{O}, \mu)$. See [4] for details and more general results.

Remark 30 Checking (90) in the examples gives some surprise, and shows important differences between the finite and the infinite dimensional case. For instance, if $\mathscr{O} = B(0, r)$ the suprema h_n of $\mathscr{H}_n$ on the spherical surfaces are bounded by a constant independent of n only if some relationship between r and the eigenvalues of Q is satisfied. In particular, (90) is satisfied only provided $r^2 < \mathrm{Tr}\, Q$.

Appendix: Hermite Polynomials in Hilbert Spaces

Let us consider the analytic function

$$F(t, \xi) = e^{-\frac{t^2}{2} + t\xi}, \quad t, \xi \in \mathbb{R}, \tag{91}$$

whose Taylor series in t is given by

$$F(t, \xi) = \sum_{n=0}^{\infty} \frac{t^n}{\sqrt{n!}} H_n(\xi), \quad t, \xi \in \mathbb{R}.$$

Then for any $n \in \{0\} \cup \mathbb{N}$ the following identity is easily checked

$$H_n(\xi) = \frac{(-1)^n}{\sqrt{n!}} e^{\frac{\xi^2}{2}} D_\xi^n \left(e^{-\frac{\xi^2}{2}} \right), \quad \xi \in \mathbb{R}. \tag{92}$$

By (92) we see that for any $n \in \mathbb{N}$, H_n is a polynomial of degree n having a positive leading coefficient. H_n are called *Hermite* polynomials. The following result is classic.

Proposition 31 *System $(H_n)_{n\in\{0\}\cup\mathbb{N}}$ is orthonormal and complete in $L^2(\mathbb{R}, \mu)$, where $\mu = N_1$.*

Now let us define a complete orthonormal system on $L^2(H, N_Q)$, where $\mu = N_Q$ is a non degenerate Gaussian measure. Let (e_k) be an orthonormal basis in H and (λ_k) a sequence of positive numbers such that $Qe_k = \lambda_k e_k$, $k \in \mathbb{N}$. We start with a basic property of Hermite polynomial composed with white noise functions.

Lemma 32 *Let $h, g \in H$ with $|h| = |g| = 1$ and let $n, m \in \mathbb{N} \cup \{0\}$. Then we have,*

$$\int_H H_n(W_h) H_m(W_g) d\mu = [\langle h, g\rangle]^n \delta_{n,m}, \tag{93}$$

where $\delta_{n,m}$ is the Kronecker symbol.

Proof For $t, s \in \mathbb{R}$ write

$$\int_H F(t, W_h) F(s, W_g) d\mu = e^{-\frac{t^2+s^2}{2}} \int_H e^{tW_h + sW_g} \, d\mu$$

$$= e^{-\frac{t^2+s^2}{2}} \int_H e^{W_{th+sg}} \, d\mu = e^{-\frac{t^2+s^2}{2}} e^{\frac{1}{2}|th+sg|^2} = e^{ts\langle h,g\rangle},$$

because $|h| = |g| = 1$. It follows that

$$e^{ts\langle h,g\rangle} = \sum_{m,n=0}^{\infty} \frac{t^n s^m}{\sqrt{n!m!}} \int_H H_n(W_h) H_m(W_g) \, d\mu,$$

which implies clearly (93).

We can now define *Hermite polynomials* in H. Let Γ be the set of all mappings $\gamma : \mathbb{N} \to \{0\} \cup \mathbb{N}$, $n \mapsto \gamma_n$, such that $|\gamma| := \sum_{k=1}^{\infty} \gamma_k < +\infty$. Notice that if $\gamma \in \Gamma$ then $\gamma_n = 0$ for all $n \in \mathbb{N}$, except at most a finite number. We denote by 0 the element $\gamma \in \Gamma$ such that $\gamma_h = 0$ for all $h \in \mathbb{N}$. Then for any $\gamma \in \Gamma$ we define the Hermite polynomial H_γ as

$$H_\gamma(x) = \prod_{k=1}^{\infty} H_{\gamma_k}(W_{e_k}(x_k)) = \prod_{k=1}^{\infty} H_{\gamma_k}(\lambda_k^{-1/2} x_k), \quad x \in H. \tag{94}$$

Notice that in the product in (94) all factors (with the possible exception of a finite number of them) are equal to 1. In particular we have $H_0 = 1$. It is obvious that $H_\gamma \in L^2(H, \mu)$. One can show the following result, see e. g. [7, Theorem 9.1.5]

Theorem 33 *System $(H_\gamma)_{\gamma\in\Gamma}$ is orthonormal and complete on $L^2(H, \mu)$.*

Setting

$$\langle \gamma, \lambda^{-1}\rangle = \sum_{h=1}^{\infty} \frac{\gamma_h}{\lambda_h}, \quad \langle \gamma, \lambda^{-2}\rangle = \sum_{h=1}^{\infty} \frac{\gamma_h}{\lambda_h^2}, \tag{95}$$

we find with a simple computation the following identities, see e. g. [7],

$$\int_H |D\varphi|^2 \, d\mu = \sum_{\gamma\in\Gamma} \langle \gamma, \lambda^{-1}\rangle \, |\varphi_\gamma|^2. \tag{96}$$

and

$$\int_H \mathrm{Tr}[(D^2\varphi)^2] d\mu = \sum_{\gamma\in\Gamma} |\varphi_\gamma|^2 \langle \gamma, \lambda^{-1}\rangle^2 - \sum_{\gamma\in\Gamma} |\varphi_\gamma|^2 \langle \gamma, \lambda^{-2}\rangle. \tag{97}$$

References

1. H. Airault, P. Malliavin, Intégration géométrique sur l'espace de Wiener. Bull. Sci. Math. **112**, 3–52 (1988)
2. V. Barbu, G. Da Prato, L. Tubaro, Kolmogorov equation associated to the stochastic reflection problem on a smooth convex set of a Hilbert space. Ann. Probab. **37**,1427–1458 (2009)
3. G. Da Prato, *Kolmogorov equations in Hilbert spaces* (Birkhäuser, Basel, 2004).
4. G. Da Prato, A. Lunardi, Maximal L^2 regularity for Dirichlet problems in Hilbert spaces. J. Math. Pures Appl. **99**(6), 741–765 (2013)
5. G. Da Prato, A. Lunardi, Sobolev regularity for a class of second order elliptic PDE's in infinite dimension. Ann. Probab. **42**(5), 2113–2160 (2014)
6. G. Da Prato, A. Lunardi, Maximal Sobolev regularity in Neumann problems for gradient systems in infinite dimensional domains. Ann. Inst. Henri Poincaré Probab. Stat. **51**(3), 1102–1123 (2015)
7. G. Da Prato, J. Zabczyk, *Second Order Partial Differential Equations in Hilbert Spaces*. London Mathematical Society Lecture Notes 293 (Cambridge University Press, Cambridge, 2002)
8. G. Da Prato, A. Lunardi, L. Tubaro, Surface measures in infinite dimension. Rend. Lincei Mat. Appl. **25**, 309–330 (2014)
9. E.B. Davies, *One Parameter Semigroups* (Academic Press, London, 1980).
10. A.N. Kolmogorov, Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. Math. Ann. **104**, 415–458 (1931)
11. A. Lunardi, G. Metafune, D. Pallara, Dirichlet boundary conditions for elliptic operators with unbounded drift. Proc. Am. Math. Soc. **133**, 2625–2635 (2005)
12. P. Malliavin, *Stochastic Analysis* (Springer, New York, 1997).

13. P.A. Meyer, Note sur les processus d'Ornstein–Uhlenbeck, in *Séminaire de Probabilités, XVI*. Lecture Notes in Mathematics, vol. 920 (Springer, New York, 1982), pp. 95–133.
14. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Applied Mathematical Sciences, vol. 44 (Springer, New York, 1983)
15. I. Shigekawa, Sobolev spaces over the Wiener space based on an Ornstein-Uhlenbeck operator. J. Math. Kyoto Univ. **32**, 731–748 (1982)
16. J. Zabczyk, *Mathematical Control Theory: An Introduction. Systems & Control: Foundations & Applications* (Birkhäuser, Basel, 1992)

On Generalized Pseudo- and Quasiconvexities for Nonsmooth Functions



Ville-Pekka Eronen, Marko M. Mäkelä, and Napsu Karmitsa

1 Introduction

Convexity plays a crucial role in mathematical optimization theory. Especially, in duality theory and when constructing optimality conditions convexity has been the most important concept since the basic reference by Rockafellar [22]. Numerous attempts to generalize the concept of convexity have been made in order to weaken the assumptions of the attained results (see e.g., [1, 2, 4, 8, 12, 14, 21, 23, 25]). Generalized convexities have proved to be the main tool when constructing optimality conditions, particularly sufficient conditions. There exist a wide amount of papers published for smooth single-objective case (see [21] and references therein and [14]). For nonsmooth and multiobjective problems necessary conditions were derived for instance in [18, 19, 24].

In this paper, we analyze the properties of the generalized pseudo- and quasiconvexities for nondifferentiable locally Lipschitz continuous functions. The treatment is based on the Clarke subdifferentials and generalized directional derivatives [6]. The paper follows the presentation in [3] with some additional results such as introduction of GNC-property. In Sect. 2 we recall some basic tools from nonsmooth analysis. Sections 3 and 4 are devoted to generalized pseudo- and quasiconvexity, respectively. Also, some relations between generalized pseudo- and quasiconvexities are considered in Sect. 4. Finally, the derived results are summarized in Sect. 5.

V.-P. Eronen (✉) · M. M. Mäkelä · N. Karmitsa
Department of Mathematics and Statistics, University of Turku, Turku, Finland
e-mail: vpoero@utu.fi

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_6

2 Nonsmooth Analysis

In this section we collect some notions and results from nonsmooth analysis. Nevertheless, we start by recalling the notion of convexity and Lipschitz continuity. The function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for all $\mathbf{x}$, $\boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$ we have

$$f\big(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}\big) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

In what follows the considered functions are assumed to be locally Lipschitz continuous. A function is *locally Lipschitz continuous at a point* $\boldsymbol{x} \in \mathbb{R}^n$ if there exist scalars $K > 0$ and $\delta > 0$ such that

$$|f(\boldsymbol{y}) - f(\boldsymbol{z})| \leq K \|\boldsymbol{y} - \boldsymbol{z}\| \quad \text{for all } \boldsymbol{y}, \boldsymbol{z} \in B(\boldsymbol{x}; \delta),$$

where $B(\boldsymbol{x}; \delta) \subset \mathbb{R}^n$ is an open ball with center $\boldsymbol{x}$ and radius δ. Function is said to be *locally Lipschitz continuous on a set* $U \subseteq \mathbb{R}^n$ if it is locally Lipschitz continuous at every point belonging to the set U. Furthermore, if $U = \mathbb{R}^n$ the function is called *locally Lipschitz continuous*. Note that both convex and smooth (continuously differentiable) functions are always locally Lipschitz continuous (see, e.g., [6]).

Function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *Lipschitz continuous on a set* $U \subseteq \mathbb{R}^n$ if there exists a scalar K such that

$$|f(\boldsymbol{y}) - f(\boldsymbol{z})| \leq K \|\boldsymbol{y} - \boldsymbol{z}\| \quad \text{for all } \boldsymbol{y}, \boldsymbol{z} \in U.$$

If $U = \mathbb{R}^n$ then f is said to be *Lipschitz continuous*.

Definition 1 ([6]) Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x} \in S \subseteq \mathbb{R}^n$. The *Clarke generalized directional derivative* of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{d} \in \mathbb{R}^n$ is defined by

$$f^\circ(\boldsymbol{x}; \boldsymbol{d}) := \limsup_{\substack{\boldsymbol{y} \to \boldsymbol{x} \\ t \downarrow 0}} \frac{f(\boldsymbol{y} + t\boldsymbol{d}) - f(\boldsymbol{y})}{t}$$

and the *Clarke subdifferential* of f at $\boldsymbol{x}$ by

$$\partial f(\boldsymbol{x}) := \{\boldsymbol{\xi} \in \mathbb{R}^n \mid f^\circ(\boldsymbol{x}; \boldsymbol{d}) \geq \boldsymbol{\xi}^T \boldsymbol{d} \text{ for all } \boldsymbol{d} \in \mathbb{R}^n\}.$$

Each element $\boldsymbol{\xi} \in \partial f(\boldsymbol{x})$ is called a *subgradient* of f at $\boldsymbol{x}$.

Note that the Clarke generalized directional derivative $f^\circ(\boldsymbol{x}; \boldsymbol{d})$ always exists for a locally Lipschitz continuous function f. If f is convex $\partial f(\boldsymbol{x})$ coincides with the classical subdifferential of convex function (cf. [22]), in other words the set of $\boldsymbol{\xi} \in \mathbb{R}^n$ satisfying

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{\xi}^T (\boldsymbol{y} - \boldsymbol{x}) \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^n.$$

Furthermore, if f is smooth $\partial f(\boldsymbol{x})$ reduces to $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. Smoothness is critical here as the function

$$g(x) = \begin{cases} 0, & x = 0 \\ x^2 \cos(\frac{1}{x}), & x \neq 0 \end{cases} \tag{1}$$

shows. Function g is locally Lipschitz continuous and differentiable everywhere but nonsmooth (not continuously differentiable) and $\partial g(0) \neq \{\nabla g(0)\}$. More detailed analysis on this function can be found in [3, pp. 68–69].

The following properties first derived in [6] are characteristic to the generalized directional derivative and subdifferential.

Theorem 1 *If $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous at $\boldsymbol{x} \in \mathbb{R}^n$, then*

(i) $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ *is positively homogeneous, subadditive and Lipschitz continuous function such that* $f^\circ(\boldsymbol{x}; -\boldsymbol{d}) = (-f)^\circ(\boldsymbol{x}; \boldsymbol{d})$.
(ii) $\partial f(\boldsymbol{x})$ *is a nonempty, convex and compact set.*
(iii) $f^\circ(\boldsymbol{x}; \boldsymbol{d}) = \max\{\boldsymbol{\xi}^T \boldsymbol{d} \mid \boldsymbol{\xi} \in \partial f(\boldsymbol{x})\}$ *for all* $\boldsymbol{d} \in \mathbb{R}^n$.

The subdifferential can be constructed as a convex hull of all possible limits of gradients at point $\boldsymbol{x}_i$ converging to $\boldsymbol{x}$. Let Ω_f be the set of points where f is not differentiable. That is,

$$\Omega_f = \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \ f \text{ is not differentiable at the point } \boldsymbol{x}\right\}.$$

By Rademacher's Theorem [16] a locally Lipschitz continuous function on $U \subseteq \mathbb{R}^n$ is differentiable almost everywhere on U, in other words, $\text{meas}(\Omega_f) = 0$.

Theorem 2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x} \in \mathbb{R}^n$. Then*

$$\partial f(\boldsymbol{x}) = \text{conv}\left\{\boldsymbol{\xi} \in \mathbb{R}^n \mid \exists(\boldsymbol{x}_i) \subset \mathbb{R}^n \backslash \Omega_f \text{ s.t. } \boldsymbol{x}_i \to \boldsymbol{x} \text{ and } \nabla f(\boldsymbol{x}_i) \to \boldsymbol{\xi}\right\},$$

where conv *denotes the convex hull of a set.*

Proof See, for example, [16, pp. 50–51]. □

Some derivation rules can be generalized for nonsmooth functions. However, these subderivations rules are valid as inclusions instead of equalities. To maintain equalities instead of inclusions in these rules we need the following regularity property.

Definition 2 The function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *subdifferentially regular* at $\boldsymbol{x} \in \mathbb{R}^n$ if it is locally Lipschitz continuous at $\boldsymbol{x}$ and for all $\boldsymbol{d} \in \mathbb{R}^n$ the classical directional derivative

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x} + t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

exists and $f'(\boldsymbol{x}; \boldsymbol{d}) = f^\circ(\boldsymbol{x}; \boldsymbol{d})$.

The equality $f'(\boldsymbol{x}; \boldsymbol{d}) = f^\circ(\boldsymbol{x}; \boldsymbol{d})$ need not to be valid in general even if $f'(\boldsymbol{x}; \boldsymbol{d})$ exists. This is the case, for instance, with concave nonsmooth functions. For example, the function $f(x) = -|x|$ has the directional derivative $f'(0; 1) = -1$, but the generalized directional derivative is $f^\circ(0; 1) = 1$. Nevertheless, convexity, as well as smoothness implies subdifferential regularity [6]. Furthermore, it is easy to show that a necessary and sufficient condition for convexity is that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &\geq f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \\ &= f'(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}). \end{aligned} \tag{2}$$

We will now present two subderivation rules of composite functions, namely the finite maximum and positive linear combination of locally Lipschitz continuous functions.

Theorem 3 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x}$ for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \max \{ f_i(\boldsymbol{x}) \mid i = 1, \ldots, m \}$$

is locally Lipschitz continuous at $\boldsymbol{x}$ and

$$\partial f(\boldsymbol{x}) \subseteq \operatorname{conv} \{ \partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x}),\ i = 1, \ldots, m \} \tag{3}$$

In addition, if f_i is subdifferentially regular at $\boldsymbol{x}$ for all $i = 1, \ldots, m$, then f is also subdifferentially regular at $\boldsymbol{x}$ and equality holds in (3).

Proof See, for example, [6, p. 47]. □

Theorem 4 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x}$ and $\lambda_i \in \mathbb{R}$ for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x})$$

is locally Lipschitz continuous at $\boldsymbol{x}$ and

$$\partial f(\boldsymbol{x}) \subseteq \sum_{i=1}^{m} \lambda_i \partial f_i(\boldsymbol{x}). \tag{4}$$

In addition, if f_i is subdifferentially regular at $\boldsymbol{x}$ and $\lambda_i \geq 0$ for all $i = 1, \ldots, m$, then f is also subdifferentially regular at $\boldsymbol{x}$ and equality holds in (4).

Proof See, for example, [6, pp. 39–40]. □

The following two results generalize the classical Mean-Value Theorem and the Chain Rule, respectively.

Theorem 5 *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be such that $\boldsymbol{x} \neq \boldsymbol{y}$ and let the function f be locally Lipschitz continuous on an open set $U \subseteq \mathbb{R}^n$ such that the line segment $[\boldsymbol{x}, \boldsymbol{y}] \subset U$. Then there exists a point $\boldsymbol{z} \in (\boldsymbol{x}, \boldsymbol{y})$ such that*

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \in \partial f(z)^T (\boldsymbol{y} - \boldsymbol{x}).$$

Proof See, for example, [6, pp. 41–42]. □

Theorem 6 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x}$ and $g : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz continuous at $h(\boldsymbol{x})$. Then the composite function $f = g \circ h : \mathbb{R}^n \to \mathbb{R}$ is also locally Lipschitz continuous at $\boldsymbol{x}$ and one has*

$$\partial f(\boldsymbol{x}) \subseteq \operatorname{conv} \{\partial g(h(\boldsymbol{x})) \partial h(\boldsymbol{x})\}.$$

Proof See, for example, [6, pp. 72–73]. □

Next we give the basic unconstrained optimality condition.

Theorem 7 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x}^*$. If f attains its local minimum at $\boldsymbol{x}^*$, then*

$$\mathbf{0} \in \partial f(\boldsymbol{x}^*).$$

If, in addition, f is convex, then the above condition is also sufficient for $\boldsymbol{x}^*$ to be a global minimum.

Proof See, for example, [16, pp. 70–71]. □

Now we present a theorem and two lemmas that are used later.

Lemma 1 *Let $\varepsilon > 0$, function $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $\boldsymbol{x}$ and $\boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{d} \neq \mathbf{0}$. Then*

$$f^\circ(\boldsymbol{x}; \boldsymbol{d}) - \varepsilon \leq \limsup \left\{ \nabla f(\boldsymbol{y})^T \boldsymbol{d} \mid \boldsymbol{y} \to \boldsymbol{x},\ \boldsymbol{y} \notin \Omega_f \right\}.$$

Proof See, for example, [16, pp. 51–52]. □

Lemma 2 *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a point, where $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and differentiable. Let K be the Lipschitz constant of the function f at the point $\boldsymbol{x}$. Then the function $\boldsymbol{d} \mapsto f'(\boldsymbol{x}; \boldsymbol{d})$ is positively homogeneous and Lipschitz continuous with the constant K.*

Proof The proof is similar to the corresponding results for the Clarke generalized directional derivative. □

The *level set* of f with a parameter $\alpha \in \mathbb{R}$ is defined as

$$\text{lev}_\alpha f := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \leq \alpha\}$$

Lemma 3 *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and on $[\boldsymbol{x}, \boldsymbol{y}]$ locally Lipschitz continuous function f be such that $f(\boldsymbol{x}) < f(\boldsymbol{y})$. Then, there exists a point $\bar{\boldsymbol{x}} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}$, $\lambda \in (0,1)$ such that $f(\bar{\boldsymbol{x}}) > f(\boldsymbol{x})$ and $f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x}) > 0$.*

Proof Consider the nonempty set $A = \text{lev}_{f(\boldsymbol{x})} f \cap [\boldsymbol{x}, \boldsymbol{y}]$. Since level sets of a continuous function are closed sets and $[\boldsymbol{x}, \boldsymbol{y}]$ is compact, the set A is a compact set. Since function $g(\boldsymbol{w}) := \|\boldsymbol{w} - \boldsymbol{y}\|$ is continuous, it has a minimum on the set A according to the well-known Weierstrass Theorem. Let this minimum point be $\boldsymbol{z}$. Then $\boldsymbol{z}$ is the nearest point to $\boldsymbol{y}$ on the set A and the continuity of function f implies $f(\boldsymbol{z}) = f(\boldsymbol{x})$. Also, $\boldsymbol{z} \neq \boldsymbol{y}$ since $f(\boldsymbol{x}) < f(\boldsymbol{y})$. The Mean-Value Theorem 5 implies that there exist $\bar{\boldsymbol{z}} \in (\boldsymbol{z}, \boldsymbol{y})$ and $\boldsymbol{\xi} \in \partial f(\bar{\boldsymbol{z}})$ such that

$$f(\boldsymbol{y}) - f(\boldsymbol{z}) = \boldsymbol{\xi}^T (\boldsymbol{y} - \boldsymbol{z}).$$

Since $f(\boldsymbol{z}) < f(\boldsymbol{y})$ we have

$$0 < f(\boldsymbol{y}) - f(\boldsymbol{z}) = \boldsymbol{\xi}^T (\boldsymbol{y} - \boldsymbol{z}) \leq f^\circ(\bar{\boldsymbol{z}}; \boldsymbol{y} - \boldsymbol{z}) \leq f^\circ(\bar{\boldsymbol{z}}; \boldsymbol{y} - \boldsymbol{x}),$$

where the last inequality follows from positive homogeneity and inequality $\|\boldsymbol{y} - \boldsymbol{z}\| \leq \|\boldsymbol{y} - \boldsymbol{x}\|$. By the choice of the point $\boldsymbol{z}$ we know that $f(\boldsymbol{z}) < f(\bar{\boldsymbol{z}})$ since $\bar{\boldsymbol{z}} \in (\boldsymbol{z}, \boldsymbol{y})$. Choosing $\bar{\boldsymbol{x}} = \bar{\boldsymbol{z}}$ we have $f(\boldsymbol{x}) = f(\boldsymbol{z}) < f(\bar{\boldsymbol{x}})$ and the lemma has been proven. □

3 Generalized Pseudoconvexity

The most famous definition of pseudoconvexity for smooth functions was introduced in [17].

Definition 3 A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *pseudoconvex*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) < f(\boldsymbol{x}) \qquad \text{implies} \qquad \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) < 0.$$

The main result for a smooth pseudoconvex function f is that the convexity assumption of Theorem 7 can be weakened. In other words, a smooth pseudoconvex function f attains a global minimum at $\boldsymbol{x}^*$, if and only if $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ (see [17]).

The concept of pseudoconvexity has been extended for nonsmooth case, for instance, in [2, 14, 21] and the references therein. One way to do this is the usage of directional derivatives. The Dini directional derivatives were used, for example, in [5, 8, 11]. In [12] this idea was generalized for lower semicontinuous (l.s.c.)

functions via h-pseudoconvexity, where $h(\boldsymbol{x}, \boldsymbol{d})$ is any real-valued bifunction, that is, for example, any directional derivative. In this paper we use the definition first given in [10] for locally Lipschitz continuous functions.

Definition 4 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *f°-pseudoconvex*, if it is locally Lipschitz continuous and for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) < f(\boldsymbol{x}) \qquad \text{implies} \qquad f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) < 0.$$

A convex function is always f°-pseudoconvex by (2). The next result shows that f°-pseudoconvexity is a natural extension of pseudoconvexity.

Theorem 8 *If f is smooth, then f is f°-pseudoconvex, if and only if f is pseudoconvex.*

Proof Follows immediately from Theorem 1 (iii), since for a smooth f we have $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = f'(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$. □

The reasoning chain in the definition of f°-pseudoconvexity may be converted.

Lemma 4 *A locally Lipschitz continuous function f is f°-pseudoconvex, if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0 \qquad \textit{implies} \qquad f(\boldsymbol{y}) \geq f(\boldsymbol{x}).$$

Proof Follows directly from the definition of f°-pseudoconvexity. □

The important sufficient extremum property of pseudoconvex functions remain also for f°-pseudoconvex functions.

Theorem 9 *An f°-pseudoconvex f attains its global minimum at $\boldsymbol{x}^*$, if and only if*

$$\mathbf{0} \in \partial f(\boldsymbol{x}^*).$$

Proof If f attains its global minimum at $\boldsymbol{x}^*$, then by Theorem 7 we have $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$. On the other hand, if $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$ and $\boldsymbol{y} \in \mathbb{R}^n$, then by Definition 1

$$f^\circ(\boldsymbol{x}^*; \boldsymbol{y} - \boldsymbol{x}^*) \geq \mathbf{0}^T(\boldsymbol{y} - \boldsymbol{x}^*) = 0$$

and, thus by Lemma 4 we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*).$$

□

The notion of monotonicity is closely related to convexity. It has been studied in e.g. [12, 20].

Definition 5 The generalized directional derivative f° is called *pseudomonotone*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0 \qquad \text{implies} \qquad f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) \leq 0$$

or, equivalently

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0 \qquad \text{implies} \qquad f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) < 0.$$

Furthermore, f° is *strictly pseudomonotone*, if

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0 \qquad \text{implies} \qquad f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) < 0.$$

Theorem 10 *If f is locally Lipschitz continuous such that f° is pseudomonotone, then f is f°-pseudoconvex.*

Proof Let us, on the contrary, assume that f is not f°-pseudoconvex. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x})$ and

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0. \tag{5}$$

Then by the Mean-Value Theorem 5 there exists $\hat{\lambda} \in (0, 1)$ such that $\hat{\boldsymbol{x}} = \boldsymbol{x} + \hat{\lambda}(\boldsymbol{y} - \boldsymbol{x})$ and

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \in \partial f(\hat{\boldsymbol{x}})^T (\boldsymbol{x} - \boldsymbol{y}).$$

This means that due to the definition of the Clarke subdifferential there exists $\hat{\boldsymbol{\xi}} \in \partial f(\hat{\boldsymbol{x}})$ such that

$$0 < f(\boldsymbol{x}) - f(\boldsymbol{y}) = \hat{\boldsymbol{\xi}}^T (\boldsymbol{x} - \boldsymbol{y}) \leq f^\circ(\hat{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y}). \tag{6}$$

On the other hand, from (5) and the positive homogeneity of $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ (see Theorem 1 (i)) we deduce that

$$f^\circ(\boldsymbol{x}; \hat{\boldsymbol{x}} - \boldsymbol{x}) = f^\circ(\boldsymbol{x}; \hat{\lambda}(\boldsymbol{y} - \boldsymbol{x})) = \hat{\lambda} f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0.$$

Then the pseudomonotonicity, the positive homogeneity of $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ and (6) imply that

$$0 \geq f^\circ(\hat{\boldsymbol{x}}; \boldsymbol{x} - \hat{\boldsymbol{x}}) = \hat{\lambda} f^\circ(\hat{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y}) > 0,$$

which is impossible. Thus f is f°-pseudoconvex. □

Theorem 10 is a special case of Theorem 4.1 in [20], where it was proven for lower semicontinuous and radially continuous pseudoconvex function.

The converse result of Theorem 10 is also true. Few lemmas are needed before the proof.

Lemma 5 *Let f be an f°-pseudoconvex function, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\bar{\lambda} \in (0, 1)$. Denote $\bar{\boldsymbol{x}} = \bar{\lambda}\boldsymbol{x} + (1 - \bar{\lambda})\boldsymbol{y}$. Then, $f(\bar{\boldsymbol{x}}) \leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$*

Proof On the contrary assume that $f(\bar{\boldsymbol{x}}) > \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$. Since f is f°-pseudoconvex and $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ is positively homogeneous by Theorem 1 (i), we have

$$0 > f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{x} - \bar{\boldsymbol{x}}) = f^\circ(\bar{\boldsymbol{x}}; (1 - \bar{\lambda})(\boldsymbol{x} - \boldsymbol{y})) = (1 - \bar{\lambda}) f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y})$$

and thus

$$f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y}) < 0.$$

Correspondingly, we obtain

$$0 > f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{y} - \bar{\boldsymbol{x}}) = f^\circ(\bar{\boldsymbol{x}}; \bar{\lambda}(\boldsymbol{y} - \boldsymbol{x})) = \bar{\lambda} f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x})$$

and thus

$$f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x}) < 0.$$

Since $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ is subadditive by Theorem 1 (i), we have

$$0 > f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y}) + f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x}) \geq f^\circ(\bar{\boldsymbol{x}}; (\boldsymbol{x} - \boldsymbol{y}) + (\boldsymbol{y} - \boldsymbol{x})) = f^\circ(\bar{\boldsymbol{x}}; \boldsymbol{0}) = 0,$$

which is impossible. In other words, $f(\bar{\boldsymbol{x}}) \leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$. □

Lemma 6 *Let f be an f°-pseudoconvex function. Then there exist no points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, satisfying*

(i) $f(\boldsymbol{x}) = f(\boldsymbol{y})$ *and*
(ii) $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0$.

Proof On the contrary, assume that there exist points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\delta > 0$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = \delta$ and $f(\boldsymbol{x}) = f(\boldsymbol{y})$. Since f is locally Lipschitz continuous there exist $\varepsilon, K > 0$ such that K is the Lipschitz constant in the ball $B(\boldsymbol{x}; \varepsilon)$. Since $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = \delta$ Lemma 1 implies that there exists a sequence $(\boldsymbol{z}^i)$ of points where f is differentiable and $I \in \mathbb{N}$ such that $\boldsymbol{z}^i \to \boldsymbol{x}$ and

$$f'(\boldsymbol{z}^i; \boldsymbol{y} - \boldsymbol{x}) = \nabla f(\boldsymbol{z}^i)^T (\boldsymbol{y} - \boldsymbol{x}) > \frac{\delta}{2} \tag{7}$$

holds when $i \geq I$. Let

$$\hat{\varepsilon} = \min\left\{\varepsilon, \frac{\delta}{2K}\right\}$$

and $z \in B(\boldsymbol{x}; \hat{\varepsilon}) \cap \left\{(z^i) \mid i \geq I\right\}$. According to Lemma 2 $f'(z; \cdot)$ is Lipschitz continuous with the constant K. Hence,

$$\begin{aligned}\left|f'(z; \boldsymbol{y} - \boldsymbol{x}) - f'(z; \boldsymbol{y} - z)\right| &\leq K \left\|\boldsymbol{y} - \boldsymbol{x} - (\boldsymbol{y} - z)\right\| \\ &= K \left\|z - \boldsymbol{x}\right\| < K \frac{\delta}{2K} = \frac{\delta}{2}. \end{aligned} \tag{8}$$

Thus, $f'(z; \boldsymbol{y} - z) > 0$ according to (7) and (8). Since $f'(z; \boldsymbol{y} - z) > 0$ there exists $\mu \in (0, 1)$ such that

$$f(\mu z + (1 - \mu)\boldsymbol{y}) > f(z). \tag{9}$$

Since $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = \delta$, Theorem 1 (i) implies that there exists $\bar{\varepsilon} > 0$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{d}) > 0$ when $\boldsymbol{d} \in B(\boldsymbol{y} - \boldsymbol{x}; \bar{\varepsilon})$. Let $\bar{z} \in B(\boldsymbol{y}; \bar{\varepsilon})$. Since

$$\left\|\bar{z} - \boldsymbol{x} - (\boldsymbol{y} - \boldsymbol{x})\right\| = \left\|\bar{z} - \boldsymbol{y}\right\| < \bar{\varepsilon},$$

it follows that $\bar{z} - \boldsymbol{x} \in B(\boldsymbol{y} - \boldsymbol{x}; \bar{\varepsilon})$. Thus, $f^\circ(\boldsymbol{x}; \bar{z} - \boldsymbol{x}) > 0$ and the f°-pseudoconvexity of the function f implies $f(\bar{z}) \geq f(\boldsymbol{x}) = f(\boldsymbol{y})$. Thus, $\boldsymbol{y}$ is a local minimum for the function f and Theorem 7 implies that $\boldsymbol{0} \in \partial f(\boldsymbol{y})$. Due to Theorem 9 $\boldsymbol{y}$ is also a global minimum. Thus, we have $f(\boldsymbol{y}) \leq f(z)$ and the inequality (9) implies that

$$f(\mu z + (1 - \mu)\boldsymbol{y}) > \max\left\{f(z), f(\boldsymbol{y})\right\},$$

which is impossible by Lemma 5. □

Remark 1 Differentiability at the point $\boldsymbol{x}$ is crucial in Lemma 2. This allows us to assume that directional derivatives $f'(\boldsymbol{x}; \boldsymbol{d})$ exist at $\boldsymbol{x}$ which was needed in Lemma 6. Unlike the convexity, the f°-pseudoconvexity does not guarantee that directional derivatives exist at every point. An example of an f°-pseudoconvex function that does not have a directional derivative at every point can be found in [3, pp. 144–145].

Now we are ready to prove the converse result of Theorem 10. Note that, the similar result has been proved in [20] (Theorem 4.1) for lower semicontinuous and radially continuous functions.

Theorem 11 *The generalized directional derivative of an f°-pseudoconvex function is pseudomonotone.*

Proof Let f be f°-pseudoconvex and, on the contrary, assume that there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \geq 0$ and $f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) > 0$. Then, by f°-pseudoconvexity $f(\boldsymbol{x}) \leq f(\boldsymbol{y})$ and $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$, hence $f(\boldsymbol{x}) = f(\boldsymbol{y})$. Thus, we have $f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) > 0$ and $f(\boldsymbol{x}) = f(\boldsymbol{y})$, which contradicts Lemma 6. □

In what follows we consider how to verify the f°-pseudoconvexity in practice. That is, we analyze how the pseudoconvexity is conserved in composite-, sum- and max-functions. Before that, we need the following result.

Lemma 7 *A locally Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ is f°-pseudoconvex and strictly increasing, if and only if $\varsigma > 0$ for all $\varsigma \in \partial g(x)$ and $x \in \mathbb{R}$.*

Proof First, let g be both f°-pseudoconvex and strictly increasing and let $v < 0$. Then for every $x \in \mathbb{R}$ we have $g(x + v) < g(x)$ and due to f°-pseudoconvexity $g^\circ(x; v) < 0$. By the definition of the subdifferential we have

$$\varsigma v \leq g^\circ(x; v) < 0,$$

for all $\varsigma \in \partial g(x)$. This implies $\varsigma > 0$.

On the other hand, let all the subgradients of g be positive. We first prove that g is strictly increasing. On the contrary, suppose that there exist $y, x \in \mathbb{R}$ such that $y < x$ and $g(y) \geq g(x)$. By the Mean-Value Theorem 5 there exists $\hat{x} \in (y, x)$ such that

$$g(x) - g(y) \in \partial g(\hat{x})(x - y).$$

This means that there exists $\hat{\varsigma} \in \partial g(\hat{x})$ such that $\hat{\varsigma} > 0$ and

$$0 \geq g(x) - g(y) = \hat{\varsigma}(x - y) > 0,$$

which is impossible. Thus, g is strictly increasing.

Since g is strictly increasing we have $g(y) < g(x)$ if and only if $y < x$, where $x, y \in \mathbb{R}$. Thus, to prove f°-pseudoconvexity we need to show that $y < x$ implies $g^\circ(x; y - x) < 0$. Let $x, y \in \mathbb{R}$ be arbitrary such that $y < x$. By Theorem 1 (iii)

$$g^\circ(x; y - x) = \max\{\varsigma(y - x) \mid \varsigma \in \partial g(x)\} < 0.$$

Thus, g is f°-pseudoconvex. □

Theorem 12 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be f°-pseudoconvex and $g : \mathbb{R} \to \mathbb{R}$ be f°-pseudoconvex and strictly increasing. Then the composite function $f := g \circ h : \mathbb{R}^n \to \mathbb{R}$ is also f°-pseudoconvex.*

Proof According to Theorem 6 function f is locally Lipschitz continuous. Suppose now that $f(\boldsymbol{y}) < f(\boldsymbol{x})$. Then $g(h(\boldsymbol{y})) = f(\boldsymbol{y}) < f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$ and since g is strictly increasing we have

$$h(\boldsymbol{y}) < h(\boldsymbol{x}). \tag{10}$$

From Theorems 1 (iii) and 6 we obtain that

$$f^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) = \max\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi} \in \partial f(\boldsymbol{x})\} \leq \max\left\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi} \in \operatorname{conv}\{\partial g(h(\boldsymbol{x}))\partial h(\boldsymbol{x})\}\right\}. \tag{11}$$

Due to the definition of a convex hull the right hand side of (11) is equivalent to

$$\begin{aligned}
&\max\left\{\Big(\sum_{i=1}^m \lambda_i \varsigma_i \boldsymbol{\zeta}_i\Big)^T(\boldsymbol{y}-\boldsymbol{x}) \mid \varsigma_i \in \partial g(h(\boldsymbol{x})),\ \boldsymbol{\zeta}_i \in \partial h(\boldsymbol{x}),\ \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1\right\} \\
&\leq \max\left\{\Big(\sum_{i=1}^m \lambda_i \varsigma_i\Big) \cdot \max_{\boldsymbol{\zeta}_i \in \partial h(\boldsymbol{x})} \boldsymbol{\zeta}_i^T(\boldsymbol{y}-\boldsymbol{x}) \mid \varsigma_i \in \partial g(h(\boldsymbol{x})),\ \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1\right\} \\
&= \max\left\{\Big(\sum_{i=1}^m \lambda_i \varsigma_i\Big) h^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) \mid \varsigma_i \in \partial g(h(\boldsymbol{x})),\ \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1\right\},
\end{aligned}$$

since by Lemma 7 we have $\varsigma_i > 0$ for all $i = 1, \ldots, m$ and thus

$$\sum_{i=1}^m \lambda_i \varsigma_i > 0.$$

On the other hand, since h is f°-pseudoconvex, (10) implies that $h^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) < 0$. Then

$$f^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) \leq \max\left\{\Big(\sum_{i=1}^m \lambda_i \varsigma_i\Big) h^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) \mid \varsigma_i \in \partial g(h(\boldsymbol{x})),\ \lambda_i \geq 0,\ \sum_{i=1}^m \lambda_i = 1\right\} < 0$$

thus f is f°-pseudoconvex. □

Theorem 13 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be f°-pseudoconvex for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \max\{f_i(\boldsymbol{x}) \mid i = 1, \ldots, m\}$$

is also f°-pseudoconvex.

Proof According to Theorem 3 f is locally Lipschitz continuous. Suppose that $f(\boldsymbol{y}) < f(\boldsymbol{x})$. Define the index set

$$I(\boldsymbol{x}) := \{i \in \{1, \ldots, m\} \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}.$$

Then for all $i \in I(\boldsymbol{x})$ we have

$$f_i(\boldsymbol{y}) \leq f(\boldsymbol{y}) < f(\boldsymbol{x}) = f_i(\boldsymbol{x}). \tag{12}$$

From Theorem 1 (iii) and 3, the definition of a convex hull, f°-pseudoconvexity of f_i, and (12) we obtain that

$$\begin{aligned} f^\circ(\boldsymbol{x};\boldsymbol{y}-\boldsymbol{x}) &= \max\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi}\in\partial f(\boldsymbol{x})\} \\ &\le \max\left\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi}\in \operatorname{conv}\{\partial f_i(\boldsymbol{x})) \mid i\in I(\boldsymbol{x})\}\right\} \\ &= \max\left\{\Big(\sum_{i\in I(\boldsymbol{x})}\lambda_i\boldsymbol{\xi}_i\Big)^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi}_i\in\partial f(\boldsymbol{x}),\ \lambda_i\ge 0,\ \sum_{i\in I(\boldsymbol{x})}\lambda_i=1\right\} \\ &\le \max\left\{\sum_{i\in I(\boldsymbol{x})}\lambda_i\cdot\max_{\boldsymbol{\xi}_i\in\partial f_i(\boldsymbol{x})}\boldsymbol{\xi}_i^T(\boldsymbol{y}-\boldsymbol{x}) \mid \lambda_i\ge 0,\ \sum_{i\in I(\boldsymbol{x})}\lambda_i=1\right\} \\ &= \max\left\{\sum_{i\in I(\boldsymbol{x})}\lambda_i f_i^\circ(\boldsymbol{x};\boldsymbol{y}-\boldsymbol{x}) \mid \lambda_i\ge 0,\ \sum_{i\in I(\boldsymbol{x})}\lambda_i=1\right\} < 0. \end{aligned}$$

Thus, f is f°-pseudoconvex. □

We can say slightly more about the subdifferential of maximum of f°-pseudoconvex functions than the relation (3) in Theorem 3. Before the results we need a lemma. Recall that ray $S := \{\lambda\boldsymbol{x} \mid \boldsymbol{x}\in S, \lambda\ge 0\}$.

Lemma 8 *Let S be a convex compact set such that $\mathbf{0}\notin S$ and $\boldsymbol{z}\notin$ ray S. Then there exists $\boldsymbol{w}\in\mathbb{R}^n$ such that $\boldsymbol{w}^T\boldsymbol{z}>0$ and $\boldsymbol{w}^T\boldsymbol{x}<0$ for all $\boldsymbol{x}\in$ ray $S\setminus\{\mathbf{0}\}$.*

Proof Since $\mathbf{0}\notin S$ and S is compact, ray S is closed by Theorem 2.12 in [3, pp. 25]. By the Separation Theorem (e.g. Theorem 2.5 in [3, pp. 18]) there exists $\boldsymbol{q}\in\mathbb{R}^n$ such that

$$\boldsymbol{q}^T\boldsymbol{z}>\alpha \text{ and } \boldsymbol{q}^T\boldsymbol{x}\le\alpha, \text{ for all } \boldsymbol{x}\in\text{ray } S.$$

Since $\mathbf{0}\in$ ray S and ray S is a cone we may choose $\alpha=0$. By the Separation Theorem, when we choose sets $\{\mathbf{0}\}$ and S, there exists $\boldsymbol{p}\in\mathbb{R}^n$ such that $\boldsymbol{p}^T\boldsymbol{x}<0$ for all $\boldsymbol{x}\in S$. This holds for all $\boldsymbol{x}\in$ ray $S\setminus\{\mathbf{0}\}$. Suppose $\boldsymbol{p}^T\boldsymbol{z}\neq 0$ and let $\boldsymbol{w}=\frac{1}{2}\frac{|\boldsymbol{q}^T\boldsymbol{z}|}{|\boldsymbol{p}^T\boldsymbol{z}|}\boldsymbol{p}+\boldsymbol{q}$. Then

$$\boldsymbol{w}^T\boldsymbol{z}=\frac{1}{2}\frac{|\boldsymbol{q}^T\boldsymbol{z}|}{|\boldsymbol{p}^T\boldsymbol{z}|}\boldsymbol{p}^T\boldsymbol{z}+\boldsymbol{q}^T\boldsymbol{z}\ge\frac{1}{2}\boldsymbol{q}^T\boldsymbol{z}>0$$

and

$$\boldsymbol{w}^T\boldsymbol{x}=\frac{1}{2}\frac{|\boldsymbol{q}^T\boldsymbol{z}|}{|\boldsymbol{p}^T\boldsymbol{z}|}\boldsymbol{p}^T\boldsymbol{x}+\boldsymbol{q}^T\boldsymbol{x}<0$$

for all $\boldsymbol{x}\in$ ray $S\setminus\{\mathbf{0}\}$. If $\boldsymbol{p}^T\boldsymbol{z}=0$, we may choose $\boldsymbol{w}=\boldsymbol{p}+\boldsymbol{q}$ to obtain the result. □

Theorem 14 *Let $\boldsymbol{x} \in \mathbb{R}^n$, $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ be f°-pseudoconvex functions and $f(\boldsymbol{x}) := \max\{f_i(\boldsymbol{x}) \mid i = 1, \ldots, m\}$. If $\mathbf{0} \notin \partial f(\boldsymbol{x})$ then*

$$\partial f(\boldsymbol{x}) \subseteq \operatorname{conv}\{\partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x}), i = 1, \ldots, m\} \subset \operatorname{ray} \partial f(\boldsymbol{x}) \setminus \{\mathbf{0}\}.$$

Proof The first inclusion is stated in Theorem 3. For the second inclusion suppose, on the contrary, that there exists

$$\boldsymbol{\eta} \in \operatorname{conv}\{\partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x}), i = 1, \ldots, m\} \setminus \{\operatorname{ray} \partial f(\boldsymbol{x}) \setminus \{\mathbf{0}\}\}.$$

Then there exists i and $\boldsymbol{\xi} \in \partial f_i(\boldsymbol{x}) \setminus \{\operatorname{ray} \partial f(\boldsymbol{x}) \setminus \{\mathbf{0}\}\}$ such that $f_i(\boldsymbol{x}) = f(\boldsymbol{x})$. Note that $\boldsymbol{\xi} \neq \mathbf{0}$ since this would imply that $\boldsymbol{x}$ is a minimizer of f_i and, thus, a minimizer of f. This would imply that $\mathbf{0} \in \partial f(\boldsymbol{x})$ contradicting the assumption. Hence, $\boldsymbol{\xi} \in \partial f_i(\boldsymbol{x}) \setminus \operatorname{ray} \partial f(\boldsymbol{x})$.

By Lemma 8 there exists $\boldsymbol{d} \in \mathbb{R}^n$ such that $\boldsymbol{d}^T \boldsymbol{\xi} > 0$ and $\boldsymbol{d}^T \boldsymbol{\zeta} < 0$ for all $\boldsymbol{\zeta} \in \operatorname{ray} \partial f(\boldsymbol{x}) \setminus \{\mathbf{0}\}$. Thus, $f_i^\circ(\boldsymbol{x}; \boldsymbol{d}) > 0$ and $f^\circ(\boldsymbol{x}; \boldsymbol{d}) < 0$. This means that there exists $t_0 > 0$ such that $f(\boldsymbol{x} + t_0 \boldsymbol{d}) < f(\boldsymbol{x})$. By f°-pseudoconvexity $f_i(\boldsymbol{x} + t_0 \boldsymbol{d}) \geq f_i(\boldsymbol{x})$. Hence $f(\boldsymbol{x} + t_0 \boldsymbol{d}) < f_i(\boldsymbol{x} + t_0 \boldsymbol{d})$. This contradiction proves the theorem. □

Due to the fact that the sum of f°-pseudoconvex functions is not necessarily f°-pseudoconvex we need the following property.

Definition 6 The functions $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ are said to be *additively strictly monotone*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda_i \geq 0$, $i = 1, \ldots, m$

$$\sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{y}) < \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) \qquad \text{implies} \qquad f_i(\boldsymbol{y}) < f_i(\boldsymbol{x}).$$

Theorem 15 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be f°-pseudoconvex and additively strictly monotone, and let $\lambda_i \geq 0$ for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x})$$

is f°-pseudoconvex.

Proof According to Theorem 4 f is locally Lipschitz continuous. Suppose that $f(\boldsymbol{y}) < f(\boldsymbol{x})$. Then the additive strict monotonicity implies that for all $i = 1, \ldots, m$ we have

$$f_i(\boldsymbol{y}) < f_i(\boldsymbol{x}). \tag{13}$$

From Theorem 1 (iii) and 4, nonnegativity of λ_i, f°-pseudoconvexity of f_i, and (13) we deduce that

$$
\begin{aligned}
f^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) &= \max\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi} \in \partial f(\boldsymbol{x})\} \\
&\leq \max\{\boldsymbol{\xi}^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi} \in \sum_{i=1}^{m} \lambda_i \partial f_i(\boldsymbol{x})\} \\
&= \max\left\{\Big(\sum_{i=1}^{m} \lambda_i \boldsymbol{\xi}_i\Big)^T(\boldsymbol{y}-\boldsymbol{x}) \mid \boldsymbol{\xi}_i \in \partial f_i(\boldsymbol{x})\right\} \\
&\leq \sum_{i=1}^{m} \lambda_i \cdot \max_{\boldsymbol{\xi}_i \in \partial f_i(\boldsymbol{x})} \boldsymbol{\xi}_i^T(\boldsymbol{y}-\boldsymbol{x}) \\
&= \sum_{i=1}^{m} \lambda_i f_i^\circ(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) < 0.
\end{aligned}
$$

Thus, f is f°-pseudoconvex. □

4 Generalized Quasiconvexity

The notion of quasiconvexity is the most widely used generalization of convexity and there exist various equivalent definitions and characterizations for it. Next we recall the most commonly used definition (see e.g. [2]).

Definition 7 The function $f : \mathbb{R}^n \to \mathbb{R}$ is *quasiconvex*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}.$$

Remark 2 Lemma 5 implies that f°-pseudoconvex function is also quasiconvex.

The previous definition of quasiconvexity does not require any differentiability assumptions (cf. pseudoconvexity). Next we will give a well-known geometrical characterization to quasiconvexity.

Theorem 16 *A function f is quasiconvex, if and only if the level set $lev_\alpha f$ is a convex set for all $\alpha \in \mathbb{R}$.*

Proof Let f be quasiconvex, $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{lev}_\alpha f$, $\lambda \in [0, 1]$ and $\alpha \in \mathbb{R}$. Then

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\} \leq \max\{\alpha, \alpha\} = \alpha,$$

thus $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathrm{lev}_\alpha f$.

On the other hand, let $\mathrm{lev}_\alpha f$ be a convex set for all $\alpha \in \mathbb{R}$. By choosing $\beta := \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$ we have $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{lev}_\beta f$. The convexity of $\mathrm{lev}_\beta f$ implies, that $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathrm{lev}_\beta f$ for all $\lambda \in [0, 1]$, in other words

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \beta = \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}. \qquad \square$$

We give also a useful result concerning a finite maximum of quasiconvex functions.

Theorem 17 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be quasiconvex at $\boldsymbol{x}$ for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \max\{f_i(\boldsymbol{x}) \mid i = 1, \ldots, m\}$$

is also quasiconvex.

Proof Follows directly from the definition of quasiconvexity. $\square$

The concept of quasiconvexity has been studied by many authors (see [21] and references therein). The Dini directional derivatives were used in the characterization of quasiconvexity for radially lower semicontinuous functions in [8]. Analogously to the Definition 4 we can define the corresponding generalized concept, which is a special case of h-quasiconvexity defined in [12] when h is the Clarke generalized directional derivative.

Definition 8 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *f°-quasiconvex*, if it is locally Lipschitz continuous and for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) \qquad \text{implies} \qquad f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \leq 0.$$

Similarly to f°-pseudoconvexity, the reasoning chain may be converted.

Lemma 9 *The locally Lipschitz continuous function f is f°-quasiconvex, if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$*

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0 \qquad \textit{implies} \qquad f(\boldsymbol{y}) > f(\boldsymbol{x}).$$

Proof Follows directly from the definition of f°-quasiconvexity. $\square$

A locally Lipschitz continuous and quasiconvex function can be expressed similar to Definition 8.

Definition 9 A function $f : \mathbb{R}^n \to \mathbb{R}$ is *l-quasiconvex*, if it is locally Lipschitz continuous and for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f(\boldsymbol{y}) < f(\boldsymbol{x}) \qquad \text{implies} \qquad f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) \leq 0.$$

Remark 3 Definitions 8 and 9 imply that an f°-quasiconvex function is l-quasiconvex.

The are different definitions for f°-quasiconvexity in the literature. For instance, if the results given in [14] are converted to the case we are considering (i.e. locally Lipschitz continuous functions and Clarke subdifferential) the f°-quasiconvexity defined in [14] is the same as our l-quasiconvexity. On the other hand, in [14] the

function class we call f°-quasiconvex is called f°-protoconvex (see also [13]). These different naming possibilities arise from the fact that for a differentiable function the definitions of l-quasiconvexity and f°-quasiconvexity are equivalent as noted, for example, in e.g. [7]). Hence, the quasiconvexity can be straightforwardly generalized by using either definition.

Next, we prove that l-quasiconvexity coincides with quasiconvexity in locally Lipschitz continuous case. This result can be found in [9, 20].

Theorem 18 *If a locally Lipschitz continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex then it is l-quasiconvex.*

Proof Let f be locally Lipschitz continuous and quasiconvex. Let $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ be such that $f(\boldsymbol{z}) < f(\boldsymbol{x})$. Since local Lipschitz continuity implies continuity there exists $\varepsilon > 0$ such that $f(\boldsymbol{z} + \boldsymbol{d}) < f(\boldsymbol{x} + \boldsymbol{d})$ for all $\boldsymbol{d} \in B(\boldsymbol{0}; \varepsilon)$. For generalized directional derivative $f^\circ(\boldsymbol{x}; \boldsymbol{z} - \boldsymbol{x})$ we have

$$\begin{aligned} f^\circ(\boldsymbol{x}; \boldsymbol{z} - \boldsymbol{x}) &= \limsup_{\substack{\boldsymbol{y} \to \boldsymbol{x} \\ t \downarrow 0}} \frac{f(\boldsymbol{y} + t(\boldsymbol{z} - \boldsymbol{x})) - f(\boldsymbol{y})}{t} \\ &= \limsup_{\substack{\boldsymbol{y} \to \boldsymbol{x} \\ t \downarrow 0}} \frac{f(\boldsymbol{y} + t(\boldsymbol{z} - \boldsymbol{x} + \boldsymbol{y} - \boldsymbol{y})) - f(\boldsymbol{y})}{t} \\ &= \limsup_{\substack{\boldsymbol{y} \to \boldsymbol{x} \\ t \downarrow 0}} \frac{f((1-t)\boldsymbol{y} + t(\boldsymbol{z} + \boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{y})}{t} \end{aligned}$$

When $t \in (0, 1)$ and $\boldsymbol{y} - \boldsymbol{x} \in B(\boldsymbol{0}; \varepsilon)$ the quasiconvexity of f implies

$$\begin{aligned} &\frac{f((1-t)\boldsymbol{y} + t(\boldsymbol{z} + \boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{y})}{t} \\ \leq\ &\frac{\max\{f(\boldsymbol{y}), f(\boldsymbol{z} + \boldsymbol{y} - \boldsymbol{x})\} - f(\boldsymbol{y} - \boldsymbol{x} + \boldsymbol{x})}{t} \\ =\ &\frac{\max\{0, f(\boldsymbol{z} + \boldsymbol{y} - \boldsymbol{x}) - f(\boldsymbol{x} + \boldsymbol{y} - \boldsymbol{x})\}}{t} = 0 \end{aligned}$$

Passing to the limit $t \to 0$ and $\boldsymbol{y} \to \boldsymbol{x}$ we get $f^\circ(\boldsymbol{x}; \boldsymbol{z} - \boldsymbol{x}) \leq 0$. Thus, f is l-quasiconvex. □

Theorem 19 *If function $f : \mathbb{R}^n \to \mathbb{R}$ is l-quasiconvex then it is quasiconvex.*

Proof On the contrary assume that an l-quasiconvex function f is not quasiconvex. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\bar{\lambda} \in (0, 1)$ such that $f(\bar{\boldsymbol{x}}) > \max\{f(\boldsymbol{x}), f(\boldsymbol{y})\}$, where $\bar{\boldsymbol{x}} = \bar{\lambda}\boldsymbol{x} + (1 - \bar{\lambda})\boldsymbol{y}$. Without a loss of generality we may assume that $f(\boldsymbol{x}) \geq f(\boldsymbol{y})$. Lemma 3 implies that there exists $\tilde{\boldsymbol{x}} \in (\boldsymbol{x}, \bar{\boldsymbol{x}})$, for which

$$f(\tilde{\boldsymbol{x}}) > f(\boldsymbol{x}) \quad \text{and} \quad f^\circ(\tilde{\boldsymbol{x}}; \bar{\boldsymbol{x}} - \boldsymbol{x}) > 0.$$

Denote $\tilde{\boldsymbol{x}} = \tilde{\lambda}\boldsymbol{x} + (1-\tilde{\lambda})\boldsymbol{y}$, where $\tilde{\lambda} \in (\bar{\lambda}, 1)$. From the definitions of points $\bar{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}$ we see that

$$\bar{\boldsymbol{x}} - \boldsymbol{x} = (1-\bar{\lambda})(\boldsymbol{y} - \boldsymbol{x}) \quad \text{and} \quad \boldsymbol{y} - \tilde{\boldsymbol{x}} = \tilde{\lambda}(\boldsymbol{y} - \boldsymbol{x})$$

Thus,

$$\bar{\boldsymbol{x}} - \boldsymbol{x} = \frac{1-\bar{\lambda}}{\tilde{\lambda}}(\boldsymbol{y} - \tilde{\boldsymbol{x}})$$

and

$$0 < f^\circ(\tilde{\boldsymbol{x}}; \bar{\boldsymbol{x}} - \boldsymbol{x}) = \frac{1-\bar{\lambda}}{\tilde{\lambda}} f^\circ(\tilde{\boldsymbol{x}}; \boldsymbol{y} - \tilde{\boldsymbol{x}}).$$

Thus, $0 < f^\circ(\tilde{\boldsymbol{x}}; \boldsymbol{y} - \tilde{\boldsymbol{x}})$ and $f(\tilde{\boldsymbol{x}}) > f(\boldsymbol{x}) \geq f(\boldsymbol{y})$ which contradicts the l-quasiconvexity of function f. Hence, f is quasiconvex. □

Corollary 1 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and quasiconvex if and only if it is l-quasiconvex.*

Proof The result follows directly from Theorems 18 and 19. □

Corollary 2 *If f is f°-quasiconvex, then f is quasiconvex.*

Proof The result follows from Remark 3 and Theorem 19. □

Similar to the pseudomonotonicity there exists a concept of quasimonotonicity (see [12, 20]).

Definition 10 The generalized directional derivative f° is called *quasimonotone*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0 \qquad \text{implies} \qquad f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) \leq 0$$

or, equivalently

$$\min\{f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}), f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y})\} \leq 0.$$

Note that analogously to the pseudomonotonicity (see Definition 5) we could define also the strict quasimonotonicity, but it would be equivalent to the pseudomonotonicity.

It turns out that the generalized directional derivative f° of the function $f : \mathbb{R}^n \to \mathbb{R}$ is quasimonotone if and only if the function is locally Lipschitz continuous and quasiconvex. The following two results have been proven for lower semicontinuous functions in [15]. The proofs use subdifferential that is applicable for l.s.c. functions and they are more complex than the ones given below.

Theorem 20 *If f° is quasimonotone, then f is quasiconvex.*

Proof Let us, on the contrary assume, that f is not quasiconvex. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\bar{\lambda} \in (0, 1)$ such that

$$f(\bar{\boldsymbol{x}}) > f(\boldsymbol{x}) \geq f(\boldsymbol{y}),$$

where $\bar{\boldsymbol{x}} = \boldsymbol{x} + \bar{\lambda}(\boldsymbol{y} - \boldsymbol{x})$. Then by the Mean-Value Theorem 5 there exist $\hat{\boldsymbol{x}}, \tilde{\boldsymbol{x}} \in \mathbb{R}^n$ such that

$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{y}) \in \partial f(\hat{\boldsymbol{x}})^T (\bar{\boldsymbol{x}} - \boldsymbol{y})$$

and

$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}) \in \partial f(\tilde{\boldsymbol{x}})^T (\bar{\boldsymbol{x}} - \boldsymbol{x}),$$

where

$$\hat{\boldsymbol{x}} = \boldsymbol{x} + \hat{\lambda}(\boldsymbol{y} - \boldsymbol{x}), \quad \tilde{\boldsymbol{x}} = \boldsymbol{x} + \tilde{\lambda}(\boldsymbol{y} - \boldsymbol{x}), \quad 0 < \tilde{\lambda} < \bar{\lambda} < \hat{\lambda} < 1.$$

This means that, due to the definition of the Clarke subdifferential, there exist $\hat{\boldsymbol{\xi}} \in \partial f(\hat{\boldsymbol{x}})$ and $\tilde{\boldsymbol{\xi}} \in \partial f(\tilde{\boldsymbol{x}})$ such that

$$0 < f(\bar{\boldsymbol{x}}) - f(\boldsymbol{y}) = \hat{\boldsymbol{\xi}}^T (\bar{\boldsymbol{x}} - \boldsymbol{y}) \leq f^\circ(\hat{\boldsymbol{x}}; \bar{\boldsymbol{x}} - \boldsymbol{y}) = (1 - \bar{\lambda}) f^\circ(\hat{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y})$$

and

$$0 < f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}) = \tilde{\boldsymbol{\xi}}^T (\bar{\boldsymbol{x}} - \boldsymbol{x}) \leq f^\circ(\tilde{\boldsymbol{x}}; \bar{\boldsymbol{x}} - \boldsymbol{x}) = \bar{\lambda} f^\circ(\tilde{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x})$$

by the positive homogeneity of $\boldsymbol{d} \mapsto f^\circ(\boldsymbol{x}; \boldsymbol{d})$ (see Theorem 1 (i)). Then we deduce that

$$f^\circ(\hat{\boldsymbol{x}}; \tilde{\boldsymbol{x}} - \hat{\boldsymbol{x}}) = (\hat{\lambda} - \tilde{\lambda}) f^\circ(\hat{\boldsymbol{x}}; \boldsymbol{x} - \boldsymbol{y}) > 0$$

and

$$f^\circ(\tilde{\boldsymbol{x}}; \hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}) = (\hat{\lambda} - \tilde{\lambda}) f^\circ(\tilde{\boldsymbol{x}}; \boldsymbol{y} - \boldsymbol{x}) > 0,$$

which contradicts the quasimonotonicity. Thus, f is quasiconvex. □

Theorem 21 *If function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous and quasiconvex then the generalized directional derivative f° is quasimonotone.*

Proof On the contrary, assume that f° is not quasimonotone. Then there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0$ and $f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y}) > 0$ Let

$$\delta = \min\left\{f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}),\ f^\circ(\boldsymbol{y}; \boldsymbol{x} - \boldsymbol{y})\right\}.$$

Let $\varepsilon_1 > 0$ be such that the local Lipschitz condition holds in the ball $B(\boldsymbol{x}; \varepsilon_1)$ with Lipschitz constant K_1. Correspondingly, let $\varepsilon_2 > 0$ be such that the local Lipschitz condition holds in the ball $B(\boldsymbol{y}; \varepsilon_2)$ with Lipschitz constant K_2. Let $K = \max\{K_1, K_2\}$ and $\varepsilon = \min\left\{\frac{\delta}{4K}, \varepsilon_1, \varepsilon_2\right\}$. According to Lemma 1 there exists a sequence (z_1^i), such that f is differentiable, $\lim_{i\to\infty} z_1^i = \boldsymbol{x}$ and an index $I \in \mathbb{N}$ such that

$$f'(z_1^i; \boldsymbol{y} - \boldsymbol{x}) = \nabla f(z_1^i)^T(\boldsymbol{y} - \boldsymbol{x}) \geq \frac{\delta}{2}$$

when $i \geq I$. Similarly, there exists a sequence (z_2^j), such that f is differentiable, $\lim_{j\to\infty} z_2^j = \boldsymbol{y}$ and an index $J \in \mathbb{N}$ such that

$$f'(z_2^j; \boldsymbol{x} - \boldsymbol{y}) = \nabla f(z_2^j)^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\delta}{2}$$

when $j \geq J$. Let $z_1 \in B(\boldsymbol{x}; \varepsilon) \cap \left\{(z_1^i) \mid i \geq I\right\}$ and $z_2 \in B(\boldsymbol{y}; \varepsilon) \cap \left\{(z_2^j) \mid j \geq J\right\}$. Due to symmetry we may assume that $f(z_1) \geq f(z_2)$ without a loss of generality. According to Lemma 2

$$\begin{aligned}&\left|f'(z_1; z_2 - z_1) - f'(z_1; \boldsymbol{y} - \boldsymbol{x})\right| \leq K\left\|z_2 - z_1 - (\boldsymbol{y} - \boldsymbol{x})\right\| \\ &\leq K\left\|\boldsymbol{x} - z_1\right\| + K\left\|z_2 - \boldsymbol{y}\right\| < 2K\frac{\delta}{4K} = \frac{\delta}{2}.\end{aligned}$$

Since $f'(z_1; \boldsymbol{y} - \boldsymbol{x}) > \frac{\delta}{2}$ also $f'(z_1; z_2 - z_1) > 0$. Thus, there exists $\lambda \in (0, 1)$ such that

$$f(z_1 + \lambda(z_2 - z_1)) > f(z_1) \geq f(z_2),$$

which contradicts the quasiconvexity. □

Corollary 3 *A function f is l-quasiconvex if and only if f° is quasimonotone.*

Proof The result follows from Corollary 1 and Theorems 20 and 21. □

Corollary 4 *If f is f°-quasiconvex, then f° is quasimonotone.*

Proof The results follows from Remark 3 and Corollary 3. □

By Corollary 2 f°-quasiconvex function is quasiconvex. The next result shows, that for a subdifferentially regular function quasiconvexity and f°-quasiconvexity coincides.

Theorem 22 *If f is both quasiconvex and subdifferentially regular, then f is f°-quasiconvex.*

Proof Due to the subdifferential regularity f is locally Lipschitz continuous. Suppose, that $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$. Then the subdifferential regularity and quasiconvexity implies, that

$$f^{\circ}(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) = f'(\boldsymbol{x};\, \boldsymbol{y}-\boldsymbol{x}) = \lim_{t \downarrow 0} \frac{f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})) - f(\boldsymbol{x})}{t}$$

$$= \lim_{t \downarrow 0} \frac{f(t\boldsymbol{y}+(1-t)\boldsymbol{x}) - f(\boldsymbol{x})}{t} \leq \lim_{t \downarrow 0} \frac{f(\boldsymbol{x}) - f(\boldsymbol{x})}{t} = 0$$

in other words, f is f°-quasiconvex. □

Corollary 5 *A subdifferentially regular l-quasiconvex function is f°-quasiconvex.*

Proof The result follows from Corollary 1 and Theorem 22. □

Corollary 6 *A subdifferentially regular function f with quasimonotone f° is f°-quasiconvex.*

Proof The result follows from Corollaries 3 and 5. □

In Theorem 22 the subdifferential regularity cannot be omitted, which can be seen, for instance, by considering the following example:

Example 1 Define $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) := \begin{cases} |x|, & x \in (-\infty, 1) \\ 1, & x \in [1, 2] \\ x-1, & x \in (2, \infty). \end{cases}$$

Then f is clearly locally Lipschitz continuous and quasiconvex. However, by taking $x := 1$ and $y := 2$ we have $f^{\circ}(x;\, y-x) = f^{\circ}(1;\, 1) = 1 > 0$, but $f(y) = f(2) = 1 \not> 1 = f(1) = f(x)$ and thus, due to Lemma 9, f is not f°-quasiconvex. Note that f is not subdifferentially regular since $f'(1;\, 1) = 0 \neq 1 = f^{\circ}(1;\, 1)$. Furthermore, f is not f°-pseudoconvex, since $0 \in \partial f(1) = [0, 1]$ although $x = 1$ is not a global minimum (cf. Theorem 9).

As stated in Corollary 5 the subdifferential regularity ensures that the l-quasiconvexity implies f°-quasiconvexity. In [12] it was show that also the following nonconstancy property has the similar consequence.

Definition 11 A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy *nonconstancy property* (in short, NC-property), if there exists no line segment $[a, b]$ along which f is constant.

Definition 11 is somewhat restrictive as, for example, $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x$ does not satisfy it. This leads us to generalize the concept to be more applicable to this kind of functions.

Definition 12 A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy *generalized nonconstancy property* (in short, GNC-property) if there does not exist $\boldsymbol{x} \in \mathbb{R}^n$ and $\varepsilon > 0$ such that $f(\boldsymbol{y}) = f(\boldsymbol{x})$ for all $\boldsymbol{y} \in B(\boldsymbol{x}; \varepsilon)$.

If $f : \mathbb{R}^n \to \mathbb{R}$ satisfies NC-property it satisfies GNC-property. The function $f(x, y) = x$ shows that converse is not true. For $\mathbb{R} \to \mathbb{R}$ functions these two concepts coincide.

It is good to note that the subdifferential regularity and the GNC-property are two separate concepts. An example of function which is subdifferentially regular but does not satisfy the GNC-property is

$$g_1(x) = \begin{cases} (x+1)^2 & , \text{ if } x \leq -1 \\ 0 & , \text{ if } -1 \leq x \leq 1 \\ (x-1)^2 & , \text{ if } x \geq 1 \end{cases}.$$

On the other hand, the function

$$g_2(x) = \begin{cases} 2x & , \text{ if } x \leq 0 \\ \frac{1}{2}x & , \text{ if } x \geq 0 \end{cases}$$

possesses the GNC-property but it is not subdifferentially regular since $g_2^\circ(0; 1) = 2 \neq \frac{1}{2} = g_2'(0; 1)$.

Next we will show that l-quasiconvex function satisfying GNC-property is f°-quasiconvex.

Theorem 23 *Let f be an l-quasiconvex function that possesses the GNC-property. Then it is f°-quasiconvex.*

Proof If f is not f° -quasiconvex then there exist $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) > 0$ but $f(\boldsymbol{x}) = f(\boldsymbol{y})$.

Suppose that $f(\boldsymbol{x})$ is the global minimum value of f. By the similar deductions used in Lemma 6 there exist $\boldsymbol{z} \in \mathbb{R}^n$ and $0 < \mu < 1$ such that

$$f(\mu \boldsymbol{z} + (1-\mu)\boldsymbol{y}) > f(\boldsymbol{z}). \tag{14}$$

Since $f(\boldsymbol{x}) = f(\boldsymbol{y})$ is the global minimum value we also have $f(\boldsymbol{z}) \geq f(\boldsymbol{y})$. This and (14) contradicts with the definition of quasiconvexity.

Suppose then that $f(\boldsymbol{x})$ is not the global minimum value of f. By continuity of f° as a function of $\boldsymbol{d}$ there exists $\varepsilon > 0$ such that $f^\circ(\boldsymbol{x}; \boldsymbol{d}) > 0$ for all $\boldsymbol{d} \in B(\boldsymbol{y} - \boldsymbol{x}; \varepsilon)$. Then, by the l-quasiconvexity of f there exists $\delta > 0$ such that $f(\overline{\boldsymbol{y}}) \geq f(\boldsymbol{x}) = f(\boldsymbol{y})$ for all $\overline{\boldsymbol{y}} \in B(\boldsymbol{y}; \delta)$. Since there exists $\boldsymbol{z}$ such that $f(\boldsymbol{z}) < f(\boldsymbol{x})$, the continuity of f implies there exists $\gamma > 0$ such that $f(\overline{\boldsymbol{z}}) < f(\boldsymbol{x})$ for all $\overline{\boldsymbol{z}} \in B(\boldsymbol{z}; \gamma)$. Furthermore, there exist $\boldsymbol{w} \in \mathbb{R}^n$ and $r > 0$ such that

$$B(\boldsymbol{w}; r) \subset B(\boldsymbol{y}; \delta) \cap \operatorname{conv}\{\{\boldsymbol{y}\}, B(\boldsymbol{z}; \gamma)\}. \tag{15}$$

Since $B(\boldsymbol{w}; r) \subset B(\boldsymbol{y}; \delta)$, we have $f(\overline{\boldsymbol{w}}) \geq f(\boldsymbol{y})$ for all $\overline{\boldsymbol{w}} \in B(\boldsymbol{w}; r)$. By inclusion (15), for any $\overline{\boldsymbol{w}} \in B(\boldsymbol{w}; r)$ there exist $\overline{\boldsymbol{z}} \in B(\boldsymbol{z}; \gamma)$ and $0 < \lambda < 1$ such that $\overline{\boldsymbol{w}} = \lambda \boldsymbol{y} + (1 - \lambda)\overline{\boldsymbol{z}}$. By the quasiconvexity of f we have $f(\overline{\boldsymbol{w}}) \leq \max\{f(\boldsymbol{y}), f(\overline{\boldsymbol{z}})\} = f(\boldsymbol{y})$. Hence, $f(\overline{\boldsymbol{w}}) = f(\boldsymbol{y})$ for all $\overline{\boldsymbol{w}} \in B(\boldsymbol{w}; r)$ contradicting the GNC-property assumption. □

Corollary 7 *If f° is quasimonotone and f possesses the GNC-property, then f is f°-quasiconvex.*

Proof The result follows from Corollary 3 and Theorem 23. □

Example 2 Consider the following function

$$f(x, y) = \begin{cases} \max\{x, y\}, & x \leq 0, y \leq 0 \\ y^2, & x \leq 0, y \geq 0 \\ x^2, & x \geq 0, y \leq 0 \\ y^2 + x^2, & x \geq, y \geq 0 \end{cases}.$$

The function f is l-quasiconvex as it is locally Lipschitz continuous and the level sets are convex. It does not possess the NC-property since $f(x, y) = 0$ for all $(x, y) \in [(0, -1), (0, 0)]$. Furthermore, f is not subdifferentially regular since $f'((0, 0); (1, 0)) = 0$, but $f^\circ((0, 0); (1, 0)) = 1$. However, the GNC-property holds and consequently f is f°-quasiconvex.

Example 3 Consider the function in Example 1. Its generalized directional derivative is quasimonotone since the function is quasiconvex and locally Lipschitz continuous. However, the function does not satisfy the (generalized) nonconstancy property and, thus, it is not guaranteed to be f°-quasiconvex. As shown in Example 1 the function is not f°-quasiconvex.

The next results concerning the verification of the f°-quasiconvexity are, analogous to those of f°-pseudoconvexity.

Lemma 10 *A locally Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ is increasing, if and only if $\varsigma \geq 0$ for all $\varsigma \in \partial g(x)$ and $x \in \mathbb{R}$.*

Proof The proof is almost similar to that of Lemma 7 by changing the symbols $<$ and $\leq$. □

Theorem 24 *Let $h : \mathbb{R}^n \to \mathbb{R}$ be f°-quasiconvex and $g : \mathbb{R} \to \mathbb{R}$ locally Lipschitz continuous and increasing. Then the composite function $f := g \circ h : \mathbb{R}^n \to \mathbb{R}$ is also f°-quasiconvex.*

Proof The proof is similar to that of Theorem 12. □

Theorem 25 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be f°-quasiconvex for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \max\{f_i(\boldsymbol{x}) \mid i = 1, \ldots, m\}$$

is also f°-quasiconvex.

Proof The proof is similar to that of Theorem 13. □

As in the case of f°-pseudoconvexity, the following property guarantees that the sum of f°-quasiconvex functions is also f°-quasiconvex.

Definition 13 The functions $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ are said to be *additively monotone*, if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\lambda_i \geq 0$, $i = 1, \ldots, m$

$$\sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{y}) \leq \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) \qquad \text{implies} \qquad f_i(\boldsymbol{y}) \leq f_i(\boldsymbol{x}).$$

Theorem 26 *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be f°-quasiconvex and additively monotone, and $\lambda_i \geq 0$ for all $i = 1, \ldots, m$. Then the function*

$$f(\boldsymbol{x}) := \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x})$$

is f°-quasiconvex.

Proof The proof is similar to that of Theorem 15. □

Finally we study the relations between pseudo- and quasiconvexity. According to [2] for differentiable functions pseudoconvexity implies quasiconvexity. Also, It turns out that f°-pseudoconvexity implies f°-quasiconvexity. A proof for quasiconvex, not necessarily locally Lipschitz continuous, functions can be found in [13].

Theorem 27 *An f°-pseudoconvex function is f°-quasiconvex.*

Proof On the contrary, assume that an f°-pseudoconvex function f is not f°-quasiconvex. Then, there exist points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $f^\circ(\boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x}) > 0$ and $f(\boldsymbol{x}) = f(\boldsymbol{y})$. According to Lemma 6 this is impossible for f°-pseudoconvex function. Thus, f is f°-quasiconvex. □

Corollary 8 *If f° is pseudomonotone then it is also quasimonotone.*

Proof The result follows from Corollary 4 and Theorems 10 and 27. □

The next example shows that the result in Theorem 27 cannot be converted.

Example 4 Define $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) := x^3$. Clearly f is quasiconvex and as a smooth function also subdifferentially regular. Thus, by Theorem 22 it is

f°-quasiconvex. However, by taking $x := 0$ and $y := -1$ we have $f^\circ(x; y - x) = f^\circ(0; -1) = 0$, but $f(y) = f(-1) = -1 \not\geq 0 = f(0) = f(x)$ and thus, due to Lemma 4, f is not f°-pseudoconvex.

If a quasiconvex function f is continuously differentiable and $\nabla f(\boldsymbol{x}) = 0$ only if $\boldsymbol{x}$ is a global minimum, f is pseudoconvex. Similar result holds for generalized convexities, which was proved for radially upper semicontinuous functions in [14]. The proof given below is quite similar to that in [14].

Lemma 11 *Let f be l-quasiconvex and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. If $f(\boldsymbol{y}) < f(\boldsymbol{x})$ and $\mathbf{0} \notin \partial f(\boldsymbol{x})$ then $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) < 0$.*

Proof Suppose $f(\boldsymbol{y}) < f(\boldsymbol{x})$ and $\mathbf{0} \notin \partial f(\boldsymbol{x})$. By continuity of f there exists $r > 0$ such that $f(\boldsymbol{z}) < f(\boldsymbol{x})$ for all $\boldsymbol{z} \in \overline{B}(\boldsymbol{y}; r)$.

Let $\boldsymbol{\xi} \in \partial f(\boldsymbol{x})$ be arbitrary. Since $\mathbf{0} \notin \partial f(\boldsymbol{x})$ we may define $\hat{\boldsymbol{y}} = \boldsymbol{y} + \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|} r$. The l-quasiconvexity of f and the inequality $f(\hat{\boldsymbol{y}}) < f(\boldsymbol{x})$ imply that

$$f^\circ(\boldsymbol{x}; \hat{\boldsymbol{y}} - \boldsymbol{x}) \leq 0. \tag{16}$$

By the basic properties of the Clarke generalized directional derivative, inequality (16) implies $\boldsymbol{\xi}^T(\hat{\boldsymbol{y}} - \boldsymbol{x}) \leq 0$. Thus,

$$\boldsymbol{\xi}^T(\boldsymbol{y} - \boldsymbol{x}) = \boldsymbol{\xi}^T(\hat{\boldsymbol{y}} - \frac{\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|} r - \boldsymbol{x}) = -r\,\|\boldsymbol{\xi}\| + \boldsymbol{\xi}^T(\hat{\boldsymbol{y}} - \boldsymbol{x}) \leq -r\,\|\boldsymbol{\xi}\|\,.$$

Since $\mathbf{0} \notin \partial f(\boldsymbol{x})$ we have $-r\,\|\boldsymbol{\xi}\| < 0$. Thus,

$$f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) = \max_{\boldsymbol{\zeta} \in \partial f(\boldsymbol{x})} \boldsymbol{\zeta}^T(\boldsymbol{y} - \boldsymbol{x}) \leq \max_{\boldsymbol{\zeta} \in \partial f(\boldsymbol{x})} -r\,\|\boldsymbol{\zeta}\| < 0,$$

proving the lemma. □

Theorem 28 *If f is l-quasiconvex and $\mathbf{0} \in \partial f(\boldsymbol{x})$ implies $\boldsymbol{x} \in \mathbb{R}^n$ is a global minimum point, then f is f°-pseudoconvex.*

Proof Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be such that $f(\boldsymbol{x}) > f(\boldsymbol{y})$. By assumption $\mathbf{0} \notin \partial f(\boldsymbol{x})$. Then, Lemma 11 implies $f^\circ(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x}) < 0$. □

5 Concluding Remarks

To the end we summarize in Fig. 1 all the considered relationships.

Acknowledgements We want to thank the Academy of Finland (Project No. 289500 and 294002).

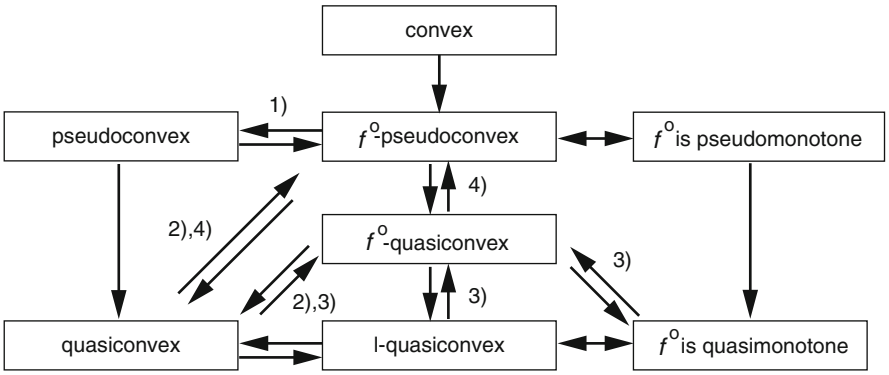


Fig. 1 The relations between considered function classes. 1) demands continuous differentiability. 2) demands local Lipschitz continuity. 3) demands GNC-property or subdifferential regularity. 4) demands that $\mathbf{0} \in \partial f(\boldsymbol{x})$ implies $\boldsymbol{x}$ is a global minimizer

References

1. D. Aussel, Subdifferential properties of quasiconvex functions: unified approach. J. Optim. Theory Appl. **97**, 29–45 (1998)
2. M. Avriel, W.E. Diewert, S. Schaible, I. Zang, *Generalized Concavity* (Plenum Press, New York, 1988)
3. A. Bagirov, M.M. Mäkelä, N. Karmitsa, *Introduction to Nonsmooth Optimization: Theory, Practice and Software* (Springer International Publishing, Cham, Heidelberg, 2014)
4. D. Bhatia, P. Jain, Generalized (F,ρ)-convexity and duality for non smooth multi-objective programs. Optimization **31**, 153–164 (1994)
5. J. Borde, J.P. Crouzeix, Continuity properties of the normal cone to the level sets of a quasiconvex function. J. Optim. Theory Appl. **66**, 415–429 (1990)
6. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley-Interscience, New York, 1983)
7. J.-P. Crouzeix, Criteria for generalized convexity and generalized monotonicity in the differentiable case, in *Handbook of Generalized Convexity and Monotonicity*, ed. by N. Hadjisavvas, S. Komlósi, S. Schaible (Springer Science+Business Media, Boston, 2005), pp. 89–119
8. W.E. Diewert, Alternative characterizations of six kinds of quasiconcavity in the nondifferentiable case with applications to nonsmooth programming, in *Generalized Concavity in Optimization and Economics*, ed. by S. Schaible, W.T. Ziemba (Academic Press, New York, 1981), pp. 51–95
9. R. Ellaia, A. Hassouni, Characterization of nonsmooth functions through their generalized gradients. Optimization **22**, 401–416 (1991)
10. J.B. Hiriart-Urruty, New concepts in nondifferentiable programming. Bull. Soc. Math. France, Mémoires **60**, 57–85 (1979)
11. S. Komlósi, Some properties of nondifferentiable pseudoconvex functions. Math. Program. **26**, 232–237 (1983)
12. S. Komlósi, Generalized monotonicity and generalized convexity. J. Optim. Theory Appl. **84**, 361–376 (1995)
13. T. Kugendran, Protoconvex functions. Optimization **44**(1), 1–24 (1998)
14. N.T.H. Linh, J.-P. Penot, Generalized convex functions and generalized differentials. Optimization **62**, 943–959 (2013)
15. D.T. Luc, Characterisation of quasiconvex functions. Bull. Aust. Math. Soc. **48**, 393–406 (1993)

16. M.M. Mäkelä, P. Neittaanmäki, *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control* (World Scientific Publishing, Singapore, 1992)
17. O.L. Mangasarian, Pseudoconvex functions. SIAM J. Control **3**, 281–290 (1965)
18. S. Nobakhtian, Infine functions and nonsmooth multiobjective optimization problems. Comput. Math. Appl. **51**, 1385–1394 (2006)
19. S. Nobakhtian, Multiobjective problems with nonsmooth equality constraints. Numer. Funct. Anal. Optim. **30**, 337–351 (2009)
20. J.-P. Penot, P.H. Quang, Generalized convexity of functions and generalized monotonicity of set-valued maps. J. Optim. Theory Appl. **92**, 343–356 (1997)
21. R. Pini, C. Singh, A survey of recent [1985–1995] advances in generalized convexity with applications to duality theory and optimality conditions. Optimization **39**, 311–360 (1997)
22. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
23. S. Schaible, Generalized monotone maps, in *Nonsmooth Optimization: Methods and Applications*, ed. by F. Giannessi (Breach Science Publishers, Amsterdam, 1992), pp. 392–408
24. T. Staib, Necessary optimality conditions for nonsmooth multicriteria optimization problem. SIAM J. Optim. **2**, 153–171 (1992)
25. X.M. Yang, S.Y. Liu, Three kinds of generalized convexity. J. Optim. Theory Appl. **86**, 501–513 (1995)

On the Equivalence of Eulerian and Lagrangian Variables for the Two-Component Camassa–Holm System



Markus Grasmair, Katrin Grunert, and Helge Holden

Dedicated to Haim Brezis and Louis Nirenberg in deep admiration.

1 Introduction

The prevalent way to analyze the ubiquitous wave breaking for the Camassa–Holm (CH) equation, is to transform the original equation from its Eulerian variables into a new coordinate system, e.g. in Lagrangian variables. The reason for the transformation is that while the solution develops singularities in Eulerian coordinates, the solution remains smooth in the Lagrangian framework. This invites the question of a closer analysis of the transformation between the Eulerian and the Lagrangian variables. That is the goal of the present paper.

A two-component generalization of the CH equation was introduced in [30, Eq. (43)], and we will study the above question in this setting. It turns out that this system, denoted the two-component Camassa–Holm (2CH) system, has a regularizing effect on the original CH equation as long as the density ρ remains positive. To set the stage, we recall that the 2CH system can be written as

$$u_t + uu_x + P_x = 0, \tag{1a}$$

$$\rho_t + (u\rho)_x = 0, \tag{1b}$$

where P is implicitly defined by

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

M. Grasmair · K. Grunert · H. Holden (✉)
Department of Mathematical Sciences, NTNU Norwegian University of Science and Technology, Trondheim, Norway
e-mail: markus.grasmair@ntnu.no; katrin.grunert@ntnu.no; helge.holden@ntnu.no

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_7

The original CH equation [4, 5] is the special case where ρ vanishes identically. The CH equation possesses many intriguing properties, and the main challenge when one considers the Cauchy problem, is that the solution develops singularities in finite time, independent of the smoothness initially. This singularity is characterized by the H^1-norm of the function u remaining finite, while the spatial derivative u_x goes to negative infinity at a specific point at the time of wave breaking. The structure of the points of wave breaking may be intricate [13]. The behavior in the proximity of the point of wave breaking, and, in particular, the prolongation of the solution past wave breaking, has been extensively studied. See, e.g., [2, 3, 6–8, 10, 12, 14–25, 27–29] and references therein. The key point here is that past wave breaking uniqueness fails, and there is a continuum of distinct solutions [19], with two extreme solutions called dissipative and conservation solutions, respectively. The various solutions can be characterized by the behavior of the total energy, as measured by the local H^1 density of the solution u. As mentioned above, the density ρ has a regularizing effect on the solution: If ρ is positive on the line initially, then the solution will not develop singularities [9, 16]. A local result, saying that if ρ initially is smooth on an interval, then the solution will remain smooth on the interval determined by the characteristics emanating from the endpoints of the original interval, can be found in [16, Thm. 6.1]. This is surprising, as the 2CH system has infinite speed of propagation [26].

In this paper we study in detail the relation between the Eulerian and the Lagrangian variables, and, in particular, the stability of solutions in the two coordinate systems. Two aspects are considered. First one may ask if the solution of the 2CH system will converge to a solution of the CH equation in the limit when the density ρ vanishes, and if so, to which of the plethora of solutions. This problem has also been studied in [16]. We show that the limit is the so-called conservative solution of the CH equation where the energy is preserved, see Theorem 7. The second question addresses the relation between stability in Eulerian variables and stability in Lagrangian variables in general. The short answer is that the two notions are equivalent. This result can hardly be considered surprising. However, as each of the norms for the variables is rather intricate, and the relation between them is highly nonlinear, the actual proofs are considerably more technical than we expected. In part, this is due to the fact that the solution does develop singularities in Eulerian coordinates, while it remains smooth in the Lagrangian framework. We have chosen to give rather detailed proofs, as we find that eases the understanding. Each proof is broken down into shorter technical arguments for the benefit of the reader.

Let us describe more precisely the content of this paper. A key role is played by the non-negative Radon measure μ with absolutely continuous part $\mu_{\mathrm{ac}} = (u_x^2 + \bar{\rho}^2)dx$. Here $\rho - \bar{\rho}$ is a real constant, and $\bar{\rho}$ is square integrable. The dynamics between the singular and absolutely continuous part of the measure encode the wave breaking. In Sect. 2 we consider the Cauchy problem for the CH equation with initial data (u, μ). We mollify these data to obtain a sequence (u_n, ρ_n, μ_n) with positive density ρ_n. The main result in this section, Theorem 1, shows that indeed $u_n \to u$ in H^1 while $\rho_n \rightharpoonup 0$, and $\mu_n((-\infty, x]) \to \mu((-\infty, x])$ at points of continuity of the limit. In Theorem 7 we prove that the same result applies to the solution of the initial

value problem. More specifically, we show (in obvious notation) that the solution $(u_n(t), \rho_n(t), \mu_n(t))$ of (1) with initial data $(u_{0,n}, \rho_{0,n}, \mu_{0,n})$ will converge to the conservative solution $(u(t), \mu(t))$ with initial data (u_0, μ_0). In Sect. 3 we study how this approximation by a mollification procedure carries over in Lagrangian coordinates. To detail this, we first need to recall the transformation between Eulerian and Lagrangian variables. We are given the pair of functions $(u, \rho) \in H^1 \times L^2$ (Eulerian variables). For simplicity we let $\rho = \bar{\rho}$. In addition, we need the energy density in the form of a positive Radon measure μ, that was introduced above, such that the absolutely continuous part equals $\mu_{\text{ac}} = (u_x^2 + \rho^2)\, dx$. The characteristic is given by $y(\xi) = \sup\{y \mid \mu((-\infty, y)) + y < \xi\}$. The Lagrangian velocity, energy density, and density read $U = u(y)$, $h = 1 - y_\xi$, and $r = \rho(y) y_\xi$, respectively. The full set of Lagrangian variables is then $X = (y, U, h, r)$. We write $X = L((u, \rho, \mu))$, and $(u, \rho, \mu) = M(X)$. There is a lack of uniqueness in this transformation, corresponding to the fact that a particle trajectory can be parametrized in several distinct ways. In our context we denote this by *relabeling*. Thus $M \circ L = \text{id}$, while $L \circ M$ is only the identity on the equivalence classes of Lagrangian functions that correspond to one and the same Eulerian solution, see [27, Thm. 3.12]. We prove that the convergence $(u_n, \rho_n, \mu_n) \to (u, 0, \mu)$ implies that $X_n \to X$ (in obvious notation) in the appropriate norm, see Theorem 3. The proof is surprisingly intricate and applies the notion of relabeling.

The situation is turned around in Sect. 4, where we consider an arbitrary sequence of Lagrangian coordinates X_n that converges to X, thus $X_n \to X$ in an appropriate norm. It is then shown that the corresponding Eulerian variables (u_n, ρ_n, μ_n) converge to (u, ρ, μ), see Theorem 4. In Sect. 5 we study how general convergence in Eulerian coordinates carries over to Lagrangian variables. To be more specific, consider a sequence (u_n, ρ_n, μ_n) that converges to (u, ρ, μ). Then we show in Theorem 5 that the corresponding Lagrangian coordinates converge. Here it is not assumed that the sequence (u_n, ρ_n, μ_n) is a mollification of (u, ρ, μ). Finally, in Sect. 6 we consider the time-dependent case. Consider a sequence of initial data $(u_{n,0}, \rho_{n,0}, \mu_{n,0})$ that converges to (u_0, ρ_0, μ_0) in $\mathscr{D}$. In Theorem 6 it is shown that the corresponding solutions converge for each fixed positive time. The proof transfers the convergence issue from Eulerian variables to Lagrangian coordinates, analyzes it in these variables, and finally translates the result back to the original variables.

2 Approximation in Eulerian Coordinates

The aim of this section is to show that any initial data $(u, 0, \mu)$ of the CH equation can be approximated by a sequence of smooth initial data (u_n, ρ_n, μ_n) of the 2CH system. We start by introducing the Banach spaces needed in this context, before recalling the definition of the set of Eulerian coordinates for the 2CH system (and hence also for the CH equation). Thereafter we state and prove the approximation theorem.

Let

$$L^2_{\text{const}}(\mathbb{R}) = \{\rho \in L^1_{\text{loc}}(\mathbb{R}) \mid \rho(x) = \bar{\rho}(x) + k,\ \bar{\rho} \in L^2(\mathbb{R}), k \in \mathbb{R}\}. \tag{3}$$

Then we can associate to any $\rho \in L^2_{\text{const}}(\mathbb{R})$ the unique pair $(\bar{\rho}, k) \in L^2(\mathbb{R}) \times \mathbb{R}$. Thus, if we equip $L^2_{\text{const}}(\mathbb{R})$ with the norm

$$\|\rho\|_{L^2_{\text{const}}} = \|\bar{\rho}\|_{L^2} + |k|, \tag{4}$$

then $L^2_{\text{const}}(\mathbb{R})$ is a Banach space.

We are now ready to introduce the set of Eulerian coordinates of the 2CH system (and hence also of the CH equation). The case of the CH equation corresponds to $\rho(x) = 0$ for all $x \in \mathbb{R}$.

Definition 1 (Eulerian Coordinates) The set $\mathscr{D}$ is composed of all triples (u, ρ, μ) such that $u \in H^1(\mathbb{R})$, $\rho \in L^2_{\text{const}}(\mathbb{R})$, and μ is a positive finite Radon measure whose absolutely continuous part μ_{ac} satisfies

$$\mu_{\text{ac}} = (u_x^2 + \bar{\rho}^2)dx.$$

We write $F(x) = \mu((-\infty, x])$.

We will need a standard *Friedrichs mollifier* $\phi \in C_c^\infty(\mathbb{R})$, chosen in such a way that $\phi(x) = \phi(-x) \geq 0$, $\|\phi\|_{L^1} = 1$, $\phi'(x) > 0$ for $x \in (-1, 0)$, and $\text{supp}(\phi) = [-1, 1]$.

Theorem 1 *Given $(u, 0, \mu) \in \mathscr{D}$, let (u_n, ρ_n, μ_n) be given through*

$$u_n(x) = \int_{\mathbb{R}} n\phi(n(x-y))u(y)dy, \tag{5a}$$

$$\rho_n(x) = \Big(\frac{1}{n^2} + \int_{\mathbb{R}} n\phi'(y)F(x - \frac{y}{n})dy - \Big(\int_{\mathbb{R}} \phi(y)u_x(x - \frac{y}{n})dy\Big)^2\Big)^{1/2} \tag{5b}$$

$$= \frac{1}{n} + \bar{\rho}_n(x),$$

$$\mu_n(x) = u_{n,x}^2(x) + \bar{\rho}_n^2(x). \tag{5c}$$

Define moreover

$$F_n(x) = \mu_n((-\infty, x]).$$

Then $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ is a sequence of smooth functions, which approximates $(u, 0, \mu)$ in the following sense:

$$u_n \to u \quad \text{in } H^1(\mathbb{R}),$$

$$F_n(x) \to F(x) \quad \text{for every } x \text{ at which } F \text{ is continuous.}$$

Proof We split the proof into several steps.

Step 1: Approximation of u by Smooth Functions u_n By assumption we have $u \in H^1(\mathbb{R})$. Thus, application of Minkowski's inequality for integrals and the dominated convergence theorem yield that u_n defined in (5a) converges to u in $H^1(\mathbb{R})$. Moreover, the smoothness of ϕ implies that $u_n \in C^\infty(\mathbb{R})$.

Step 2: Construction of Some Auxiliary Functions and Measures We start by defining the auxiliary function

$$\hat{F}_n(x) = \int_{\mathbb{R}} n\phi(n(x-y))F(y)dy. \tag{6}$$

Then $\hat{F}_n$ is smooth and converges pointwise to F at every point x at which F is continuous. Now recall that $F(x) = \mu((-\infty, x])$ and denote by μ_{d} the purely discrete part of the finite Radon measure μ. Then μ_{d} can be written as an at most countable sum of Dirac measures, the positions of which coincide with the set of discontinuities of F. In particular, F is continuous almost everywhere, and thus $\hat{F}_n$ converges to F pointwise almost everywhere. Define moreover

$$\hat{\mu}_n(x) = \int_{\mathbb{R}} n^2\phi'(n(x-y))F(y)dy = \int_{\mathbb{R}} n\phi(n(x-y))d\mu(y) \geq 0. \tag{7}$$

Then we obtain by Fubini's theorem that $\|\hat{\mu}_n\|_{L^1} = \|\mu\|$ for all $n \in \mathbb{N}$.

As a next step, we will associate a sequence of densities $\hat{\rho}_n$ to $(u_n, \hat{\mu}_n)$. To that end, we note, using the Cauchy–Schwarz inequality and the fact that $\|\phi\|_{L^1} = 1$, that

$$\begin{aligned}
&\int_{\mathbb{R}} \phi(z)u_x^2\Big(x - \frac{z}{n}\Big)dz - \Big(\int_{\mathbb{R}} \phi(z)u_x\Big(x - \frac{z}{n}\Big)dz\Big)^2 \\
&\qquad \geq \int_{\mathbb{R}} \phi(z)u_x^2\Big(x - \frac{z}{n}\Big)dz - \Big(\int_{\mathbb{R}} \phi(z)dz\Big)\Big(\int_{\mathbb{R}} \phi(z)u_x^2\Big(x - \frac{z}{n}\Big)dz\Big) = 0.
\end{aligned}$$

As a consequence, as μ is a positive Radon measure and $\mu_{\mathrm{ac}} = u_x^2 dx$, we see that

$$\begin{aligned}
\hat{\mu}_n(x) &= \int_{\mathbb{R}} n\phi(n(x-y))d\mu(y) \geq \int_{\mathbb{R}} n\phi(n(x-y))d\mu_{\mathrm{ac}}(y) \\
&= \int_{\mathbb{R}} n\phi(n(x-y))u_x^2(y)dy \geq \Big(\int_{\mathbb{R}} n\phi(n(x-y))u_x(y)dy\Big)^2,
\end{aligned}$$

and we may define $\hat{\rho}_n$ to be the non-negative root of

$$\hat{\rho}_n^2(x) = \hat{\mu}_n(x) - u_{n,x}^2(x). \tag{8}$$

Note that by construction $\hat{\rho}_n^2 \in L^1(\mathbb{R})$ and $\hat{\rho}_n^2 \in C^\infty(\mathbb{R})$. The function $\hat{\rho}_n$ itself need not be smooth, though.

Step 3: Smooth, Approximating Sequences ρ_n and μ_n Let ρ_n be defined by (5b), then

$$\rho_n^2(x) = \hat{\rho}_n^2(x) + \frac{1}{n^2}. \tag{9}$$

In particular, ρ_n is well-defined, since the term within the square root is always positive. Furthermore, we can decompose ρ_n as

$$\rho_n(x) = \frac{1}{n} + \bar{\rho}_n(x).$$

Then

$$\bar{\rho}_n(x) = -\frac{1}{n} + \left(\frac{1}{n^2} + \hat{\rho}_n^2(x)\right)^{1/2} \geq 0,$$

where we always take the positive root on the right-hand side. Since $\hat{\rho}_n^2$ is smooth and the term within the square root is bounded away from zero, it follows that $\bar{\rho}_n \in C^\infty(\mathbb{R})$ and consequently also $\mu_n \in C^\infty(\mathbb{R})$. Note also that this implies that

$$\rho_n(x) \geq \frac{1}{n} \quad \text{for all } x \in \mathbb{R}. \tag{10}$$

Moreover, we have that

$$\bar{\rho}_n^2(x) + \frac{2}{n}\bar{\rho}_n(x) = \hat{\rho}_n^2(x), \tag{11}$$

which in particular implies that

$$\bar{\rho}_n^2(x) \leq \hat{\rho}_n^2(x) \quad \text{for all } x \in \mathbb{R}. \tag{12}$$

Next, we see, using the definition of μ_n in (5c) and the Eqs. (12) and (8), that

$$\mu_n(x) = u_{n,x}^2(x) + \bar{\rho}_n^2(x) \leq u_{n,x}^2(x) + \hat{\rho}_n^2(x) = \hat{\mu}_n(x) \tag{13}$$

for all $x \in \mathbb{R}$. As a consequence,

$$\|u_{n,x}\|_{L^2}^2 + \|\bar{\rho}_n\|_{L^2}^2 = \|\mu_n\| \leq \|\hat{\mu}_n\| = \|\mu\|, \tag{14}$$

which in particular shows that μ_n is a finite Radon measure, but also that $\bar{\rho}_n \in L^2(\mathbb{R})$ and therefore $\rho_n \in L^2_{\text{const}}(\mathbb{R})$.

So far, we have shown that (u_n, ρ_n, μ_n) is a sequence of smooth functions contained in $\mathscr{D}$, and that $u_n \to u$ in $H^1(\mathbb{R})$. It now remains to show that $F_n(x) \to F(x)$ at every point x at which F is continuous, which is in this case equivalent to $\mu_n \to \mu$ weakly, cf. [11, Props. 7.19 and 8.17]. This means we have to prove that

$$\int_{\mathbb{R}} \psi(x) d\mu_n(x) \to \int_{\mathbb{R}} \psi(x) d\mu \quad \text{as } n \to \infty \tag{15}$$

for all $\psi \in C_c^\infty(\mathbb{R})$. To that end observe first that, due to (5c), (8), and (11), we have

$$\int_{\mathbb{R}} \psi(x) d\mu_n(x) dx = \int_{\mathbb{R}} \psi(x) d\hat{\mu}_n(x) - \frac{2}{n} \int_{\mathbb{R}} \psi(x) \bar{\rho}_n(x) dx. \tag{16}$$

We already know that $\hat{\mu}_n \to \mu$ weakly, that is,

$$\int_{\mathbb{R}} \psi(x) d\hat{\mu}_n(x) \to \int_{\mathbb{R}} \psi(x) d\mu(x) \quad \text{as } n \to \infty, \tag{17}$$

for all $\psi \in C_c^\infty(\mathbb{R})$. Moreover we obtain from (14) and the Cauchy–Schwarz inequality that

$$\left| \frac{2}{n} \int_{\mathbb{R}} \psi(x) \bar{\rho}_n(x) dx \right| \leq \frac{2}{n} \|\psi\|_{L^2} \|\bar{\rho}_n\|_{L^2} \leq \frac{2}{n} \|\psi\|_{L^2} \|\mu\|^{1/2} \to 0$$

for all $\psi \in C_c^\infty(\mathbb{R})$, which concludes the proof.

Remark 1 Note that one can show that the function $\frac{1}{n}\bar{\rho}_n$ converges pointwise to 0. Indeed, according to (8) and (11), we have

$$0 \leq \bar{\rho}_n^2(x) \leq \hat{\rho}_n^2(x) = \hat{\mu}_n(x) - u_{n,x}^2(x) \leq \hat{\mu}_n(x). \tag{18}$$

Moreover, from (7) we get

$$\frac{1}{n^2} \hat{\mu}_n(x) = \frac{1}{n} \int_{-1}^{0} \phi'(z) \Big(F\Big(x - \frac{z}{n}\Big) - F\Big(x + \frac{z}{n}\Big) \Big) dz \leq \frac{1}{n} \|\mu\| \, \phi(0). \tag{19}$$

Thus combining (18) and (19) yields that the sequence $\frac{1}{n}\bar{\rho}_n$ is uniformly bounded and that

$$\frac{1}{n} \bar{\rho}_n \to 0 \quad \text{pointwise as } n \to \infty.$$

Remark 2 In the next section we are not only going to use the splitting of $\hat{\mu}_n(x)$ into $u_{n,x}^2(x)$ and $\hat{\rho}_n^2(x)$ as introduced in (8), but also a second one, which we are introducing next. Namely, let $F_s(x) = \mu_s((-\infty, x])$, where μ_s denotes the singular part of the measure μ, and let ϕ be the Friedrichs mollifier. Define

$$\tilde{u}_{n,x}^2(x) = \int_{\mathbb{R}} \phi(z) u_x^2\Big(x - \frac{z}{n}\Big) dz \tag{20}$$

and

$$\tilde{\rho}_n^2(x) = \int_{\mathbb{R}} n\phi'(z) F_{\mathrm{s}}\Big(x - \frac{z}{n}\Big) dz. \tag{21}$$

Then

$$\begin{aligned} \tilde{u}_{n,x}^2(x) + \tilde{\rho}_n^2(x) &= \int_{\mathbb{R}} n\phi'(z) F_{\mathrm{s}}\Big(x - \frac{z}{n}\Big) dz + \int_{\mathbb{R}} \phi(z) u_x^2\Big(x - \frac{z}{n}\Big) dz \\ &= \int_{\mathbb{R}} n\phi'(z) F\Big(x - \frac{z}{n}\Big) dz = \hat{\mu}_n(x). \end{aligned} \tag{22}$$

Remark 3 Let $(u, 0, \mu) \in \mathscr{D}$, $n \in \mathbb{N}$, and $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ be defined as in Theorem 1. By construction we then have that u_n, $\rho_n \in C^\infty(\mathbb{R})$, μ_n is absolutely continuous, and, according to (10), that $\rho_n(x) \geq \frac{1}{n}$ for all $x \in \mathbb{R}$. Hence [16, Cor. 6.2] implies that the corresponding solution $(u_n(t), \rho_n(t), \mu_n(t))$ has the same regularity for all times t, and, in particular, no wave breaking occurs.

3 Convergence in Lagrangian Coordinates

The aim of this section is to show that the smooth approximating sequence constructed in Theorem 1 not only converges in the set of Eulerian coordinates $\mathscr{D}$ but also in the set of Lagrangian coordinates $\mathscr{F}$. Hence, we are first going to introduce the set of Lagrangian coordinates $\mathscr{F}$ and the mapping L from $\mathscr{D}$ to $\mathscr{F}$, before stating and proving the outlined convergence theorem.

Let V be the Banach space defined by

$$V = \{f \in C_b(\mathbb{R}) \mid f_\xi \in L^2(\mathbb{R})\},$$

where $C_b(\mathbb{R}) = C(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and the norm is given by $\|f\|_V = \|f\|_{L^\infty} + \|f_\xi\|_{L^2}$. Let moreover

$$E = V \times H^1(\mathbb{R}) \times L^2(\mathbb{R}) \times L^2(\mathbb{R}) \times \mathbb{R},$$

then E equipped with the norm

$$\|(\zeta, U, h, \bar{r}, k)\|_E = \|\zeta\|_V + \|U\|_{H^1} + \|h\|_{L^2} + \|\bar{r}\|_{L^2} + |k| \tag{23}$$

is a Banach space. Note that we can associate to each $(\zeta, U, h, \bar{r}, k) \in E$ the tuple (y, U, h, r) by setting

$$y = \zeta + \mathrm{id} \quad \text{and} \quad r = \bar{r} + k y_\xi. \tag{24}$$

Conversely, for any pair (y, r) such that $y - \mathrm{id} \in V$ and $r \in L^2_{\mathrm{const}}(\mathbb{R})$ there exists a unique triplet $(\zeta, \bar{r}, k) \in [L^2(\mathbb{R})]^2 \times \mathbb{R}$ such that (24) holds. For more details we

refer to [16, Sect. 3]. In what follows we will slightly abuse the notation by writing $(y, U, h, r) \in E$ instead of $(\zeta, U, h, \bar{r}, k) \in E$.

In addition we have to introduce the set of relabeling functions, which are not only needed for identifying equivalence classes in Lagrangian coordinates, but also for determining the set of Lagrangian coordinates.

Definition 2 (Relabeling Functions) We denote by G the subgroup of the group of homeomorphisms f of $\mathbb{R}$ such that

$$f - \mathrm{id} \text{ and } f^{-1} - \mathrm{id} \text{ both belong to } W^{1,\infty}(\mathbb{R}), \tag{25a}$$

$$f_\xi - 1 \text{ belongs to } L^2(\mathbb{R}), \tag{25b}$$

where id denotes the identity function.

Given $\kappa \geq 0$, we denote by G_κ the subset of G defined by

$$G_\kappa = \{f \in G \mid \|f - \mathrm{id}\|_{W^{1,\infty}} + \left\|f^{-1} - \mathrm{id}\right\|_{W^{1,\infty}} \leq \kappa\}.$$

We are now ready to introduce the set of Lagrangian coordinates of the 2CH system (and hence also of the CH equation). The case of the CH equation corresponds to $r(\xi) = 0$ for all $\xi \in \mathbb{R}$.

Definition 3 (Lagrangian Coordinates) The set $\mathscr{F}$ is composed of all tuples $X = (y, U, h, r) \in E$, such that

$$(\zeta, U, h, r) \in [W^{1,\infty}(\mathbb{R})]^2 \times [L^\infty(\mathbb{R})]^2, \tag{26a}$$

$$y_\xi \geq 0, \quad h \geq 0, \quad y_\xi + h > 0 \text{ almost everywhere}, \tag{26b}$$

$$y_\xi h = U_\xi^2 + \bar{r}^2 \text{ almost everywhere}, \tag{26c}$$

$$y + H \in G, \tag{26d}$$

where we denote $y(\xi) = \zeta(\xi) + \xi$ and $H(\xi) = \int_{-\infty}^{\xi} h(\eta) d\eta$.

Moreover, we set

$$\mathscr{F}_\kappa = \{X \in \mathscr{F} \mid y + H \in G_\kappa\}.$$

Observe that

$$\mathscr{F}_0 = \{X \in \mathscr{F} \mid y(\xi) + H(\xi) = \xi \text{ for all } \xi \in \mathbb{R}\}. \tag{27}$$

We note that the group G acts on $\mathscr{F}$ by means of right composition of the form

$$X = (y, U, h, r) \mapsto X \circ g := (y \circ g, U \circ g, (h \circ g) g_\xi, (r \circ g) g_\xi).$$

This group action then allows us to define equivalence classes of Lagrangian coordinates, where we say that two coordinates X and $\hat{X}$ are equivalent, if there exists some $g \in G$ such that $\hat{X} = X \circ g$.

Given an arbitrary $X = (y, U, h, r)$, we note that $y + H \in G$ and hence also $(y + H)^{-1} \in G$. In particular, if we introduce

$$\Gamma(X) = X \circ (y + H)^{-1}, \tag{28}$$

then a short computation yields that $\Gamma(X) \in \mathscr{F}_0$. This shows that every equivalence class $X \circ G$ of Lagrangian coordinates has a unique canonical representative $\Gamma(X)$ in $\mathscr{F}_0$. Moreover, it has been shown in [16, Lem. 4.6] that the mapping $\Gamma|_{\mathscr{F}_\kappa} : \mathscr{F}_\kappa \to \mathscr{F}_0$ is continuous for each $\kappa > 0$.

Finally we can introduce the mapping L from Eulerian to Lagrangian coordinates.

Theorem 2 ([16, Thm. 4.9]) *For any (u, ρ, μ) in $\mathscr{D}$, let*

$$y(\xi) = \sup\{y \mid \mu((-\infty, y)) + y < \xi\}, \tag{29a}$$

$$h(\xi) = 1 - y_\xi(\xi), \tag{29b}$$

$$U(\xi) = u \circ y(\xi), \tag{29c}$$

$$r(\xi) = \rho \circ y(\xi) y_\xi(\xi). \tag{29d}$$

Then $(y, U, h, r) \in \mathscr{F}_0$. We denote by $L : \mathscr{D} \to \mathscr{F}_0$ the mapping which to any element $(u, \rho, \mu) \in \mathscr{D}$ associates $X = (y, U, h, r) \in \mathscr{F}_0$ given by (29).

In the case of the CH equation, we have $r(\xi) = 0$ for all $\xi \in \mathbb{R}$.

Theorem 3 *Let $(u, 0, \mu) \in \mathscr{D}$, and let $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ be the corresponding approximating sequence defined in Theorem 1. Moreover, let $(y, U, h, 0) = L((u, 0, \mu))$ and $(y_n, U_n, h_n, r_n) = L((u_n, \rho_n, \mu_n))$. Then*

$$(y_n, U_n, h_n, r_n) \to (y, U, h, 0) \quad \text{in } E.$$

Proof Let $(u, 0, \mu) \in \mathscr{D}$ and $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ be the approximating series defined in Theorem 1. Furthermore, let $X = (y, U, h, 0) = L((u, 0, \mu))$ and $X_n = (y_n, U_n, h_n, r_n) = L((u_n, \rho_n, \mu_n))$, which yields a smooth sequence in Lagrangian coordinates, cf. [16, Proof of Thm. 6.1]. However, due to the construction of our approximating sequence (u_n, ρ_n, μ_n), it turns out that in order to prove that $X_n \to X$ in Lagrangian coordinates, it is better to introduce another sequence $\hat{X}_n = (\hat{y}_n, \hat{U}_n, \hat{h}_n, \hat{r}_n)$ which is linked to the sequence X_n via relabeling. For better understanding, we split the proof into several steps. After first defining the new sequence $\hat{X}_n$, we show that for every $n \in \mathbb{N}$ there exists $g_n \in G$ such that $\hat{X}_n = X_n \circ g_n$ (Step 1). Thereafter, we establish that $\hat{X}_n \to X$ in E (Steps 2–9). Finally, we show that $\hat{X}_n \to X$ implies $X_n \to X$ in E (Step 10). The situation is also depicted in Fig. 1.

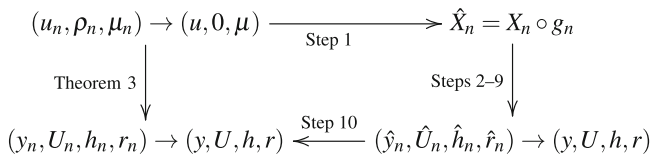


Fig. 1 Outline of the proof of Theorem 3

Step 1: Definition of the Sequence $\hat{X}_n$ and Proof That $\hat{X}_n = X_n \circ g_n$ Define[1] $\hat{F}_n(x)$ by

$$\hat{F}_n(x) = \int_{\mathbb{R}} n\phi(n(x-y))F(y)dy = \hat{\mu}_n((-\infty, x]), \tag{30}$$

such that

$$\hat{\mu}_n(x) = \int_{-1}^{0} n\phi'(z)\Big(F\Big(x - \frac{z}{n}\Big) - F\Big(x + \frac{z}{n}\Big)\Big)dz.$$

Introduce $\hat{\rho}_n^2 = \hat{\mu}_n - u_{n,x}^2$. Then $\hat{\mu}_n = u_{n,x}^2 + \bar{\rho}_n^2 + \frac{2}{n}\bar{\rho}_n = u_{n,x}^2 + \hat{\rho}_n^2$, as in (8)–(11). Let now $\hat{X}_n = (\hat{y}_n, \hat{U}_n, \hat{h}_n, \hat{r}_n) \in \mathscr{F}$, where

$$\hat{y}_n(\xi) = \sup\left\{y \mid \hat{\mu}_n((-\infty, y)) + y < \xi\right\}, \tag{31a}$$

$$\hat{h}_n(\xi) = 1 - \frac{2}{n}\bar{\hat{r}}_n(\xi) - \hat{y}_{n,\xi}(\xi), \tag{31b}$$

$$\hat{U}_n(\xi) = u_n \circ \hat{y}_n(\xi), \tag{31c}$$

$$\bar{\hat{r}}_n(\xi) = \bar{\rho}_n \circ \hat{y}_n(\xi)\hat{y}_{n,\xi}(\xi), \tag{31d}$$

$$\hat{r}_n(\xi) = \rho_n \circ \hat{y}_n(\xi)\hat{y}_{n,\xi}(\xi). \tag{31e}$$

We are going to show that we can write $\hat{X}_n = X_n \circ g_n$ for some $g_n \in G$, that is,[2]

$$\begin{aligned}\hat{X}_n(\xi) &= (\hat{y}_n(\xi), \hat{U}_n(\xi), \hat{h}_n(\xi), \hat{r}_n(\xi)) \\ &= \big(y_n(g_n(\xi)), U_n(g_n(\xi)), h_n(g_n(\xi))g_{n,\xi}(\xi), r_n(g_n(\xi))g_{n,\xi}(\xi)\big) \\ &= X_n \circ g_n(\xi),\end{aligned} \tag{32}$$

[1] This construction resembles the one used in Step 2 of the proof of Theorem 1. However, here we perform the construction in Lagrangian variables.

[2] Note the factors $g_{n,\xi}(\xi)$.

which implies immediately that $\hat{X}_n \in \mathscr{F}$ and that it belongs to the same equivalence class as X_n. Additionally, we will show that there exists some κ independent of n such that $g_n \in G_\kappa$ for all $n \in \mathbb{N}$.

Since both μ_n and $\hat{\mu}_n$ are smooth and purely absolutely continuous, we have that

$$y_n(\xi) + F_n(y_n(\xi)) = \xi, \tag{33}$$

and

$$\hat{y}_n(\xi) + \hat{F}_n(\hat{y}_n(\xi)) = \xi, \tag{34}$$

for all $\xi \in \mathbb{R}$. Moreover, recall that

$$\mu_n(x) + \frac{2}{n}\bar{\rho}_n(x) = \hat{\mu}_n(x) \quad \text{for all } x \in \mathbb{R}$$

according to (11), (8) and (5c), and $\frac{2}{n}\bar{\rho}_n \in L^1(\mathbb{R})$. Hence we can rewrite (34) as

$$\hat{y}_n(\xi) + \hat{H}_n(\xi) = \hat{y}_n(\xi) + F_n(\hat{y}_n(\xi)) = \xi - \frac{2}{n}\int_{-\infty}^{\hat{y}_n(\xi)} \bar{\rho}_n(x)dx = g_n(\xi), \tag{35}$$

which defines $g_n(\xi)$. Here $\hat{H}_n(\xi) = \int_{-\infty}^{\xi} \hat{h}_n(\eta)d\eta$. Moreover, using (33) we have

$$y_n(g_n(\xi)) + F_n(y_n(g_n(\xi))) = g_n(\xi) \quad \text{for all } \xi \in \mathbb{R},$$

and, since $\mathrm{id} + F_n$ is strictly increasing, we conclude that

$$\hat{y}_n(\xi) = y_n(g_n(\xi)) \quad \text{for all } \xi \in \mathbb{R}, \tag{36}$$

which immediately implies that

$$\hat{U}_n(\xi) = u_n(\hat{y}_n(\xi)) = u_n(y_n(g_n(\xi))) = U_n(g_n(\xi)) \quad \text{for all } \xi \in \mathbb{R}.$$

Using (31b), (31d), (35), (36), and (29d) we infer that

$$\begin{aligned}\hat{h}_n(\xi) &= 1 - \frac{2}{n}\hat{\bar{r}}_n(\xi) - \hat{y}_{n,\xi}(\xi) = g_{n,\xi}(\xi) - \hat{y}_{n,\xi}(\xi) \\ &= \big(1 - y_{n,\xi}(g_n(\xi))\big)g_{n,\xi}(\xi) = h_n(g_n(\xi))g_{n,\xi}(\xi).\end{aligned} \tag{37}$$

In addition, we see that

$$\hat{r}_n(\xi) = \rho_n(\hat{y}_n(\xi))\hat{y}_{n,\xi}(\xi) = r_n(g_n(\xi))g_{n,\xi}(\xi).$$

Thus we conclude that $\hat{X}_n = X_n \circ g_n$, and it remains to show that $g_n \in G_\kappa$ for some κ independent of n.

Instead of checking that g_n satisfies all the properties listed in Definition 2, we are going to apply [27, Lem. 3.2]. Namely, if g_n is absolutely continuous, $g_{n,\xi} - 1 \in L^2(\mathbb{R})$, and there exist $c_1 \geq 1$ and $c_2 > 0$ such that $\frac{1}{c_1} \leq g_{n,\xi}(\xi) \leq c_1$ almost everywhere and $\|g_n - \mathrm{id}\|_{L^\infty} \leq c_2$, then $g_n \in G_\kappa$ for some $\kappa > 0$ depending only on c_1 and c_2. By construction, $\hat{X}_n$ is smooth and therefore g_n is smooth and, in particular, absolutely continuous. Since $\frac{2}{n}\bar{\rho}_n \in L^1(\mathbb{R})$ and $\bar{\rho}_n$ is strictly positive, we get from (35) that $g_n - \mathrm{id} \in L^\infty(\mathbb{R})$, and from (8), (11), and (14) that $\|g_n - \mathrm{id}\|_{L^\infty} \leq \|\mu\|$. Moreover, using the notation in the proof of Theorem 1, we obtain from (34) and (8) that

$$\hat{y}_{n,\xi}(\xi) = \frac{1}{1 + u_{n,x}^2(\hat{y}_n(\xi)) + \hat{\rho}_n^2(\hat{y}_n(\xi))} \leq 1. \tag{38}$$

Thus

$$\begin{aligned}
\frac{1}{2} &\leq \frac{1 + u_{n,x}^2(\hat{y}_n(\xi)) + \bar{\rho}_n^2(\hat{y}_n(\xi))}{1 + \frac{1}{n^2} + u_{n,x}^2(\hat{y}_n(\xi)) + 2\bar{\rho}_n^2(\hat{y}_n(\xi))} \\
&\leq \frac{1 + u_{n,x}^2(\hat{y}_n(\xi)) + \bar{\rho}_n^2(\hat{y}_n(\xi))}{1 + u_{n,x}^2(\hat{y}_n(\xi)) + \bar{\rho}_n^2(\hat{y}_n(\xi)) + \frac{2}{n}\bar{\rho}_n(\hat{y}_n(\xi))} \\
&= 1 - \frac{\frac{2}{n}\bar{\rho}_n(\hat{y}_n(\xi))}{1 + u_{n,x}^2(\hat{y}_n(\xi)) + \hat{\rho}_n^2(\hat{y}_n(\xi))} \\
&= g_{n,\xi}(\xi) \leq 1,
\end{aligned} \tag{39}$$

due to (11) and (35). Finally, we have to check that $g_{n,\xi} - 1 \in L^2(\mathbb{R})$. Direct computations, using (35) and (38), yield

$$\int_{\mathbb{R}} (g_{n,\xi}(\xi) - 1)^2 d\xi = \frac{4}{n^2} \int_{\mathbb{R}} \bar{\rho}_n^2(\hat{y}_n(\xi)) \hat{y}_{n,\xi}^2(\xi) d\xi \leq \frac{4}{n^2} \int_{\mathbb{R}} \bar{\rho}_n^2(x) dx.$$

Thus $g_{n,\xi} - 1 \in L^2(\mathbb{R})$, since $\bar{\rho}_n \in L^2(\mathbb{R})$. Thus [27, Lem. 3.2] implies that g_n is a relabeling function and that there exists $\kappa > 0$ independent of n such that $g_n \in G_\kappa$ for all $n \in \mathbb{N}$.

Step 2: The Sequence $\hat{y}_n - \mathrm{id}$ Converges to $y - \mathrm{id}$ in $L^\infty(\mathbb{R})$ Recall that we have by definition that

$$y(\xi) + F(y(\xi)-) \leq \xi \leq y(\xi) + F(y(\xi)) \quad \text{for all } \xi \in \mathbb{R}, \tag{40}$$

where $F(x) = \mu((-\infty, x])$. Moreover, since $\hat{\mu}_n(x)$ is smooth and purely absolutely continuous, we have

$$\hat{y}_n(\xi) + \hat{F}_n(\hat{y}_n(\xi)) = \xi \quad \text{for all } \xi \in \mathbb{R}, \tag{41}$$

where $\hat{F}_n(x) = \hat{\mu}_n((-\infty, x])$. Introducing

$$G(x) := x + F(x) \quad \text{and} \quad \hat{G}_n(x) := x + \hat{F}_n(x), \tag{42}$$

we conclude that

$$\begin{aligned}\hat{G}_n(\hat{y}_n(\xi)) &= \hat{y}_n(\xi) + \hat{F}_n(\hat{y}_n(\xi)) \\ &= \int_{-1}^{1} \phi(z)\Big(\hat{y}_n(\xi) - \frac{z}{n} + F\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)\Big)dz \\ &= \int_{-1}^{1} \phi(z) G\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz,\end{aligned}$$

where we used (30). Moreover, since $G(x)$ is strictly increasing and due to (41) and (42), one has that

$$G\Big(\hat{y}_n(\xi) - \frac{1}{n}\Big) < \hat{G}_n(\hat{y}_n(\xi)) = \xi < G\Big(\hat{y}_n(\xi) + \frac{1}{n}-\Big), \tag{43}$$

and by (40) that

$$G(y(\xi)-) \le \xi \le G(y(\xi)). \tag{44}$$

Combining (43) and (44) yields on the one hand that

$$G\Big(\hat{y}_n(\xi) - \frac{1}{n}\Big) < \xi \le G(y(\xi))$$

and on the other hand that

$$G(y(\xi)-) \le \xi < G\Big(\hat{y}_n(\xi) + \frac{1}{n}-\Big).$$

Recalling that $G(x)$, and hence also $G(x-)$, is strictly increasing, we obtain that

$$\hat{y}_n(\xi) - \frac{1}{n} < y(\xi) < \hat{y}_n(\xi) + \frac{1}{n}$$

or, equivalently,

$$-\frac{1}{n} < y(\xi) - \hat{y}_n(\xi) < \frac{1}{n} \quad \text{for all } \xi \in \mathbb{R}.$$

In particular, this shows that $\|\hat{y}_n - y\|_{L^\infty} \to 0$ as $n \to \infty$.

Step 3: Convergence of $\tilde{h}_n$ to h in $L^1(\mathbb{R})$ Let

$$\tilde{h}_n = 1 - \hat{y}_{n,\xi}. \tag{45}$$

To show that $\tilde{h}_n \to h$ in $L^1(\mathbb{R})$ is the main (and most difficult) step.

Due to our change from Eulerian to Lagrangian coordinates, it is not clear at first sight that $\tilde{h}_n$ and h belong to $L^1(\mathbb{R})$. We know that $hy_\xi = U_\xi^2$, or, equivalently, $h = U_\xi^2 + h^2$, because $y_\xi + h = 1$. However, since U_ξ and h both belong to $L^2(\mathbb{R})$, it follows that $h \in L^1(\mathbb{R})$. Combining (45) and (31b), and recalling that $\hat{X} \in \mathscr{F}$ which implies that (26c) is satisfied, one obtains

$$\tilde{h}_n = \tilde{h}_n \hat{y}_{n,\xi} + \tilde{h}_n^2 = \hat{h}_n \hat{y}_{n,\xi} + \frac{2}{n}\bar{\hat{r}}_n \hat{y}_{n,\xi} + \tilde{h}_n^2 = \hat{U}_{n,\xi}^2 + \bar{\hat{r}}_n^2 + \frac{2}{n}\bar{\hat{r}}_n \hat{y}_{n,\xi} + \tilde{h}_n^2, \tag{46}$$

where $\bar{\hat{r}}_n$, $\tilde{h}_n$, $\hat{U}_{n,\xi} \in L^2(\mathbb{R})$, $\frac{2}{n}\bar{\hat{r}}_n \in L^1(\mathbb{R})$ and $\hat{y}_{n,\xi} \in L^\infty(\mathbb{R})$. Thus also $\tilde{h}_n \in L^1(\mathbb{R})$. Define

$$\tilde{H}_n(\xi) = \int_{-\infty}^{\xi} \tilde{h}_n(\eta)d\eta \quad \text{and} \quad H(\xi) = \int_{-\infty}^{\xi} h(\eta)d\eta. \tag{47}$$

Then the identities $\tilde{H}_n + \hat{y}_n = \mathrm{id}$ and $H + y = \mathrm{id}$ together with the pointwise convergence of $\hat{y}_n \to y$ imply that $\tilde{H}_n$ converges pointwise to H and $\tilde{H}_n(\infty) = \|\hat{\mu}_n\|_{L^1}$ and $H(\infty) = \|\mu\|$. In particular, this means that $\tilde{H}_n(\infty) = H(\infty)$ for all $n \in \mathbb{N}$ and hence

$$\|\tilde{h}_n\|_{L^1} = \|h\|_{L^1}, \quad n \in \mathbb{N}. \tag{48}$$

Next we will prove that $\tilde{h}_n$ converges to h pointwise almost everywhere, which will imply that $\tilde{h}_n \to h$ in $L^1(\mathbb{R})$, see [1, Prop. 1.33]. To that end, observe first that $\tilde{h}_n(\xi) - h(\xi) = y_\xi(\xi) - \hat{y}_{n,\xi}(\xi)$ for all $\xi \in \mathbb{R}$. Thus it suffices to show that $\hat{y}_{n,\xi}$ converges pointwise to y_ξ almost everywhere. Recalling (41), (42), and that $\hat{y}_n$ is smooth, we see that this is equivalent to showing that

$$\hat{G}_n'(\hat{y}_n(\xi)) \to \frac{1}{y_\xi(\xi)} \quad \text{for almost every } \xi \in \mathbb{R}. \tag{49}$$

Moreover, note that

$$\begin{aligned}\hat{G}_n'(\hat{y}_n(\xi)) &= n\int_{-1}^{1}\phi'(z)G\Big(\hat{y}_n(\xi)-\frac{z}{n}\Big)dz\\ &= n\int_{-1}^{1}\phi'(z)\Big(G\Big(\hat{y}_n(\xi)-\frac{z}{n}\Big)-\xi\Big)dz\\ &= n^2\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}}\phi'(n(\hat{y}_n(\xi)-z))(G(z)-\xi)dz. \end{aligned} \tag{50}$$

Introducing the strictly increasing function $\tilde{G}(z) = G(z) - \xi$, it follows that we have to show that

$$\hat{G}_n'(\hat{y}_n(\xi)) = n^2\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}}\phi'(n(\hat{y}_n(\xi)-z))\tilde{G}(z)dz \to \frac{1}{y_\xi(\xi)} \tag{51}$$

for almost every $\xi \in \mathbb{R}$.

In fact, we will show below that (51) holds at every $\xi \in \mathbb{R}$ where the function y is differentiable. Since y is Lipschitz continuous and therefore differentiable almost everywhere, this will prove the convergence of $\tilde{h}_n$ to h in $L^1(\mathbb{R})$. In the proof of (51), we will consider separately the cases where the derivative of y is zero, and where it is strictly positive.

(a) The Case $y_\xi(\xi) = 0$ We have to show that $\hat{G}_n'(\hat{y}_n(\xi)) \to \infty$ as $n \to \infty$. By assumption $y_\xi(\xi) = 0$ and hence for every $\varepsilon > 0$ there exists some $\delta_\varepsilon > 0$ such that

$$0 \le \frac{y(\eta)-y(\xi)}{\eta-\xi} < \varepsilon \quad \text{whenever } |\xi-\eta| < \delta_\varepsilon. \tag{52}$$

Define $\gamma_\varepsilon := \varepsilon\delta_\varepsilon$ and let $z \in \mathbb{R}$ such that $|z - y(\xi)| < \gamma_\varepsilon$. In addition, recall (40) and (42), and observe that $y(G(z)) = z$ for all $z \in \mathbb{R}$. If $|G(z) - \xi| < \delta_\varepsilon$, we have by (52) that

$$\frac{G(z)-\xi}{z-y(\xi)} = \frac{G(z)-\xi}{y(G(z))-y(\xi)} > \frac{1}{\varepsilon}.$$

On the other hand, if $|G(z) - \xi| \ge \delta_\varepsilon$, then

$$\frac{G(z)-\xi}{z-y(\xi)} = \frac{|G(z)-\xi|}{|z-y(\xi)|} > \frac{\delta_\varepsilon}{\gamma_\varepsilon} = \frac{1}{\varepsilon}.$$

Thus

$$\frac{G(z)-\xi}{z-y(\xi)} > \frac{1}{\varepsilon} \quad \text{whenever } |z-y(\xi)| < \gamma_\varepsilon. \tag{53}$$

In the remainder of this section we are going to show that there exists a constant $C > 0$ independent of n and ε such that

$$\hat{G}_n'(\hat{y}_n(\xi)) > \frac{C}{\varepsilon} \quad \text{for all } n \text{ such that } \frac{2}{n} < \gamma_\varepsilon, \tag{54}$$

which will prove the claim. Let

$$I_0 := \{z \mid |z-\hat{y}_n(\xi)| \le |y(\xi)-\hat{y}_n(\xi)|\} = \hat{y}_n(\xi)+\{t \mid |t| \le |y(\xi)-\hat{y}_n(\xi)|\}. \tag{55}$$

Direct computations show that for all $z \in \mathbb{R} \setminus I_0$

$$\phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z) \ge 0 \tag{56}$$

and that

$$\begin{aligned}
&\int_{I_0} \phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\quad = \int_{-|\hat{y}_n(\xi)-y(\xi)|}^{|\hat{y}_n(\xi)-y(\xi)|} \phi'(nz)\tilde{G}(\hat{y}_n(\xi) - z)dz \\
&\quad = \int_{0}^{|\hat{y}_n(\xi)-y(\xi)|} \phi'(nz)\big(\tilde{G}(\hat{y}_n(\xi) - z) - \tilde{G}(\hat{y}_n(\xi) + z)\big)dz \ge 0,
\end{aligned} \tag{57}$$

since both terms in the last integral are non-positive on the interval of integration.

Again, we have to consider two situations separately depending on the difference of $y(\xi)$ and $\hat{y}_n(\xi)$.

(a.I) The Case $|\hat{y}_n(\xi) - y(\xi)| \le \frac{1}{2n}$ We only prove (54) in the case $y(\xi) \le \hat{y}_n(\xi) \le y(\xi) + \frac{1}{2n}$ and leave the other case, which follows the same lines, to the interested reader. Using (50), (53), (56), and (57) we have

$$\begin{aligned}
\hat{G}_n'(\hat{y}_n(\xi)) &= n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\quad + n^2 \int_{I_0} \phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\quad + n^2 \int_{2\hat{y}_n(\xi)-y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\ge n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi'(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz
\end{aligned}$$

$$\geq n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi'(n(\hat{y}_n(\xi)-z))\frac{z-y(\xi)}{\varepsilon}dz$$

$$= \frac{n}{\varepsilon}\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi(n(\hat{y}_n(\xi)-z))dz$$

$$\geq \frac{n}{\varepsilon}\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)-\frac{1}{2n}} \phi(n(\hat{y}_n(\xi)-z))dz$$

$$= \frac{1}{\varepsilon}\int_{\frac{1}{2}}^{1} \phi(z)dz = \frac{C}{\varepsilon}$$

with $C = \int_{1/2}^{1}\phi(z)dz$. Here we applied (53) to $\tilde{G}(z) = G(z) - \xi$, which is possible since we assume that $\frac{2}{n} < \gamma_\varepsilon$.

(a.II) The Case $\frac{1}{2n} < |\hat{y}_n(\xi) - y(\xi)| < \frac{1}{n}$ We only prove (54) in the case $y(\xi) + \frac{1}{2n} < \hat{y}_n(\xi) < y(\xi) + \frac{1}{n}$ and leave the other case, which follows the same lines, to the interested reader. Due to (56) and (57) we have

$$\hat{G}_n'(\hat{y}_n(\xi)) = n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi'(n(\hat{y}_n(\xi)-z))\tilde{G}(z)dz$$

$$+ n^2 \int_{I_0} \phi'(n(\hat{y}_n(\xi)-z))\tilde{G}(z)dz$$

$$+ n^2 \int_{2\hat{y}_n(\xi)-y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z))\tilde{G}(z)dz$$

$$\geq n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi'(n(\hat{y}_n(\xi)-z))\tilde{G}(z)dz$$

$$= n\int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi'(z)\tilde{G}\Big(\hat{y}_n(\xi)-\frac{z}{n}\Big)dz.$$

Let us turn our attention to the last integral

$$n\int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi'(z)\tilde{G}\Big(\hat{y}_n(\xi)-\frac{z}{n}\Big)dz,$$

where $n(\hat{y}_n(\xi) - y(\xi)) \in (\frac{1}{2}, 1)$. Since $\tilde{G}(\hat{y}_n(\xi) - \frac{z}{n})$ is strictly decreasing and $\tilde{G}(\hat{y}_n(\xi) - \frac{z}{n}-) \leq 0$ for all $z \in [n(\hat{y}_n(\xi) - y(\xi)), 1]$, we have

$$\begin{aligned} 0 &\leq n \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi'(z)\tilde{G}\big(\hat{y}_n(\xi) - \frac{z}{n}\big)dz \\ &= -n \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{z}{n})}^{0} \phi'(z)dp\,dz. \end{aligned}$$

Since the area of integration has finite measure and the integrand is uniformly bounded, we can interchange the order of integration and get

$$\begin{aligned} -n &\int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{z}{n})}^{0} \phi'(z)dp\,dz \\ &= -n \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{1}{n})}^{0} \int_{\max(n(\hat{y}_n(\xi)-y(\xi)),n(\hat{y}_n(\xi)-\tilde{G}^{-1}(p)))}^{1} \phi'(z)dz\,dp. \end{aligned}$$

Evaluating the inner integral and using that $\phi(z)$ is decreasing on $[0, 1]$, we end up with

$$\begin{aligned} -n &\int_{\tilde{G}(\hat{y}_n(\xi)-\frac{1}{n})}^{0} \int_{\max(n(\hat{y}_n(\xi)-y(\xi)),n(\hat{y}_n(\xi)-\tilde{G}^{-1}(p)))}^{1} \phi'(z)dz\,dp \\ &= n \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{1}{n})}^{0} \phi\big(\max(n(\hat{y}_n(\xi) - y(\xi)), n(\hat{y}_n(\xi) - \tilde{G}^{-1}(p)))\big)dp \\ &\geq n \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{1}{n})}^{0} \int_{\max(n(\hat{y}_n(\xi)-y(\xi)),n(\hat{y}_n(\xi)-\tilde{G}^{-1}(p)))}^{1} \\ &\qquad \frac{\phi(z)}{1 - \max(n(\hat{y}_n(\xi) - y(\xi)), n(\hat{y}_n(\xi) - \tilde{G}^{-1}(p)))}dz\,dp \\ &\geq \frac{n}{1 - n(\hat{y}_n(\xi) - y(\xi))} \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{1}{n})}^{0} \int_{\max(n(\hat{y}_n(\xi)-y(\xi)),n(\hat{y}_n(\xi)-\tilde{G}^{-1}(p)))}^{1} \phi(z)dz\,dp \\ &= \frac{n}{1 - n(\hat{y}_n(\xi) - y(\xi))} \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \int_{\tilde{G}(\hat{y}_n(\xi)-\frac{z}{n})}^{0} \phi(z)dp\,dz \\ &= -\frac{n}{1 - n(\hat{y}_n(\xi) - y(\xi))} \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz. \end{aligned}$$

In the last step we used once more that both the area of integration and the integrand are bounded, which justifies once more the interchange of the order of integration. Thus we showed, so far, that

$$\begin{aligned}\hat{G}_n'(\hat{y}_n(\xi)) &\geq n \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi'(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz \\ &\geq -\frac{n}{1-n(\hat{y}_n(\xi)-y(\xi))} \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz.\end{aligned}$$

The last step towards (54) is to replace the interval of integration $[n(\hat{y}_n(\xi)-y(\xi)), 1]$ by $[-1, n(\hat{y}_n(\xi) - y(\xi))]$ and to use (53). To that end observe that we have

$$\int_{-1}^{1} \phi(z)G\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz = \hat{G}_n(\hat{y}_n(\xi)) = \xi = \int_{-1}^{1} \phi(z)\xi dz \quad \text{for all } \xi \in \mathbb{R}. \tag{58}$$

Since $\tilde{G}(\hat{y}_n(\xi) - \frac{z}{n}) = G(\hat{y}_n(\xi) - \frac{z}{n}) - \xi$, we have

$$\int_{-1}^{n(\hat{y}_n(\xi)-y(\xi))} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz = -\int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz,$$

and, accordingly,

$$\begin{aligned}\hat{G}_n'(\hat{y}_n(\xi)) &\geq -\frac{n}{1-n(\hat{y}_n(\xi)-y(\xi))} \int_{n(\hat{y}_n(\xi)-y(\xi))}^{1} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz \\ &= \frac{n}{1-n(\hat{y}_n(\xi)-y(\xi))} \int_{-1}^{n(\hat{y}_n(\xi)-y(\xi))} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz \\ &\geq \frac{n}{1-n(\hat{y}_n(\xi)-y(\xi))} \int_{-1}^{0} \phi(z)\frac{\hat{y}_n(\xi) - \frac{z}{n} - y(\xi)}{\varepsilon}dz \\ &\geq \frac{n(\hat{y}_n(\xi)-y(\xi))}{1-n(\hat{y}_n(\xi)-y(\xi))}\frac{1}{\varepsilon}\int_{-1}^{0} \phi(z)dz \\ &\geq \frac{1}{2\varepsilon},\end{aligned}$$

where we used in the last step that $\frac{1}{2} \leq n(\hat{y}_n(\xi) - y(\xi)) \leq 1$. This finishes the proof of (54).

(b) The Case $y_\xi(\xi) = c > 0$ By assumption $y_\xi(\xi) = c > 0$ and hence for every $\varepsilon > 0$ there exists some $\delta_\varepsilon > 0$ such that

$$(1-\varepsilon)c < \frac{y(\eta) - y(\xi)}{\eta - \xi} < (1+\varepsilon)c \quad \text{whenever } |\eta - \xi| \leq \delta_\varepsilon. \tag{59}$$

Let $0 < \varepsilon < 1$ be fixed and define $\gamma_\varepsilon := (1-\varepsilon)c\delta_\varepsilon$. In addition, let $z \in \mathbb{R}$ be such that $|y(\xi) - z| < \gamma_\varepsilon$. We will first show that $|G(z) - \xi| < \delta_\varepsilon$. Indeed, assume the opposite. Then, due to (59), if $G(z) \geq \xi + \delta_\varepsilon$, we have

$$z = y(G(z)) \geq y(\xi + \delta_\varepsilon) \geq y(\xi) + \delta_\varepsilon(1-\varepsilon)c = y(\xi) + \gamma_\varepsilon,$$

and, if $G(z) \leq \xi - \delta_\varepsilon$, then

$$z = y(G(z)) \leq y(\xi - \delta_\varepsilon) \leq y(\xi) - \delta_\varepsilon(1-\varepsilon)c = y(\xi) - \gamma_\varepsilon.$$

Together, these estimates contradict $|y(\xi) - z| < \gamma_\varepsilon$, and hence prove that $|G(z) - \xi| < \delta_\varepsilon$.

As an immediate consequence, we obtain

$$(1-\varepsilon)c < \frac{y(G(z)) - y(\xi)}{G(z) - \xi} < (1+\varepsilon)c \quad \text{whenever } |z - y(\xi)| < \gamma_\varepsilon,$$

and thus, as $z = y(G(z))$ for all $z \in \mathbb{R}$,

$$\frac{1}{(1+\varepsilon)c} < \frac{G(z) - \xi}{z - y(\xi)} < \frac{1}{(1-\varepsilon)c} \quad \text{whenever } |z - y(\xi)| < \gamma_\varepsilon. \tag{60}$$

In view of the above inequality (60), which will play a key role, we assume without loss of generality that $\frac{2}{n} < \gamma_\varepsilon$ for the rest of this section.

The other main ingredient is to establish that $\lim_{n\to\infty} |n(\hat{y}_n(\xi) - y(\xi))| = 0$. We note here that this fast convergence of $\hat{y}_n(\xi)$ to $y(\xi)$ need not necessarily hold in points ξ where $y_\xi(\xi) = 0$, cf. the remark after this proof. We will only consider the case $\hat{y}_n(\xi) \leq y(\xi)$ and leave the other case, which follows the same lines, to the interested reader. From (58), we can deduce that

$$\begin{aligned}
0 &= \int_{-1}^{1} \phi(z)\tilde{G}\Big(\hat{y}_n(\xi) - \frac{z}{n}\Big)dz \\
&= n\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&= n\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\quad + n\int_{y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi) - z))\tilde{G}(z)dz \\
&\leq n\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{y(\xi)} \phi(n(\hat{y}_n(\xi) - z))\frac{z - y(\xi)}{(1+\varepsilon)c}dz \\
&\quad + n\int_{y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi) - z))\frac{z - y(\xi)}{(1-\varepsilon)c}dz
\end{aligned}$$

$$
\begin{aligned}
&= n \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi)-z)) \frac{z-y(\xi)}{(1+\varepsilon)c} dz \\
&\quad + n \int_{y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi)-z)) \frac{z-y(\xi)}{c} \left[\frac{1}{1-\varepsilon} - \frac{1}{1+\varepsilon} \right] dz \\
&= \frac{\hat{y}_n(\xi)-y(\xi)}{(1+\varepsilon)c} + n \frac{2\varepsilon}{(1-\varepsilon^2)c} \int_{y(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi)-z))(z-y(\xi)) dz \\
&\leq \frac{\hat{y}_n(\xi)-y(\xi)}{(1+\varepsilon)c} + n \frac{2\varepsilon}{(1-\varepsilon^2)c} \int_{\hat{y}_n(\xi)}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi(n(\hat{y}_n(\xi)-z)) \frac{2}{n} dz \\
&\leq \frac{\hat{y}_n(\xi)-y(\xi)}{(1+\varepsilon)c} + \frac{2\varepsilon}{n(1-\varepsilon^2)c},
\end{aligned}
$$

where we used (60). Thus

$$
0 \leq y(\xi) - \hat{y}_n(\xi) \leq \frac{2\varepsilon}{n(1-\varepsilon)}, \tag{61}
$$

which implies that $\lim_{n\to\infty} n(y(\xi) - \hat{y}_n(\xi)) = 0$.

Let us return to the term $\hat{G}_n'(\hat{y}_n(\xi))$. We have from (50) that

$$
\begin{aligned}
\hat{G}_n'(\hat{y}_n(\xi)) &= n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z)) \tilde{G}(z) dz \\
&= n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z)) \frac{z-y(\xi)}{c} dz \\
&\quad + n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z)) \left[\tilde{G}(z) - \frac{z-y(\xi)}{c} \right] dz \\
&= \frac{1}{c} + n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z)) \left[\tilde{G}(z) - \frac{z-y(\xi)}{c} \right] dz.
\end{aligned}
$$

Thus (49) will follow if we can show that the last term on the right-hand side tends to 0 as $n \to \infty$. Now observe that (60) implies that

$$
\left| \tilde{G}(z) - \frac{z-y(\xi)}{c} \right| = \frac{|z-y(\xi)|}{c} \left| 1 - \frac{c\tilde{G}(z)}{z-y(\xi)} \right| \leq \frac{\varepsilon}{(1-\varepsilon)c} |z-y(\xi)|,
$$

and hence

$$\begin{aligned}
&\left|n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z))\Big[\tilde{G}(z)-\frac{z-y(\xi)}{c}\Big]dz\right| \\
&\qquad \le \frac{n^2}{c}\frac{\varepsilon}{1-\varepsilon}\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} |\phi'(n(\hat{y}_n(\xi)-z))||z-y(\xi)|dz \\
&\qquad \le n^2\frac{\varepsilon}{(1-\varepsilon)c}\int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z))(z-y(\xi))dz \\
&\qquad\quad + 2n^2\frac{\varepsilon}{(1-\varepsilon)c}\int_{I_0} |\phi'(n(\hat{y}_n(\xi)-z))||z-y(\xi)|dz \\
&\qquad = \frac{\varepsilon}{(1-\varepsilon)c}\left(1+2n^2\int_{I_0}|\phi'(n(\hat{y}_n(\xi)-z))||z-y(\xi)|dz\right),
\end{aligned}$$

where we used (56).

Recall from (55) that for $z \in I_0$ one has

$$\hat{y}_n(\xi) - |\hat{y}_n(\xi) - y(\xi)| \le z \le \hat{y}_n(\xi) + |\hat{y}_n(\xi) - y(\xi)|$$

and therefore

$$|z - y(\xi)| \le 2|\hat{y}_n(\xi) - y(\xi)| \le \frac{4\varepsilon}{n(1-\varepsilon)},$$

where we used (61).

This implies that

$$\begin{aligned}
\left|\hat{G}_n'(\hat{y}_n(\xi)) - \frac{1}{c}\right| &\le \left|n^2 \int_{\hat{y}_n(\xi)-\frac{1}{n}}^{\hat{y}_n(\xi)+\frac{1}{n}} \phi'(n(\hat{y}_n(\xi)-z))\Big[\tilde{G}(z)-\frac{z-y(\xi)}{c}\Big]dz\right| \\
&\le \frac{\varepsilon}{(1-\varepsilon)c}\left(1+\frac{8\varepsilon}{(1-\varepsilon)}n\int_{I_0}|\phi'(n(\hat{y}_n(\xi)-z))|dz\right) \\
&\le \frac{\varepsilon}{(1-\varepsilon)c}\left(1+\frac{16\varepsilon}{1-\varepsilon}\phi(0)\right).
\end{aligned}$$

Since $\varepsilon > 0$ can be chosen arbitrarily small, this implies that $\hat{G}_n'(\hat{y}_n(\xi)) \to \frac{1}{c}$ as $n \to \infty$.

To summarise, we have in this step shown that $\hat{G}_n'(\hat{y}_n(\xi))$ converges to $1/y_\xi(\xi)$ in every point $\xi \in \mathbb{R}$ where y is differentiable. Thus also $\tilde{h}_n(\xi)$ converges to $h(\xi)$ in all of these points. Together with the fact that $\|\tilde{h}_n\|_{L^1} = \|h\|_{L^1}$ for all n (see (48)), this shows that $\|\tilde{h}_n - h\|_{L^1} \to 0$.

Step 4: Convergence of $\tilde{h}_n$ to h in $L^2(\mathbb{R})$ Recall that

$$\tilde{h}_n = 1 - \hat{y}_{n,\xi} \quad \text{and} \quad h = 1 - y_\xi. \tag{62}$$

Since $\tilde{h}_n$, $\hat{y}_{n,\xi}$, h, and y_ξ all are non-negative, it follows that

$$0 \leq \tilde{h}_n(\xi) \leq 1 \quad \text{and} \quad 0 \leq h(\xi) \leq 1 \quad \text{for all } \xi \in \mathbb{R}.$$

Thus we have

$$\|\tilde{h}_n - h\|_{L^2}^2 \leq \|\tilde{h}_n - h\|_{L^1}.$$

Since we already know that $\tilde{h}_n \to h$ in $L^1(\mathbb{R})$, the claim follows.

Step 5: Convergence of $\hat{y}_{n,\xi} - 1$ to $y_\xi - 1$ in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ By definition we have

$$\tilde{h}_n = 1 - \hat{y}_{n,\xi} \quad \text{and} \quad h = 1 - y_\xi.$$

Since $\tilde{h}_n \to h$ both in $L^1(\mathbb{R})$ and $L^2(\mathbb{R})$ the claim follows.

Step 6: Convergence of $\hat{U}_n$ to U in $L^2(\mathbb{R})$ A proof can be found in [27, Prop. 5.1].

Step 7: Convergence of $\hat{U}_{n,\xi}$ to U_ξ in $L^2(\mathbb{R})$ Let $S = \{\xi \in \mathbb{R} \mid y_\xi(\xi) = 0\}$. Then $U_\xi(\xi) = 0$ for almost all $\xi \in S$, since $U_\xi^2 = h y_\xi$ almost everywhere. Thus we have

$$\|\hat{U}_{n,\xi} - U_\xi\|_{L^2}^2 = \int_S \hat{U}_{n,\xi}^2(\xi) d\xi + \int_{S^c} (\hat{U}_{n,\xi} - U_\xi)^2(\xi) d\xi. \tag{63}$$

From (46) and the fact that $\bar{\tilde{r}}_n \geq 0$, it follows that we have for almost every $\xi \in S$ that

$$\hat{U}_{n,\xi}^2(\xi) \leq \tilde{h}_n(\xi)\hat{y}_{n,\xi}(\xi) = \tilde{h}_n(\xi)(\hat{y}_{n,\xi} - y_\xi)(\xi) = \tilde{h}_n(\xi)(h - \tilde{h}_n)(\xi),$$

and, therefore,

$$\int_S \hat{U}_{n,\xi}^2(\xi) d\xi \leq \|h - \tilde{h}_n\|_{L^1},$$

since $\|\tilde{h}_n\|_{L^\infty} \leq 1$. Thus the first integral in (63) tends to 0 as $n \to \infty$.

As far as the integral over S^c is concerned, the proof of the convergence follows closely the one of $\bar{r}_n \to \bar{r}$ in $L^2(\mathbb{R})$ as $n \to \infty$ in [16, Lemma 6.4], which we reproduce here for completeness. Note that by definition we have $\hat{U}_{n,\xi}(\xi) = u_{n,x}(\hat{y}_n(\xi))\hat{y}_{n,\xi}(\xi)$ and $U_\xi(\xi) = u_x(y(\xi))y_\xi(\xi)$ for almost every $\xi \in S^c$, so that

$$\begin{aligned}\left\|U_{n,\xi}-U_\xi\right\|^2_{L^2(S^c)} &= \left\|(u_{n,x}\circ\hat{y}_n)\hat{y}_{n,\xi}-(u_x\circ y)y_\xi\right\|^2_{L^2(S^c)}\\ &= \int_{S^c}(u_{n,x}\circ\hat{y}_n)^2\hat{y}_{n,\xi}(\hat{y}_{n,\xi}-y_\xi)d\xi\\ &\quad+\int_{S^c}(u_{n,x}\circ\hat{y}_n)\hat{y}_{n,\xi}(u_{n,x}\circ\hat{y}_n-u_{n,x}\circ y)y_\xi d\xi\\ &\quad+\int_{S^c}(u_{n,x}\circ\hat{y}_n)\hat{y}_{n,\xi}(u_{n,x}\circ y-u_x\circ y)y_\xi d\xi\\ &\quad+\int_{S^c}(u_x\circ y)^2y_\xi(y_\xi-\hat{y}_{n,\xi})d\xi\\ &\quad+\int_{S^c}(u_x\circ y)y_\xi(u_x\circ y-u_x\circ\hat{y}_n)\hat{y}_{n,\xi}d\xi\\ &\quad+\int_{S^c}(u_x\circ y)y_\xi(u_x\circ\hat{y}_n-u_{n,x}\circ\hat{y}_n)\hat{y}_{n,\xi}d\xi. \end{aligned}\tag{64}$$

The first and the fourth term have the same structure, and we therefore only treat the first one. Since $(u_{n,x}\circ\hat{y}_n)^2\hat{y}_{n,\xi}\le\tilde{h}_n\le 1$, we have

$$\left\|(u_{n,x}\circ\hat{y}_n)^2\hat{y}_{n,\xi}(\hat{y}_{n,\xi}-y_\xi)\right\|_{L^1(S^c)}\le\left\|y_\xi-\hat{y}_{n,\xi}\right\|_{L^1},$$

and thus this term tends to 0 as $n\to\infty$. In order to investigate the fifth term we will use that $u_x\in L^2(\mathbb{R})$ and therefore there exists for any $\varepsilon>0$ a continuous function $\tilde{l}$ with compact support such that $\left\|u_x-\tilde{l}\right\|_{L^2}\le\varepsilon/(3\max(1,\|u_x\|_{L^2}))$. Thus we can write

$$\begin{aligned}&\left\|(u_x\circ y)y_\xi(u_x\circ y-u_x\circ\hat{y}_n)\hat{y}_{n,\xi}\right\|_{L^1(S^c)}\\ &\qquad\le\left\|(u_x\circ y)y_\xi(u_x\circ y-\tilde{l}\circ y)\hat{y}_{n,\xi}\right\|_{L^1(S^c)}\\ &\qquad\quad+\left\|(u_x\circ y)y_\xi(\tilde{l}\circ y-\tilde{l}\circ\hat{y}_n)\hat{y}_{n,\xi}\right\|_{L^1(S^c)}\\ &\qquad\quad+\left\|(u_x\circ y)y_\xi(\tilde{l}\circ\hat{y}_n-u_x\circ\hat{y}_n)\hat{y}_{n,\xi}\right\|_{L^1(S^c)}\\ &\qquad\le\|u_x\|_{L^2}\left(2\left\|u_x-\tilde{l}\right\|_{L^2}+\left\|\tilde{l}\circ\hat{y}_n-\tilde{l}\circ y\right\|_{L^2}\right).\end{aligned}\tag{65}$$

Here we have used in the last inequality that both y_ξ and $\hat{y}_{n,\xi}$ are non-negative and bounded above by 1. Since $\hat{y}_n-y\to 0$ in $L^\infty(\mathbb{R})$ and $\tilde{l}$ is continuous with compact support, we obtain by Lebesgue's dominated convergence theorem that $\tilde{l}\circ\hat{y}_n\to\tilde{l}\circ y$ in $L^2(\mathbb{R})$. In particular, we can choose n large enough so that

$$\left\|(u_x\circ y)y_\xi(u_x\circ y-u_x\circ\hat{y}_n)\hat{y}_{n,\xi}\right\|_{L^1(S^c)}\le\varepsilon.$$

Since ε can be chosen arbitrarily small, we obtain in particular that

$$\lim_{n\to\infty} \left\| (u_x \circ y) y_\xi (u_x \circ y - u_x \circ \hat{y}_n) \hat{y}_{n,\xi} \right\|_{L^1(S^c)} = 0. \tag{66}$$

For the convergence of the second term, we estimate, using again that y_ξ is bounded by 1,

$$\begin{aligned}
&\left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_{n,x} \circ \hat{y}_n - u_{n,x} \circ y) y_\xi \right\|_{L^1(S^c)} \\
&\quad \le \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_{n,x} \circ \hat{y}_n - u_x \circ \hat{y}_n) y_\xi \right\|_{L^1(S^c)} \\
&\qquad + \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_x \circ \hat{y}_n - u_x \circ y) y_\xi \right\|_{L^1(S^c)} \\
&\qquad + \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_x \circ y - u_{n,x} \circ y) y_\xi \right\|_{L^1(S^c)} \\
&\quad \le \left(\int_{S^c} (u_{n,x} \circ \hat{y}_n)^2 \hat{y}_{n,\xi} \, d\xi \right)^{1/2} \left(\int_{S^c} (u_{n,x} \circ \hat{y}_n - u_x \circ \hat{y}_n)^2 \hat{y}_{n,\xi} \, d\xi \right)^{1/2} \\
&\qquad + \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_x \circ \hat{y}_n - u_x \circ y) y_\xi \right\|_{L^1(S^c)} \\
&\qquad + \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} \right\|_{L^2} \left\| (u_x \circ y - u_{n,x} \circ y) y_\xi \right\|_{L^2}.
\end{aligned}$$

The first and third term in this last estimate tend to zero because $u_{n,x} \to u_x \in L^2(\mathbb{R})$ and both y_ξ and $\hat{y}_{n,\xi}$ are uniformly bounded, and for the convergence of the second term we can use the same method as in (65). Thus also the second term in (64) tends to zero. As far as the third (and, similarly, the last) term in (64) is concerned, we have that

$$\begin{aligned}
&\left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} (u_{n,x} \circ y - u_x \circ y) y_\xi \right\|_{L^1(S^c)} \\
&\qquad \le \left\| (u_{n,x} \circ \hat{y}_n) \hat{y}_{n,\xi} \right\|_{L^2(S^c)} \left\| (u_{n,x} \circ y - u_x \circ y) y_\xi \right\|_{L^2(S^c)} \\
&\qquad \le \left\| u_{n,x} \right\|_{L^2} \left\| u_{n,x} - u_x \right\|_{L^2},
\end{aligned}$$

which again tends to zero since by assumption $u_{n,x} \to u_x \in L^2(\mathbb{R})$. Hence all terms in (64) tend to 0 as $n \to \infty$ and therefore $\hat{U}_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$.

Step 8: Convergence of $\bar{\hat{r}}_n$ to Zero in $L^2(\mathbb{R})$ By construction, we have $\bar{\hat{r}}_n \ge 0$ since $\bar{\rho}_n \ge 0$ and $\hat{y}_{n,\xi} \ge 0$. Hence, by (26c), (29b), (46), and (62), we have

$$\begin{aligned}
\left\| \bar{\hat{r}}_n \right\|_{L^2}^2 &\le \left\| \bar{\hat{r}}_n^2 + \frac{2}{n} \bar{\hat{r}}_n \hat{y}_{n,\xi} \right\|_{L^1} \\
&= \left\| \tilde{h}_n \hat{y}_{n,\xi} - h y_\xi - \hat{U}_{n,\xi}^2 + U_\xi^2 \right\|_{L^1} \\
&= \left\| \tilde{h}_n - h - \tilde{h}_n^2 + h^2 - \hat{U}_{n,\xi}^2 + U_\xi^2 \right\|_{L^1}
\end{aligned}$$

$$\begin{aligned}&\leq 3\left\|h-\tilde{h}_n\right\|_{L^1}+\left(\left\|\hat{U}_{n,\xi}\right\|_{L^2}+\left\|U_\xi\right\|_{L^2}\right)\left\|\hat{U}_{n,\xi}-U_\xi\right\|_{L^2}\\&\leq 3\left\|h-\tilde{h}_n\right\|_{L^1}+2\|\mu\|\left\|\hat{U}_{n,\xi}-U_\xi\right\|_{L^2}.\end{aligned}$$

Since $\tilde{h}_n \to h$ in $L^1(\mathbb{R})$ and $\hat{U}_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$, the above estimate implies that $\bar{\bar{r}}_n \to 0$ in $L^2(\mathbb{R})$ as $n \to \infty$.

Step 9: Convergence of $\hat{h}_n$ to h in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ According to (29b) and (31b), we have

$$\left\|\hat{h}_n-h\right\|_{L^2} \leq \frac{2}{n}\left\|\bar{\bar{r}}_n\right\|_{L^2}+\left\|\hat{y}_{n,\xi}-y_\xi\right\|_{L^2}.$$

Since $\bar{\bar{r}}_n \to 0$ and $\hat{y}_{n,\xi}-y_\xi \to 0$ in $L^2(\mathbb{R})$, this inequality implies that $\hat{h}_n \to h$ in $L^2(\mathbb{R})$. As far as the $L^1(\mathbb{R})$ convergence is concerned, observe that

$$\begin{aligned}\hat{h}_n-h&=\hat{h}_n(\hat{y}_{n,\xi}+\tilde{h}_n)-h(y_\xi+h)\\&=\hat{h}_n\hat{y}_{n,\xi}-hy_\xi+\hat{h}_n\tilde{h}_n-h^2\\&=\hat{U}_{n,\xi}^2-U_\xi^2+\bar{\bar{r}}_n^2+\hat{h}_n\tilde{h}_n-h^2.\end{aligned}$$

Here the first equality follows from (29b) and (45), and the last equality follows from (26c) and (46). Thus

$$\begin{aligned}\left\|\hat{h}_n-h\right\|_{L^1} \leq{}& \left\|\hat{U}_{n,\xi}+U_\xi\right\|_{L^2}\left\|\hat{U}_{n,\xi}-U_\xi\right\|_{L^2}+\left\|\bar{\bar{r}}_n\right\|_{L^2}^2\\&+\left\|\tilde{h}_n\right\|_{L^2}\left\|\hat{h}_n-h\right\|_{L^2}+\|h\|_{L^2}\left\|\tilde{h}_n-h\right\|_{L^2},\end{aligned}$$

which implies that $\hat{h}_n \to h$ in $L^1(\mathbb{R})$.

Step 10: Convergence of X_n to X in E So far we have shown that $\hat{X}_n \to X$ in E. In addition we showed in Step 1 for all $n \in \mathbb{N}$ that we can write $\hat{X}_n = X_n \circ g_n$ with $g_n \in G_\kappa$ for some $\kappa > 0$ independent of n, or, equivalently that $\hat{X}_n \in \mathscr{F}_\kappa$ for all $n \in \mathbb{N}$. Moreover, it is known, see, e.g., [16, Lemma 4.6], that the mapping $\Gamma\colon \mathscr{F}_\kappa \to \mathscr{F}_0$ defined in (28) is continuous. Thus we also have that $X_n = \Gamma(\hat{X}_n) \to X$ in E, which completes the proof.

Remark 4 A closer look at the proof of Theorem 3 reveals that we showed that for every $\xi \in \mathbb{R}$ where y is differentiable and $y_\xi > 0$, we have

$$\lim_{n\to\infty} n(\hat{y}_n(\xi)-y(\xi))=0.$$

As the following example illustrates, we cannot expect this convergence to hold for almost every $\xi \in \mathbb{R}$ such that $y_\xi(\xi)=0$.

Consider the following initial data for the CH equation which corresponds to a symmetric/antisymmetric peakon-antipeakon solution, which vanishes at breaking time $t = 0$, i.e.,

$$u(x) = 0 \text{ for all } x \in \mathbb{R} \quad \text{and} \quad F(x) = \mu((-\infty, x]) = \begin{cases} 0, & x < 0, \\ \alpha, & 0 \leq x, \end{cases}$$

where $\alpha > 0$. Then

$$y(\xi) = \begin{cases} \xi, & \xi < 0, \\ 0, & 0 \leq \xi \leq \alpha, \\ \xi - \alpha, & \alpha < \xi. \end{cases}$$

and especially $y_\xi(\xi) = 0$ for all $\xi \in (0, \alpha)$. For the approximating sequence we know that

$$\hat{y}_n(\xi) + \hat{F}_n(\hat{y}_n(\xi)) = \xi \quad \text{for all } \xi \in \mathbb{R},$$

where

$$\hat{F}_n(x) = \int_{\mathbb{R}} n\phi(n(x - y))F(y)dy.$$

We are going to show that $\lim_{n\to\infty} n(\hat{y}_n(\xi) - y(\xi)) \neq 0$ for any $\xi \in (0, \alpha)$ except $\xi = \frac{\alpha}{2}$.

Indeed, if we denote

$$\Phi(x) := \int_{-\infty}^{x} \phi(y)\, dy,$$

then we see that

$$\hat{F}_n(x) = \int_{\mathbb{R}} n\phi(n(x - y))F(y)dy = \alpha \int_0^\infty n\phi(n(x - y))dy = \alpha\Phi(nx)$$

for all $x \in \mathbb{R}$. Now assume that $0 < \xi < \alpha$. Then $y(\xi) = 0$ and thus

$$\begin{aligned} \xi = \hat{y}_n(\xi) + \hat{F}_n(\hat{y}_n(\xi)) &= \hat{y}_n(\xi) + \alpha\Phi(n\hat{y}_n(\xi)) \\ &= \hat{y}_n(\xi) - y(\xi) + \alpha\Phi(n(\hat{y}_n(\xi) - y(\xi))). \end{aligned}$$

In Step 2 in the proof of Theorem 3 we have shown that $\hat{y}_n(\xi) \to y(\xi)$. Taking the limit $n \to \infty$ in the previous equation therefore implies that

$$\frac{\xi}{\alpha} = \lim_{n\to\infty} \Phi(n(\hat{y}_n(\xi) - y(\xi))).$$

Using that $\left|n(\hat{y}_n(\xi) - y(\xi))\right| < 1$ for all n and that Φ is continuously invertible on $(-1, 1)$, it follows that

$$\lim_{n\to\infty} n(\hat{y}_n(\xi) - y(\xi)) = \Phi^{-1}(\xi/\alpha).$$

Since $\Phi(0) = 1/2$ and therefore $\Phi^{-1}(1/2) = 0$, this shows in particular that the sequence $n(\hat{y}_n(\xi) - y(\xi))$ only converges to 0 for $\xi = \alpha/2$.

4 Convergence in Lagrangian Coordinates Implies Convergence in Eulerian Coordinates

In the previous two sections, we saw that we can approximate any given initial data (u, μ) for the CH equation by a sequence of smooth initial data (u_n, ρ_n, μ_n) for the 2CH system where the measures μ_n are purely absolutely continuous. Afterwards we saw that this convergence in Eulerian coordinates is transported, via the mapping L, to convergence in Lagrangian coordinates.

In this section we consider the case when we are given a sequence $X_n \in \mathscr{F}_0$ and $X \in \mathscr{F}_0$, such that $X_n \to X$ in E. Does $M(X_n) \to M(X)$ in some sense in Eulerian coordinates? Here $M\colon \mathscr{F}_0 \to \mathscr{D}$ denotes the mapping from Lagrangian to Eulerian coordinates, which is defined as follows.

Definition 4 ([16, Thm. 4.10]) Given any element $X = (y, U, h, r) \in \mathscr{F}_0$, we define (u, ρ, μ) as follows[3]

$$u(x) = U(\xi) \text{ for any } \xi \text{ such that } x = y(\xi), \tag{67a}$$

$$\mu = y_\#(h(\xi)\, d\xi), \tag{67b}$$

$$\bar{\rho}(x)\, dx = y_\#(\bar{r}(\xi)\, d\xi), \tag{67c}$$

$$\rho(x) = k + \bar{\rho}(x), \tag{67d}$$

where k is implicitly given through the relation $r(\xi) = \bar{r}(\xi) + k y_\xi(\xi)$ for all $\xi \in \mathbb{R}$. We have that (u, ρ, μ) belongs to $\mathscr{D}$ and, in particular, that the measure $y_\#(\bar{r}(\xi)\, d\xi)$ is absolutely continuous. We denote by $M\colon \mathscr{F}_0 \to \mathscr{D}$ the mapping which to any X in $\mathscr{F}_0$ associates the element $(u, \rho, \mu) \in \mathscr{D}$ as given by (67).

[3]Here we denote by $y_\#(h\, d\xi)$ the *push-forward* of the measure $h\, d\xi$ by y, defined by $y_\#(h\, d\xi)(A) = \int_{y^{-1}(A)} h(\xi) d\xi$ for all Borel sets $A \subset \mathbb{R}$.

We recall from Definition 3 that for any $(y, U, h, r) \in \mathscr{F}_0$ we have that $y_\xi \geq 0$ and $U_\xi = 0$ whenever $y_\xi = 0$, or, in other words, that U is constant whenever the increasing function y is constant. As a consequence, the value $U(\xi)$ is uniquely determined by $y(\xi)$, which means that the definition of the function u in (67a) is independent of the choice of ξ satisfying $x = y(\xi)$. Also, the fact that y is Lipschitz continuous (see (26a)) implies that the push-forward of the absolutely continuous measure $\bar{r}(\xi)\, d\xi$ under y is again absolutely continuous, cf. [16, Thm. 4.10].

Moreover, we consider the following notion of sequential convergence on $\mathscr{D}$.

Definition 5 We say that a sequence $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ converges to $(u, \rho, \mu) \in \mathscr{D}$ as $n \to \infty$ if[4]

$$u_n \to u \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R}), \tag{68a}$$

$$u_{n,x} \rightharpoonup u_x, \tag{68b}$$

$$\bar{\rho}_n \rightharpoonup \bar{\rho}, \tag{68c}$$

$$k_n \to k, \tag{68d}$$

$$\int_{\mathbb{R}} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx \to \int_{\mathbb{R}} \frac{u_x^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)} dx, \tag{68e}$$

$$\int_{\mathbb{R}} \frac{\bar{\rho}_n^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx \to \int_{\mathbb{R}} \frac{\bar{\rho}^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)} dx, \tag{68f}$$

$$F_n(x) \to F(x) \text{ for every } x \text{ at which } F \text{ is continuous}, \tag{68g}$$

$$F_n(\infty) \to F(\infty), \tag{68h}$$

where $F_n(x) = \mu_n((-\infty, x])$ for all $n \in \mathbb{N}$ and $F(x) = \mu((-\infty, x])$.

With this definition, the convergence result can be stated as follows.

Theorem 4 *Given a sequence $X_n = (y_n, U_n, h_n, r_n) \in \mathscr{F}_0$ and $X = (y, U, h, r) \in \mathscr{F}_0$ such that $X_n \to X$ in E as $n \to \infty$, then $(u_n, \rho_n, \mu_n) = M(X_n)$ converges to $(u, \rho, \mu) = M(X)$ in the sense of Definition 5.*

Proof The proof is divided into 8 steps for convenience.

Step 1: Convergence of u_n to u in $L^\infty(\mathbb{R})$ For a proof we refer the interested reader to [16, Thm. 6.5].

Step 2: Convergence of u_n to u in $L^2(\mathbb{R})$ If we can show that the assumptions of the Radon–Riesz theorem are fulfilled, see, e.g., [1, Thm. 1.37], the claim follows. Thus we have to show that $\|u_n\|_{L^2} \to \|u\|_{L^2}$ and that u_n converges weakly to u as $n \to \infty$.

[4] We say that $f_n \rightharpoonup f$ if $\int_{\mathbb{R}} f_n(x) g(x) dx \to \int_{\mathbb{R}} f(x) g(x) dx$ for every $g \in L^2(\mathbb{R})$.

A straightforward computation using (67a) yields

$$\|u_n\|_{L^2}^2 = \int_{\mathbb{R}} (U_n^2 - U^2) y_{n,\xi}(\xi) d\xi + \int_{\mathbb{R}} U^2 (y_{n,\xi} - y_\xi)(\xi) d\xi + \|u\|_{L^2}^2 , \tag{69}$$

where we have used that $U^2 y_\xi(\xi) = 0$ whenever $y_\xi(\xi) = 0$, and similarly that $U_n^2 y_{n,\xi}(\xi) = 0$ whenever $y_{n,\xi}(\xi) = 0$. Applying the Cauchy–Schwarz inequality to the first and second term on the right-hand side of (69) yields that

$$\begin{aligned} &\left| \int_{\mathbb{R}} (U_n^2 - U^2) y_{n,\xi}(\xi) d\xi \right| + \left| \int_{\mathbb{R}} U^2 (y_{n,\xi} - y_\xi)(\xi) d\xi \right| \\ &\qquad \leq (\|U_n - U\|_{L^2} + 2\|U\|_{L^2}) \|U_n - U\|_{L^2} \\ &\qquad + \|U\|_{L^\infty} \|U\|_{L^2} \left\| y_{n,\xi} - y_\xi \right\|_{L^2} , \end{aligned}$$

where we used that $0 \leq y_{n,\xi} \leq 1$. Since $U_n \to U$ in $L^2(\mathbb{R})$ and $\zeta_{n,\xi} \to \zeta_\xi$ in $L^2(\mathbb{R})$, we obtain from (69) that $\|u_n\|_{L^2} \to \|u\|_{L^2}$ as $n \to \infty$.

Since $C_c^\infty(\mathbb{R})$ is dense in $L^2(\mathbb{R})$ and $\|u_n\|_{L^2} \to \|u\|_{L^2}$, it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} u_n \psi(x) dx = \int_{\mathbb{R}} u \psi(x) dx \tag{70}$$

for all $\psi \in C_c^\infty(\mathbb{R})$. This however follows immediately, as

$$\left| \int_{\mathbb{R}} (u_n - u) \psi(x) dx \right| \leq \|u_n - u\|_{L^\infty} \|\psi\|_{L^1} \to 0$$

according to Step 1.

Step 3: Convergence of $F_n(x)$ to $F(x)$ for all x at Which $F(x)$ is Continuous According to [11, Props. 7.19 and 8.17], this is equivalent to showing that

$$\lim_{n\to\infty} \int_{\mathbb{R}} \psi(x) d\mu_n(x) = \int_{\mathbb{R}} \psi(x) d\mu(x) \tag{71}$$

for all $\psi \in C_c^\infty(\mathbb{R})$. It follows from (67b) that

$$\int_{\mathbb{R}} \psi d\mu_n(x) = \int_{\mathbb{R}} \psi(y_n(\xi)) h_n(\xi) d\xi, \tag{72}$$

and

$$\int_{\mathbb{R}} \psi d\mu(x) = \int_{\mathbb{R}} \psi(y(\xi)) h(\xi) d\xi. \tag{73}$$

Since $y_n - \mathrm{id} \to y - \mathrm{id}$ in $L^\infty(\mathbb{R})$, the support of $\psi \circ y_n$ is contained in some compact set which can be chosen independently of n, and, from Lebesgue's dominated convergence theorem, we have that $\psi \circ y_n \to \psi \circ y$ in $L^2(\mathbb{R})$. Hence, since $h_n \to h$ in $L^2(\mathbb{R})$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \psi(y_n(\xi)) h_n(\xi) d\xi = \int_{\mathbb{R}} \psi(y(\xi)) h(\xi) d\xi,$$

and (71) follows from (72) and (73).

Note that, in particular, $\mu_n(\mathbb{R}) \to \mu(\mathbb{R})$ as $n \to \infty$, since $\|h_n\|_{L^1} \to \|h\|_{L^1}$ by assumption. Moreover,

$$\mu_{n,\mathrm{ac}} = (u_{n,x}^2 + \bar{\rho}_n^2) dx \quad \text{and} \quad \mu_{\mathrm{ac}} = (u_x^2 + \bar{\rho}^2) dx,$$

which implies

$$\|u_{n,x}\|_{L^2}^2 + \|\bar{\rho}_n\|_{L^2}^2 \le \mu_n(\mathbb{R}) \quad \text{and} \quad \|u_x\|_{L^2}^2 + \|\bar{\rho}\|_{L^2}^2 \le \mu(\mathbb{R}), \tag{74}$$

and hence $u_{n,x}$, u_x, $\bar{\rho}_n$ and $\bar{\rho}$ belong to $L^2(\mathbb{R})$.

Step 4: Weak Convergence of $u_{n,x}$ to u_x Since $C_c^\infty(\mathbb{R})$ is dense in $L^2(\mathbb{R})$ and $\|u_{n,x}\|_{L^2}$ and $\|u_x\|_{L^2}$ can be uniformly bounded according to (74), it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} u_{n,x}(x) \psi(x) dx = \int_{\mathbb{R}} u_x(x) \psi(x) dx$$

for all $\psi \in C_c^\infty(\mathbb{R})$. To that end, observe that

$$\int_{\mathbb{R}} u_{n,x}(x) \psi(x) dx = \int_{\mathbb{R}} U_{n,\xi}(\xi) \psi(y_n(\xi)) d\xi,$$

since $U_{n,\xi}(\xi) = 0$ for all $\xi \in \mathbb{R}$ such that $y_{n,\xi}(\xi) = 0$, and

$$\int_{\mathbb{R}} u_x(x) \psi(x) dx = \int_{\mathbb{R}} U_\xi(\xi) \psi(y(\xi)) d\xi.$$

Thus it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} U_{n,\xi}(\xi) \psi(y_n(\xi)) d\xi = \int_{\mathbb{R}} U_\xi(\xi) \psi(y(\xi)) d\xi, \tag{75}$$

for all $\psi \in C_c^\infty(\mathbb{R})$. By assumption we have that $U_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$ and $y_n - \mathrm{id} \to y - \mathrm{id}$ in $L^\infty(\mathbb{R})$ and hence the support of $\psi(y_n(\xi))$ and $\psi(y(\xi))$ is contained in some compact set that can be chosen independent of n. Thus

$\psi(y_n(\xi)) \to \psi(y(\xi))$ in $L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$, and Lebesgue's dominated convergence theorem implies (75).

Step 5: Weak Convergence of $\bar{\rho}_n$ to $\bar{\rho}$ The argument closely follows the argument for the weak convergence of $u_{n,x}$ to u_x in Step 4.

Step 6: Convergence of $\int_{\mathbb{R}} \frac{u_{n,x}^2(x)}{1+u_{n,x}^2(x)+\bar{\rho}_n^2(x)} dx$ to $\int_{\mathbb{R}} \frac{u_x^2(x)}{1+u_x^2(x)+\bar{\rho}^2(x)} dx$ Let $S = \{\xi \in \mathbb{R} \mid y_\xi(\xi) = 0\}$ and $S_n = \{\xi \in \mathbb{R} \mid y_{n,\xi}(\xi) = 0\}$. Furthermore, let $B_n = y_n(S)$. Then we claim that $\text{meas}(B_n) \to 0$ as $n \to \infty$.

By definition, we have that $S = \{\xi \in \mathbb{R} \mid y_\xi(\xi) = 0\} = \{\xi \in \mathbb{R} \mid h(\xi) = 1\}$, which implies that $\text{meas}(S) \le \|h\|_{L^1}$. Thus

$$\begin{aligned}
\text{meas}(B_n) = \text{meas}(y_n(S)) &= \int_{y_n(S)} dx \\
&= \int_S y_{n,\xi}(\xi) d\xi = \int_S \big(y_{n,\xi}(\xi) - y_\xi(\xi)\big) d\xi \\
&\le \text{meas}(S)^{1/2} \left\| y_{n,\xi} - y_\xi \right\|_{L^2} \\
&\le \|h\|_{L^1}^{1/2} \left\| y_{n,\xi} - y_\xi \right\|_{L^2}.
\end{aligned}$$

By assumption $y_{n,\xi} - y_\xi \to 0$ in $L^2(\mathbb{R})$, and hence $\text{meas}(B_n)$ tends to 0 as $n \to \infty$. Moreover,

$$U_{n,\xi}(\xi) = 0 \text{ for } \xi \in S_n \quad \text{and} \quad U_\xi(\xi) = 0 \text{ for } \xi \in S.$$

As far as $y_{n,\xi}(\xi)$ and $y_\xi(\xi)$ are concerned, we have the representations

$$y_{n,\xi}(\xi) = \frac{1}{1 + u_{n,x}^2(y_n(\xi)) + \bar{\rho}_n^2(y_n(\xi))} \quad \text{for almost every } \xi \in S_n^c$$

and

$$y_\xi(\xi) = \frac{1}{1 + u_x^2(y(\xi)) + \bar{\rho}^2(y(\xi))} \quad \text{for almost every } \xi \in S^c.$$

This means, in particular, that

$$\begin{aligned}
\int_{S^c} U_{n,\xi}^2(\xi) d\xi &= \int_{S^c \cap S_n^c} U_{n,\xi}^2(\xi) d\xi \\
&= \int_{S^c \cap S_n^c} u_{n,x}^2(y_n(\xi)) y_{n,\xi}^2(\xi) d\xi \\
&= \int_{S^c \cap S_n^c} \frac{u_{n,x}^2(y_n(\xi))}{1 + u_{n,x}^2(y_n(\xi)) + \bar{\rho}_n^2(y_n(\xi))} y_{n,\xi}(\xi) d\xi
\end{aligned}$$

$$= \int_{B_n^c \cap y_n(S_n^c)} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx$$
$$= \int_{B_n^c} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx,$$

since y_n is Lipschitz continuous and therefore $\operatorname{meas}(y_n(S_n)) = 0$ for all $n \in \mathbb{N}$. Similarly, one obtains

$$\int_{S^c} U_\xi^2(\xi) d\xi = \int_{\mathbb{R}} \frac{u_x^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)} dx,$$

as $\operatorname{meas}(y(S)) = 0$. Since $U_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$ we find that

$$\lim_{n\to\infty} \int_{B_n^c} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx = \int_{\mathbb{R}} \frac{u_x^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)} dx.$$

Furthermore, note that $\frac{u_{n,x}^2(x)}{1+u_{n,x}^2(x)+\bar{\rho}_n^2(x))}$ is uniformly bounded by 1 for all $n \in \mathbb{N}$ and $x \in \mathbb{R}$. This means, in particular, that

$$\int_{B_n} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx \leq \operatorname{meas}(B_n),$$

and thus the term on the left-hand side converges to 0 as $n \to \infty$ since $\operatorname{meas}(B_n) \to 0$. Thus we get

$$\lim_{n\to\infty} \int_{\mathbb{R}} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx = \int_{\mathbb{R}} \frac{u_x^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)} dx.$$

Step 7: Convergence of $\int_{\mathbb{R}} \frac{\bar{\rho}_n^2(x)}{1+u_{n,x}^2(x)+\bar{\rho}_n^2(x)} dx$ to $\int_{\mathbb{R}} \frac{\bar{\rho}^2(x)}{1+u_x^2(x)+\bar{\rho}^2(x)} dx$ The argument is similar to the one in Step 6.

Step 8: Convergence of k_n to k By definition we have

$$r(\xi) = \bar{r}(\xi) + k y_\xi(\xi)$$

and

$$r_n(\xi) = \bar{r}_n(\xi) + k_n y_{n,\xi}(\xi).$$

By assumption X_n converges to X in E, and thus according to (23), we infer that $k_n \to k$.

Remark 5 Note that the convergence in Lagrangian coordinates implies in particular that $\bar{\rho}_n$ converges to $\bar{\rho}$ weakly. Thus, in the special case $k = 0$ and $r(\xi) = \bar{r}(\xi) = 0$ for all $\xi \in \mathbb{R}$, we infer that $\bar{\rho}_n$ converges weakly to zero and $k_n \to 0$. Thus $\rho(x) = 0$ for all $x \in \mathbb{R}$, and hence $(u, \rho, \mu) = (u, 0, \mu)$ belongs to the set of Eulerian coordinates for the CH equation. Thus the sequence $\bar{\rho}_n$ in Theorem 1 converges to zero in the weak sense, since all the assumptions in Theorem 4 are satisfied due to Theorem 3.

5 Convergence in Eulerian Coordinates Implies Convergence in Lagrangian Coordinates

In this section we want to show that convergence in Eulerian coordinates implies convergence in Lagrangian coordinates. Due to the definition of Eulerian coordinates, one might guess that it is natural to impose only the condition $u_n \to u$ in $H^1(\mathbb{R})$. However, due to Theorem 4 we will require a somewhat stronger mode of convergence for $u_{n,x}$ to u_x, which in the end yields an equivalence between convergence in Eulerian and Lagrangian coordinates.

Theorem 5 *Given a sequence $(u_n, \rho_n, \mu_n) \in \mathscr{D}$ and some $(u, \rho, \mu) \in \mathscr{D}$ such that (u_n, ρ_n, μ_n) converges to (u, ρ, μ) as $n \to \infty$ in the sense of Definition 5, let $X_n = (y_n, U_n, h_n, r_n) = L((u_n, \rho_n, \mu_n)) \in \mathscr{F}_0$ and $X = (y, U, h, r) = L((u, \rho, \mu)) \in \mathscr{F}_0$. Then $X_n \to X$ in E as $n \to \infty$.*

Proof The proof is divided into 7 steps.

Step 1: The Sequence y_n Converges Pointwise to y Denote $D = \{\xi \in \mathbb{R} \mid F \text{ is discontinuous at } y(\xi)\}$. By construction we have for all $\xi \in D^c$ that

$$F(y(\xi)) + y(\xi) = \xi, \tag{76}$$

and, in particular,

$$F(y(\xi)) = \mu((-\infty, y(\xi)]) = \mu((-\infty, y(\xi))). \tag{77}$$

As far as $y_n(\xi)$ is concerned, we have by (29a) that

$$\mu_n((-\infty, y_n(\xi))) + y_n(\xi) \leq \xi \leq \mu_n((-\infty, y_n(\xi)]) + y_n(\xi) = F_n(y_n(\xi)) + y_n(\xi). \tag{78}$$

To show the pointwise convergence of $y_n(\xi)$ to $y(\xi)$ for $\xi \in D^c$, we have to distinguish two cases.
For $y_n(\xi) \leq y(\xi)$, combining (76)–(78) yields

$$\begin{aligned} F(y(\xi)) + y(\xi) = \xi &\leq F_n(y_n(\xi)) + y_n(\xi) \\ &= F_n(y(\xi)) - \mu_n((y_n(\xi), y(\xi)]) + y_n(\xi). \end{aligned}$$

Thus

$$0 \leq y(\xi) - y_n(\xi) + \mu_n((y_n(\xi), y(\xi)]) \leq F_n(y(\xi)) - F(y(\xi)).$$

For $y(\xi) < y_n(\xi)$, combining again (76)–(78) yields

$$\begin{aligned}\mu_n((-\infty, y_n(\xi))) + y_n(\xi) &= F_n(y(\xi)) + \mu_n((y(\xi), y_n(\xi))) + y_n(\xi) \\ &\leq \xi = F(y(\xi)) + y(\xi).\end{aligned}$$

Hence

$$0 \leq y_n(\xi) - y(\xi) + \mu_n((y(\xi), y_n(\xi))) \leq F(y(\xi)) - F_n(y(\xi)).$$

Since μ_n and μ are positive finite Radon measures for all $n \in \mathbb{N}$, we get that

$$|y_n(\xi) - y(\xi)| \leq |F_n(y(\xi)) - F(y(\xi))|, \quad \xi \in D^c. \tag{79}$$

Since by assumption $\xi \in D^c$, we have that F is continuous at the point $y(\xi)$, which in turn implies that $|F_n(y(\xi)) - F(y(\xi))|$ converges to zero. Thus $y_n(\xi) \to y(\xi)$ for every $\xi \in D^c$.

For $\xi \in D$, we argue as follows. Any point x at which F is discontinuous in Eulerian coordinates, corresponds to a maximal interval $[\xi_1, \xi_2]$ in Lagrangian coordinates such that $y(\xi) = x$ for all $\xi \in [\xi_1, \xi_2]$ and $\xi_2 - \xi_1 = \mu(\{x\})$. In particular, there exists an increasing sequence $\xi_i \in D^c$ such that ξ_i converges to ξ_1. We may write

$$y_n(\xi_1) - y(\xi_1) = (y_n(\xi_1) - y_n(\xi_i)) + (y_n(\xi_i) - y(\xi_i)) + (y(\xi_i) - y(\xi_1)).$$

Because y_n and y are Lipschitz continuous with Lipschitz constant at most 1 due to (29a), we can thus estimate

$$|y_n(\xi_1) - y(\xi_1)| \leq 2\,|\xi_i - \xi_1| + |y_n(\xi_i) - y(\xi_i)|\,.$$

Since y is continuous at ξ_i (cf. (79)), the second term on the right-hand side tends to 0 as $n \to \infty$, which shows that $|y_n(\xi_1) - y(\xi_1)|$ can be made arbitrarily small and thus $y_n(\xi_1) \to y(\xi_1)$. A similar argument shows that $y_n(\xi_2) \to y(\xi_2)$ by taking a decreasing sequence $\xi_i \in D^c$ such that ξ_i converges to ξ_2.

We can now show that $y_n(\xi) \to y(\xi)$ for all $\xi \in [\xi_1, \xi_2]$, By definition y_n is an increasing function, and $y(\xi)$ is constant on $[\xi_1, \xi_2]$. Thus

$$|y_n(\xi) - y(\xi)| \leq \max\big(|y_n(\xi_1) - y(\xi_1)|, |y_n(\xi_2) - y(\xi_2)|\big) \quad \text{for all } \xi \in [\xi_1, \xi_2].$$

Since both $|y_n(\xi_1) - y(\xi_1)|$ and $|y_n(\xi_2) - y(\xi_2)|$ tend to zero as $n \to \infty$, it follows immediately that $y_n(\xi) \to y(\xi)$ for all $\xi \in [\xi_1, \xi_2]$. Thus $y_n \to y$ pointwise.

Step 2: Convergence of h_n to h and $\zeta_{n,\xi}$ to ζ_ξ in $L^2(\mathbb{R})$ By definition, we have that $X_n \in \mathscr{F}_0$ for all $n \in \mathbb{N}$ and $X \in \mathscr{F}_0$. Thus $H_n(\xi) = \xi - y_n(\xi)$, $n \in \mathbb{N}$, and $H(\xi) = \xi - y(\xi)$ for almost every $\xi \in \mathbb{R}$. As $y_n(\xi)$ converges pointwise to $y(\xi)$, we infer that $H_n(\xi) \to H(\xi)$ pointwise almost everywhere as $n \to \infty$. Moreover, $H_n(\xi)$, $n \in \mathbb{N}$, and $H(\xi)$ are all continuous, and hence we conclude that, actually, we have pointwise convergence of $H_n(\xi) \to H(\xi)$ for every $\xi \in \mathbb{R}$. Moreover, since

$$H_n(\xi) = \int_{-\infty}^{\xi} h_n(\eta)d\eta \quad \text{and} \quad H(\xi) = \int_{-\infty}^{\xi} h(\eta)d\eta,$$

h_n and h can be seen as positive finite Radon measures, and hence

$$\lim_{n\to\infty} \int_{\mathbb{R}} h_n(\xi)\psi(\xi)d\xi = \int_{\mathbb{R}} h(\xi)\psi(\xi)d\xi \tag{80}$$

for all $\psi \in C_c^\infty(\mathbb{R})$ according to [11, Props. 7.19 and 8.17]. If we can show that $\|h_n\|_{L^2} \to \|h\|_{L^2}$, (80) will remain true for all $\psi \in L^2(\mathbb{R})$ by a density argument and hence all assumptions of the Radon–Riesz theorem are satisfied. Thus $h_n \to h$ in $L^2(\mathbb{R})$, provided $\|h_n\|_{L^2} \to \|h\|_{L^2}$.

In order to show this convergence, observe that

$$\|h_n\|_{L^1} = \|H_n\|_{L^\infty} = \mu_n(\mathbb{R}) \to \mu(\mathbb{R}) = \|H\|_{L^\infty} = \|h\|_{L^1} \quad \text{as } n \to \infty. \tag{81}$$

Since $X_n \in \mathscr{F}_0$ and $X \in \mathscr{F}_0$, we have because of (26c) that

$$h_n^2(\xi) = h_n(\xi) - U_{n,\xi}^2(\xi) - \bar{r}_n^2(\xi) \quad \text{and} \quad h^2(\xi) = h(\xi) - U_\xi^2(\xi) - \bar{r}^2(\xi), \tag{82}$$

respectively. Moreover, let $S = \{\xi \in \mathbb{R} \mid y_\xi(\xi) = 0\}$ and $S_n = \{\xi \in \mathbb{R} \mid y_{n,\xi}(\xi) = 0\}$. Then

$$y_{n,\xi}(\xi) = \frac{1}{1 + u_{n,x}^2(y_n(\xi)) + \bar{\rho}_n^2(y_n(\xi))} \quad \text{for almost every } \xi \in S_n^c,$$

and

$$y_\xi(\xi) = \frac{1}{1 + u_x^2(y(\xi)) + \bar{\rho}^2(y(\xi))} \quad \text{for almost every } \xi \in S^c.$$

Hence we get

$$\begin{aligned}\int_{\mathbb{R}} U_{n,\xi}^2(\xi)d\xi &= \int_{S_n^c} U_{n,\xi}^2(\xi)d\xi = \int_{S_n^c} \frac{u_{n,x}^2(y_n(\xi))}{1 + u_{n,x}^2(y_n(\xi)) + \bar{\rho}_n^2(y_n(\xi))} y_{n,\xi}(\xi)d\xi \\ &= \int_{\mathbb{R}} \frac{u_{n,x}^2(x)}{1 + u_{n,x}^2(x) + \bar{\rho}_n^2(x)} dx,\end{aligned}$$

where we used that $U_{n,\xi}(\xi) = y_{n,\xi}(\xi) = 0$ for all $\xi \in S_n$ and that $\mathrm{meas}(y_n(S_n)) = 0$. Similar arguments yield

$$\int_{\mathbb{R}} U_\xi^2(\xi)d\xi = \int_{\mathbb{R}} \frac{u_x^2(x)}{1 + u_x^2(x) + \bar{\rho}^2(x)}dx.$$

Thus, according to (68e),

$$\lim_{n\to\infty} \|U_{n,\xi}\|_{L^2}^2 = \lim_{n\to\infty} \int_{\mathbb{R}} U_{n,\xi}^2(\xi)d\xi = \int_{\mathbb{R}} U_\xi^2(\xi)d\xi = \|U_\xi\|_{L^2}^2. \tag{83}$$

Following the same argument, this time using (68f), we obtain

$$\lim_{n\to\infty} \|\bar{r}_n\|_{L^2}^2 = \|\bar{r}\|_{L^2}^2. \tag{84}$$

Hence combining (81)–(82) and (83)–(84) yields that $\|h_n\|_{L^2} \to \|h\|_{L^2}$, and, in particular, $h_n \to h$ in $L^2(\mathbb{R})$ and $y_{n,\xi} \to y_\xi$ in $L^2(\mathbb{R})$, since both X_n and X belong to $\mathscr{F}_0$.

Step 3: Convergence of $U_{n,\xi}$ to U_ξ in $L^2(\mathbb{R})$ In order to conclude that $U_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$, it suffices to show, according to the Radon–Riesz theorem, that $U_{n,\xi}(\xi) \rightharpoonup U_\xi(\xi)$ since we have convergence of the corresponding norms, cf. (83). Due to the fact that $C_c^\infty(\mathbb{R})$ is dense in $L^2(\mathbb{R})$, it suffices to show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} U_{n,\xi}(\xi)\psi(\xi)d\xi = \int_{\mathbb{R}} U_\xi(\xi)\psi(\xi)d\xi$$

for all $\psi \in C_c^\infty(\mathbb{R})$. Observe that we have for any $\psi \in C_c^\infty(\mathbb{R})$

$$\int_{\mathbb{R}} U_{n,\xi}(\xi)\psi(\xi)d\xi = \int_{S_n^c} U_{n,\xi}(\xi)\psi(\xi)d\xi = \int_{\mathbb{R}} u_{n,x}(x)\psi(y_n^{-1}(x))dx.$$

Here $S_n = \{\xi \in \mathbb{R} \mid y_{n,\xi}(\xi) = 0\}$ and hence, according to (26c), $U_{n,\xi}(\xi) = 0$ for almost every $\xi \in S_n$, and $y_n^{-1}(x)$ denotes the pseudo inverse to $y_n(\xi)$ defined as

$$y_n^{-1}(x) = \inf\{\xi \in \mathbb{R} \mid y_n(\xi) > x\}.$$

Similarly,

$$\int_{\mathbb{R}} U_\xi(\xi)\psi(\xi)d\xi = \int_{\mathbb{R}} u_x(x)\psi(y^{-1}(x))dx,$$

where $y^{-1}(x)$ denotes the pseudo inverse to $y(\xi)$, i.e.,

$$y^{-1}(x) = \inf\{\xi \in \mathbb{R} \mid y(\xi) > x\}.$$

Thus it suffices to show that

$$\lim_{n\to\infty}\int_{\mathbb{R}} u_{n,x}(x)\psi(y_n^{-1}(x))dx = \int_{\mathbb{R}} u_x(x)\psi(y^{-1}(x))dx \tag{85}$$

for all $\psi \in C_c^\infty(\mathbb{R})$.

Let $\psi \in C_c^\infty(\mathbb{R})$. By assumption, there exist $c, d \in \mathbb{R}$ such that $\operatorname{supp}(\psi) \subset [c, d]$. Then $\operatorname{supp}(\psi \circ y^{-1}) \subset [\xi_1, \xi_2]$, where

$$\xi_1 = \min\{\xi \in \mathbb{R} \mid y(\xi) \geq c\} \quad \text{and} \quad \xi_2 = \max\{\xi \in \mathbb{R} \mid y(\xi) \leq d\}.$$

Since $y(\xi) + H(\xi) = \xi$ for all $\xi \in \mathbb{R}$ and $H(\xi) \leq \|\mu\|$, we have

$$c - \|\mu\| \leq \xi_1 < \xi_2 \leq d + \|\mu\|$$

or, in other words, $\operatorname{supp}(\psi \circ y^{-1}) \subset [c - \|\mu\|, d + \|\mu\|]$. Similarly one obtains that $\operatorname{supp}(\psi \circ y_n^{-1}) \subset [c - \|\mu_n\|, d + \|\mu_n\|]$. In particular, cf. (68g), there exists $N \in \mathbb{N}$ such that

$$\operatorname{supp}(\psi \circ y_n^{-1}) \subset [c - 2\|\mu\|, d + 2\|\mu\|], \qquad \text{for all } n \geq N.$$

Moreover, to any $x \in \mathbb{R}$ we can assign a unique $y^{-1}(x)$ and $y_n^{-1}(x)$ using the pseudo inverse to $y(\xi)$ and $y_n(\xi)$, respectively. Thus we have from (29a)

$$y^{-1}(x) = x + \mu((-\infty, x]) = x + F(x).$$

and

$$y_n^{-1}(x) = x + \mu_n((-\infty, x]) = x + F_n(x).$$

In particular,

$$|y_n^{-1}(x) - y^{-1}(x)| = |F_n(x) - F(x)|.$$

Thus $y_n^{-1}(x)$ converges to $y^{-1}(x)$ for any $x \in \mathbb{R}$ at which $F(x)$ is continuous. In particular, $y_n^{-1}(x)$ converges to $y^{-1}(x)$ for almost every $x \in \mathbb{R}$, since $F(x)$ has at most countably many discontinuities. Hence, after using Lebesgue's dominated convergence theorem, we obtain that $\psi \circ y_n^{-1} \to \psi \circ y^{-1}$ in $L^2(\mathbb{R})$. Moreover, we have by assumption that $u_{n,x}$ converges weakly to u_x. Thus $u_{n,x}\, \psi \circ y_n^{-1}$ is the product of a weakly convergent sequence and a strongly convergent sequence, which implies that its integral converges to the integral of the limit $u_x\, \psi \circ y^{-1}$, which in turn proves (85).

Step 4: Convergence of $\bar{r}_n$ to $\bar{r}$ in $L^2(\mathbb{R})$ The proof follows exactly the same lines as the one of $U_{n,\xi} \to U_\xi$ in $L^2(\mathbb{R})$ in Step 3.

Step 5: Convergence of ζ_n to ζ in $L^\infty(\mathbb{R})$ Since both X_n and X belong to $\mathscr{F}_0$, we have

$$|y(\xi) - y_n(\xi)| = |H(\xi) - H_n(\xi)| \leq \|h_n - h\|_{L^1}. \tag{86}$$

Moreover, by (82),

$$h_n(\xi) - h(\xi) = h_n^2(\xi) - h^2(\xi) + U_{n,\xi}^2(\xi) - U_\xi^2(\xi) + \bar{r}_n^2(\xi) - \bar{r}^2(\xi),$$

which together with (86) implies that $\|y_n - y\|_{L^\infty} \to 0$ as $n \to \infty$, since $h_n \to h$, $U_{n,\xi} \to U_\xi$ and $\bar{r}_n \to \bar{r}$ in $L^2(\mathbb{R})$.

Step 6: Convergence of U_n to U in $L^2(\mathbb{R})$ For a proof we refer the interested reader to [27, Prop. 5.1].

Step 7: Convergence of k_n to k By definition we have

$$r(\xi) = \bar{r}(\xi) + k y_\xi(\xi) = \bar{\rho}(y(\xi)) y_\xi(\xi) + k y_\xi(\xi) = \rho(y(\xi)) y_\xi$$

and

$$r_n(\xi) = \bar{r}_n(\xi) + k_n y_{n,\xi}(\xi) + \bar{\rho}_n(y_n(\xi)) y_{n,\xi}(\xi) + k_n y_{n,\xi}(\xi) = \rho_n(y_n(\xi)) y_{n,\xi}(\xi).$$

Thus, the constants k_n in Eulerian and Lagrangian coordinates coincide and the same is true for k, and the claim is an immediate consequence of (68d).

6 Convergence for the Initial Data Implies Convergence for the Solution of the 2CH System

Finally, we would like to turn our attention to the 2CH system. In particular, we are going to study the consequences of the results derived so far in the context of the global weak conservative solutions of the 2CH system.

Theorem 6 *Given (u_0, ρ_0, μ_0) in $\mathscr{D}$, let $(u_{n,0}, \rho_{n,0}, \mu_{n,0})$ in $\mathscr{D}$ be a sequence satisfying $(u_{n,0}, \rho_{n,0}, \mu_{n,0}) \to (u_0, \rho_0, \mu_0)$ in the sense of Definition 5. Consider the weak conservative solutions $(u(t,\cdot), \rho(t,\cdot), \mu(t,\cdot))$ and $(u_n(t,\cdot), \rho_n(t,\cdot), \mu_n(t,\cdot))$ of the 2CH system with initial data $(u, \rho, \mu)|_{t=0} = (u_0, \rho_0, \mu_0)$ and $(u_n, \rho_n, \mu_n)|_{t=0} = (u_{n,0}, \rho_{n,0}, \mu_{n,0})$, respectively. Then we have, for all $t \in [0, \infty)$,*

$$\begin{aligned} u_n(t,\cdot) &\to u(t,\cdot) \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R}), \\ u_{n,x}(t,\cdot) &\rightharpoonup u_x(t,\cdot), \\ \bar{\rho}_n(t,\cdot) &\rightharpoonup \bar{\rho}(t,\cdot), \\ k_n &\to k, \end{aligned}$$

$$\int_{\mathbb{R}} \frac{u_{n,x}^2(t,x)}{1+u_{n,x}^2(t,x)+\bar{\rho}_n^2(t,x)}dx \to \int_{\mathbb{R}} \frac{u_x^2(t,x)}{1+u_x^2(t,x)+\bar{\rho}^2(t,x)}dx,$$

$$\int_{\mathbb{R}} \frac{\bar{\rho}_n^2(t,x)}{1+u_{n,x}^2(t,x)+\bar{\rho}_n^2(t,x)}dx \to \int_{\mathbb{R}} \frac{\bar{\rho}^2(t,x)}{1+u_x^2(t,x)+\bar{\rho}^2(t,x)}dx,$$

$$F_n(t,x) \to F(t,x) \text{ for every } x \text{ at which } F(t,x) \text{ is continuous},$$

$$F_n(t,\infty) \to F(t,\infty),$$

where $F_n(t,x) = \mu_n(t,(-\infty,x])$ *for all* $n \in \mathbb{N}$ *and* $F(t,x) = \mu(t,(-\infty,x])$. *That is, for every* $t \geq 0$ *we have that the sequence* $(u_n(t,\cdot), \rho_n(t,\cdot), \mu_n(t,\cdot))$ *converges to* $(u(t,\cdot), \rho(t,\cdot), \mu(t,\cdot))$ *in the sense of Definition 5.*

Proof Again, we are going to split the proof into several steps.

Step 1: Convergence in Eulerian Coordinates Implies Convergence in Lagrangian Coordinates for the Initial Data Let $X_0 = (y_0, U_0, h_0, r_0) = L((u_0, \rho_0, \mu_0))$ and $X_{n,0} = (y_{n,0}, U_{n,0}, h_{n,0}, r_{n,0}) = L((u_{n,0}, \rho_{n,0}, \mu_{0,n}))$. Then according to Theorem 5, $X_{n,0} \to X_0$ in E.

Step 2: Convergence at Initial Time Implies Convergence at Any Later Time for the Solution in Lagrangian Coordinates Consider the following semilinear system of ordinary differential equations, which describes weak conservative solutions of the 2CH system in Lagrangian coordinates, cf. [16],

$$\zeta_t = U, \tag{87a}$$

$$U_t = -Q(X), \tag{87b}$$

$$h_t = 2(U^2 + \frac{1}{2}k^2 - P)U_\xi, \tag{87c}$$

$$\bar{r}_t = -kU_\xi, \tag{87d}$$

$$k_t = 0, \tag{87e}$$

where $y(t,\xi) = \zeta(t,\xi) + \xi$, and

$$P(X)(t,\xi) = \frac{1}{4}\int_{\mathbb{R}} e^{-|y(t,\xi)-y(t,\eta)|}(2U^2 y_\xi + 2k\bar{r} + h)(t,\eta)d\eta + \frac{1}{2}k^2,$$

and

$$Q(X)(t,\xi) = -\frac{1}{4}\int_{\mathbb{R}} \operatorname{sgn}(\xi-\eta)e^{-|y(t,\xi)-y(t,\eta)|}(2U^2 y_\xi + 2k\bar{r} + h)(t,\eta)d\eta.$$

Then to $X(0) = X_0$ and $X_n(0) = X_{n,0} \in \mathscr{F}$, there exists a unique global solution to (87) in $\mathscr{F}$, which we denote $X(t)$ and $X_n(t)$, respectively. Moreover,

the mappings $X \mapsto P(X) - \frac{1}{2}k^2$ and $X \mapsto Q(X)$ are Lipschitz continuous on bounded sets as mappings from E to $H^1(\mathbb{R})$. In particular, one has, cf. [27, Lemma 2.1], that

$$\|P(X_n(t)) - P(X(t))\|_{L^2\cap L^\infty} \leq C_{t,n} \|X_n(t) - X(t)\|_E \,, \tag{88}$$

where $C_{t,n}$ is dependent on $\|X_n(t)\|_E$, $\|X(t)\|_E$. Similarly, we have

$$\|Q(X_n(t)) - Q(X(t))\|_{L^2\cap L^\infty} \leq C_{t,n} \|X_n(t) - X(t)\|_E \,. \tag{89}$$

Furthermore, following closely the proof of [18, Thm. 3.5] and [27, Thm. 2.8], we get from (87) and (27) that

$$\begin{aligned}\Sigma(X(t)) &= \int_{\mathbb{R}} (U^2 y_\xi + h)(t,\eta)d\eta = \int_{\mathbb{R}} (U^2 y_\xi + h)(0,\eta)d\eta \\ &= \int_{\mathbb{R}} (U^2 y_\xi + h^2 + U_\xi^2 + \bar{r}^2)(0,\eta)d\eta = \Sigma(X(0)),\end{aligned} \tag{90}$$

and, in particular,

$$\|X(t)\|_E \leq D(t, \Sigma(X(0))) \|X(0)\|_E \,,$$

where $D(t, \Sigma(X(0)))$ depends on t and $\Sigma(X(0))$. A similar estimate holds for $X_n(t)$ with $n \in \mathbb{N}$. Thus $C_{t,n}$ in (88) and (89) only depends on t, $\|X_n(0)\|_E$, and $\|X(0)\|_E$, due to (90). Furthermore, since $X_n(0) \to X(0)$ in E, there exists $M > 0$ such that $\max(\|X_n(0)\|_E, \|X(0)\|_E) \leq M$. Thus, (88) and (89) imply that the right-hand side of (87) is Lipschitz continuous on bounded sets, and, in particular, applying Gronwall's inequality yields

$$\|X_n(t) - X(t)\|_E \leq C_t \|X_n(0) - X(0)\|_E \,, \tag{91}$$

where C_t only depends on M and t.

Step 3: Convergence Independent of Relabeling in $\mathscr{F}$ As in [27, Lemma 3.3], one can show, given $T \geq 0$ and $\bar{X}(0) \in \mathscr{F}_0$, one has

$$\bar{X}(t) \in \mathscr{F}_\alpha$$

for all $t \in [0, T]$, where α only depends on t and $\|\bar{X}(0)\|_E$. In our case, since $X_n(0) \to X(0)$ in E, there exists $M > 0$ such that $M \geq \|X(0)\|_E$ and $M \geq \|X_n(0)\|_E$ for all $n \in \mathbb{N}$. Thus there exists $\beta(t) > 0$ independent of n, such that

$$X(t) \in \mathscr{F}_{\beta(t)} \quad \text{and} \quad X_n(t) \in \mathscr{F}_{\beta(t)} \quad \text{for all } n \in \mathbb{N}.$$

Moreover, it is known, see, e.g., [16, Lemma 4.6], that for $\beta(t) \geq 0$, the mapping $\Gamma\colon \mathscr{F}_{\beta(t)} \to \mathscr{F}_0$ with $X \mapsto \Gamma(X) = X \circ (y + H)^{-1}$ is continuous. Let $\tilde{X}_n(t) = \Gamma(X_n(t))$ and $\tilde{X}(t) = \Gamma(X(t))$. Then for each $t \geq 0$ the convergence $X_n(t) \to X(t)$ in E implies $\tilde{X}_n(t) \to \tilde{X}(t)$ in E.

Step 4: Convergence of the Solutions in Eulerian Coordinates Since $\tilde{X}_n(t) \to \tilde{X}(t)$ in E for all $t \geq 0$ and $\tilde{X}_n(t), \tilde{X}(t) \in \mathscr{F}_0$ for all $t \geq 0$, applying Theorem 4 finishes the proof.

The next result gives the corresponding result in the case where the approximation is constructed using the mollifier.

Theorem 7 *Given $(u_0, 0, \mu_0)$ in $\mathscr{D}$ and let $(u_{n,0}, \rho_{n,0}, \mu_{n,0})$ in $\mathscr{D}$ be the smooth approximation given by (5) in Theorem 1. Consider the weak, conservative solutions $(u(t,\cdot), 0, \mu(t,\cdot))$ and $(u_n(t,\cdot), \rho_n(t,\cdot), \mu_n(t,\cdot))$ of the 2CH system with initial data $(u, 0, \mu)|_{t=0} = (u_0, 0, \mu_0)$ and $(u_n, \rho_n, \mu_n)|_{t=0} = (u_{n,0}, \rho_{n,0}, \mu_{n,0})$, respectively. Then we have, for all $t \in [0, \infty)$,*

$$
\begin{aligned}
u_n(t,\cdot) &\to u(t,\cdot) \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R}),\\
u_{n,x}(t,\cdot) &\rightharpoonup u_x(t,\cdot),\\
\bar{\rho}_n(t\cdot) &\overset{\star}{\rightharpoonup} \bar{\rho}(t,\cdot),\\
k_n &\to 0,\\
\int_{\mathbb{R}} \frac{u_{n,x}^2(t,x)}{1 + u_{n,x}^2(t,x) + \bar{\rho}_n^2(t,x)}dx &\to \int_{\mathbb{R}} \frac{u_x^2(t,x)}{1 + u_x^2(t,x)}dx,\\
\int_{\mathbb{R}} \frac{\bar{\rho}_n^2(t,x)}{1 + u_{n,x}^2(t,x) + \bar{\rho}_n^2(t,x)}dx &\to 0,\\
F_n(t,x) &\to F(t,x) \text{ for every } x \text{ at which } F(t,x)\\
&\qquad \text{is continuous},\\
F_n(t,\infty) &\to F(t,\infty),
\end{aligned}
$$

where $F_n(t,x) = \mu_n(t,(-\infty,x])$ for all $n \in \mathbb{N}$ and $F(t,x) = \mu(t,(-\infty,x])$.

Moreover, for each $n \in \mathbb{N}$, $(u_n(t,\cdot), \rho_n(t,\cdot), \mu_n(t,\cdot))$ is a smooth solution to the 2CH system, that is, $u_n(t,x)$ and $\rho_n(t,x)$ belong to $C^\infty(\mathbb{R}\times\mathbb{R})$ and $\mu_n(t,x) = \mu_{n,\mathrm{ac}}(t,x) = (u_x^2(t,x) + \bar{\rho}^2(t,x))dx$ for all $t \geq 0$, and, in particular, no wave breaking occurs.

Proof Since we showed in Theorem 3 that $X_n = (y_n, U_n, h_n, r_n)$ converges to $X = (y, U, h, 0)$ in E, and hence according to Theorem 4, the sequence $(u_{n,0}, \rho_{n,0}, \mu_{n,0})$ converges to $(u_0, 0, \mu_0)$ in the sense of Definition 5, the first part of the theorem is an immediate consequence of Theorem 6.

As far as the smoothness of the solution $(u_n(t,\,\cdot\,), \rho_n(t,\,\cdot\,), \mu_n(t,\,\cdot\,))$ for any $t \geq 0$ and $n \in \mathbb{N}$ is concerned, we refer the interested reader to [16, Sect. 6].

Acknowledgements Research supported in part by the Research Council of Norway projects NoPiMa and WaNP, and by the Austrian Science Fund (FWF) under Grant No. J3147. KG and HH are grateful to Institut Mittag-Leffler, Stockholm, for the generous hospitality during the fall of 2016, when part of this paper was written.

References

1. L. Ambrosio, N. Fusco, 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, A. Constantin, Global conservative solutions of the Camassa–Holm equation. Arch. Ration. Mech. Anal. **183**, 215–239 (2007)
3. A. Bressan, A. Constantin, Global dissipative solutions of the Camassa–Holm equation. Anal. Appl. **5**, 1–27 (2007)
4. R. Camassa, D.D. Holm, An integrable shallow water equation with peaked solutions. Phys. Rev. Lett. **71**(11), 1661–1664 (1993)
5. R. Camassa, D.D. Holm, J. Hyman, A new integrable shallow water equation. Adv. Appl. Mech. **31**, 1–33 (1994)
6. R.M. Chen, Y. Liu, Wave breaking and global existence for a generalized two-component Camassa–Holm system. Int. Math. Res. Not. Article ID rnq118, 36 pp. (2010)
7. M. Chen, S.-Q. Liu, Y. Zhang, A two-component generalization of the Camassa–Holm equation and its solutions. Lett. Math. Phys. **75**, 1–15 (2006)
8. A. Constantin, J. Escher, Wave breaking for nonlinear nonlocal shallow water equations. Acta Math. **181**, 229–243 (1998)
9. A. Constantin, R.I. Ivanov, On an integrable two-component Camassa–Holm shallow water system. Phys. Lett. A **372**, 7129–7132 (2008)
10. J. Escher, O. Lechtenfeld, Z. Yin, Well-posedness and blow-up phenomena for the 2-component Camassa–Holm equation. Discrete Contin. Dyn. Syst. **19**(3), 493–513 (2007)
11. G.B. Folland, *Real Analysis*, 2nd edn. (Wiley, New York, 1999)
12. Y. Fu, C. Qu, Well posedness and blow-up solution for a new coupled Camassa–Holm equations with peakons. J. Math. Phys. **50**, 012906 (2009)
13. K. Grunert, Solutions of the Camassa–Holm equation with accumulating breaking times. Dyn. Partial Differ. Equ. **13**, 91–105 (2016)
14. K. Grunert, H. Holden, X. Raynaud, Lipschitz metric for the periodic Camassa–Holm equation. J. Differ. Equ. **250**, 1460–1492 (2011)
15. K. Grunert, H. Holden, X. Raynaud, Global conservative solutions to the Camassa–Holm equation for initial data with nonvanishing asymptotics. Discrete Cont. Dyn. Syst. Ser. A **32**(12), 4209–4227 (2012)
16. K. Grunert, H. Holden, X. Raynaud, Global solutions for the two-component Camassa–Holm system. Commun. Partial Differ. Equ. **37**, 2245–2271 (2012)
17. K. Grunert, H. Holden, X. Raynaud, Lipschitz metric for the Camassa–Holm equation on the line. Discrete Cont. Dyn. Syst. Ser. A **33**, 2809–2827 (2013)
18. K. Grunert, H. Holden, 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)

19. K. Grunert, H. Holden, X. Raynaud, A continuous interpolation between conservative and dissipative solutions for the two-component Camassa–Holm system. Forum Math. Sigma **1**, e1, 70 pp. (2014). https://doi.org/10.1017/fms.2014.29
20. C. Guan, Z. Yin, Global existence and blow-up phenomena for an integrable two-component Camassa–Holm water system. J. Differ. Equ. **248**, 2003–2014 (2010)
21. C. Guan, Z. Yin, Global weak solutions for a two-component Camassa–Holm shallow water system. J. Funct. Anal. **260**, 1132–1154 (2011)
22. C. Guan, K.H. Karlsen, Z. Yin, Well-posedness and blow-up phenomenal for a modified two-component Camassa–Holm equation, in *Nonlinear Partial Differential Equations and Hyperbolic Wave Phenomena*, ed. by H. Holden, K.H. Karlsen. Contemporary Mathematics, vol. 526 (American Mathematical Society, Providence, 2010), pp. 199–220
23. G. Gui, Y. Liu, On the global existence and wave breaking criteria for the two-component Camassa–Holm system. J. Funct. Anal. **258**, 4251–4278 (2010)
24. G. Gui, Y. Liu, On the Cauchy problem for the two-component Camassa–Holm system. Math. Z. **268**, 45–66 (2011)
25. Z. Guo, Y. Zhou, On solutions to a two-component generalized Camassa–Holm equation. Stud. Appl. Math. **124**, 307–322 (2010)
26. D. Henry, Infinite propagation speed for a two component Camassa–Holm equation. Discrete Contin. Dyn. Syst. Ser. B **12**(3), 597–606 (2009)
27. H. Holden, X. Raynaud, Global conservative solutions for the Camassa–Holm equation—a Lagrangian point of view. Commun. Partial Differ. Equ. **32**, 1511–1549 (2007)
28. H. Holden, X. Raynaud, Global dissipative multipeakon solutions for the Camassa–Holm equation. Commun. Partial Differ. Equ. **33**, 2040–2063 (2008)
29. H. Holden, X. Raynaud, Dissipative solutions for the Camassa–Holm equation. Discrete Cont. Dyn. Syst. **24**, 1047–1112 (2009)
30. P.J. Olver, P. Rosenau, Tri-hamiltonian duality between solitons and solitary-wave solutions having compact support. Phys. Rev. B **53**(2), 1900–1906 (1996)

Sum-Intersection Property of Sobolev Spaces



Petru Mironescu

Dedicated to Haim Brezis and Louis Nirenberg, with deep esteem and affection

1 Introduction

In connection with the factorization of unimodular Sobolev maps, Haim Brezis and the author observed the following property of Sobolev spaces [5]. Let $1 < p < \infty$ and $0 < \lambda < 1$. Then every function $f \in W^{1,p}(\mathbb{R}^N)$ can be decomposed as

$$f = g + h, \text{ with } g \in (W^{\lambda,p/\lambda} \cap W^{1,p})(\mathbb{R}^N) \text{ and } h \in (W^{p,1} \cap W^{1,p})(\mathbb{R}^N). \quad (1)$$

We will present in appendix a proof of this fact using factorization. We will also explain there how (1) is related to functional calculus (superposition operators) in Sobolev spaces.

Decomposition (1) has a flavor of interpolation, and indeed we have for example when $p = 2$ the equality [20, Section 2.4.3, Theorem, p. 66]

$$W^{1,2} = [W^{\lambda,2/\lambda}, F^2_{1,1}]_{\theta,2}, \text{ with } \theta := 1/(2-\lambda). \quad (2)$$

[We will recall in the next section the definition of the Triebel-Lizorkin spaces $F^s_{p,q}$.] Using (2) and the embedding $F^2_{1,1} \hookrightarrow W^{2,1}$ (see the next section), we find that $W^{1,2} \subset W^{\lambda,2/\lambda} + W^{2,1}$. However, this does not yield the stronger conclusion $W^{1,2} \subset (W^{\lambda,2/\lambda} \cap W^{1,2}) + (W^{2,1} \cap W^{1,2})$. Actually, one cannot derive the equality $Z = (X \cap Z) + (Y \cap Z)$ merely from the inclusion $Z \subset X + Y$ (take e.g. $X = \mathbb{R} \times \{0\}$, $Y = \{0\} \times \mathbb{R}$ and $Z = \{(x, x);\ x \in \mathbb{R}\}$).

P. Mironescu (✉)
Université de Lyon, Université Lyon 1, CNRS UMR 5208 Institut Camille Jordan, 43, boulevard du 11 novembre 1918, F-69622 Villeurbanne Cedex, France
e-mail: mironescu@math.univ-lyon1.fr

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_8

We address here the following question. Let $0 \le s, s_1, s_2 < \infty$, and $1 \le p_1, p, p_2 \le \infty$. Assume that

$$W^{s,p}(\mathbb{R}^N) \subset W^{s_1,p_1}(\mathbb{R}^N) + W^{s_2,p_2}(\mathbb{R}^N) \text{ for any } N. \tag{3}$$

Is it true that

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (W^{s_2,p_2} \cap W^{s,p})(\mathbb{R}^N) \text{ for any } N? \tag{4}$$

We emphasize the fact that we ask for N-independent properties. For example, by the Sobolev embeddings we have $W^{1,1} \subset L^2$ when $N = 1$ or 2, but not for $N \ge 3$, and thus (3) does not hold for $s_1 = s_2 = 0$, $s = 1$, $p_1 = p_2 = 2$, $p = 1$.

Our first results characterize *most* of the triples $T = (W^{s_1,p_1}, W^{s,p}, W^{s_2,p_2})$ such that (3) and (4) hold.

Proposition 1 *Assume that* (3) *holds. Then there exists some* $\theta \in [0, 1]$ *such that*

$$s \ge \theta s_1 + (1-\theta) s_2, \tag{5}$$

$$\frac{1}{p} = \frac{\theta}{p_1} + \frac{(1-\theta)}{p_2}. \tag{6}$$

Proposition 2 *Assume that for some* $\theta \in [0, 1]$ *we have* (6) *and* $s > \theta s_1 + (1-\theta) s_2$. *Then both* (3) *and* (4) *hold.*

On the other hand, (3) and (4) trivially hold when (5)–(6) are satisfied with $\theta = 0$ or 1, since we then have either $W^{s,p} \hookrightarrow W^{s_2,p_2}$, or $W^{s,p} \hookrightarrow W^{s_1,p_1}$. We next investigate the case where

$$s = \theta s_1 + (1-\theta) s_2, \ \frac{1}{p} = \frac{\theta}{p_1} + \frac{1-\theta}{p_2} \text{ for some } \theta \in (0, 1). \tag{7}$$

In this case, (3) holds *most* of the time, but *not always*. For example, when $N = 1$ we have

$$W^{1/2,2}(\mathbb{R}) \not\subset W^{1,1}(\mathbb{R}) + L^\infty(\mathbb{R}), \tag{8}$$

i.e., (3) does not hold for the triple $T = (W^{1,1}, W^{1/2,2}, L^\infty)$. Indeed, for $N = 1$ we have $W^{1,1} \hookrightarrow L^\infty$, and thus $W^{1,1} + L^\infty = L^\infty$. However, $W^{1/2,2} \not\subset L^\infty$.

Definition 1 A triple $T = (W^{s_1,p_1}, W^{s,p}, W^{s_2,p_2})$ is *admissible* if it satisfies (7).

An admissible triple T is *irregular* if $s_1 \ne s_2$, $1 < p < \infty$ and (exactly) one of the spaces W^{s_1,p_1}, W^{s_2,p_2} is of the form $W^{k,\infty}$ with $k \in \mathbb{N}$. T is *regular* otherwise.

Thus $T = (W^{1,1}, W^{1/2,2}, L^\infty)$ (which corresponds to the example occurring in (8)) is irregular.

Our main result is the following

Theorem 1 *Let T be a regular triple. Then both (3) and (4) hold.*

Equivalently, for every regular triple T we have

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (W^{s_2,p_2} \cap W^{s,p})(\mathbb{R}^N), \ \forall N. \tag{9}$$

For most of the regular triples, (4) follows *automatically* from (3), as explained in Proposition 3 below. Thus, in particular, the conclusion of the theorem follows whenever T is as in Proposition 3 and $W^{s,p}$ can be obtained by interpolation from W^{s_1,p_1} and W^{s_2,p_2}. However, when T is admissible $W^{s,p}$ need not be an interpolation space between W^{s_1,p_1} and W^{s_2,p_2}, at least for the standard real and complex methods [20, Sections 2.4.2–2.4.7, p. 64–73]; thus one cannot derive Theorem 1 directly from Proposition 3. We will present, in Sect. 3, a proof of Theorem 1 which does not rely on interpolation and establishes simultaneously (3) and (4).

Definition 2 A Sobolev space $W^{s,p}$ is *exceptional* if $s \in \mathbb{N}$ and either $p = 1$ or $p = \infty$. It is *ordinary* otherwise.

Proposition 3 *Assume that $W^{s,p}$, W^{s_1,p_1} and W^{s_2,p_2} are all three ordinary Sobolev spaces. Assume that for some (fixed) N we have $W^{s,p}(\mathbb{R}^N) \subset W^{s_1,p_1}(\mathbb{R}^N) + W^{s_2,p_2}(\mathbb{R}^N)$. Then for such N we have*

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (W^{s_2,p_2} \cap W^{s,p})(\mathbb{R}^N).$$

We now turn to irregular T's. At least in some special cases (see (8) and, more generally, the triples $T = (W^{1,1}, W^{1/p,p}, L^\infty)$, with $1 < p < \infty$), (3) does not hold for such triples. We do not know the characterization of irregular triples T for which (3) and/or (4) do not hold. For irregular triples, we were only able to establish a weaker form of (4), in which the space $W^{k,\infty}$ is replaced by a slightly larger space, modeled on *bmo* (the *local BMO* space whose definition will be recalled in the next section).

Theorem 2 *Let T be an irregular triple, and assume e.g. that $p_2 = \infty$ (and thus s_2 is an integer). Let $1 < q_2 < \infty$. Then*

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (F^{s_2}_{\infty,q_2} \cap W^{s,p})(\mathbb{R}^N). \tag{10}$$

In particular, when $s_2 = 0$ (and thus $W^{s_2,p_2} = L^\infty$) we have

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (bmo \cap W^{s,p})(\mathbb{R}^N). \tag{11}$$

When $s_2 > 0$, we have

$$\begin{aligned} W^{s,p}(\mathbb{R}^N) =& (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) \\ &+ (\{f \in W^{s_2-1,\infty};\ D^{s_2-1} f \in bmo\} \cap W^{s,p})(\mathbb{R}^N). \end{aligned} \tag{12}$$

In the special case $s_2 = 0$, $s \notin \mathbb{N}$, $p_1 = 1$, Theorem 2 was established in [5, Chapter 6].

Remark 1 The question of the validity of (3)–(4) is somewhat *dual* to the one of the validity of the Gagliardo-Nirenberg inequalities. There, one asks whether the inclusion

$$W^{s_1,p_1}(\mathbb{R}^N) \cap W^{s_2,p_2}(\mathbb{R}^N) \subset W^{s,p}(\mathbb{R}^N) \tag{13}$$

leads, for some appropriate $\theta \in [0, 1]$, to the estimate

$$\|f\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|^{\theta}_{W^{s_1,p_1}(\mathbb{R}^N)} \|f\|^{1-\theta}_{W^{s_2,p_2}(\mathbb{R}^N)}. \tag{14}$$

In the spirit of our Proposition 1, one may prove that the validity of (13) for every N requires

$$s \leq \theta s_1 + (1-\theta) s_2, \tag{15}$$

$$\frac{1}{p} = \frac{\theta}{p_1} + \frac{(1-\theta)}{p_2} \tag{16}$$

for some $\theta \in [0, 1]$. If we have either "$<$" in (15) or $\theta \in \{0, 1\}$, then we have both (13) and (14); this follows from the main result in [4]. As in our situation, the interesting case is the one of admissible triples. In that case, (15) and (16) hold when s_1, s, s_2 are integers, as established in the seminal contributions of Gagliardo [11] and Nirenberg [15]. It turns out that (15) and (16) hold for *most* of the admissible triples, but not all of them. A characterization of the admissible triples for which (15) and (16) hold has been obtained in [4]; see also [8, 12, 16] for older partial results.

Remark 2 As one may expect, whenever it is possible to decompose $f = f_1 + f_2$ with $f_1 \in (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N)$ and $f_2 \in (W^{s_2,p_2} \cap W^{s,p})(\mathbb{R}^N)$, we also have a norm control for f_1 and f_2 in terms of $\|f\|_{W^{s,p}}$. A simple example of such decomposition with norm control is the following. For $f \in L^2(\mathbb{R}^N)$, set $f_1 := f\, \mathbb{1}_{\{x;\, |f(x)| > \|f\|_{L^2(\mathbb{R}^N)}\}}$ and $f_2 := f\, \mathbb{1}_{\{x;\, |f(x)| \leq \|f\|_{L^2(\mathbb{R}^N)}\}}$. Then clearly $f_1 \in (L^1 \cap L^2)(\mathbb{R}^N)$ and $f_2 \in (L^\infty \cap L^2)(\mathbb{R}^N)$, and in addition we have the norm controls

$$\|f_1\|_{L^1(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)},\ \|f_1\|_{L^2(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)},$$

$$\|f_2\|_{L^\infty(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)},\ \|f_2\|_{L^2(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)}.$$

Note however that the map $f \mapsto (f_1, f_2)$ is not linear. Likewise, in general we will construct nonlinear decompositions.

Our text is organized as follows. In Sect. 2, we recall some basic facts on function spaces, instrumental for our purposes. The proofs of Propositions 1, 2 and 3 and of Theorems 1 and 2 are presented in Sect. 3. A final appendix presents the

factorization theory and its connections with the sum-intersection property and with the functional calculus in Sobolev spaces.

2 Basic Properties of Triebel-Lizorkin Spaces

Definition 3 Let $\psi \in C_c^\infty(\mathbb{R}^N)$ be such that $\psi = 1$ in $B_1(0)$ and $\operatorname{supp}\psi \subset B_2(0)$. Define $\psi_0 = \psi$ and, for $j \geq 1$, $\psi_j(x) := \psi(x/2^j) - \psi(x/2^{j-1})$. Set $\varphi_j := \mathscr{F}^{-1}\psi_j \in \mathscr{S}$.[1] Then for each temperate distribution f we have

$$f = \sum_j f_j \quad \text{in } \mathscr{S}', \text{ with } f_j := f * \varphi_j. \tag{17}$$

$f = \sum f_j$ is *"the" Littlewood-Paley* decomposition of $f \in \mathscr{S}'$.

Note that $\mathscr{F} f_j = \psi_j \mathscr{F} f$ is compactly supported, and therefore $f_j \in C^\infty$ for each j.

Definition 4 Starting from for Littlewood-Paley decomposition, we define the *Triebel-Lizorkin spaces* $F^s_{p,q}$ as follows: for $-\infty < s < \infty$, $0 < p < \infty$ and $0 < q \leq \infty$, we let

$$\|f\|_{F^s_{p,q}} := \left\| \left\| \left(2^{sj} f_j(x)\right)_{j\geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)}, \quad F^s_{p,q} := \{f \in \mathscr{S}';\ \|f\|_{F^s_{p,q}} < \infty\}.$$

Same definition when $p = q = \infty$.

This definition has to be changed when $p = \infty$ and $1 < q < \infty$ [20, Section 2.3.4, p. 50]: we let

$$\|f\|_{F^s_{\infty,q}} = \inf \left\{ \operatorname*{esssup}_{x\in\mathbb{R}^N} \left\| \left\| \left(2^{sj} f_j(x)\right)_{j\geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^\infty(\mathbb{R}^N)};\ f_j \in L^\infty(\mathbb{R}^N),\ f = \sum f_j * \varphi_j \right\},$$

the latter equality being in the sense of $\mathscr{S}'$.

Most of the Sobolev spaces can be identified with Triebel-Lizorkin spaces [20, Section 2.3.5] and [17, Section 2.1.2].

Theorem 3 *The following equalities of spaces hold, with equivalence of norms:*

1. *If $s > 0$ is not an integer and $1 \leq p \leq \infty$, then $W^{s,p}(\mathbb{R}^N) = F^s_{p,p}$.*
2. *If $s \geq 0$ is an integer and $1 < p < \infty$, then $W^{s,p}(\mathbb{R}^N) = F^s_{p,2}$.*

[1] Equivalently, we have $\varphi_0 = \mathscr{F}^{-1}\psi$ and, for $j \geq 1$, $\varphi_j(x) = 2^{Nj}\varphi_0(2^j x) - 2^{N(j-1)}\varphi_0(2^{j-1}x)$.

When $s \geq 0$ is an integer and either $p = 1$ or $p = \infty$, the Sobolev space $W^{s,p}$ cannot be identified with a Triebel-Lizorkin space.

Theorem 3 is usually used in conjunction with Lemma 1 below. The reason is that, in practice, we do not know the Littlewood-Paley decomposition of f, but only a Nikol'skij decomposition of f.

Definition 5 A *Nikol'skij decomposition* of $f \in \mathscr{S}'$ is a representation of the form $f = \sum f^j$ in $\mathscr{S}'$, with $\operatorname{supp} \mathscr{F} f^j \subset \begin{cases} B_{2^{j+1}}(0) \setminus B_{2^{j-1}}(0), & \text{if } j \geq 1 \\ B_2(0), & \text{if } j = 0 \end{cases}$.

Note that in particular the Littlewood-Paley decomposition $f = \sum f_j$ is a Nikol'skij decomposition.

Lemma 1 *1. Let $1 < p < \infty$, $1 < q \leq \infty$ and $s \in \mathbb{R}$. Consider a sequence (f^j) such that $\left\| \left\| \left(2^{sj} f^j(x)\right)_{j \geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)} < \infty$. Then $f = \sum f^j * \varphi_j$ converges in $\mathscr{S}'$ and*

$$\|f\|_{F^s_{p,q}} \lesssim \left\| \left\| \left(2^{sj} f^j(x)\right)_{j \geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)}. \tag{18}$$

2. Same conclusion if $1 \leq p = q \leq \infty$.

Proof It suffices to consider finite sums, and to establish (18) in this case. We start with item *2*, which is easier. Note that $f \in L^p(\mathbb{R}^N)$, and thus $f \in \mathscr{S}'$.

Let $f = \sum_{j \geq 0} f_j$ be the Littlewood-Paley decomposition of f. Since $\varphi_j * \varphi_k = 0$ if $|j - k| \geq 2$, we find that

$$f_j = f * \varphi_j = \sum_k f^k * \varphi_k * \varphi_j = \sum_{|k-j| \leq 1} f^k * \varphi_k * \varphi_j, \tag{19}$$

and thus

$$\begin{aligned} \|f_j\|_{L^p(\mathbb{R}^N)} &\leq \sum_{|k-j| \leq 1} \|f^k * \varphi_k * \varphi_j\|_{L^p(\mathbb{R}^N)} \\ &\leq \sum_{|k-j| \leq 1} \|f^k\|_{L^p(\mathbb{R}^N)} \|\varphi_k * \varphi_j\|_{L^1(\mathbb{R}^N)} \leq C \sum_{|k-j| \leq 1} \|f^k\|_{L^p(\mathbb{R}^N)}. \end{aligned} \tag{20}$$

We obtain (18) with $p = q$ from (20).

We now consider item *1*. From (19), we find that

$$|f_j(x)| \leq \sum_{|k-j| \leq 1} |f^k * \varphi_k * \varphi_j(x)| \leq C \sum_{|k-j| \leq 1} \mathscr{M} f^k(x). \tag{21}$$

Here, $\mathscr{M}$ is the standard maximal operator, and we used the inequality [18, Proposition, p. 24]

$$|f * \rho_\varepsilon(x)| \le C_\rho \mathscr{M} f(x), \ \forall \rho \in \mathscr{S}, \ \forall \varepsilon > 0.$$

Using (21), we find that

$$\|f\|_{F^s_{p,q}} \lesssim \left\| \left\| \left(2^{sj} \mathscr{M} f^j(x)\right)_{j\ge 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)} \lesssim \left\| \left\| \left(2^{sj} f^j(x)\right)_{j\ge 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)},$$

the latter inequality being the Fefferman-Stein vectorial maximal inequality [10].

Definition 6 We define, for $f \in L^1_{loc}(\mathbb{R}^N)$,

$$\|f\|_{bmo} := \sup_{|B|\le 1} \int_B |f| + \sup_{|B|\le 1} \fint_B \fint_B |f(x) - f(y)|\, dxdy,$$

the sup being taken over the balls of volume ≤ 1. We set

$$bmo := \{f \in L^1_{loc}(\mathbb{R}^N); \ \|f\|_{bmo} < \infty\}.$$

With its natural norm, *bmo* is the *local BMO space*.

Then we have [21, Theorem, p. 47] $bmo = F^0_{\infty,2}$. Using this equality, Definition 4 and the embedding $\ell^q \hookrightarrow \ell^2$, $0 < q < 2$, we obtain the following

Corollary 1 *If $f = \sum_{j\ge 0} f^j * \varphi_j$ in $\mathscr{S}'$ and $0 < q < 2$, then*

$$\|f\|^2_{bmo} \le C \operatorname*{esssup}_{x\in\mathbb{R}^N} \sum_j |f^j(x)|^2 \le C \operatorname*{esssup}_{x\in\mathbb{R}^N} \left(\sum_j |f^j(x)|^q\right)^{2/q}, \tag{22}$$

for some C independent of the f^j's.

Corollary 2 *For $1 < q < 2$, we have $F^0_{\infty,q} \hookrightarrow bmo$.*

As we noticed above, when $s \in \mathbb{N}$ the space $W^{s,1}$ is not a Triebel-Lizorkin space. However, we have the following

Lemma 2

1. *When $s \ge 0$, we have $F^s_{1,1} \hookrightarrow W^{s,1}(\mathbb{R}^N)$.*
2. *More generally, for every $s \ge 0$ and $1 \le p < \infty$ we have $F^s_{p,1} \hookrightarrow W^{s,p}(\mathbb{R}^N)$. The same holds when $p = \infty$ and $s > 0$ is not an integer.*
3. *When $k > 0$ is an integer and $1 < q \le 2$, we have*

$$F^k_{\infty,q} \hookrightarrow \{f \in W^{k-1,\infty}(\mathbb{R}^N); \ D^{k-1} f \in bmo\}.$$

Proof We start with $p = 1$. When s is not an integer, we actually have equality. When $s = 0$ and $f \in F^0_{1,1}$, we have $\|f\|_{L^1(\mathbb{R}^N)} \leq \sum_{j\geq 0} \|f_j\|_{L^1(\mathbb{R}^N)} = \|f\|_{F^0_{1,1}} < \infty$. When $s \geq 1$ is an integer, we use the fact that [20, Section 2.3.8, Theorem (ii), pp. 58–59]

$$\|f\|_{F^s_{1,1}} \sim \sum_{j=0}^{s} \|D^j f\|_{F^{s-j}_{1,1}} \geq \sum_{j=0}^{s} \|D^j f\|_{F^0_{1,1}} \geq \sum_{j=0}^{s} \|D^j f\|_{L^1(\mathbb{R}^N)} = \|f\|_{W^{s,1}(\mathbb{R}^N)}.$$

When $1 < p < \infty$, the desired inclusion follows from

$$F^s_{p,1} \hookrightarrow F^s_{p,q} = W^{s,p}(\mathbb{R}^N) \text{ (with } q = 2 \text{ or } q = p, \text{ according to } s).$$

Similarly if $p = \infty$ and s is not an integer.

Finally, if $p = \infty$ and s is an integer, we argue as for $p = 1$, relying on Corollary 2 and [20, Section 2.3.8, Remark 2, p. 60].

We now briefly recall the characterization of Triebel-Lizorkin spaces in terms of wavelets.

Let ψ_0, ψ_1 be respectively a father and mother (sufficiently smooth) wavelets. For $G \in \{0,1\}^N$, $j \in \mathbb{N}$ and $m \in \mathbb{Z}^N$, let $\psi^j_{G,m}(x) := 2^{Nj/2} \prod_{r=1}^{N} \psi_{G_r}(2^j x_r - m_r)$, $x \in \mathbb{R}^N$. Let, for $f \in \mathscr{S}'$,

$$\lambda^j_{G,m} := \begin{cases} 0, & \text{if } j > 0 \text{ and } G = \{0\}^N \\ 2^{Nj/2}\,(f, \psi^j_{G,m}), & \text{otherwise} \end{cases}.$$

Recall [22, Section 3.1.3] that $f = \sum_{j,G,m} 2^{-Nj/2} \lambda^j_{G,m} \psi^j_{G,m}$ in the sense of $\mathscr{S}'$. Conversely, if

$$f = \sum_{G,m} \mu^0_{G,m} \psi^0_{G,m} + \sum_{j>0,\, G\neq\{0\}^N,\, m} 2^{-Nj/2} \mu^j_{G,m} \psi^j_{G,m} \text{ in the sense of } \mathscr{S}',$$

then the wavelet coefficients $\lambda^j_{G,m}$ of f are given by

$$\lambda^j_{G,m} = \begin{cases} 0, & \text{if } j > 0 \text{ and } G = \{0\}^N \\ \mu^j_{G,m}, & \text{otherwise} \end{cases}.$$

Let, for $j \in \mathbb{N}$ and $m \in \mathbb{Z}^N$, $Q_{j,m}$ be the cube $\prod_{r=1}^{N} [2^{-j}(m_r - 1), 2^{-j}(m_r + 1)]$. Set, for $0 < q < \infty$, $s \in \mathbb{R}$,

$$g(x) = g^s_{p,q}(x) := \Big(\sum 2^{sqj} |\lambda^j_{G,m}|^q \mathbb{1}_{Q_{j,m}}(x)\Big)^{1/q}. \tag{23}$$

When $q = \infty$, we replace the ℓ^q norm by the sup norm.

Then one may read the smoothness of f in terms of the integrability properties of g. The following statement is a rephrasing of [22, Theorem 1.64, p. 33].

Theorem 4

1. *Let $-\infty < s < \infty$, $1 \le p < \infty$, $0 < q \le \infty$. Then $\|f\|_{F^s_{p,q}} \sim \|g^s_{p,q}\|_{L^p(\mathbb{R}^N)}$.*
2. *Same conclusion if $p = q = \infty$.*
3. *In particular, if $s > 0$ is not an integer and $1 \le p \le \infty$, then $\|f\|_{W^{s,p}(\mathbb{R}^N)} \sim \|g^s_{p,p}\|_{L^p}$.*
4. *If $s \ge 0$ is an integer and $1 < p < \infty$, then $\|f\|_{W^{s,p}(\mathbb{R}^N)} \sim \|g_{s,2}\|_{L^p}$.*

Let us note that when $p = q$, this norm equivalence takes a particularly simple form. More specifically, we have

$$\|f\|^p_{F^s_{p,p}} \sim \sum_{j,G,m} 2^{(sp-N)j} |\lambda^j_{G,m}|^p, \ -\infty < s < \infty, \ 1 \le p < \infty \tag{24}$$

$$\|f\|_{F^s_{\infty,\infty}} \sim \sup_{j,G,m} 2^{sj} |\lambda^j_{G,m}|, \ -\infty < s < \infty. \tag{25}$$

Our next result relies on properties of the Besov spaces $B^s_{p,q}$. In order to keep this section short, we will be rather sketchy.

Lemma 3 *Let $0 \le s < \infty$, $1 \le p \le \infty$ and $\varepsilon > 0$. Then, with $g^s_{p,q}$ as in (23), we have*

$$\|f\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|g^{s+\varepsilon}_{p,p}\|_{L^p(\mathbb{R}^N)}, \tag{26}$$

$$\|g^{s-\varepsilon}_{p,p}\|_{L^p(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}. \tag{27}$$

Proof (Sketch of Proof) The above estimates are equivalent to the embeddings

$$F^{s+\varepsilon}_{p,p} \hookrightarrow W^{s,p}(\mathbb{R}^N) \hookrightarrow F^{s-\varepsilon}_{p,p}. \tag{28}$$

When s is not an in an integer, we have $W^{s,p}(\mathbb{R}^N) = F^s_{p,p}$, and the conclusion is clear.

When s is an integer and $1 \le p \le \infty$, the Littlewood-Paley decomposition $f = \sum_j f_j$ of f satisfies [7, Lemma 2.1.1]

$$\|f_0\|_{L^p(\mathbb{R}^N)} \le \|f\|_{L^p(\mathbb{R}^N)}, \ 2^{sj}\|f_j\|_{L^p(\mathbb{R}^N)} \lesssim \|D^s f\|_{L^p(\mathbb{R}^N)}, \ \forall\, j \ge 1. \tag{29}$$

Thus $\sup_j 2^{sj}\|f_j\|_{L^p(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}$, i.e., we have the embedding

$$W^{s,p}(\mathbb{R}^N) \hookrightarrow B^s_{p,\infty}. \tag{30}$$

On the other hand, we have [19, Chapter 5, Lemma 3.14]

$$\|D^s f_j\|_{L^p(\mathbb{R}^N)} \lesssim 2^{sj} \|f_j\|_{L^p(\mathbb{R}^N)}, \ \forall\, j \geq 0,$$

and thus

$$\|f\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \sum_j \left(\|f_j\|_{L^p(\mathbb{R}^N)} + \|D^s f_j\|_{L^p(\mathbb{R}^N)}\right) \lesssim \sum_j 2^{sj} \|f_j\|_{L^p(\mathbb{R}^N)}.$$

Equivalently, we have the embedding

$$B^s_{p,1} \hookrightarrow W^{s,p}(\mathbb{R}^N). \tag{31}$$

We obtain (28) via (30)–(31) and the following elementary embeddings [20, Section 2.3.2, Proposition 2, p. 47]

$$F^{s+\varepsilon}_{p,p} = B^{s+\varepsilon}_{p,p} \hookrightarrow B^s_{p,1} \hookrightarrow W^{s,p}(\mathbb{R}^N) \hookrightarrow B^s_{p,\infty} \hookrightarrow B^{s-\varepsilon}_{p,p} = F^{s-\varepsilon}_{p,p}.$$

□

3 Proofs

Proof (Proof of Proposition 1) In order to prove the existence of some θ such that (6) holds, we have to establish the double inequality

$$\min\{p_1, p_2\} \leq p \leq \max\{p_1, p_2\}. \tag{32}$$

We argue by contradiction. Assume first that $p > \max\{p_1, p_2\}$. Let

$$f(x) = \frac{2}{(1+x^2)^{(1+\varepsilon)/(2p)}}, \ \forall\, x \in \mathbb{R}.$$

Clearly, $f \in L^p(\mathbb{R})$, and more generally $f \in W^{k,p}(\mathbb{R})$ for every integer k. It follows that $f \in W^{s,p}(\mathbb{R})$ for every $s \geq 0$. On the other hand, for every f_1, f_2 such that $f = f_1 + f_2$ and every x we have either $|f_1(x)| \geq f(x)/2$ or $|f_2(x)| \geq f(x)/2$. We find that

$$|f_1(x)|^{p_1} + |f_2(x)|^{p_2} \gtrsim f(x)^{p_1} + f(x)^{p_2} := g(x).$$

Since, for sufficiently small ε, we have $g \notin L^1(\mathbb{R})$, we find that $f \notin L^{p_1}(\mathbb{R}) + L^{p_2}(\mathbb{R})$. Therefore, $f \notin W^{s_1,p_1}(\mathbb{R}) + W^{s_2,p_2}(\mathbb{R})$, which is a contradiction.

Assume next that $p < \min\{p_1, p_2\}$. Let $p < r < \min\{p_1, p_2\}$. Let N be sufficiently large such that $W^{s,p}(B) \not\subset L^r(B)$; here, B is a ball in $\mathbb{R}^N$. By a standard extension argument, there exists some $f \in W_c^{s,p}(\mathbb{R}^N)$ such that $f \notin L^r(\mathbb{R}^N)$. Such an f does not belong to $L^r_{loc}(\mathbb{R}^N)$, and thus does not belong to $L^{p_1}(\mathbb{R}^N)+L^{p_2}(\mathbb{R}^N)$. We find that $f \notin W^{s_1,p_1}(\mathbb{R}^N) + W^{s_2,p_2}(\mathbb{R}^N)$, again a contradiction.

We thus know that (32) holds, or equivalently, that (6) holds for some θ.

We next proceed to the proof of (5). Assume first that $p_1 = p_2 = p$. Then θ is not determined by (6), and its existence is equivalent to $s \geq \min\{s_1, s_2\}$. Arguing by contradiction, assume that $s < \min\{s_1, s_2\}$. Let $s < \rho < \min\{s_1, s_2\}$. If $f \in W^{s,p}(\mathbb{R}) \setminus W^{\rho,p}(\mathbb{R})$, then

$$f \notin W^{s_1,p}(\mathbb{R}) + W^{s_2,p}(\mathbb{R}) = W^{\min\{s_1,s_2\},p}(\mathbb{R}) \subset W^{\rho,p}(\mathbb{R}),$$

a contradiction.

Assume next that $p_1 \neq p_2$. Then θ is determined by (6). Argue again by contradiction and assume that $s < \theta s_1 + (1-\theta)s_2$. Set $\sigma := \theta s_1 + (1-\theta)s_2 > s$. Consider some $\varepsilon > 0$ such that $s + \varepsilon < \sigma - \varepsilon$. In view of Lemma 3, in order to contradict (3) it suffices to establish, for some appropriate N, the non inclusion

$$F^{s+\varepsilon}_{p,p} \not\subset F^{s_1-\varepsilon}_{p_1,p_1} + F^{s_1-\varepsilon}_{p_2,p_2}. \tag{33}$$

With no loss of generality, we may assume that

$$1 \leq p_1 < p_2 \leq \infty. \tag{34}$$

We will treat separately the cases $p_2 < \infty$ and $p_2 = \infty$.

Set, in all cases,

$$\alpha := \frac{\frac{s_1-\varepsilon}{p_2} - \frac{s_2-\varepsilon}{p_1}}{\frac{1}{p_1} - \frac{1}{p_2}} = \frac{\frac{s_1}{p_2} - \frac{s_2}{p_1}}{\frac{1}{p_1} - \frac{1}{p_2}} + \varepsilon. \tag{35}$$

Proof of (33) *When* $p_2 < \infty$. We rely on the following

Claim. For appropriate $C_1, C_2 > 0$, we have

$$[a+b = S, S \geq C_1 2^{\alpha j}] \implies [2^{(s_1-\varepsilon)jp_1}|a|^{p_1} + 2^{(s_2-\varepsilon)jp_2}|b|^{p_2} \geq C_2 2^{(\sigma-\varepsilon)jp}]. \tag{36}$$

Granted the claim, we conclude as follows. Consider some $f \in \mathscr{S}'$ such that for every j, G and m we have either $\lambda^j_{G,m} = 0$ or $|\lambda^j_{G,m}| \geq C_1 2^{\alpha j}$, with C_1 as in (36). The claim combined with (24) implies that for *every possible* decomposition $f = f_1 + f_2$ we have

$$\|f_1\|^{p_1}_{F^{s_1-\varepsilon}_{p_1,p_1}} + \|f_2\|^{p_1}_{F^{s_2-\varepsilon}_{p_2,p_2}} \gtrsim \|f\|^{p}_{F^{\sigma-\varepsilon}_{p,p}}. \tag{37}$$

We are now in position to obtain a contradiction. Let N be sufficiently large such that $(\sigma-\varepsilon+\alpha)p < N$. Let $\delta := N-(\sigma-\varepsilon+\alpha)p > 0$. Fix some $G_0 \in \{0,1\}^N \setminus \{0\}^N$. For every $j \in \mathbb{N}$, consider a set $M_j \subset \mathbb{Z}^N$ such that $\#M_j \sim 2^{\delta j}$. Set

$$f := \sum_{j,\, m\in M_j} 2^{-Nj/2} C_1 2^{\alpha j} \psi^j_{G_0,m}.$$

By (24), we have

$$\|f\|^p_{F^{s+\varepsilon}_{p,p}} \sim \sum_j 2^{((s+\varepsilon+\alpha)p-N+\delta)j} = \sum_j 2^{-((\sigma-\varepsilon)-(s+\varepsilon))jp} < \infty,$$

while

$$\|f\|^p_{F^{\sigma-\varepsilon}_{p,p}} \sim \sum_j 2^{((\sigma-\varepsilon+\alpha)p-N+\delta)j} = \sum_j 1 = \infty.$$

We complete the proof of (33) when $p_2 < \infty$ using the two above inequalities and (37).

Proof of (33) *When* $p_2 = \infty$ *and* $\theta \in (0,1]$. This time we have $\alpha = -(s_2-\varepsilon)$. We modify the definition of f by setting

$$f := \sum_{j,\, m\in M_j} 2^{-Nj/2} j 2^{\alpha j} \psi^j_{G_0,m}.$$

Assume, by contradiction, that $f = f_1 + f_2$ for some $f_1 \in F^{s_1-\varepsilon}_{p_1,p_1}$ and $f_2 \in F^{s_2-\varepsilon}_{\infty,\infty}$. Write $f_1 = \sum_{j,G,m} 2^{-Nj/2} a^j_{G,m} \psi^j_{G,m}$, $f_2 = \sum_{j,G,m} 2^{-Nj/2} b^j_{G,m} \psi^j_{G,m}$.

Since $f_2 \in F^{s_2-\varepsilon}_{\infty,\infty}$, we have

$$|b^j_{G,m}| \le C 2^{-(s_2-\varepsilon)j} = C 2^{\alpha j},\ \forall\, j, G, m.$$

Since $a^j_{G_0,m} + b^j_{G_0,m} = j2^{\alpha j}$, $\forall\, j$, $\forall\, m \in M_j$, for sufficiently large j_0 we have

$$|a^j_{G_0,m}| \ge \frac{1}{2} j 2^{\alpha j},\ \forall\, j \ge j_0,\ \forall\, m \in M_j.$$

Inserting this into (24) and using the fact that

$$(s_1-\varepsilon+\alpha)p_1 - N + \delta = (s_1-s_2)p_1 - \theta(s_1-s_2)p = (s_1-s_2)(p_1-\theta p) = 0$$

(since $p_1 = \theta p$), we find that

$$\|f\|^{p_1}_{F^{s_1-\varepsilon}_{p_1,p_1}} \gtrsim \sum_{j\geq j_0,\, m\in M_j} j^{p_1} 2^{((s_1-\varepsilon+\alpha)p_1-N)j}$$
$$\sim \sum_{j\geq j_0} j^{p_1} 2^{((s_1-\varepsilon+\alpha)p_1-N+\delta)j} = \sum_{j\geq j_0,\, m\in M_j} j^{p_1} = \infty.$$

On the other hand, we have

$$\|f\|^{p}_{F^{s+\varepsilon}_{p,p}} \sim \sum j^{p} 2^{((s+\varepsilon+\alpha)p-N+\delta)j} = \sum j^{p} 2^{-((\sigma-\varepsilon)-(s+\varepsilon))jp} < \infty.$$

This leads to a contradiction and completes the proof of (33) when $p_2 = \infty$ and $\theta \in (0, 1]$.

Proof of (33) *When* $p_2 = \infty$ *and* $\theta = 0$. This is similar to the case $p_2 = \infty$ and $\theta \in (0, 1]$. We have $\alpha = -(s_2 - \varepsilon) = -(\sigma - \varepsilon) < -(s + \varepsilon)$. Consider $f := \sum_{j,m} 2^{-Nj/2} j 2^{\alpha j} \psi^j_{G_0,m}$. [This time, the sum in m is over all $m \in \mathbb{Z}^N$.] We then have $f \in F^{s+\varepsilon}_{\infty,\infty}$. Arguing by contradiction, we obtain that f cannot be decomposed as $f = f_1 + f_2$ with $f_1 \in F^{s_1-\varepsilon}_{p_1,p_1}$ and $f_2 \in F^{s_2-\varepsilon}_{\infty,\infty}$. Indeed, as in the previous case, if $f_2 \in F^{s_2-\varepsilon}_{\infty,\infty}$ then for large j_0 we have

$$\|f_1\|^{p_1}_{F^{s_1-\varepsilon}_{p_1,p_1}} \gtrsim \sum_{j\geq j_0} \sum_{m\in\mathbb{Z}^N} j^{p_1} 2^{(s_1\alpha p_1-N)j} = \infty.$$

Proof of the Claim. Let $S > 0$. The function

$$[0, \infty) \ni t \mapsto g(t) := 2^{(s_1-\varepsilon)j}(1 - t)S + 2^{(s_2-\varepsilon)jp_2/p_1} t^{p_2/p_1} S^{p_2/p_1}$$

is convex, and its derivative at the origin is negative. Thus g has a global minimum at the point t_0 where $g'(t_0) = 0$. Solving the equation $g'(t) = 0$, we find that $t_0 = C_1 2^{\alpha j} S^{-1}$, with $C_1 > 0$ independent of j. Provided that $S \geq C_1 2^{\alpha j}$, we have $t_0 \leq 1$, and therefore the first term in $g(t)$ is non negative. For such S, we thus have

$$g(t) \geq g(t_0) \geq 2^{(s_2-\varepsilon)jp_2/p_1} (t_0)^{p_2/p_1} S^{p_2/p_1} = c\, 2^{(s_2-\varepsilon+\alpha)p_2/p_1 j} = c 2^{(\sigma-\varepsilon)p/p_1 j},$$
$$\forall\, t \geq 0,$$

with $c > 0$ independent of S.

Let now a, b be such that $a + b = S \geq C_1 2^{\alpha j}$. Then

$$2^{s_1 jp_1} |a|^{p_1} + 2^{s_2 jp_2} |b|^{p_2} \geq 2^{s_1 jp_1} \underline{a}^{p_1} + 2^{s_2 jp_2} \underline{b}^{p_2},$$

where

$$\underline{a} := \begin{cases} a, & \text{if } 0 \leq a, b \leq S \\ 0, & \text{if } a < 0 \text{ and } b > S \\ S, & \text{if } a > S \text{ and } b < 0 \end{cases}, \quad \underline{b} := \begin{cases} b, & \text{if } 0 \leq a, b \leq S \\ S, & \text{if } a < 0 \text{ and } b > S \\ 0, & \text{if } a > S \text{ and } b < 0 \end{cases}. \tag{38}$$

Therefore, it suffices to prove (36) under the extra assumption that $0 \le a, b \le S$. Write $a = (1-t)S$, $b = tS$, with $t \in [0,1]$. We then have

$$2^{(s_1-\varepsilon)jp_1}a^{p_1} + 2^{(s_2-\varepsilon)jp_2}b^{p_2} \sim \left(2^{(s_1-\varepsilon)j}a + 2^{(s_2-\varepsilon)jp_2/p_1}b^{p_2/p_1}\right)^{p_1}$$
$$= [g(t)]^{p_1} \ge [g(t_0)]^{p_1} \ge c^{p_1}2^{(\sigma-\varepsilon)jp}.$$

□

Proof of Proposition 2 Assuming Theorem 1. As already noticed in the proof of Proposition 1, when $p_1 = p_2 = p$ or when $\theta \in \{0,1\}$, properties (3) and (4) are trivially true. We may thus assume that $p_1 \neq p_2$ and $\theta \in (0,1)$. Set $\lambda := s - (\theta s_1 + (1-\theta)s_2) > 0$. For $0 < \varepsilon < \dfrac{\lambda}{\theta}$, let $\delta > 0$ satisfy $\theta\varepsilon + (1-\theta)\delta = \lambda$. Then we may pick ε such that neither $s_1 + \varepsilon$ nor $s_2 + \delta$ is an integer. Thus the triple $T := (W^{s_1+\varepsilon,p_1}, W^{s,p}, W^{s_2+\delta,p_2})$ is regular. Granted Theorem 1, this implies

$$W^{s,p}(\mathbb{R}^N) = (W^{s_1+\varepsilon,p_1} \cap W^{s,p})(\mathbb{R}^N) + (W^{s_2+\delta,p_2} \cap W^{s,p})(\mathbb{R}^N)$$
$$\subset (W^{s_1,p_1} \cap W^{s,p})(\mathbb{R}^N) + (W^{s_2,p_2} \cap W^{s,p})(\mathbb{R}^N).$$

□

Proof of Proposition 3. Decompose $f \in W^{s,p}(\mathbb{R}^N)$ as $f = f_1 + f_2$, with $f_1 \in W^{s_1,p_1}(\mathbb{R}^N)$ and $f_2 \in W^{s_2,p_2}(\mathbb{R}^N)$. Write, in the sense of $\mathscr{S}'$,

$$f = \sum_{j,G,m} 2^{-Nj/2}\lambda^j_{G,m}\psi^j_{G,m},$$
$$f_1 = \sum_{j,G,m} 2^{-Nj/2}a^j_{G,m}\psi^j_{G,m},\ f_2 = \sum_{j,G,m} 2^{-Nj/2}b^j_{G,m}\psi^j_{G,m}.$$

In the spirit of (38), define

$$\underline{a}^j_{G,m} := \begin{cases} a^j_{G,m}, & \text{if } 0 \le a^j_{G,m}, b^j_{G,m} \le \lambda^j_{G,m} \\ 0, & \text{if } a^j_{G,m} < 0 \text{ and } b^j_{G,m} > \lambda^j_{G,m} \\ \lambda^j_{G,m}, & \text{if } a^j_{G,m} > \lambda^j_{G,m} \text{ and } b^j_{G,m} < 0 \end{cases},\ \underline{f}_1 := \sum_{j,G,m} 2^{-Nj/2}\underline{a}^j_{G,m}\psi^j_{G,m}$$

$$\underline{b}^j_{G,m} := \begin{cases} b^j_{G,m}, & \text{if } 0 \le a^j_{G,m}, b^j_{G,m} \le \lambda^j_{G,m} \\ \lambda^j_{G,m}, & \text{if } a^j_{G,m} < 0 \text{ and } b^j_{G,m} > \lambda^j_{G,m} \\ 0, & \text{if } a^j_{G,m} > \lambda^j_{G,m} \text{ and } b^j_{G,m} < 0 \end{cases},\ \underline{f}_2 := \sum_{j,G,m} 2^{-Nj/2}\underline{b}^j_{G,m}\psi^j_{G,m}.$$

Then $f = \underline{f}_1 + \underline{f}_2$, and Theorem 4 implies that

$$\left\|\underline{f}_1\right\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)} < \infty, \quad \left\|\underline{f}_1\right\|_{W^{s_1,p_1}(\mathbb{R}^N)} \lesssim \|f_1\|_{W^{s_1,p_1}(\mathbb{R}^N)} < \infty,$$
$$\left\|\underline{f}_2\right\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)} < \infty, \quad \left\|\underline{f}_2\right\|_{W^{s_2,p_2}(\mathbb{R}^N)} \lesssim \|f_2\|_{W^{s_2,p_2}(\mathbb{R}^N)} < \infty.$$

□

Proof of Theorem 1. The case where $p_1 = p_2$ is trivial, since we then have $W^{s,p} \subset W^{\min\{s_1,s_2\},p}$.

We may thus assume that

$$1 \le p_1 < p < p_2 \le \infty. \tag{39}$$

We further distinguish between the cases $s_1 = s_2$ and $s_1 \neq s_2$, and also between $p_2 < \infty$ and $p_2 = \infty$.

Given $f \in W^{s,p}(\mathbb{R}^N)$, we write $f = \sum_{j,G,m} 2^{-Nj/2}\lambda^j_{G,m}\psi^j_{G,m}$.

Case 1. $s_1 = s_2 = s \notin \mathbb{N}$. Set

$$f_1 := \sum_{j,G,m} 2^{-Nj/2} a^j_{G,m}\psi^j_{G,m}, \ f_2 := \sum_{j,G,m} 2^{-Nj/2} b^j_{G,m}\psi^j_{G,m},$$

with

$$a^j_{G,m} := \begin{cases} \lambda^j_{G,m}, & \text{if } |\lambda^j_{G,m}| \ge 2^{-sj} \\ 0, & \text{if } |\lambda^j_{G,m}| < 2^{-sj} \end{cases}, \ b^j_{G,m} := \begin{cases} 0, & \text{if } |\lambda^j_{G,m}| \ge 2^{-sj} \\ \lambda^j_{G,m}, & \text{if } |\lambda^j_{G,m}| < 2^{-sj} \end{cases}.$$

Since $p_1 < p$, we have

$$|a^j_{G,m}|^{p_1} \le 2^{sj(p_1-p)}|\lambda^j_{G,m}|^p. \tag{40}$$

Using (40), the fact that s is not an integer and (24), we find that

$$\|f_1\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_1\|^{p_1}_{W^{s_1,p_1}(\mathbb{R}^N)} \lesssim \|f\|^p_{W^{s,p}(\mathbb{R}^N)}. \tag{41}$$

Similarly, if $p_2 < \infty$ then we have

$$\|f_2\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_2\|^{p_2}_{W^{s_2,p_2}(\mathbb{R}^N)} \lesssim \|f\|^p_{W^{s,p}(\mathbb{R}^N)}. \tag{42}$$

On the other hand, if $p_2 = \infty$ then

$$\|f_2\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_2\|_{W^{s_2,\infty}(\mathbb{R}^N)} \lesssim 1. \tag{43}$$

We complete this step via (41)–(43).

Remark 3 The estimates (41)–(43) are nonlinear, while one would expect linear estimates. Actually, it is possible to obtain linear estimates by cutting the coefficients $\lambda_{G,m}^j$ at height $A2^{-sj}$ instead of 2^{-sj}, with $A := \|f\|_{W^{s,p}(\mathbb{R}^N)}$. The corresponding decomposition satisfies

$$\|f_1\|_{W^{s,p}(\mathbb{R}^N)} + \|f_1\|_{W^{s_1,p_1}(\mathbb{R}^N)} + \|f_2\|_{W^{s,p}(\mathbb{R}^N)} + \|f_2\|_{W^{s_2,p_2}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}.$$

Similar observations apply to all the other cases.

Case 2. $s_1 = s_2 = s \in \mathbb{N}$. In this case, we follow the ideas of DeVore and Scherer [9] concerning the interpolation theory of classical spaces, in the form presented in Bennett and Sharpley [1, Section 5.5, pp. 347–362].

We claim that it suffices to decompose every $f \in (W^{s,p} \cap C^\infty)(\mathbb{R}^N)$ as $f = f_1 + f_2$, with

$$\|f_1\|_{W^{s,1}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_1\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \tag{44}$$

$$\|f_2\|_{W^{s,\infty}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_2\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}. \tag{45}$$

Indeed, if this holds then Hölder's inequality implies that

$$\|f_1\|_{W^{s,p_1}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ \|f_2\|_{W^{s,p_2}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}, \ 1 \le p_1 < p_2 \le \infty, \tag{46}$$

and then a density argument shows that (44)–(46) hold without the extra assumption $f \in C^\infty$; this settles this case.

We next proceed to the construction of f_1 and f_2. Let $\mathscr{M}$ denote the standard maximal (uncentered) operator. Set $H_\ell(x) := \sum_{|\alpha|=\ell} |\partial^\alpha f(x)|$ and $H(x) := \sum_{\ell=0}^k H_\ell(x)$. Let $\Omega := \{x \in \mathbb{R}^N; \mathscr{M}H(x) > \tau\}$ and $M := \mathbb{R}^N \setminus \Omega$. Thus M is closed and $H(x) \le \tau, \forall\, x \in M$.

Let c be such that $\|\mathscr{M}g\|_{L^p(\mathbb{R}^N)} \le c\|g\|_{L^p(\mathbb{R}^N)}, \forall\, g \in L^p(\mathbb{R}^N)$. If $\tau := c\|H\|_{L^p(\mathbb{R}^N)} \sim \|f\|_{W^{s,p}(\mathbb{R}^N)}$, then

$$|\Omega| \le \frac{1}{\tau^p} \int_\Omega (\mathscr{M}H)^p(x)\, dx \le \frac{1}{\tau^p} \|\mathscr{M}H\|_{L^p(\mathbb{R}^N)}^p \le 1. \tag{47}$$

We then let f_2 be the Whitney extension of $f_{|M}$ and set $f_1 := f - f_2$. More specifically, let (Q_j) be a Whitney covering of Ω with cubes of size ℓ_j and centers y_j. Let $Q_{j,t}$ denote the cube of center y_j and size $t\ell_j$. Recall the following properties of the Whitney covering:

$$(Q_{j,9/8}) \text{ is a covering of } \Omega, \ Q_{j,4} \text{ intersects } M, \ \forall\, j, \ \sum_j \mathbb{1}_{Q_j}(x) \le C(N), \ \forall\, x. \tag{48}$$

Let (ϕ_j) be an adapted Whitney partition of unity in Ω, i.e.,

$$\operatorname{supp}\phi_j \subset Q_{j,9/8},\ \forall\, j,\ \text{and}\ |\partial^\alpha \phi_j| \lesssim (\ell_j)^{-\alpha},\ \forall\, \alpha \in \mathbb{N}^N. \tag{49}$$

Let $x_j \in M \cap Q_{j,4}$ and set

$$T_j(x) := \sum_{|\alpha|\le s-1} \partial^\alpha f(x_j)\frac{(x-x_j)^\alpha}{\alpha!},$$

the Taylor expansion of order $s-1$ of f around x_j. Then we set $f_2 := \begin{cases} f, & \text{in } M \\ \sum T_j\phi_j, & \text{in } \Omega \end{cases}$.

This f_2 satisfies [1, Theorem 5.10, p. 355] $f_2 \in W^{s,\infty}(\mathbb{R}^N)$ and

$$\|f_2\|_{W^{s,\infty}(\mathbb{R}^N)} \lesssim \tau \sim \|f\|_{W^{s,p}(\mathbb{R}^N)}. \tag{50}$$

On the other hand, using the fact that $|\Omega| \le 1$ (by (47)), we find that for every $1 \le r \le p$ the function f_1 satisfies

$$\begin{aligned} \|f_1\|_{W^{s,r}(\mathbb{R}^N)} &= \|f - f_2\|_{W^{s,r}(\Omega)} \le \|f\|_{W^{s,r}(\Omega)} + \|f_2\|_{W^{s,r}(\Omega)} \\ &\lesssim \|f\|_{W^{s,r}(\Omega)} + \tau \lesssim \|f\|_{W^{s,p}(\Omega)} + \tau \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}. \end{aligned} \tag{51}$$

Combining (50)–(51), we also have

$$\|f_2\|_{W^{s,p}(\mathbb{R}^N)} \lesssim \|f\|_{W^{s,p}(\mathbb{R}^N)}. \tag{52}$$

We obtain (44) (and complete this case) from (50)–(52).

Case 3. $s_1 \ne s_2$ *and* $p_2 < \infty$. This is somewhat the *general* case. We will prove below that

$$F^s_{p,q} = (F^{s_1}_{p_1,q_1} \cap F^s_{p,q}) + (F^{s_2}_{p_2,q_2} \cap F^s_{p,q}), \tag{53}$$

under the assumptions

$$\begin{aligned} &-\infty < s_1, s, s_2 < \infty,\ s_1 \ne s_2,\ 0 < p_1 < p < p_2 < \infty \\ &\text{such that (7) holds,}\ 0 < q_1, q, q_2 < \infty. \end{aligned} \tag{54}$$

In view of Theorem 3 and Lemma 2, this is stronger than the conclusion of Theorem 1.

We now proceed to the proof of (53). Throughout the calculations we perform in this case, we assume (54).

Define, in the spirit of (35),

$$\alpha := \frac{\frac{s_1}{p_2} - \frac{s_2}{p_1}}{\frac{1}{p_1} - \frac{1}{p_2}}. \tag{55}$$

Let us first note that the proportionality condition (7) leads to the following identities

$$\alpha = \frac{\frac{s}{p_2} - \frac{s_2}{p}}{\frac{1}{p} - \frac{1}{p_2}} = \frac{\frac{s_1}{p} - \frac{s}{p_1}}{\frac{1}{p_1} - \frac{1}{p}} \tag{56}$$

and

$$(s_1 + \alpha)p_1 = (s + \alpha)p = (s_2 + \alpha)p_2. \tag{57}$$

In addition, we have

$$\text{either } s_1 + \alpha,\ s + \alpha,\ s_2 + \alpha > 0, \text{ or } s_1 + \alpha,\ s + \alpha,\ s_2 + \alpha < 0. \tag{58}$$

Given a sequence (x_j) of nonnegative numbers, set, for $i = 1, 2$,

$$\begin{aligned} S_i(x) &:= \sum 2^{s_i jq_i} (x_j)^{q_i},\ g_i(x) := [S_i(x)]^{p_i/q_i}, \\ T(x) &:= \sum 2^{sjq} (x_j)^q,\ h(x) := [T(x)]^{p/q}. \end{aligned} \tag{59}$$

Lemma 4 *There exists some finite constant C such that*

$$[x_j \le 2^{\alpha j},\ \forall\, j] \implies g_2(x) \le Ch(x).$$

Lemma 5 *There exists some finite constant C such that*

$$[\forall\, j,\ x_j \ge 2^{\alpha j} \text{ or } x_j = 0] \implies g_1(x) \le Ch(x).$$

Granted the two lemmas, we proceed to the proof of (53).
Let $f \in F^s_{p,q}$ and write, in the sense of $\mathscr{S}'$, $f = \sum_{j,G,m} 2^{-Nj/2} \lambda^j_{G,m} \psi^j_{G,m}$. Set

$$f_1 := \sum_{|\lambda^j_{G,m}| > 2^{\alpha j}} 2^{-Nj/2} \lambda^j_{G,m} \psi^j_{G,m},\ f_2 := \sum_{|\lambda^j_{G,m}| \le 2^{\alpha j}} 2^{-Nj/2} \lambda^j_{G,m} \psi^j_{G,m}. \tag{60}$$

Clearly, $f_1, f_2 \in F^s_{p,q}$.

We next note that, for each x and j, there exists some subset $M(j,x)$ of $\mathbb{Z}^N$, say $M(j,x) = \{m_{j,x}^\ell\}_{\ell=1}^k$ (with $k := 3^N$ independent of j and x), such that $m \notin M(j,k) \implies x \notin Q_{j,m}$. This implies that for all $x \in \mathbb{R}^N$ we have

$$\sum_{j,G,m} 2^{\sigma j\rho} \left|a_{G,m}^j\right|^\rho \mathbb{1}_{Q_{j,m}}(x) \sim \sum_{j,G,\ell} 2^{\sigma j\rho} \left|a_{G,m_{j,x}^\ell}^j\right|^\rho \mathbb{1}_{Q_{j,m_{j,x}^\ell}}(x),\ \forall\sigma,\ \forall\rho,\ \forall a_{G,m}^j. \tag{61}$$

Applying Lemmas 4 and 5 with $x_j := \left|\lambda_{G,m_{j,x}^\ell}^j\right| \mathbb{1}_{Q_{j,m_{j,x}^\ell}}(x)$ and using (60)–(61), we find that

$$\|f_1\|_{F_{p_1,q_1}^{s_1}}^{p_1} \lesssim \|f\|_{F_{p,q}^s}^p,\ \|f_2\|_{F_{p_2,q_2}^{s_2}}^{p_2} \lesssim \|f\|_{F_{p,q}^s}^p. \tag{62}$$

It thus remains to prove Lemmas 4 and 5.

Proof of Lemma 4. Define $A := (s_2+\alpha)q_2$, $B := (s+\alpha)q$. By (57), we have either $A, B > 0$, or $A, B < 0$.

Set $a_j := 2^{-\alpha j} x_j \in [0,1]$. Then

$$S_2(x) = \widetilde{S}_2(a) := \sum 2^{Aj}(a_j)^{q_2},\ g_2(x) = \widetilde{g}_2(a) := \left[\widetilde{S}_2(a)\right]^{p_2/q_2},$$
$$T(x) = \widetilde{T}(a) := \sum 2^{Bj}(a_j)^q,\ h(x) = \widetilde{h}(a) := \left[\widetilde{T}(a)\right]^{p/q}.$$

Let J be an arbitrary nonnegative integer, and set

$$A_J^2 := \{a = (a_j)_{j\geq 0};\ a_j \in [0,1],\ \forall\, j,\ \text{and}\ a_j = 0,\ \forall\, j > J\}. \tag{63}$$

In order to establish the lemma, it suffices to prove that

$$\widetilde{g}_2(a) \leq C\widetilde{h}(a),\ \forall a \in A_J^2, \tag{64}$$

provided C does not depend on J.

Fix J. For $a \in A_J^2$, $a \not\equiv 0$, set $\widetilde{f}_2(a) := \dfrac{\widetilde{g}_2(a)}{\widetilde{h}(a)}$. Since $\widetilde{f}_2$ is homogeneous of degree $p_2 - p > 0$, it attains its maximum at some a such that at least one of the a_j's equals 1. For this a, set

$$\Lambda_1 := \{j \leq J;\ a_j = 0\},\ \Lambda_2 := \{j \leq J;\ a_j = 1\},\ \Lambda_2 := \{j \leq J;\ 0 < a_j < 1\}.$$

By the above, we have $\Lambda_2 \neq \emptyset$. Set $m := \min \Lambda_2$ and $M := \max \Lambda_2$.

Step 1. Proof of the lemma when $\Lambda_3 = \emptyset$. Assume first that $A, B > 0$. Then

$$\widetilde{S}_2(a) = \sum_{j\in\Lambda_2} 2^{Aj} \leq \sum_{j\leq M} 2^{Aj} \lesssim 2^{AM},\ \widetilde{T}(a) = \sum_{j\in\Lambda_2} 2^{Bj} \geq 2^{BM}.$$

We find that

$$\widetilde{f}_2(a) \lesssim \frac{(2^{AM})^{p_2/q_2}}{(2^{BM})^{p/q}} = 1,$$

since $A\dfrac{p_2}{q_2} = B\dfrac{p}{q}$ (by (57)).

If $A, B < 0$, we have similarly $\widetilde{S}_2(a) \lesssim 2^{Am}$ and $\widetilde{T}(a) \geq 2^{Bm}$, and therefore $\widetilde{f}_2(a) \lesssim 1$.

Step 2. Proof of the lemma when $\Lambda_3 \neq \emptyset$. Set $\ell := \min \Lambda_3$, $L := \max \Lambda_3$.

If $j \in \Lambda_3$, then $\dfrac{\partial}{\partial a_j}\widetilde{f}_2(a) = 0$, and thus

$$p_2 2^{Aj}[\widetilde{S}_2(a)]^{p_2/q_2-1}(a_j)^{q_2-1}[\widetilde{T}(a)]^{p/q} = p2^{Bj}[\widetilde{T}(a)]^{p/q-1}(a_j)^{q-1}[\widetilde{S}_2(a)]^{p_2/q_2},$$

which implies that

$$(a_j)^{q_2-q} = C_1 2^{(B-A)j}, \ \forall\, j \in \Lambda_2, \text{ with } C_1 = C_1(a) \text{ constant.} \tag{65}$$

Step 2.1. Proof of the lemma when $\Lambda_3 \neq \emptyset$ *and* $q_2 = q$. By (65), the quantity $2^{(B-A)j}$ does not depend on $j \in \Lambda_3$. On the other hand, since $q_2 = q$ we have $B - A = (s - s_2)q \neq 0$. Thus Λ_3 contains only one element, $\Lambda_3 = \{\ell\} = \{L\}$. We find that

$$\widetilde{g}_2(a) = \sum_{j\in\Lambda_2} 2^{Aj} + 2^{A\ell}(a_\ell)^q, \ \widetilde{h}(a) = \sum_{j\in\Lambda_2} 2^{Bj} + 2^{B\ell}(a_\ell)^q.$$

As in Step 1, when $A, B > 0$ we find that

$$\widetilde{f}_2(a) \lesssim \frac{\left(2^{AM} + 2^{A\ell}(a_\ell)^q\right)^{p_2/q}}{\left(2^{BM} + 2^{B\ell}(a_\ell)^q\right)^{p/q}} \lesssim \frac{2^{Ap_2/qM} + 2^{Ap_2/q}(a_\ell)^{p_2}}{2^{Bp/qM} + 2^{Bp/q}(a_\ell)^p} \leq 1,$$

the latter inequality following from $A\dfrac{p_2}{q_2} = B\dfrac{p}{q}$, $p_2 > p$ and $0 < a_\ell < 1$.

The case where $A, B < 0$ is handled similarly.

Step 2.2. Proof of the lemma when $\Lambda_3 \neq \emptyset$ *and* $q_2 \neq q$. Define $\gamma := \dfrac{B - A}{q_2 - q}$. It follows from (65) that

$$a_j = C_2 2^{\gamma j}, \ \forall\, j \in \Lambda_3. \tag{66}$$

Let us note that

$$A + \gamma q_2 = A + \frac{B - A}{q_2 - q}q_2 = \frac{Bq_2 - Aq}{q_2 - q} = qq_2\frac{s - s_2}{q_2 - q} \neq 0.$$

We therefore have the following four possibilities:

1. $A, B > 0$, $A + \gamma q_2 > 0$.
2. $A, B > 0$, $A + \gamma q_2 < 0$.
3. $A, B < 0$, $A + \gamma q_2 > 0$.
4. $A, B < 0$, $A + \gamma q_2 < 0$.

We complete Step 2.2 in one of these cases, and let to the reader the three other ones, which are similar. Assume e.g. that $A, B > 0$ and $A + \gamma q_2 < 0$. In this case we obtain an information on C_2 by letting, in (66), $j = \ell$. [If $A + \gamma q_2 > 0$, we take $j = L$.] Since $0 < a_\ell < 1$, we have $0 < C_2 2^{\gamma \ell} < 1$, and thus $C_2 = C_3 2^{-\gamma \ell}$, with $0 < C_3 < 1$. We find that

$$a_j = C_3 2^{\gamma(j-\ell)}, \ \forall\, j \in \Lambda_3, \ \text{for some } C_3 \in (0,1). \tag{67}$$

Since $A > 0$ and $A + \gamma q_2 < 0$, we find that

$$\begin{aligned}\widetilde{S}_2(a) &\le \sum_{j \le M} 2^{Aj} + \sum_{j \ge \ell} 2^{(A+\gamma q_2)j} (C_3)^{q_2} 2^{-\gamma q_2 \ell} \\ &\lesssim 2^{AM} + 2^{(A+\gamma q_2)\ell} (C_3)^{q_2} 2^{-\gamma q_2 \ell} = 2^{AM} + 2^{A\ell} (C_3)^{q_2},\end{aligned}$$

while

$$\widetilde{T}(a) \ge 2^{BM} + 2^{B\ell} (C_3)^q.$$

We find that

$$\widetilde{f}_2(a) \lesssim \frac{\left(2^{AM} + 2^{A\ell}(C_3)^{q_2}\right)^{p_2/q_2}}{\left(2^{BM} + 2^{B\ell}(C_3)^q\right)^{p/q}} \lesssim \frac{2^{Ap_2/q_2 M} + 2^{Ap_2/q_2 \ell}(C_3)^{p_2}}{2^{Bp/qM} + 2^{Bp/q\ell}(C_3)^p} \le 1,$$

since $A\dfrac{p_2}{q_2} = B\dfrac{p}{q}$, $0 < C_3 < 1$ and $p_2 > p$.

The proof of Lemma 4 is complete. □

Sketch of Proof of Lemma 5. This is very much similar to the proof of Lemma 4. This time, we have $a_j \in \{0\} \cup [1, \infty)$. With $C := (s_1 + \alpha) q_1$, we set $\widetilde{S}_1(a) := \sum 2^{Cj} (a_j)^{q_1}$ and

$$A_J^1 := \{a = (a_j)_{j \ge 0};\ a_j = 0 \text{ or } a_j \ge 2^{\alpha j}, \ \forall\, j, \ \text{and } a_j = 0, \ \forall\, j > J\}.$$

If $a \in A_J^1$, $a \not\equiv 0$, we set $\widetilde{f}_1(a) := \dfrac{[\widetilde{S}_1(a)]^{p_1/q_1}}{[T(a)]^{p/q}}$. We have to prove that $\widetilde{f}_1(a) \lesssim 1, \forall\, J, \forall\, a \in A_J^1$, $a \not\equiv 0$. This is obtained following the same strategy as in the proof of Lemma 4, considering, for a maximum point a of $\widetilde{f}_1$, the sets

$$\Lambda_1 := \{j \le J;\ a_j = 0\},\ \Lambda_2 := \{j \le J;\ a_j = 1\},\ \Lambda_3 := \{j \le J;\ 1 < a_j < \infty\}.$$

The key ingredients are that C and B are either both positive or both negative, respectively the fact that, when $q \neq q_1$, the quantity $C + \dfrac{B-C}{q_1 - q} q_1$ does not vanish.

Details are left to the reader. □

Case 4. $s_1 \neq s_2$ *and* $p_2 = \infty$. This is very much similar to Case 3. We prove the equality

$$F^s_{p,q} = (F^{s_1}_{p_1,q_1} \cap F^s_{p,q}) + (F^{s_2}_{\infty,\infty} \cap F^s_{p,q}) \tag{68}$$

under the assumptions

$$\begin{gathered} -\infty < s_1, s, s_2 < \infty,\ s_1 \neq s_2,\ 0 < p_1 < p < p_2 = \infty \\ \text{such that (7) holds},\ 0 < q_1, q < \infty. \end{gathered} \tag{69}$$

[For an improvement of (68) under more restrictive conditions of p_1, see the proof of Theorem 2.]

In view of Theorem 3 and Lemma 2, this implies Case 4. In order to prove (68), we decompose $f \in F^s_{p,q}$ as in (60). By Theorem 4 and Lemma 5, we have $f_1 \in F^{s_1}_{p_1,q_1} \cap F^s_{p,q}$. On the other hand, since $p_2 = \infty$ we have $\alpha = -s_2$, and then clearly (60) implies that $f_2 \in F^{s_2}_{\infty,\infty} \cap F^{s,p}$.

The proof of Theorem 1 is complete. □

Proof of Theorem 2. We will prove the following version of (68): we have

$$F^s_{p,q} = (F^{s_1}_{p_1,q_1} \cap F^s_{p,q}) + (F^{s_2}_{\infty,q_2} \cap F^s_{p,q}) \tag{70}$$

under one of the following assumptions

$$\begin{gathered} -\infty < s_1, s, s_2 < \infty,\ s_1 \neq s_2,\ 1 < p_1 < p < p_2 = \infty \\ \text{such that (7) holds},\ 0 < q < \infty,\ 1 < q_1, q_2 < \infty \end{gathered} \tag{71}$$

or

$$\begin{gathered} -\infty < s_1, s, s_2 < \infty,\ s_1 \neq s_2,\ 1 = p_1 < p < p_2 = \infty \\ \text{such that (7) holds},\ 0 < q < \infty,\ q_1 = 1,\ 1 < q_2 < \infty. \end{gathered} \tag{72}$$

Granted (70), we obtain the conclusion of Theorem 2 via Theorem 3, Corollary 2 and Lemma 2.

We now proceed to the proof of (70).

Let $f \in F^s_{p,q}$, and let $f = \sum f_j$ be the Littlewood-Paley decomposition of f. Set $f^j := \sum_{|k-j|\le 1} f_k = \sum_{|k-j|\le 1} f * \varphi_k * \varphi_j$. Taking into account the fact that

$\varphi_j * \varphi_k = 0$ if $|j-k| \geq 2$ and that $\sum_k \varphi_k = \delta$ in the sense of $\mathscr{S}'$, we find that

$$\sum_j f^j * \varphi_j = \sum_{j,k} f * \varphi_k * \varphi_j = \sum_j f * \varphi_j = f. \tag{73}$$

On the other hand, we clearly have

$$\left\| \left\| \left(2^{sj} f^j(x)\right)_{j\geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)} \lesssim \left\| \left\| \left(2^{sj} f_j(x)\right)_{j\geq 0} \right\|_{l^q(\mathbb{N})} \right\|_{L^p(\mathbb{R}^N)} = \|f\|_{F^s_{p,q}}. \tag{74}$$

Define

$$\delta := \frac{1}{p(s-s_2)} \neq 0. \tag{75}$$

Let us note that (7) and (75) imply the identity

$$\frac{p_1}{p} + (s_1 - s)\delta p_1 = 1. \tag{76}$$

Given $x \in \mathbb{R}^N$, let $h(x) := \left(\sum 2^{sjq} |f^j(x)|^q\right)^{p/q}$, so that $h < \infty$ a.e. Whenever $h(x) < \infty$, define $J = J(x)$ as follows: J is the least non negative integer such that $2^J \geq [h(x)]^\delta$.

Lemma 6 *Let δ and J be as above.*

1. *If $\delta > 0$, then*

$$\left(\sum_{j<J} 2^{s_1 j q_1} |f^j(x)|^{q_1}\right)^{p_1/q_1} \lesssim h(x) \text{ and } \left(\sum_{j\geq J} 2^{s_2 j q_2} |f^j(x)|^{q_2}\right)^{1/q_2} \lesssim 1. \tag{77}$$

2. *If $\delta < 0$, then*

$$\left(\sum_{j>J} 2^{s_1 j q_1} |f^j(x)|^{q_1}\right)^{p_1/q_1} \lesssim h(x) \text{ and } \left(\sum_{j\leq J} 2^{s_2 j q_2} |f^j(x)|^{q_2}\right)^{1/q_2} \lesssim 1. \tag{78}$$

Granted Lemma 6, we complete the proof of Theorem 2 as follows. Assume e.g. that $\delta > 0$, the case $\delta < 0$ being similar. Define, for a.e. $x \in \mathbb{R}^N$,

$$g^j(x) := \begin{cases} f^j(x), & \text{if } j < J(x) \\ 0, & \text{if } j \geq J(x) \end{cases}, \quad h^j(x) := \begin{cases} 0, & \text{if } j < J(x) \\ f^j(x), & \text{if } j \geq J(x) \end{cases}. \tag{79}$$

Combining (73), (74), Lemma 6 and Lemma 1, we find that the series $f_1 := \sum g^j * \varphi_j$ and $f_2 := h^j * \varphi_j$ converge in $\mathscr{S}'$, that $f = f_1 + f_2$, and that $f_1 \in F^{s_1}_{p_1,q_1} \cap F^{s}_{p,q}$, $f_2 \in F^{s_2}_{\infty,q_2} \cap F^{s}_{p,q}$. □

Proof of Lemma 6. We consider only the case $\delta > 0$, the case $\delta < 0$ being similar. Set $M := [h(x)]^\delta$. We let to the reader the case where $M < 1$ and thus $J = 0$ and the first sum in (77) vanishes. Assuming that $M \geq 1$, we have $2^J \sim M$ and

$$|f^j(x)| \leq 2^{-sj}[h(x)]^{1/p} = 2^{-sj} M^{1/(\delta p)}, \ \forall\, j \geq 0. \tag{80}$$

Since $\delta > 0$, we have $s > s_2$, and thus $s_1 > s > s_2$. Using (80), we find that

$$\sum_{j<J} 2^{s_1 j q_1} |f^j(x)|^{q_1} \lesssim M^{q_1/(\delta p)} \sum_{j<J} 2^{(s_1-s) j q_1} \lesssim M^{q_1/(\delta p)} 2^{(s_1-s) J q_1}$$
$$\sim M^{q_1[1/(\delta p)+(s_1-s)]}. \tag{81}$$

Combining (76) and (81), we find that

$$\left(\sum_{j<J} 2^{s_1 j q_1} |f^j(x)|^{q_1} \right)^{p_1/q_1} \lesssim \left[M^{1/\delta} \right]^{p_1/p+(s-1-s)\delta} = M^{1/\delta} = h(x),$$

i.e., the first inequality in (77) holds.

For the second inequality, we note that (80) leads to

$$\sum_{j \geq J} 2^{s_2 j q_2} |f^j(x)|_2^q \lesssim M^{q_2/(\delta p)} \sum_{j \geq J} 2^{(s_2-s) j q_2}$$
$$\lesssim M^{q_2/(\delta p)} 2^{(s_2-s) J q_2} \sim M^{q_2[1/(\delta p)+(s_2-s)]} = 1,$$

the latter equality following from the definition of δ. □

Appendix: Factorization, Functional Calculus, Sum-Intersection

The lifting problem for $\mathbb{S}^1$-valued Sobolev maps is the following. Let B be a ball in $\mathbb{R}^N$. Let $s > 0$ and $1 \leq p \leq \infty$. Is it possible to lift *every* map $u \in W^{s,p}(B; \mathbb{S}^1)$ as $u = e^{\imath\varphi}$ with $\varphi \in W^{s,p}(\mathbb{B}; \mathbb{R})$? This question has been completely answered in [3]. The answer depends on s, p and N. For example, in $W^{1,p}(B; \mathbb{S}^1)$ the answer is positive if $N = 1$ or [$N \geq 2$ and $p \geq 2$], but negative if [$N \geq 2$ and $1 \leq p < 2$]. Factorization is a substitute to lifting, but is also valid and relevant if the answer to the lifting problem is positive. Special cases of factorization were announced in [13]. The general case is presented in [5] and asserts the following. Let $s > 0$ and

$1 \le p < \infty$. Then *every* map $u \in W^{s,p}(B;\mathbb{S}^1)$ can be *factorized* as $u = e^{\imath\varphi}\, v$, with $\varphi \in W^{s,p}(B;\mathbb{R})$ and $v \in F^{sp}_{1,1}(B;\mathbb{S}^1)$.

Factorization has the following application announced in the introduction. Let $p > 1$ and consider some $f \in W^{1,p}(B;\mathbb{S}^1)$. Set $u := e^{\imath f} \in W^{1,p}(B;\mathbb{S}^1)$. Let $0 < \lambda < 1$. Since $u \in W^{1,p} \cap L^\infty$, we also have $u \in W^{\lambda,p/\lambda}$ (by Gagliardo-Nirenberg). Factorization implies that $u = e^{\imath\varphi}\, v$, with $\varphi \in W^{\lambda,p/\lambda}$ and $v \in F^{p}_{1,1} \hookrightarrow W^{p,1}$.

We note that

$$W^{p,1}(B;\mathbb{R}) \ni v = e^{\imath(f-\varphi)}, \text{ with } f \in W^{1,p}(B;\mathbb{R}) \text{ and } \varphi \in W^{\lambda,p/\lambda}(B;\mathbb{R}). \tag{82}$$

We next invoke the following delicate result [5]. If $f_1 \in W^{s_1,p_1}(B;\mathbb{R})$, $f_2 \in W^{s_2,p_2}(B;\mathbb{R})$ are such that

$$s_1 p_1 \ge 1,\ s_2 p_2 \ge 1,\ e^{\imath(f_1+f_2)} \in W^{s_3,p_3},\ \text{with } s_3 \ge 1,$$

then $f_1 + f_2 \in W^{s_3,p_3} \cap W^{s_3 p_3,1}$. In our case, this implies that $\psi := f - \varphi \in W^{p,1} \cap W^{1,p}$, and thus $\varphi = f - \psi \in W^{1,p}$. Finally, $f = \varphi + \psi$, with $\varphi \in W^{\lambda,p/\lambda} \cap W^{1,p}$ and $\psi \in W^{p,1} \cap W^{1,p}$.

Our Theorem 1 yields the same conclusion without factorization.

Let us also note that not only factorization leads to a sum-intersection property, but sum-intersection is *necessary* for factorization to hold. Indeed, let $p \ge 2$ and $u \in W^{1,p}(B;\mathbb{S}^1)$. Then we may write $u = e^{\imath f}$ with $f \in W^{1,p}(B;\mathbb{R})$ [2]. Assume that we want to factorize $u = e^{\imath\varphi}\, v$ with $\varphi \in W^{\lambda,p/\lambda}(B;\mathbb{R})$ and $v \in W^{p,1}(B;\mathbb{S}^1)$. The first step consists of splitting (assuming this is possible) $f = \varphi + \psi$, with $\varphi \in W^{\lambda,p/\lambda}$ and $\psi \in W^{p,1}$. However, this decomposition does not imply that $v := e^{\imath\psi} \in W^{p,1}(B;\mathbb{S}^1)$. Indeed, if $s > 1$ and $\rho > 1$, then a map $g \in W^{\sigma,\rho}$ satisfies $e^{\imath g} \in W^{\sigma,\rho}$ if and only if g satisfies the extra-assumption $g \in W^{1,\sigma\rho}$ [6]. In our case, this implies that factorization in $W^{\lambda,p/\lambda}(B;\mathbb{S}^1)$ *requires* the sum-intersection property of the triple $T = (W^{1,p}, W^{\lambda,p/\lambda}, W^{p,1})$. However, one cannot reduce factorization to sum-intersection, since in general $W^{s,p}(B;\mathbb{S}^1)$ does not have the lifting property.

Sum-intersection property has the following implication related to lifting, presented in [5]. If $sp < 1$, then maps in $W^{s,p}(B;\mathbb{S}^1)$ can be lifted within $W^{s,p}$ [3]. Factorization leads to a better result. Indeed, let $u \in W^{s,p}(B;\mathbb{S}^1)$ and let $\varphi \in W^{s,p}(B;\mathbb{R})$ be a lifting of u. Write, as in Theorem 2, $\varphi = \varphi_1 + \varphi_2$, with $\varphi_1 \in BMO \cap W^{s,p}$ and $\varphi_2 \in W^{sp,1} \cap W^{s,p}$. Set $v := e^{\imath\varphi_2} \in W^{sp,1}$. Then v has a lifting $\varphi_3 \in W^{sp,1} \cap L^\infty$ [14]. By Gagliardo-Nirenberg, we also have $\varphi_3 \in W^{s,p}$, and clearly $\varphi_3 \in BMO$ (since $\varphi_3 \in L^\infty$). Finally, $u = e^{\imath\psi}$, where $\psi := \varphi_1 + \varphi_3$ satisfies the improved regularity $\psi \in W^{s,p} \cap BMO$.

Acknowledgements The author thanks N. Badr, H. Brezis and B. Matei for useful discussions. He was supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

1. C. Bennett, R. Sharpley, *Interpolation of Operators*. Pure and Applied Mathematics, vol. 129 (Academic, Boston, MA, 1988)
2. F. Bethuel, X.M. Zheng, Density of smooth functions between two manifolds in Sobolev spaces. J. Funct. Anal. **80**(1), 60–75 (1988)
3. J. Bourgain, H. Brezis, P. Mironescu, Lifting in Sobolev spaces. J. Anal. Math. **80**, 37–86 (2000)
4. H. Brezis, P. Mironescu, Gagliardo-Nirenberg inequalities and non-inequalities: the full story. Ann. Inst. H. Poincaré Anal. Non Linéaire (2018, to appear). https://hal.archives-ouvertes.fr/hal-01626613
5. H. Brezis, P. Mironescu, *Sobolev Maps with Values Into the Circle* (Birkhäuser, Basel) (in preparation)
6. H. Brezis, P. Mironescu, Gagliardo-Nirenberg, composition and products in fractional Sobolev spaces. J. Evol. Equ. **1**(4), 387–404 (2001). Dedicated to the memory of Tosio Kato
7. J.-Y. Chemin, *Perfect Incompressible Fluids*. Oxford Lecture Series in Mathematics and Its Applications, vol. 14 (The Clarendon Press/Oxford University Press, New York, 1998). Translated from the 1995 French original by Isabelle Gallagher and Dragos Iftimie
8. A. Cohen, W. Dahmen, I. Daubechies, R. DeVore, Harmonic analysis of the space BV. Rev. Mat. Iberoamericana **19**(1), 235–263 (2003)
9. R. DeVore, K. Scherer, Interpolation of linear operators on Sobolev spaces. Ann. Math. (2) **109**(3), 583–599 (1979)
10. C. Fefferman, E.M. Stein, Some maximal inequalities. Am. J. Math. **93**, 107–115 (1971)
11. E. Gagliardo, Ulteriori proprietà di alcune classi di funzioni in più variabili. Ricerche Mat. **8**, 24–51 (1959)
12. B.S. Kašin, Remarks on the estimation of Lebesgue functions of orthonormal systems. Mat. Sb. (N.S.) **106**(148) no 3, 380–385 (495) (1978)
13. P. Mironescu, Decomposition of $\mathbb{S}^1$-valued maps in Sobolev spaces. C. R. Math. Acad. Sci. Paris **348**(13/14), 743–746 (2010)
14. P. Mironescu, I. Molnar, Phases of unimodular complex valued maps: optimal estimates, the factorization method, and the sum-intersection property of Sobolev spaces, Ann. Inst. H. Poincaré Anal. Non Linéaire **32**(5), 965–1013 (2015)
15. L. Nirenberg, On elliptic partial differential equations. Ann. Scuola Norm. Sup. Pisa (3), **13**, 115–162 (1959)
16. F. Oru, Rôle des oscillations dans quelques problèmes d'analyse non-linéaire, Ph.D. thesis (ENS Cachan, 1998). Unpublished
17. T. Runst, W. Sickel, *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*. de Gruyter Series in Nonlinear Analysis and Applications, vol. 3 (Walter de Gruyter & Co., Berlin, 1996)
18. E.M. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*. Princeton Mathematical Series, vol. 43 (Princeton University Press, Princeton, NJ, 1993). With the assistance of Timothy S. Murphy. Monographs in Harmonic Analysis, III
19. E.M. Stein, G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton Mathematical Series, vol. 32 (Princeton University Press, Princeton, NJ, 1971)
20. H. Triebel, *Theory of Function Spaces*. Monographs in Mathematics, vol. 78 (Birkhäuser, Basel, 1983)
21. H. Triebel, *Theory of Function Spaces. II*. Monographs in Mathematics, vol. 84 (Birkhäuser, Basel, 1992)
22. H. Triebel, *Theory of Function Spaces. III*. Monographs in Mathematics, vol. 100 (Birkhäuser, Basel, 2006)

Operator-Norm Convergence of the Trotter Product Formula on Hilbert and Banach Spaces: A Short Survey



Hagen Neidhardt, Artur Stephan, and Valentin A. Zagrebnov

1 Introduction

Recall that the product formula

$$e^{-\tau C} = \lim_{n\to\infty} \left(e^{-\tau A/n} e^{-\tau B/n}\right)^n, \quad \tau \geq 0, \tag{1.1}$$

was established by S. Lie (in 1875) for matrices where $C := A + B$. The proof of formula (1.1) can be carried over easily to bounded operators on Banach spaces. Moreover, a straightforward computation shows that the convergence rate is $O(1/n)$, i.e.

$$\sup_{\tau\in[0,T]} \|e^{-\tau A/n} e^{-\tau B/n} - e^{-\tau C/n}\| = O(1/n). \tag{1.2}$$

Trotter [25] has extended this result to unbounded operators A and B on Banach spaces, but in the strong operator topology. He proved that if A and B are generators of contractions semigroups on a separable Banach space such that the algebraic sum $A + B$ is a densely defined closable operator and the closure $C = \overline{A + B}$ is a

H. Neidhardt
WIAS Berlin, Berlin, Germany
e-mail: hagen.neidhardt@wias-berlin.de

A. Stephan
Humboldt-Universität zu Berlin, Institut für Mathematik, Berlin, Germany
e-mail: stephan@math.hu-berlin.de

V. A. Zagrebnov (✉)
Université d'Aix-Marseille - Institut de Mathématiques de Marseille (UMR 7373), CMI-Technopôle Château-Gombert, Marseille, France
e-mail: valentin.zagrebnov@univ-amu.fr

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_9

generator of a contraction semigroup, then

$$e^{-\tau C} = \mathrm{s}-\lim_{n\to\infty} \left(e^{-\tau A/n} e^{-\tau B/n}\right)^n , \tag{1.3}$$

uniformly in $\tau \in [0, T]$ for any $T > 0$.

Formula (1.3) is often called the Trotter or the Lie-Trotter product formula. It was a long-time belief that this formula is valid only in the *strong* operator topology. But in nineties it was discovered that under certain quite standard assumptions the strong convergence of the Trotter product formula can be improved to the *operator-norm* convergence. In the following we give a review of these results.

The paper is organised as follows. In Sect. 2.1 we give an overview on operator-norm convergence of the Trotter product formula if the generators A and B are non-negative *self-adjoint* operators. Section 2.2 summarises the case when one of the generator is only a *maximal accretive* operator. Section 2.3 is devoted to the *evolution* case, which arises in the theory of the abstract *non-autonomous* Cauchy problem. These results are commented in Sect. 2.4.

Section 3 is concerned with the operator-norm convergence of the Trotter product formula on the Banach spaces. Section 3.1 presents the results under the assumption that one of generators is for holomorphic semigroup. Section 3.2 considers again the evolution case but on Banach spaces. The relation between *evolution semigroups* and propagators is explained in Sect. 3.3. We comment the results in Sect. 3.4.

In Sect. 4 we collect some examples and counterexamples. They show what is expectable and what is not and even surprising.

We use below the following notations and definitions.

1. We use a definition of the semigroup *generator* C (1.3), which differs from the standard one by a *minus*, as it is in the book [11].
2. Furthermore, we widely use the so-called *Landau symbols*:

$$g(n) = O(f(n)) \Longleftrightarrow \limsup_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| < \infty ,$$

$$g(n) = o(f(n)) \Longleftrightarrow \limsup_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| = 0 ,$$

$$g(n) = \Theta(f(n)) \Longleftrightarrow 0 < \liminf_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| \le \limsup_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| < \infty ,$$

$$g(n) = \omega(f(n)) \Longleftrightarrow \limsup_{n\to\infty} \left|\frac{g(n)}{f(n)}\right| = \infty .$$

3. We use the notation $C^{0,\beta}([0, T])$ for the Hölder ($\beta \in (0, 1)$) and, respectively, for the Lipschitz ($\beta = 1$) continuous functions.

2 Trotter Product Formula on Hilbert Spaces

2.1 Self-Adjoint Case

Considering the Trotter product formula on a separable Hilbert space H, T. Kato has shown in [9, 10] that for non-negative operators A and B the Trotter formula (1.3) holds in the *strong* operator topology if $\mathrm{dom}(\sqrt{A})\cap\mathrm{dom}(\sqrt{B})$ is dense in the Hilbert space and $C = A\dot{+}B$ is the form-sum of operators A and B.

Naturally the problem arises whether Kato's result can be extended to the operator-norm convergence. A first attempt in this direction was undertaken by Rogava [23]. He proved that if A and B are non-negative self-adjoint operators such that $\mathrm{dom}(A) \subseteq \mathrm{dom}(B)$ and the operator-sum: $C = A + B$, is *self-adjoint*, then

$$\|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O(\ln(n)/\sqrt{n}), \quad n \to \infty, \tag{2.4}$$

holds. In [15] it was shown that if one substitutes in above conditions the self-adjointness of the operator-sum by the A-smallness of B with a relative bound less then *one*, then (2.4) is true with the rate of convergence improved to

$$\|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O(\ln(n)/n), \quad n \to \infty.$$

The problem in its original formulation was finally solved in [8]. There it was shown that the best possible in this general setup rate (1.2) holds if the operator sum: $C = A + B$, is already a self-adjoint operator. Obviously, Rogava's result, as well as many other results (including [15]), when the operator sum of generators is self-adjoint, follow from [8] .

A new direction comes due to results for the fractional-power conditions. In [20], with elucidation in [6], it was proven that assuming

$$\mathrm{dom}(C^\alpha) \subseteq \mathrm{dom}(A^\alpha) \cap \mathrm{dom}(B^\alpha), \quad \alpha \in (1/2, 1), \quad C = A\dot{+}B, \tag{2.5}$$

and

$$\mathrm{dom}(A^{1/2}) \subseteq \mathrm{dom}(B^{1/2}) \tag{2.6}$$

one obtains that

$$\sup_{\tau\in[0,T]} \|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O(n^{-(2\alpha-1)}).$$

Notice that formally $\alpha = 1$ yields the rate obtained in [8].

We remark also that the results of [6, 20] do *not* cover the case $\alpha = 1/2$. Although, it turns out that in this case the Trotter product formula converges on the operator norm:

$$\sup_{\tau\in[0,T]} \|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = o(1) ,$$

if $\sqrt{B}$ is *relatively compact* with respect to $\sqrt{A}$, i.e. $\sqrt{B}(I + A)^{-1/2}$ is compact, see [16].

2.2 Nonself-Adjoint Case

Another direction was related with attempts to extend the Trotter, and the Trotter-Kato, product formulae to the case of nonself-adjoint *sectorial* generators [3]. Let A be a non-negative self-adjoint operator and let B be a maximal *accretive* ($\Re e(Bf, f) \geq 0$ for $f \in \mathrm{dom}(B)$) operator, such that

$$\mathrm{dom}(A) \subseteq \mathrm{dom}(B) \quad \text{and} \quad \mathrm{dom}(A) \subseteq \mathrm{dom}(B^*).$$

If B is A-small with a relative bound less than one, then the rate estimate (2.4) holds, for generator C, which is the well-defined maximal accretive operator-sum: $C = A + B$, see [1].

In [2] this result was generalised as follows. Let A be a non-negative self-adjoint operator and let B be a maximal accretive operator such that $\mathrm{dom}(A) \subseteq \mathrm{dom}(B)$ and B is A-small with relative bound less than one. If the condition

$$\mathrm{dom}((C^*)^\alpha) \subseteq \mathrm{dom}(A^\alpha) \cap \mathrm{dom}((B^*)^\alpha), \quad C = A + B ,$$

is satisfied for some $\alpha \in (0, 1]$, then the norm-convergent Trotter product formula:

$$\sup_{\tau\in[0,T]} \|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O(\ln(n)/n^\alpha) ,$$

holds as $n \to \infty$.

In fact, more results are known about the operator-norm Trotter product formula convergence for nonself-adjoint semigroups, but *without* the rate estimates, see [4].

2.3 Evolution Case

At the first glance a very different result about a Trotter-type product formula was obtained in [7]. The authors consider instead of the self-adjoint operator B a family $\{B(t)\}_{t\in[0,T]}$ of self-adjoint operators on the separable Hilbert space $\mathfrak{H}$ such that the condition

$$\mathrm{dom}(A^{\alpha}) \subseteq \mathrm{dom}(B(t)), \quad t \in [0, T], \tag{2.7}$$

is satisfied for some $\alpha \in [0, 1)$ for A, which is a non-negative self-adjoint operator. Then the operator sum $C(t) = A + B(t)$ defines a family of self-adjoint operators in $\mathfrak{H}$ space such that $\mathrm{dom}(C(t)) = \mathrm{dom}(A)$, $t \in [0, T]$. With the family $\{C(t)\}_{t\in[0,T]}$ one associates the *evolution* equation:

$$\partial_t u(t) = -C(t)u(t), \quad t \in [0, T]\,, \tag{2.8}$$

corresponding to the non-autonomous Cauchy problem with initial condition $u_0 = u(0)$ for $t = 0$.

It turns out that equation (2.8) admits a *propagator* $\{U(t, s)\}_{(t,s)\in\bar{\Delta}}$, $\bar{\Delta} = \{(t, s) \in [0, T] \times [0, T] : 0 \leq s \leq t \leq T\}$, which solves this problem. We remind that a family $\{U(t, s)\}_{(t,s)\in\Delta}$ of bounded operators is called a propagator if the operator-valued function $U(\cdot, \cdot) : \bar{\Delta} \longrightarrow \mathscr{B}(\mathfrak{H})$ is strongly continuous and verifies the conditions:

$$U(t, t) = I \quad \text{for} \quad t \in [0, T], \tag{2.9}$$

$$U(t, s) = U(t, r)U(r, s) \quad \text{for} \quad t, r, s \in [0, T] \quad \text{with} \quad s \leq r \leq t\,. \tag{2.10}$$

Let $\{t_j\}_{j=0}^{n}$ be a partition of the closed interval $[0, t]$:

$$0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = t, \quad t_j = \tau j, \tau = t/n\,,$$

for any $0 < t \leq T$. Further, let

$$Q_j(t, s; n) = e^{-\frac{t-s}{n}A} e^{-\frac{t-s}{n}B(s+j\frac{t-s}{n})}, \quad j = 0, 1, \ldots, n\,, \tag{2.11}$$

and

$$\begin{aligned} E(t, s; n) &:= Q_{n-1}(t, s; n)Q_{n-2}(t, s; n) \times \cdots \times Q_1(t, s; n)Q_0(t, s; n) \\ &:= \prod_{j=0}^{(n-1)\leftarrow} Q_j(t, s\,; n)\,, \end{aligned} \tag{2.12}$$

where the symbol $\prod_{j=1}^{n\leftarrow}$ means that the product is increasingly ordered in j from the right to the left. If in addition to assumption (2.7), the condition

$$\|A^{-\alpha}(B(t) - B(s))A^{-\alpha}\| \leq L_1|t - s|, \quad t, s \in [0, T], \quad L_1 > 0, \tag{2.13}$$

is satisfied, then in [7] it was proved that the propagator $\{U(t, s)\}_{(t,s)\in\bar{\Delta}}$, which solves the Cauchy problem (2.8), admits the approximation

$$\sup_{t\in[0,T]} \|U(t,0) - E_n(t,0;n)\| = O(\ln(n)/n) \quad \text{as} \quad n \to \infty. \tag{2.14}$$

Scrutinising the proof in [7] one finds that in fact the claim (2.14) can be slightly generalised to any interval $\bar{\Delta}$

$$\sup_{(t,s)\in\bar{\Delta},t-s\geq\tau_0>0} \|U(t,s) - E_n(t,s;n)\| = O(\ln(n)/n) \quad \text{as} \quad n \to \infty, \tag{2.15}$$

for every $\tau_0 > 0$. At the first glance, it seems that the result (2.15) is quite far from the Trotter product formula. However, this is not the case. To show this we follow the evolution semigroup approach to evolution equations developed in [12, 21]. Let us introduce the Hilbert space $L^2([0,T],\mathfrak{H})$ and consider on this space the semigroup

$$(\mathscr{U}(\tau)f)(t) := U(t,t-\tau)\chi_{[0,T]}(t-\tau)f(t-\tau), \quad f \in L^2([0,T],\mathfrak{H}), \tag{2.16}$$

$t \in [0,T]$. It turns out that $\{\mathscr{U}(\tau)\}_{\tau\in\mathbb{R}_+}$ is a C_0-semigroup on the Hilbert space of the $\mathfrak{H}$-valued trajectories $L^2([0,T],\mathfrak{H})$. By $\mathscr{K}$ we denote its generator, then $\mathscr{U}(\tau) = e^{-\tau\mathscr{K}}$, $\tau \geq 0$.

Let us introduce two multiplication operators

$$\begin{aligned}
&(\mathscr{A}f)(t) = Af(t), \quad f \in \mathrm{dom}(\mathscr{A}),\\
&\mathrm{dom}(\mathscr{A}) := \left\{ f \in L^2([0,T],\mathfrak{H}) : \begin{matrix} f(t) \in \mathrm{dom}(A) \quad \text{for a.e. } t \in [0,T] \\ Af(t) \in L^2([0,T],\mathfrak{H}) \end{matrix} \right\},
\end{aligned} \tag{2.17}$$

and

$$\begin{aligned}
&(\mathscr{B}f)(t) = B(t)f(t), \quad f \in \mathrm{dom}(\mathscr{B}),\\
&\mathrm{dom}(\mathscr{B}) := \left\{ f \in L^2([0,T],\mathfrak{H}) : \begin{matrix} f(t) \in \mathrm{dom}(B(t)) \quad \text{for a.e. } t \in [0,T] \\ B(t)f(t) \in L^2([0,T],\mathfrak{H}) \end{matrix} \right\}.
\end{aligned} \tag{2.18}$$

Note that both operators are well-defined and self-adjoint in $L^2([0,T],\mathfrak{H})$. Moreover, condition (2.7) yields: $\mathrm{dom}(\mathscr{A}) \subseteq \mathrm{dom}(\mathscr{B})$.

By D_0 we define in $L^2([0,T],\mathfrak{H})$ the generator of the right-shift semigroup $e^{-\tau D_0}$ given by

$$(e^{-\tau D_0}f)(t) = \chi_{[0,T]}(t-\tau)f(t-\tau), \quad f \in L^2([0,T],\mathfrak{H}).$$

Notice that the operator D_0 is defined by

$$\begin{aligned}
(D_0 f)(t) &= \frac{\partial}{\partial t} f(t),\\
f \in \mathrm{dom}(D_0) &:= \{ f \in W^{2,2}([0,T],\mathfrak{H}) : f(0) = 0 \}.
\end{aligned}$$

Collecting these definitions we introduce the operator

$$\begin{aligned}(\widetilde{\mathscr{K}} f) &= D_0 f + \mathscr{A} f + \mathscr{B} f,\\ f \in \mathrm{dom}(\widetilde{\mathscr{K}}) &:= \mathrm{dom}(D_0) \cap \mathrm{dom}(\mathscr{A}) \cap \mathrm{dom}(\mathscr{B}).\end{aligned} \tag{2.19}$$

Since $\{U(t,s)\}_{(t,s)\in\bar{\Delta}}$ is a propagator solving the evolution equation (2.8) one deduces that by virtue of assumptions (2.7) and (2.13) the operator $\widetilde{\mathscr{K}}$ is closable and that its closure coincides with generator $\mathscr{K}$, see [12, Theorem 4.5].

Furthermore, let us define the operator

$$(\mathscr{K}_0 f)(t) = D_0 f + \mathscr{A} f, \quad f \in \mathrm{dom}(\mathscr{K}_0) := \mathrm{dom}(D_0) \cap \mathrm{dom}(\mathscr{A}).$$

Then operator $\mathscr{K}_0$ is a generator of the C_0-semigroup, which has the form

$$(e^{-\tau\mathscr{K}_0} f)(t) = e^{-\tau A}\chi_{t\in[0,T]}(t-\tau) f(t-\tau), \quad f \in L^2([0,T],\mathfrak{H}).$$

Note that we obviously get that $\mathscr{K} = \overline{\mathscr{K}_0 + \mathscr{B}} = \mathscr{K}_0 + \mathscr{B}$.

For the pair $\{\mathscr{K}_0, \mathscr{B}\}$ one obtains the following result.

Proposition 2.1 *Let A be a non-negative self-adjoint operator on the separable Hilbert space $\mathfrak{H}$ and let $\{B(t)\}_{t\in[0,T]}$ be a family of non-negative self-adjoint operators. If the assumptions (2.7) and (2.13) are satisfied, then*

$$\sup_{\tau\geq\tau_0>0} \|(e^{-\tau\mathscr{K}_0/n} e^{-\tau\mathscr{B}/n})^n - e^{-\tau\mathscr{K}}\| = O(\ln(n)/n) \quad \text{for } n\to\infty\,, \tag{2.20}$$

and

$$\sup_{\tau\geq\tau_0>0} \|(e^{-\tau\mathscr{B}/n} e^{-\tau\mathscr{K}_0/n})^n - e^{-\tau\mathscr{K}}\| = O(\ln(n)/n) \quad \text{for } n\to\infty, \tag{2.21}$$

for every $\tau_0 > 0$.

Proof A straightforward computation shows that for $f \in L^2([0,T],\mathfrak{H})$ one gets

$$((e^{-\tau\mathscr{K}_0/n} e^{-\tau\mathscr{B}/n})^n f)(t) = E(t, t-\tau\,; n)\chi_{[0,T]}(t-\tau) f(t-\tau),$$

where $E(t,s\,;n)$ is given by (2.12). Taking into account (2.16) we obtain

$$\begin{aligned}&((e^{-\tau\mathscr{K}_0/n} e^{-\tau\mathscr{B}/n})^n f)(t) - (e^{-\tau\mathscr{K}} f)(t)\\ &\qquad = (E(t,t-\tau\,;n) - U(t,t-\tau))\chi_{[0,T]}(t-\tau) f(t-\tau)\,,\end{aligned}$$

which yields the estimate

$$\|((e^{-\tau\mathscr{K}_0/n}e^{-\tau\mathscr{B}/n})^n f)(t) - (e^{-\tau\mathscr{K}} f)(t)\| \\ \leq \chi_{[0,T]}(t-\tau)\|E(t,t-\tau\,;n) - U(t,t-\tau)\|\,\|f(t-\tau)\|.$$

Hence, this implies

$$\sup_{\tau\geq\tau_0>0}\|(e^{-\tau\mathscr{K}_0/n}e^{-\tau\mathscr{B}/n})^n f - e^{-\tau\mathscr{K}} f\|^2 \\ \leq \sup_{\tau\geq\tau_0>0}\int_0^T \chi_{[0,T]}(t-\tau)\|E(t,t-\tau\,;n) - U(t,t-\tau)\|^2\,\|f(t-\tau)\|^2 dt \\ \leq \int_0^T \sup_{(t,s)\in\vec{\Delta},t-s\geq\tau_0}\|E(t,s\,;n) - U(t,s)\|^2\,\|f(r)\|^2 dr.$$

Using (2.15) we immediately obtain (2.20). Similarly (2.21) follows from [7]. □

2.4 Comments

Section 2.1

The operator-norm convergence rate $O(1/n)$ of [8] for pairs of non-negative self-adjoint operators is *sharp* and *ultimate optimal* due to observations in [24]. The same remark concerns the sharpness and optimality of the rate $O(1/n^{2\alpha-1})$ obtained first in [20] under assumption (2.5) together with the A^α-smallness of B^α with relative bound less then one. Then the same ultimate sharp rate was proven in [6], when the smallness condition is relaxed to the mild subordination (2.6). It is an open problem whether the assumption (2.6) is really necessary.

Section 2.2

It is unclear whether the convergence rates $O(\ln(n)/n)$ and $O(\ln(n)/n^\alpha)$ are sharp. One expects convergence rates identical to that in Sect. 2.1.

Section 2.3

The approach used here was developed in [12, 17–19, 21, 22]. The idea is to transform a *time-dependent* evolution problem to a *time-independent* problem, see also the next section.

Let us add some remarks. One easily checks that the operator $\mathscr{K}_0$ is not self-adjoint whereas the operator $\mathscr{B}$ is self-adjoint. However, $\mathscr{K}_0$ is maximal accretive. This is in some sense in contrast to Sect. 2.2, where the pair $\{A, B\}$ consists of a

self-adjoint operator A and an maximal accretive operator B such that $\mathrm{dom}(A) \subseteq \mathrm{dom}(B)$ and $\mathrm{dom}(A) \subseteq \mathrm{dom}(B^*)$. In the evolution case the conditions $\mathrm{dom}(\mathscr{K}_0) \subseteq \mathrm{dom}(\mathscr{B})$ and $\mathrm{dom}(\mathscr{K}_0^*) \subseteq \mathrm{dom}(\mathscr{B})$ are satisfied but in the reversed order with respect to Sect. 2.2. However, the convergence rate $O(\ln(n)/n)$ is not affected by this.

The proof of the estimate (2.14) and (2.15) are very involved. Naturally the problem arises whether one can give a direct proof the estimate (2.14) avoiding those propagator estimates.

3 Trotter Product Formula on Banach Spaces

3.1 Holomorphic Case

There are only few generalisations of the results of Sect. 2 to Banach spaces. The main obstacle for that is the fact that the concept of self-adjointness is missing in the Banach spaces. One of solution is to relax the self-adjointness replacing the non-negative self-adjoint generator A by a generator of the holomorphic semigroup. The following result was proved in [5].

Theorem 3.1 ([5, Theorem 3.6 and Corollary 3.7]) *Let A be a generator of a holomorphic contraction semigroup on the separable Banach space $\mathfrak{X}$ and let B a generator of a contraction semigroup on $\mathfrak{X}$.*

(i) *If for some $\alpha \in (0,1)$ the condition*

$$\mathrm{dom}(A^\alpha) \subseteq \mathrm{dom}(B)\,,$$

holds and $\mathrm{dom}(A^*) \subseteq \mathrm{dom}(B^*)$ is satisfied, then the operator sum $C = A + B$ is a generator of a contraction semigroup and

$$\sup_{\tau\in[0,T]} \|(e^{-\tau B/n}e^{-\tau A/n})^n - e^{-\tau C}\| = O(\ln(n)/n^{1-\alpha})\,, \tag{3.22}$$

for any $T > 0$.

(ii) *If for some $\alpha \in (0,1)$ the condition*

$$\mathrm{dom}((A^\alpha)^*) \subseteq \mathrm{dom}(B^*)\,,$$

is satisfied and $\mathrm{dom}(A) \subseteq \mathrm{dom}(B)$ is valid, then $C = A + B$ is the generator of a contraction semigroup and

$$\sup_{\tau\in[0,T]} \|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O(\ln(n)/n^{1-\alpha})\,, \tag{3.23}$$

for any $T > 0$.

Theorem 3.2 ([5, Theorem 3.6 and Corollary 3.7]) *Let A be a generator of a holomorphic contraction semigroup on $\mathfrak{X}$ and let B a generator of a contraction semigroup on $\mathfrak{X}$. If B is in addition a bounded operator, then*

$$\sup_{\tau\in[0,T]} \|(e^{-\tau B/n}e^{-\tau A/n})^n - e^{-\tau C}\| = O((\ln(n))^2/n),$$

and

$$\sup_{\tau\in[0,T]} \|(e^{-\tau A/n}e^{-\tau B/n})^n - e^{-\tau C}\| = O((\ln(n))^2/n),$$

for any $T > 0$.

3.2 Evolution Case

Similarly of Sect. 2.3 let us consider a generator A of a holomorphic semigroup on the separable Banach space $\mathfrak{X}$ and a family of $\{B(t)\}_{t\in[0,T]}$ of generators of holomorphic semigroups on $\mathfrak{X}$. We make the following assumptions:

Assumption 3.3

(A1) The operator A is a generator of a holomorphic contraction semigroup on $\mathfrak{X}$ such that $0 \in \rho(A)$.

(A2) Let $\{B(t)\}_{t\in[0,T]}$ be a family of closed operators such that for a.e. $t \in [0, T]$ and some $\alpha \in (0, 1)$ the condition $\mathrm{dom}(A^\alpha) \subset \mathrm{dom}(B(t))$ is satisfied such that

$$C_\alpha := \mathrm{ess\,sup}_{t\in[0,T]} \|B(t)A^{-\alpha}\|_{\mathscr{B}(X)} < \infty\,.$$

(A3) Let $\{B(t)\}_{t\in[0,T]}$ be a family of generators of contraction semigroups in $\mathfrak{X}$ such that the function $[0, T] \ni t \mapsto (B(t)+\xi)^{-1}x \in \mathfrak{X}$ is strongly measurable for any $x \in \mathfrak{X}$ and any $\xi > b > 0$.

(A4) We assume that $\mathrm{dom}(A^*) \subset \mathrm{dom}(B(t)^*)$ and

$$C_1^* := \mathrm{ess\,sup}_{t\in[0,T]} \|B(t)^*(A^*)^{-1}\|_{\mathscr{B}(X^*)} < \infty,$$

where A^* and $B(t)^*$ denote operators which are adjoint of A and $B(t)$, respectively.

(A5) There exists $\beta \in (\alpha, 1)$ and a constant $L_\beta > 0$ such that for a.e. $t, s \in [0, T]$ one has the estimate:

$$\|A^{-1}(B(t) - B(s))A^{-\alpha}\| \le L_\beta |t - s|^\beta\,.$$

(A6) There exists a constant $L_1 > 0$ such that for a.e. $t, s \in [0, T]$ one has the estimate:

$$\|A^{-\alpha}(B(t) - B(s))A^{-\alpha}\| \leq L_1|t - s| \,.$$

The assumption $0 \in \rho(A)$ in (A1) is made for simplicity. We note that assumption (A3) is similar to (2.7). Assumption (A4) is automatically satisfied for self-adjoint operators. Assumption (A5) is a modification of (2.13) while assumption (A6) coincides with (2.13).

With the family $\{C(t)\}_{t\in[0,T]}$, $C(t) = A + B(t)$, one associates the evolution equation (2.8). It turns out that under the assumptions (A1) and (A2) the family $\{C(t)\}_{t\in[0,T]}$ consists of generators of contraction semigroups.

In accordance with Sect. 2.3 we consider the Banach space $L^p([0, T], \mathfrak{X})$ for some fixed $p \in [1, \infty)$ and introduce the multiplication operators $\mathscr{A}$ and $\mathscr{B}$, cf. (2.17) and (2.18). Under the assumptions (A1)–(A3) the operators $\mathscr{A}$ and $\mathscr{B}$ are generators of contraction semigroups such that $\mathrm{dom}(\mathscr{A}) \subseteq \mathrm{dom}(\mathscr{B})$, in particular, $\mathscr{A}$ is a generator of holomorphic semigroup. Similarly, one can introduce the multiplication operator $\mathscr{C}$ induced by the family $\{C(t)\}_{t\in[0,T]}$ which is also a generator of a holomorphic semigroup. Notice that $\mathscr{C} = \mathscr{A} + \mathscr{B}$ and $\mathrm{dom}(\mathscr{C}) = \mathrm{dom}(\mathscr{A})$.

Let D_0 the generator of the right-shift semigroup on $L^p([0, T], \mathfrak{X})$, i.e.

$$(e^{-\tau D_0} f)(t) = f(t - \tau)\chi_{[0,T]}(t - \tau) f(t - \tau), \quad f \in L^p([0, T], \mathfrak{X}),$$

cf. (2.19). We consider the operator

$$\widetilde{\mathscr{K}} f = D_0 f + \mathscr{A} f + \mathscr{B} f,$$
$$f \in \mathrm{dom}(\widetilde{\mathscr{K}}) = \mathrm{dom}(D_0) \cap \mathrm{dom}(\mathscr{A}) \cap \mathrm{dom}(\mathscr{B}),$$

cf. (2.19). Assuming (A1)–(A3) it was shown in [12] that the operator $\widetilde{K}$ is closable and its closure $\mathscr{K}$ is the generator of a semigroup.

Furthermore, we set

$$\widetilde{\mathscr{K}_0} f = D_0 f + \mathscr{A} f, \quad f \in \mathrm{dom}(\widetilde{\mathscr{K}_0}) = \mathrm{dom}(D_0) \cap \mathrm{dom}(\mathscr{A}),$$

In contrast to the Hilbert space the operator $\widetilde{\mathscr{K}_0}$ is not necessary a generator of a semigroup. However, the operator $\mathscr{K}_0$ closable and its closure $\mathscr{K}_0$ is a generator. Notice that $\mathscr{K}$ coincides with the algebraic sum of $\mathscr{K}_0$ and $\mathscr{B}$, i.e $\mathscr{K} = \mathscr{K}_0 + \mathscr{B}$.

In [12] the following theorem was proved.

Theorem 3.4 ([12, Theorem 7.8]) *Let the assumptions (A1)–(A4) be satisfied for some $\alpha \in (0, 1)$. If (A5) holds, then*

$$\sup_{\tau\geq 0} \|(e^{-\tau B/n} e^{-\tau \mathscr{K}_0/n})^n - e^{-\tau \mathscr{K}}\| = O(1/n^{\beta-\alpha}), \tag{3.24}$$

when $n \to \infty$.

Assuming instead of assumption (A5) the assumption (A6) the result slightly modifies, see [14].

Theorem 3.5 ([14, Theorem 5.4]) *Let the assumptions (A1)–(A4) be satisfied for some* $\alpha \in (1/2, 1)$. *If (A6) is valid, then for* $n \to \infty$ *one gets the asymptotic:*

$$\sup_{\tau \geq 0} \| (e^{-\tau B/n} e^{-\tau \mathscr{K}_0/n})^n - e^{-\tau \mathscr{K}} \| = O(1/n^{1-\alpha}). \tag{3.25}$$

3.3 Convergence Rates for Propagators

The proof of both theorems does not use propagator approximations of type (2.14) or (2.15). However, Theorems 3.4 and 3.5 can be used to prove propagator approximations. To this end one has to introduce the notion of a evolution semigroup.

Definition 3.6 A generator $\mathscr{K}$ in $L^p([0, T], X)$, $p \in [1, \infty)$, is called a evolution generator if

(i) $\mathrm{dom}(\mathscr{K}) \subset C([0, T], X)$ and $M(\phi)\mathrm{dom}(\mathscr{K}) \subset \mathrm{dom}(\mathscr{K})$ for $\phi \in W^{1,\infty}([0, T])$,
(ii) $\mathscr{K} M(\phi) f - M(\phi)\mathscr{K} f = M(\dot{\phi}) f$ for $f \in \mathrm{dom}(\mathscr{K})$ and $\phi \in W^{1,\infty}([0, T])$, where $\dot{\phi} = \partial_t \phi$,
(iii) the domain $\mathrm{dom}(\mathscr{K})$ has a dense cross-section, i.e. for each $t \in (0, T]$ the set

$$[\mathrm{dom}(\mathscr{K})]_t := \{x \in \mathfrak{X} : \exists f \in \mathrm{dom}(\mathscr{K}) \text{ such that } x \in f(t)\},$$

is dense in $\mathfrak{X}$.

By $M(\phi)$, $\phi \in L^\infty([0, 1])$, the bounded operator

$$(M(\phi) f)(t) = \phi(t) f(t), \quad f \in L^p([0, T], \mathfrak{X}).$$

is meant.

One can check that the operator $\mathscr{K}$ defined as the closure of $\widetilde{\mathscr{K}}$ is an evolution generator, cf. [12, Theorem 1.2]. Evolution generators a directly related to propagators. For this purpose one has slightly weaken the notion of a propagator defined in Sect. 2.3.

Definition 3.7 Let $\{U(t, s)\}_{(t,s)\in\Delta}$, $\Delta = \{(t, s) \in (0, T] \times (0, T] : s \leq t \leq T\}$, be a strongly continuous family of bounded operators on $\mathfrak{X}$. If the conditions

$$U(t, t) = I \quad \text{for} \quad t \in (0, T], \tag{3.26}$$

$$U(t, r)U(r, s) = U(t, s) \quad \text{for} \quad t, r, s \in (0, T] \quad \text{with} \quad s \leq r \leq t, \tag{3.27}$$

$$\|U\|_{\mathscr{B}(X)} := \sup_{(t,s)\in\Delta} \|U(t,s)\| < \infty \tag{3.28}$$

are satisfied, then $\{U(t,s)\}_{(t,s)\in\Delta}$ is called a propagator.

Comparing with Sect. 2.3 we note that Δ slightly differs from $\bar{\Delta}$. Indeed, $\Delta \subseteq \bar{\Delta}$ but $\overline{\Delta} = \bar{\Delta}$. Restricting (2.9) and (2.10) to $(0, T]$ we get (3.26) and (3.27), respectively. Condition (3.28) is necessary because the set Δ is not closed.

It is known that there is an one-to-one correspondence between the set of all evolution generators on $L^p([0,T],\mathfrak{X})$ and the set of all propagators in the sense of Definition 3.7 established by

$$(e^{-\tau\mathscr{K}} f)(t) = U(t, t-\tau)\chi_{[0,T]}(t-\tau) f(t-\tau), \quad f \in L^p([0,T],\mathfrak{X}),$$

cf. [12, Theorem 3.3] or [17, Theorem 4.12].

Let $\mathscr{K}_0$ be the generator of an evolution semigroup $\{\mathscr{U}_0(\tau)\}_{\tau\geq 0}$ and let $\mathscr{B}$ be a multiplication operator induced by a measurable family $\{B(t)\}_{t\in[0,T]}$ of generators of contraction semigroups. Note that in this case the multiplication operator $\mathscr{B}$ is a generator of a contraction semigroup $(e^{-\tau\mathscr{B}} f)(t) = e^{-\tau B(t)} f(t)$, on the Banach space $L^p([0,T], X)$. Since $\{\mathscr{U}_0(\tau)\}_{\tau\geq 0}$ is an evolution semigroup, then there is a propagator $\{U_0(t,s)\}_{(t,s)\in\Delta}$ such that the representation

$$(\mathscr{U}_0(\tau) f)(t) = U_0(t, t-\tau)\chi_{[0,T]}(t-\tau) f(t-\tau), \quad f \in L^p([0,T], X),$$

is valid for a.e. $t \in [0,T]$ and $\tau \geq 0$. Then we define

$$Q_j(t,s;n) := U_0(s + j\tfrac{(t-s)}{n}, s + (j-1)\tfrac{(t-s)}{n}) e^{-\frac{(t-s)}{n} B\left(s+(j-1)\frac{(t-s)}{n}\right)}$$

where $j \in \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$, $(t,s) \in \Delta$, and we set

$$V_n(t,s) := \prod_{j=1}^{n\leftarrow} Q_j(t,s;n), \quad n \in \mathbb{N}, \quad (t,s) \in \Delta,$$

where the product is increasingly ordered in j from the right to the left. Then a straightforward computation shows that the representation

$$\left(\left(e^{-\tau\mathscr{K}_0/n} e^{-\tau\mathscr{B}/n}\right)^n f\right)(t) = V_n(t, t-\tau)\chi_{[0,T]}(t-\tau) f(t-\tau)\,,$$

$f \in L^p([0,T], X)$, holds for each $\tau \geq 0$ and a.e. $t \in [0,T]$. Similarly we can introduce

$$G_j(t,s;n) = e^{-\frac{t-s}{n} B(s+j\frac{t-s}{n})} U_0(s + j\tfrac{t-s}{n}, s + (j-1)\tfrac{t-s}{n})$$

where $j \in \{1, 2, \ldots, n\}$, $n \in \mathbb{N}$, $(t,s) \in \Delta$. Let

$$U_n(t,s) := \prod_{j=1}^{n\leftarrow} G_j(t,s;n), \quad n \in \mathbb{N}, \quad (t,s) \in \Delta,$$

where the product is again increasingly ordered in j from the right to the left. We verify that

$$\left(\left(e^{-\tau\mathscr{B}/n}e^{-\tau\mathscr{K}_0/n}\right)^n f\right)(t) = U_n(t,t-\tau)\chi_{[0,T]}(t-\tau)f(t-\tau)\,,$$

$f \in L^p([0,T],X)$, holds for each $\tau \geq 0$ and a.e. $t \in [0,T]$.

Proposition 3.8 ([13, Proposition 2.1]) *Let $\mathscr{K}$ and $\mathscr{K}_0$ be generators of evolution semigroups on the Banach space $L^p([0,T],X)$ for some $p \in [1,\infty)$. Further, let $\{B(t)\}_{t\in[0,T]}$ be a strongly measurable family of generators of contraction on $\mathfrak{X}$. Then*

$$\sup_{\tau\geq 0}\left\|e^{-\tau\mathscr{K}} - \left(e^{-\tau\mathscr{K}_0/n}e^{-\tau\mathscr{B}/n}\right)^n\right\| = \operatorname*{ess\,sup}_{(t,s)\in\Delta}\|U(t,s)-V_n(t,s)\|, \quad n \in \mathbb{N},$$

and

$$\sup_{\tau\geq 0}\left\|e^{-\tau\mathscr{K}} - \left(e^{-\tau\mathscr{B}/n}e^{-\tau\mathscr{K}_0/n}\right)^n\right\| = \operatorname*{ess\,sup}_{(t,s)\in\Delta}\|U(t,s)-U_n(t,s)\|, \quad n \in \mathbb{N}.$$

Let us introduce the approximations

$$U_n(t,s) := \prod_{j=1}^{n\leftarrow} G_j(t,s\,;n), \quad n = 1,2,\dots\,,$$

$$G_j(t,s\,;n) := e^{-\frac{t-s}{n}B(s+j\frac{t-s}{n})}e^{-\frac{t-s}{n}A}, \quad j = 0,1,2,\dots,n,$$

$(t,s) \in \Delta$, with increasingly ordered product in j from the right to the left. From Theorem 3.4 and Proposition 3.8 one immediately obtains the following theorem.

Theorem 3.9 ([12, Theorem 1.4]) *Let the assumptions (A1)–(A4) be satisfied. If (A5) holds, then for $n \to \infty$ the rate:*

$$\operatorname*{ess\,sup}_{(t,s)\in\Delta}\|U_n(t,s)-U(t,s)\| = O(1/n^{\beta-\alpha})\,. \tag{3.29}$$

From Theorem 3.5 and Proposition 3.8 we get

Theorem 3.10 ([14, Theorem 5.6]) *Let the assumptions (A1)–(A4) be satisfied for some $\alpha \in (1/2,1)$. If (A6) is valid, then for $n \to \infty$ one obtains a better rate:*

$$\operatorname*{ess\,sup}_{(t,s)\in\Delta}\|U_n(t,s)-U(t,s)\| = O(1/n^{1-\alpha})\,.$$

3.4 Comments

Section 3.1

It is unclear whether Theorem 3.1 is sharp. Theorem 3.1 should be valid if the contractivity of the involved semigroups is replaced Trotter-stability. A pair of generators $\{A, B\}$ is called Trotter-stable if the Trotter product is uniformly bounded in $n \in \mathbb{N}$, i.e., if

$$\sup_{n\in\mathbb{N}} \sup_{\tau\geq 0} \left\| \left(e^{-\tau A/n} e^{-\tau B/n} \right)^n \right\| < \infty.$$

It turns out that the pair $\{A, B\}$ is Trotter-stable if and only $\{B, A\}$ is Trotter-stable.

Section 3.2

Theorems 3.4 and 3.5 are proved in [12] and [14] under the assumption that the pair $\{\mathscr{K}_0, \mathscr{B}\}$ is Trotter-stable. The convergence rates (3.24) and (3.25) differ significantly from the convergence rate $O(\ln(n)/n)$ of Proposition 2.1. It is an open problem whether the convergences rates (3.24) and (3.25) can be improved to $O(\ln(n)/n)$. One has to mention that the convergence (3.25) coincides with that one of (3.23) despite the fact that $\mathscr{K}_0$ is not a generator of a holomorphic semigroup. Indeed, the generator $\mathscr{K}_0$ is not holomorphic since $e^{-\tau\mathscr{K}_0} = 0$ for $\tau \geq T$.

Section 3.3

It is a bit surprising that the operator-norm convergence of the Trotter product formula for the pairs: generator of evolution semigroup and multiplication operator, is equivalent to the operator-norm convergence of a certain approximation of the corresponding propagator, see Proposition 3.8. In particular, this yields that two convergences: (2.15) and (2.20), are equivalent.

4 Sharpness

4.1 Example

Let us consider a "solvable" example. We study bounded perturbations of the evolution generator D_0. To do this aim we consider $\mathfrak{X} = \mathbb{C}$ and we denote by $L^2([0, 1])$ the Hilbert space $L^2([0, T], \mathbb{C})$.

For $t \in [0,1]$, let $q : t \mapsto q(t) \in L^\infty([0,1])$. Then, q induces a bounded multiplication operator Q in the Hilber space $L^2([0,1])$:

$$(Qf)(t) = q(t)f(t), \quad f \in L^2([0,1]).$$

For simplicity we assume that $q \geq 0$. Then Q generates on $L^2([0,1])$ a contraction semigroup $\{e^{-\tau Q}\}_{\tau \geq 0}$. Since generator Q is bounded, the closed operator $\mathscr{K} := D_0 + Q$, with domain $\mathrm{dom}(\mathscr{K}) = \mathrm{dom}(D_0)$, is generator of a semigroup on $L^2([0,1])$. From [25] one gets immediately

$$\text{s-}\lim_{n\to\infty} \left(e^{-\tau D_0/n} e^{-\tau Q/n}\right)^n = e^{-\tau(D_0+Q)}$$

uniformly in $\tau \in [0,T]$ for any $T > 0$. One easily checks that $\mathscr{K}$ is an evolution generator. A straightforward computation shows that

$$\left(e^{-\tau(D_0+Q)} f\right)(t) = e^{-\int_{t-\tau}^t q(y)dy} \chi_{[0,1]}(t-\tau) f(t-\tau)$$

which yields that the propagator corresponding to $\mathscr{K}$ is given by

$$U(t,s) = e^{-\int_s^t q(y)dy}, \quad (t,s) \in \Delta.$$

A simple computation shows that

$$\left(\left(e^{-\tau D_0/n} e^{-\tau Q/n}\right)^n f\right)(t) =: V_n(t, t-\tau)\chi_{[0,T]}(t-\tau) f(t-\tau)\,.$$

Then by straightforward calculations one finds that

$$V_n(t,s) = e^{-\frac{t-s}{n}\sum_{k=0}^{n-1} q(s+k\frac{t-s}{n})}, \quad (t,s) \in \Delta\,.$$

Proposition 4.1 ([13, Proposition 3.1]) *Let $q \in L^\infty([0,T])$ be non-negative. Then*

$$\sup_{\tau\geq 0} \left\| e^{-\tau(D_0+Q)} - \left(e^{-\tau D_0/n} e^{-\tau Q/n}\right)^n \right\|_{\mathscr{B}(L^p([0,1]))}$$
$$= \Theta\left(\operatorname*{ess\,sup}_{(t,s)\in\Delta} \left| \int_s^t q(y)dy - \frac{t-s}{n}\sum_{k=0}^{n-1} q(s+k\tfrac{t-s}{n}) \right| \right)$$

as $n \to \infty$, where Θ is the Landau symbol defined in the introduction.

Note that by Proposition 3.8 the operator-norm convergence rate of the Trotter product formula for the pair $\{D_0, Q\}$ coincides with the convergence rate of the integral Darboux-Riemann sum approximation of the Lebesgue integral.

4.2 Results

Below we give a series of examples which show the dependence of the convergence rate on the smoothness of the function $q \in L^\infty([0,T])$. First we consider the Hölder and Lipschitz continuous cases.

Theorem 4.2 ([13, Theorem 3.2]) *If the function: $q \in C^{0,\beta}([0,T])$, $\beta \in (0,1]$, is non-negative, then for $n \to \infty$ one gets*

$$\sup_{\tau \geq 0} \left\| e^{-\tau(D_0+Q)} - \left(e^{-\tau D_0/n} e^{-\tau Q/n} \right)^n \right\| = O(1/n^\beta) \, .$$

Now a natural question that one may to ask is: what happens, when q is simply *continuous*?

Theorem 4.3 ([13, Theorem 3.3]) *If $q : [0,1] \to \mathbb{C}$, is continuous and non-negative, then for $n \to \infty$*

$$\left\| e^{-\tau(D_0+Q)} - \left(e^{-\tau D_0/n} e^{-\tau Q/n} \right)^n \right\| = o(1) \, . \tag{4.30}$$

We comment that for a general continuous q one can say nothing about the Trotter product formula convergence rate. Indeed, as it follows from the next theorem the convergence to zero in (4.30) may be *arbitrary* slow.

Theorem 4.4 ([13, Theorem 3.4]) *Let $\delta_n > 0$ be a sequence with $\delta_n \to 0$ as $n \to \infty$. Then there exists a continuous function $q : [0,1] \to \mathbb{R}$ such that*

$$\sup_{\tau \geq 0} \left\| e^{-\tau(D_0+Q)} - \left(e^{-\tau D_0/n} e^{-\tau Q/n} \right)^n \right\|_{\mathscr{B}(L^p([0,1]))} = \omega(\delta_n) \, ,$$

as $n \to \infty$, where ω is the Landau symbol defined in the Introduction.

Our final comment concerns the case when q is only *measurable*. Then it can happen that the Trotter product formula for that pair $\{D_0, Q\}$ does *not* converge in the *operator-norm* topology:

Theorem 4.5 ([13, Theorem 3.5]) *There is a non-negative measurable function $q \in L^\infty([0,1])$, such that*

$$\liminf_{n \to \infty} \sup_{\tau \geq 0} \left\| e^{-\tau(D_0+Q)} - \left(e^{-\tau D_0/n} e^{-\tau Q/n} \right)^n \right\|_{\mathscr{B}(L^p([0,1]))} > 0 \, .$$

We note that Theorem 4.5 does not exclude the convergence of the Trotter product formula for the pair $\{D_0, Q\}$ in the *strong* operator topology.

4.3 Comments

Section 4.1

Our example can be considered as a kind of solvable model. It fits into the evolution cases considered in Sects. 2.3 and 3.2. Indeed, one has to set $\mathfrak{X} = \mathbb{C}$, $p = 2$, $L^p([0,T],\mathfrak{X}) = L^2([0,T])$, $A = I$ and $B(t) = q(t)$ and $\mathscr{B} = Q$. One easily checks that

$$(e^{-\tau/n\mathscr{K}_0}e^{-\tau\mathscr{B}/n})^n = e^{-\tau}(e^{-\tau/nD_0}e^{-\tau Q/n})^n$$

which shows that the convergence rate of $(e^{-\tau/nD_0}e^{-\tau Q/n})^n$ coincides with that one $(e^{-\tau/n\mathscr{K}_0}e^{-\tau\mathscr{B}/n})^n$. One easily checks that the choice $A = I$, $B(t) = q(t)$, $q(t) \geq 0$, guarantees the assumptions (A1)–(A4) If $q \in C^{0,\beta}([0,T])$, then the assumption (A5) is satisfied. If Q is Lipschitz continuous, i.e. $q \in C^{0,1}([0,T])$, then the assumption (A6) is valid.

Section 4.2

Theorem 4.2 shows that for $\beta \in (0,1)$ the convergence rate is $O(1/n^\beta)$, which is better than the convergence rates $O(1/n^{\beta-\alpha})$ and $O(1/n^{1-\alpha})$ in Theorems 3.4 and 3.5. Hence, they are not sharp.

Theorems 4.3 and 4.4 demonstrate that the convergence rate can be arbitrary slow if the smoothness of Q is weaker and weaker.

Finally, Theorem 4.5 shows that there is a bounded operator such that the Trotter product formula does not converge in the operator norm. This makes clear that Theorem 3.2 becomes false if the condition that A is a holomorphic generator is dropped. Indeed the operator D_0 which plays the role of A of Theorem 3.2 is not a generator of holomorphic semigroup.

References

1. V. Cachia, H. Neidhardt, V.A. Zagrebnov, Accretive perturbations and error estimates for the Trotter product formula. Integr. Equ. Oper. Theory **39**(4), 396–412 (2001)
2. V. Cachia, H. Neidhardt, V.A. Zagrebnov, Comments on the Trotter product formula error-bound estimates for nonself-adjoint semigroups. Integr. Equ. Oper. Theory **42**(4), 425–448 (2002)
3. V. Cachia, V.A. Zagrebnov, Operator-norm approximation of semigroups by quasi-sectorial contractions. J. Funct. Anal. **180**(1), 176–194 (2001)
4. V. Cachia, V.A. Zagrebnov, Operator-norm convergence of the Trotter product formula for sectorial generators. Lett. Math. Phys. **50**, 203–211 (1999)
5. V. Cachia, V.A. Zagrebnov, Operator-norm convergence of the Trotter product formula for holomorphic semigroups. J. Oper. Theory **46**(1), 199–213 (2001)

6. T. Ichinose, H. Neidhardt, V.A. Zagrebnov, Trotter-Kato product formula and fractional powers of self-adjoint generators. J. Funct. Anal. **207**(1), 33–57 (2004)
7. T. Ichinose, H. Tamura, Error estimate in operator norm of exponential product formulas for propagators of parabolic evolution equations. Osaka J. Math. **35**(4), 751–770 (1998)
8. T. Ichinose, H. Tamura, H. Tamura, V.A. Zagrebnov, Note on the paper: the norm convergence of the Trotter-Kato product formula with error bound by T. Ichinose and H. Tamura. Comm. Math. Phys. **221**(3), 499–510 (2001)
9. T. Kato, On the Trotter-Lie product formula. Proc. Jpn. Acad. **50**, 694–698 (1974)
10. T. Kato, Trotter's product formula for an arbitrary pair of self-adjoint contraction semigroups, in *Topics in Functional Analysis: Essays Dedicated to M.G. Kreĭn on the Occasion of his 70th Birthday*. Adv. in Math. Suppl. Stud., vol. 3 (Academic Press, New York, 1978), pp. 185–195
11. T. Kato, Perturbation theory for linear operators. *Classics in Mathematics* (Springer, Berlin, 1995), Reprint of the 1980 edition
12. H. Neidhardt, A. Stephan, V.A. Zagrebnov, Convergence rate estimates for Trotter product approximations of solution operators for non-autonomous Cauchy problems (December 2016). arXiv:1612.06147 [math.FA]
13. H. Neidhardt, A. Stephan, V.A. Zagrebnov, Remarks on the operator-norm convergence of the Trotter product formula. Integr. Equ. Oper. Theory **90**(15) (2018)
14. H. Neidhardt, A. Stephan, V.A. Zagrebnov, On convergence rate estimates for approximations of solution operators for linear non-autonomous evolution equations. Nanosyst. Phys. Chem. Math. **8**(2), 202–215 (2017)
15. H. Neidhardt, V.A. Zagrebnov, On error estimates for the Trotter-Kato product formula. Lett. Math. Phys. **44**(3), 169–186 (1998)
16. H. Neidhardt, V.A. Zagrebnov, Trotter-Kato product formula and operator-norm convergence. Comm. Math. Phys. **205**(1), 129–159 (1999)
17. H. Neidhardt, On abstract linear evolution equations I. Math. Nachr. **103**, 283–298 (1981)
18. H. Neidhardt, On abstract linear evolution equations II, Prepr. Akad. Wiss. DDR, Inst. Math. P-MATH-07/81, 56 p. (1981)
19. H. Neidhardt, On linear evolution equations III: hyperbolic case. Prepr., Akad. Wiss. DDR, Inst. Math. p-MATH-05/82, 74 p. (1982)
20. H. Neidhardt, V.A. Zagrebnov, Fractional powers of self-adjoint operators and Trotter-Kato product formula. Integr. Equ. Oper. Theory **35**(2), 209–231 (1999)
21. H. Neidhardt, V.A. Zagrebnov, Linear non-autonomous Cauchy problems and evolution semigroups. Adv. Differ. Equ. **14**(3–4), 289–340 (2009)
22. G. Nickel, On evolution semigroups and nonautonomous Cauchy problems. Diss. Summ. Math. **1**(1–2), 195–202 (1996)
23. Dzh.L. Rogava, Error bounds for Trotter-type formulas for self-adjoint operators. Funct. Anal. Appl. **27**, 217–219 (1993)
24. H. Tamura, A remark on operator-norm convergence of Trotter-Kato product formula. Integr. Equ. Oper. Theory **37**(3), 350–356 (2000)
25. H.F. Trotter, On the product of semi-groups of operators. Proc. Amer. Math. Soc. **10**, 545–551 (1959)

Multiple Periodic Solutions of Lagrangian Systems of Relativistic Oscillators



Biagio Ricceri

1 Introduction

As the reader can notice, the title of the present paper is, intentionally, almost identical to the one of [1].

Actually, it is our aim to show how to obtain the multiplicity of periodic solutions for the systems mentioned in the title making a joint use of the theory developed by Brezis and Mawhin in [1] with that we developed in [3].

To be more precise, we now fix some notations that we will keep throughout the paper and recall the main result of [1].

L, T are two fixed positive numbers. For each $r > 0$, we set $B_r = \{x \in \mathbf{R}^n : |x| < r\}$ ($|\cdot|$ being the Euclidean norm on $\mathbf{R}^n$) and $\overline{B_r}$ is the closure of B_r.

We denote by $\mathscr{A}$ the family of all homeomorphisms ϕ from B_L onto $\mathbf{R}^n$ such that $\phi(0) = 0$ and $\phi = \nabla\Phi$, where the function $\Phi : \overline{B_L} \to]-\infty, 0]$ is continuous and strictly convex in $\overline{B_L}$, and of class C^1 in B_L. Notice that 0 is the unique global minimum of Φ in $\overline{B_L}$.

We denote by $\mathscr{B}$ the family of all functions $F : [0, T] \times \mathbf{R}^n \to \mathbf{R}$ which are measurable in $[0, T]$, of class C^1 in $\mathbf{R}^n$ and such that $\nabla_x F$ is measurable in $[0, T]$ and, for each $r > 0$, one has $\sup_{x \in B_r} |\nabla_x F(\cdot, x)| \in L^1([0, T])$, with $F(\cdot, 0) \in L^1([0, T])$.

Given $\phi \in \mathscr{A}$ and $F \in \mathscr{B}$, we consider the problem

$$\begin{cases} (\phi(u'))' = \nabla_x F(t, u) & \text{in } [0, \mathrm{T}] \\ \\ u(0) = u(T)\ ,\ u'(0) = u'(T)\ . \end{cases} \qquad (P_{\phi,F})$$

B. Ricceri (✉)
Department of Mathematics, University of Catania, Catania, Italy
e-mail: ricceri@dmi.unict.it

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_10

A solution of this problem is any function $u : [0, T] \to \mathbf{R}^n$ of class C^1, with $u'([0, T]) \subset B_L$, $u(0) = u(T)$, $u'(0) = u'(T)$, such that the composite function $\phi \circ u'$ is absolutely continuous in $[0, T]$ and one has $(\phi \circ u')'(t) = \nabla_x F(t, u(t))$ for a.e. $t \in [0, T]$.

Now, we set

$$K = \{u \in \mathrm{Lip}([0, T], \mathbf{R}^n) : |u'(t)| \leq L \ \ for\ a.e.\ t \in [0, T], u(0) = u(T)\} ,$$

$\mathrm{Lip}([0, T], \mathbf{R}^n)$ being the space of all Lipschitzian functions from $[0, T]$ into $\mathbf{R}^n$.

Clearly, one has

$$\sup_{[0,T]} |u| \leq LT + \inf_{[0,T]} |u| \tag{1.1}$$

for all $u \in K$. To see this, take $t_0 \in [0, T]$ such that $|u(t_0)| = \inf_{[0,T]} |u|$ and observe that, for each $t \in [0, T]$, one has

$$|u(t) - u(t_0)| = \left| \int_{t_0}^{t} u'(\tau) d\tau \right| \leq LT .$$

Next, consider the functional $I : K \to \mathbf{R}$ defined by

$$I(u) = \int_0^T (\Phi(u'(t)) + F(t, u(t)))dt$$

for all $u \in K$.

The basic result of the theory developed in [1] is as follows:

Theorem 1.1 ([1], Theorem 5.2) *Any global minimum of I in K is a solution of problem ($P_{\phi,F}$).*

Well, the aim of the present paper is to provide a set of conditions under which the functional I has at least two global minima in K.

As far as we know, this is the first result of this kind, and so we cannot do any proper comparison with previous ones.

Notice that some multiplicity results for problem ($P_{\phi,F}$) are already available in the literature. In this connection, we refer to the numerous references contained in the very recent survey by Mawhin [2] and in [4]. But, as we repeat, in those papers the multiple solutions of problem ($P_{\phi,F}$) are not shown to be global minima of the functional I in K.

As we said at the beginning, our main tool is provided by the main result obtained in [3] which is recalled in the next section.

Finally, Sect. 3 contains the statement of our multiplicity result, its proof and various related remarks.

2 A Well-Posedness Theorem

In this section, we summarize the theory developed in [3].

So, let X be a Hausdorff topological space, J, Ψ two real-valued functions defined in X, and a, b two numbers in $[-\infty, +\infty]$, with $a < b$.

If $a \in \mathbf{R}$ (resp. $b \in \mathbf{R}$), we denote by M_a (resp. M_b) the set of all global minima of the function $J + a\Psi$ (resp. $J + b\Psi$), while if $a = -\infty$ (resp. $b = +\infty$), M_a (resp. M_b) stands for the empty set. We adopt the conventions $\inf \emptyset = +\infty$, $\sup \emptyset = -\infty$. We also set

$$\alpha = \max \left\{ \inf_X \Psi, \sup_{M_b} \Psi \right\} ,$$

$$\beta = \min \left\{ \sup_X \Psi, \inf_{M_a} \Psi \right\} .$$

One proves that $\alpha \leq \beta$.

As usual, given a function $f : X \to \mathbf{R}$ and a set $C \subseteq X$, we say that the problem of minimizing f over C is well-posed if the following two conditions hold:

- the restriction of f to C has a unique global minimum, say $\hat{x}$;
- every sequence $\{x_n\}$ in C such that $\lim_{n\to\infty} f(x_n) = \inf_C f$, converges to $\hat{x}$.

A set of the type $\{x \in X : f(x) \leq r\}$ is said to be a sub-level set of f. The main result of [3] is as follows:

Theorem 2.1 ([3, Theorem 1]) *Assume that $\alpha < \beta$ and that, for each $\lambda \in]a, b[$, the function $J + \lambda\Psi$ has sequentially compact sub-level sets and admits a unique global minimum in X.*

Then, for each $r \in]\alpha, \beta[$, the problem of minimizing J over $\Psi^{-1}(r)$ is well-posed.

Moreover, if we denote by $\hat{x}_r$ the unique global minimum of $J_{|\Psi^{-1}(r)}$ $(r \in]\alpha, \beta[)$, the functions $r \to \hat{x}_r$ and $r \to J(\hat{x}_r)$ are continuous in $]\alpha, \beta[$, and, for some $\hat{\lambda}_r \in]a, b[$, $\hat{x}_r$ is the global minimum in X of the function $J + \hat{\lambda}_r\Psi$

3 The Main Result

Here is our main result:

Theorem 3.1 *Let $\phi \in \mathscr{A}$, $F \in \mathscr{B}$, $G \in C^1(\mathbf{R}^n)$, $\psi \in L^1([0, T]) \setminus \{0\}$, with $\psi \geq 0$. Moreover, let $\gamma : [0, +\infty[\to \mathbf{R}$ be a convex strictly increasing function such that $\lim_{s\to+\infty} \frac{\gamma(s)}{s} = +\infty$. Assume that the following assumptions are satisfied:*

(i_1) *for a.e.* $t \in [0, T]$ *and for every* $x \in \mathbf{R}^n$, *one has*

$$\gamma(|x|) \leq F(t, x) ;$$

(i_2) $\liminf_{|x|\to+\infty} \frac{G(x)}{|x|} > -\infty$;

(i_3) *the function* G *has no global minima in* $\mathbf{R}^n$;

(i_4) *there exist two distinct points* $x_1, x_2 \in \mathbf{R}^n$ *such that*

$$\inf_{x\in\mathbf{R}^n} \int_0^T F(t, x)dt < \max\left\{\int_0^T F(t, x_1)dt, \int_0^T F(t, x_2)dt\right\}$$

and

$$G(x_1) = G(x_2) = \inf_{B_c} G$$

where

$$c = LT + \gamma^{-1}\left(\frac{1}{T}\max\left\{\int_0^T F(t, x_1)dt, \int_0^T F(t, x_2)dt\right\}\right) .$$

Then, there exist $\tilde{\lambda} > 0$ *such that the problem*

$$\begin{cases} (\phi(u'))' = \nabla_x(F(t, u) + \tilde{\lambda}\psi(t)G(u)) \ \ in\ [0, \mathrm{T}] \\ u(0) = u(T) \ , \ u'(0) = u'(T) \end{cases}$$

has at least two solutions which are global minima in K *of the functional*

$$u \to \int_0^T (\Phi(u'(t)) + F(t, u(t)) + \tilde{\lambda}\psi(t)G(u(t)))dt .$$

Proof Let $C^0([0, T], \mathbf{R}^n)$ be the space of all continuous functions from $[0, T]$ into $\mathbf{R}^n$, with the norm $\sup_{[0,T]} |u|$. We are going to apply Theorem 2.1 taking: $a = 0$, $b = +\infty$, $X = K$ regarded as a subset of $C^0([0, T], \mathbf{R}^n)$ with the relative topology and

$$J(u) = \int_0^T \psi(t)G(u(t))dt ,$$

$$\Psi(u) = \int_0^T (\Phi(u'(t)) + F(t, u(t)))dt$$

for all $u \in K$. Fix $\lambda > 0$. By (i_2), for a suitable constant $\delta > 0$, we have

$$-\delta(|x| + 1) \leq G(x)$$

for all $x \in \mathbf{R}^n$. For each $u \in K$ with $\sup_{[0,T]} |u| \geq LT$, in view of (i_1), (1.1) and of the convexity of γ, using Jensen inequality, we get

$$\int_0^T \psi(t)G(u(t))dt + \lambda \int_0^T (\Phi(u'(t)) + F(t, u(t)))dt \geq$$

$$-\delta \int_0^T \psi(t)|u(t)|dt + \lambda \int_0^T \gamma(|u(t)|)dt - \delta \int_0^T \psi(t)dt + \lambda\Phi(0)T \geq$$

$$-\delta \int_0^T \psi(t)dt \sup_{[0,T]} |u| + \lambda T\gamma\left(\frac{1}{T}\int_0^T |u(t)|dt\right) - \delta \int_0^T \psi(t)dt + \lambda\Phi(0)T \geq$$

$$-\delta \int_0^T \psi(t)dt \sup_{[0,T]} |u| + \lambda T\gamma\left(\inf_{[0,T]} |u|\right) - \delta \int_0^T \psi(t)dt + \lambda\Phi(0)T \geq$$

$$-\delta \int_0^T \psi(t)dt \sup_{[0,T]} |u| + \lambda T\gamma\left(\sup_{[0,T]} |u| - LT\right) - \delta \int_0^T \psi(t)dt + \lambda\Phi(0)T \,. \tag{3.1}$$

In turn, since $\lim_{s\to+\infty} \frac{\gamma(s-LT)}{s} = +\infty$, we infer from (3.1) that, for every $\rho \in \mathbf{R}$, there is $M > 0$ such that

$$\left\{u \in K : \int_0^T \psi(t)G(u(t))dt + \lambda \int_0^T (\Phi(u'(t)) + F(t, u(t)))dt \leq \rho\right\} \subseteq$$

$$\left\{u \in K : \sup_{[0,T]} |u| \leq M\right\} \,. \tag{3.2}$$

Now, observe that K is a closed subset of $C^0([0, T], \mathbf{R}^n)$. On the other hand, from Lemma 4.1 of [1], it follows that the functional $J + \lambda\Psi$ is lower semicontinuous in K. Summarizing: the functions belonging to any sub-level set of $J + \lambda\Psi$ are equi-continuous (since they are in K) and equi-bounded in view of (3.2). Hence, by the Ascoli-Arzelà theorem, any sub-level set of $J + \lambda\Psi$ is relatively sequentially compact in $C^0([0, T], \mathbf{R}^n)$. But, for the remarks above, the same set is closed in $C^0([0, T], \mathbf{R}^n)$, and so it is sequentially compact in K. Next, observe that the functional J ha no global minima in K. Since the constant functions lie in K, it is clear that

$$\inf_K J = \inf_{\mathbf{R}^n} G \int_0^T \psi(t)dt \,.$$

Hence, if G is unbounded below, so J is too. Now, suppose that G is bounded below. Arguing by contradiction, assume that $\hat{u} \in K$ is a global minimum of J. Then, we would have

$$\int_0^T \psi(t)\left(G(\hat{u}(t)) - \inf_{\mathbf{R}^n} G\right) dt = 0\,,$$

and so, since the integrand is non-negative, it would follow

$$\psi(t)\left(G(\hat{u}(t)) - \inf_{\mathbf{R}^n} G\right) = 0$$

for a.e. $t \in [0,T]$. Therefore, since $\psi \neq 0$, for some $t \in [0,T]$, we would have $G(\hat{u}(t)) = \inf_{\mathbf{R}^n} G$, against (i_3). Notice that the absence of global minima for J implies that

$$\beta = \sup_K \Psi\ .$$

Moreover, since $\lim_{s\to+\infty} \gamma(s) = +\infty$, from (i_1) it follows that

$$\sup_K \Psi = +\infty\ .$$

Furthermore, since $b = +\infty$, we have

$$\alpha = \inf_K \Psi\ .$$

Clearly

$$\inf_K \Psi \leq \inf_{x\in\mathbf{R}^n} \int_0^T F(t,x)dt + \Phi(0)T\ .$$

Now, put

$$r = \max\left\{\int_0^T F(t,x_1)dt, \int_0^T F(t,x_2)dt\right\} + \Phi(0)T\ .$$

By the above remarks and by the inequality in (i_4), we have

$$\alpha < r < \beta\ .$$

Fix $u \in \Psi^{-1}(]-\infty, r])$. By (i_1) and Jensen inequality again, we have

$$r \geq \int_0^T (\Phi(u'(t)) + F(t,u(t)))dt \geq \int_0^T \gamma(|u(t)|)dt + \Phi(0)T \geq$$

$$T\gamma\left(\frac{1}{T}\int_0^T |u(t)|dt\right) + \Phi(0)T\ ,$$

and so

$$\gamma\left(\frac{1}{T}\int_0^T |u(t)|dt\right) \leq \frac{1}{T}\max\left\{\int_0^T F(t,x_1)dt, \int_0^T F(t,x_2)dt\right\} .$$

Applying γ^{-1}, we get

$$\frac{1}{T}\int_0^T |u(t)|dt \leq \gamma^{-1}\left(\frac{1}{T}\max\left\{\int_0^T F(t,x_1)dt, \int_0^T F(t,x_2)dt\right\}\right)$$

and hence

$$\inf_{[0,T]} |u| \leq \gamma^{-1}\left(\frac{1}{T}\max\left\{\int_0^T F(t,x_1)dt, \int_0^T F(t,x_2)dt\right\}\right) .$$

In view of (1.1), we then infer that

$$\sup_{[0,T]} |u| \leq LT + \gamma^{-1}\left(\frac{1}{T}\max\left\{\int_0^T F(t,x_1)dt, \int_0^T F(t,x_2)dt\right\}\right) . \tag{3.3}$$

In turn, in view of (i_4), (3.3) implies that

$$J(x_1) = J(x_2) \leq J(u) .$$

Since $x_1, x_2 \in \Psi^{-1}(]-\infty, r])$, we then conclude that x_1, x_2 are two distinct global minima of $J_{|\Psi^{-1}(]-\infty,r])}$. Now, arguing by contradiction, assume that, for every $\lambda > 0$, the functional $J + \lambda\Psi$ has a unique global minimum in K. Then, by Theorem 2.1 (recall that $J + \lambda\Psi$ has sequentially compact sub-level sets), there would exist $\hat{\lambda}_r > 0$ and $\hat{u}_r \in \Psi^{-1}(r)$ such that $\hat{u}_r$ is the unique global minimum of $J + \hat{\lambda}_r\Psi$ in K. Then, for $i = 1, 2$, we would have

$$\inf_{u\in K}(J(u) + \hat{\lambda}_r\Psi(u)) \leq J(x_i) + \hat{\lambda}_r\Psi(x_i) \leq J(\hat{u}_r) + \hat{\lambda}_r\Psi(\hat{u}_r) =$$

$$\inf_{u\in K}(J(u) + \hat{\lambda}_r\Psi(u)) .$$

That is to say, x_1 and x_2 would be two distinct global minima in K of the functional $J + \hat{\lambda}_r\Psi$, a contradiction. So, there exists some $\hat{\lambda} > 0$ such that the functional $J + \hat{\lambda}\Psi$ has at least two global minima in K. To conclude the proof, take $\tilde{\lambda} = \frac{1}{\hat{\lambda}}$ and apply Theorem 1.1. □

In the sequel, the following further result from [1] will be useful:

Proposition 3.1 ([1, Proposition 3.2]) *Let $\phi \in \mathscr{A}$, $p > 1$ and $\mu > 0$. Then, for every $\omega \in L^1([0,T], \mathbf{R}^n)$, the problem*

$$\begin{cases} (\phi(u'))' = \mu|u|^{p-2}u + \omega(t) \quad in\ [0, T] \\ \\ u(0) = u(T)\ ,\ u'(0) = u'(T) \end{cases}$$

has a unique solution.

The next three examples (where $\phi \in \mathscr{A}$) show that, in Theorem 3.1, none of $(i_2) - (i_4)$ can be removed at all.

Example 3.1 Take: $F(x) = \frac{|x|^2}{2}$, $G(x) = \langle z, x\rangle$, with $z \in \mathbf{R}^n \setminus \{0\}$, $\psi = 1$, $\gamma(s) = \frac{s^2}{2}$. Clearly, $(i_1) - (i_3)$ are satisfied, but, for every $\lambda \in \mathbf{R}$, the problem

$$\begin{cases} (\phi(u'))' = u + \lambda z & \text{in } [0, T] \\ \\ u(0) = u(T)\ ,\ u'(0) = u'(T) \end{cases}$$

has a unique solution by Proposition 3.1.

Example 3.2 Take: $F(x) = \frac{|x|^2}{2}$, $G = 0$, $\gamma(s) = \frac{s^2}{2}$. Clearly, (i_1), (i_2) and (i_4) are satisfied, but, by Proposition 3.1, 0 is the unique solution of the problem

$$\begin{cases} (\phi(u'))' = u & \text{in } [0, T] \\ \\ u(0) = u(T)\ ,\ u'(0) = u'(T)\ . \end{cases}$$

Example 3.3 Take: $F(x) = |x|^2$,

$$G(x) = \begin{cases} 0 & \text{if } |x| \le LT + 1 \\ \\ -(|x| - LT - 1)^3 & \text{if } |x| > LT + 1\ , \end{cases}$$

$\psi = 1$, $\gamma(s) = s^2$. Clearly, (i_1), (i_3) and (i_4) are satisfied. In particular, (i_4) is satisfied by any pair of distinct points $x_1, x_2 \in \mathbf{R}^n$ such that $|x_1| = |x_2| \le 1$. However, for each $\lambda > 0$, the functional $u \to \int_0^T (\Phi(u'(t)) + |u(t)|^2 + \lambda G(u(t)))dt$ is unbounded below in K.

We conclude with a joint consequence of Theorem 3.1 and Proposition 3.1.

Theorem 3.2 *Let $\phi \in \mathscr{A}$, $p > 1$, $G \in C^1(\mathbf{R}^n)$, $\psi \in L^1([0, T])\setminus\{0\}$, with $\psi \ge 0$. Assume that G satisfies assumptions (i_2) , (i_3) of Theorem 3.1 and the following:*

(j_1) there is $\rho > LT$ such that G is constant in B_ρ .

Then, there exists $\tilde{\lambda} > 0$ such that the problem

$$\begin{cases} (\phi(u'))' = |u|^{p-2}u + \tilde{\lambda}\psi(t)\nabla G(u) \ \ in\ [0, T] \\ \\ u(0) = u(T)\ ,\ \ u'(0) = u'(T) \end{cases}$$

has at least one solution which is a global minimum in K of the functional

$$u \to \int_0^T \left(\Phi(u'(t)) + \frac{|u(t)|^p}{p} + \tilde{\lambda}\psi(t)G(u(t)) \right) dt$$

and whose range is contained in $\mathbf{R}^n \setminus \overline{B_{\rho - LT}}$.

Proof Apply Theorem 3.1 with $F(t,x) = \frac{|x|^p}{p}$ and $\gamma(s) = \frac{s^p}{p}$. Concerning (i_4), notice that it is satisfied by any pair of distinct points $x_1, x_2 \in \mathbf{R}^n$, such that $|x_1| = |x_2| \leq \rho - LT$. This comes from (j_1) after observing that $\gamma^{-1}\left(\frac{1}{T}\int_0^T F(t,x_1)dt\right) = |x_1|$. So, by Theorem 3.1, there exists $\tilde{\lambda} > 0$ such that the problem

$$\begin{cases} (\phi(u'))' = |u|^{p-2}u + \tilde{\lambda}\psi(t)\nabla G(u) \ \ \text{in } [0, T] \\ \\ u(0) = u(T)\ ,\ \ u'(0) = u'(T) \end{cases}$$

has at least one non-zero solution which is a global minimum in K of the functional

$$u \to \int_0^T \left(\Phi(u'(t)) + \frac{|u(t)|^p}{p} + \tilde{\lambda}\psi(t)G(u(t)) \right) dt\ .$$

Denote by w such a solution. To complete the proof, we have to show that $\inf_{[0,T]} |w| > \rho - LT$. Arguing by contradiction, assume that $\inf_{[0,T]} |w| \leq \rho - LT$. Then, by (1.1), we would have

$$\sup_{[0,T]} |w| \leq \rho\ .$$

By (j_1), this would imply that w is a solution of the problem

$$\begin{cases} (\phi(u'))' = |u|^{p-2}u & \text{in } [0, T] \\ \\ u(0) = u(T)\ ,\ \ u'(0) = u'(T) \end{cases}$$

and hence $w = 0$ by Proposition 3.1, which is a contradiction. The proof is complete. □

References

1. H. Brezis, J. Mawhin, Periodic solutions of Lagrangian systems of relativistic oscillators. Commun. Appl. Anal. **15**, 235–250 (2011)
2. J. Mawhin, Multiplicity of solutions of relativistic-type systems with periodic nonlinearities: a survey. Electron. J. Differ. Equ. Conf. **23**, 77–86 (2016)
3. B. Ricceri, Well-posedness of constrained minimization problems via saddle-points. J. Global Optim. **40**, 389–397 (2008)
4. X. Wang, Q. Liu, D. Qian, Existence and multiplicity results for some nonlinear problems with singular ϕ-Laplacian via a geometric approach. Bound. Value Probl. **2016**, 47 (2016)

Applications of a Theorem of Brezis-Nirenberg



Martin Schechter

1 Introduction

For problems arising in the sciences it is expected that there should be a unique solution. Otherwise, any mathematical formulation of the problem comes under question. If the formulation is linear, this expectation is reasonable. However, scientists realize that the world is not linear, and in nonlinear formulations, uniqueness cannot be expected. On the other hand, it is almost as difficult to show that a problem does not have a unique solution as it is to prove uniqueness. In this presentation we show how a theorem by Brezis and Nirenberg can be used to prove non-uniqueness for a class of problems involving partial differential equations.

In [1] Brezis and Nirenberg prove the following:

Theorem 1 *Let M, N be closed subspaces of a Hilbert space E such that $0 < \dim N < \infty$ and $M = N^{\perp}$. Let $G \in C^1(E, \mathbb{R})$ satisfy the PS condition and*

$$G(v) \leq 0, \quad v \in N \cap B_R,$$

$$G(w) \geq 0, \quad w \in M \cap B_R,$$

for some $R > 0$, where B_R is the closed ball in E with center 0 and radius R. Assume that

$$-\infty < \alpha = \inf_E G < 0.$$

Then G has at least three critical points.

M. Schechter (✉)
Department of Mathematics, University of California, Irvine, CA, USA
e-mail: mschecht@math.uci.edu

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_11

Two applications can be stated as follows: Let Ω be a smooth, bounded domain in $\mathbb{R}^n$, and let A be a self-adjoint operator on $L^2(\Omega)$. We assume that

$$C_0^\infty(\Omega) \subset D := D(|A|^{1/2}) \subset H^{T,2}(\Omega) \tag{1}$$

holds for some $T > 0$ (T need not be an integer), and the eigenvalues of A satisfy

$$0 < \lambda_0 < \lambda_1 < \cdots < \lambda_k < \cdots .$$

Let $f(x,t)$ be a Carathéodory function on $\Omega \times \mathbb{R}$. We assume that the function $f(x,t)$ satisfies

$$|f(x,t)| \le C(|t|+1), \quad x \in \Omega,\ t \in \mathbb{R}. \tag{2}$$

We have

Theorem 2 *Assume that for some integers $l < m$, the following inequalities hold:*

$$t[f(x,t_1) - f(x,t_0)] \le at^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{3}$$

where $a < \lambda_{m+1}$,

$$a_0 t^2 \le 2F(x,t) \le a_1 t^2, \quad |t| < \delta, \tag{4}$$

for some $\delta > 0$, with $\lambda_l < a_0 \le a_1 < \lambda_{l+1}$, where

$$F(x,t) := \int_0^t f(x,s)ds,$$

and

$$a_2 t^2 - W_1(x) \le 2F(x,t), \quad |t| > K, \tag{5}$$

for some $K \ge 0$, where $a_2 > \lambda_m$ and $W_1 \in L^1(\Omega)$. Then the equation

$$Au = f(x,u), \quad u \in D, \tag{6}$$

has at least two nontrivial solutions.

Theorem 3 *Equation (6) will have at least two nontrivial solutions if we assume that for some integers $l > m$, the following inequalities hold:*

$$t[f(x,t_1) - f(x,t_0)] \ge at^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{7}$$

where $a > \lambda_m$,

$$a_0 t^2 \le 2F(x,t) \le a_1 t^2, \quad |t| < \delta, \tag{8}$$

for some $\delta > 0$, *with* $\lambda_l < a_0 \le a_1 < \lambda_{l+1}$,

$$2F(x,t) \le a_2 t^2 + W_2(x), \quad |t| > K, \tag{9}$$

for some $K \ge 0$, *where* $a_2 < \lambda_{m+1}$ *and* $W_2 \in L^1(\Omega)$.

More comprehensive applications will be presented in the next section. Proofs will be given in Sect. 4.

2 Further Applications

We use the notation

$$a(u,v) = (Au,v), \quad a(u) = a(u,u), \quad u, v \in D.$$

We define

$$\|u\|_D := \|A^{1/2}u\| \tag{10}$$

and

$$G(u) := \|u\|_D^2 - 2\int_\Omega F(x,u)dx.$$

It is known that G is a continuously differentiable functional on the whole of D (cf., e.g., [17]) and

$$(G'(u),v)_D = 2(u,v)_D - 2(f(u),v),$$

where we write $f(u)$ in place of $f(x,u(x))$. In connection with the operator A, the following quantities are very useful. For each fixed positive integer ℓ, we let N_ℓ denote the subspace of D spanned by the eigenfunctions corresponding to $\lambda_0, \ldots, \lambda_\ell$, and let $M_\ell = N_\ell^\perp \cap D$. Then $D = M_\ell \oplus N_\ell$. For real a, b, we define

$$\begin{aligned}
I(u,a,b) &= (Au,u) - a\|u^-\|^2 - b\|u^+\|^2,\\
\gamma_\ell(a) &= \sup\{I(v,a,0) : v \in N_\ell, \|v^+\| = 1\},\\
\Gamma_\ell(a) &= \inf\{I(w,a,0) : w \in M_\ell, \|w^+\| = 1\},\\
F_{1\ell}(w,a,b) &= \sup\{I(v+w,a,b) : v \in N_\ell\},\\
F_{2\ell}(v,a,b) &= \inf\{I(v+w,a,b) : w \in M_\ell\},
\end{aligned}$$

$$\begin{aligned} M_\ell(a,b) &= \inf\{F_{1\ell}(w,a,b) : w \in M_\ell, \|w\|_D = 1\},\\ m_\ell(a,b) &= \sup\{F_{2\ell}(v,a,b) : v \in N_\ell, \|v\|_D = 1\},\\ \nu_\ell(a) &= \sup\{b : M_\ell(a,b) \geq 0\},\\ \mu_\ell(a) &= \inf\{b : m_\ell(a,b) \leq 0\}, \end{aligned}$$

where $u^{\pm}(x) = \max\{\pm u(x), 0\}$. Our first result is

Theorem 4 *Assume that for some integers $l < m$, the following inequalities hold:*

$$t[f(x,t_1) - f(x,t_0)] \leq a(t^-)^2 + b(t^+)^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{11}$$

where $b < \Gamma_m(a)$,

$$a_0(t^-)^2 + b_0(t^+)^2 \leq 2F(x,t) \leq a_1(t^-)^2 + b_1(t^+)^2, \quad |t| < \delta, \tag{12}$$

for some $\delta > 0$, with $a_0, b_0 < \lambda_{l+1}$, $a_1, b_1 > \lambda_l$, $b_0 > \mu_l(a_0)$, and $b_1 < \nu_l(a_1)$,

$$a_2(t^-)^2 + b_2(t^+)^2 - W_1(x) \leq 2F(x,t), \quad |t| > K, \tag{13}$$

for some $K \geq 0$, where $a_2, b_2 < \lambda_{m+1}$, $b_2 > \mu_m(a_2)$, and $W_1 \in L^1(\Omega)$. Then (6) has at least two nontrivial solutions.

In contrast to this, we have

Theorem 5 *Equation (6) will have at least two nontrivial solutions if we assume that for some integers $l > m$, the following inequalities hold:*

$$t[f(x,t_1) - f(x,t_0)] \geq a(t^-)^2 + b(t^+)^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{14}$$

where $b > \gamma_m(a)$,

$$a_0(t^-)^2 + b_0(t^+)^2 \leq 2F(x,t) \leq a_1(t^-)^2 + b_1(t^+)^2, \quad |t| < \delta \tag{15}$$

for some $\delta > 0$, with $a_0, b_0 < \lambda_{l+1}$, $b_0 > \mu_l(a_0)$ and $a_1, b_1 > \lambda_l$, $b_1 < \nu_l(a_1)$,

$$2F(x,t) \leq a_2(t^-)^2 + b_2(t^+)^2 + W_2(x), \quad |t| > K, \tag{16}$$

for some $K \geq 0$, where $a_2, b_2 > \lambda_m$, $b_2 < \nu_m(a_2)$, and $W_2 \in L^1(\Omega)$.

It was shown in [16] that the functions γ_l, μ_l, ν_{l-1}, Γ_{l-1} all emanate from the point (λ_l, λ_l) and satisfy

$$\Gamma_{l-1}(a) \leq \nu_{l-1}(a) \leq \mu_l(a) \leq \gamma_l(a)$$

on their common domains. It would therefore give a weaker result if we assumed in Theorems 4 and 5 that $b_0 > \gamma_l(a_0)$ and $b_1 < \Gamma_l(a_1)$. However, the functions γ_l, Γ_l are defined on the whole of $\mathbb{R}$, while the others are not. For cases in which the other functions are not defined, we state the following.

Theorem 6 *Theorems 4 and 5 remain true if we assume that (12) holds with* $b_0 > \gamma_l(a_0)$, *and* $b_1 < \Gamma_l(a_1)$ *for some* $a_0,\ a_1 \in \mathbb{R}$.

3 Some Lemmas

The proofs of the theorems of the previous section will be based on a series of lemmas.

Lemma 1 *If* $b < \Gamma_l(a)$, *then there is an* $\varepsilon > 0$ *such that*

$$I(w,a,b) \geq \varepsilon \|w\|_D^2, \quad w \in M_l. \tag{17}$$

Proof By the continuity of Γ_l, there is a $t < 1$ such that $b/t < \Gamma_l(a/t)$. Then

$$I(w,a/t,b/t) = \|w\|_D^2 - \frac{a}{t}\|w^-\|^2 - \frac{b}{t}\|w^+\|^2 \geq 0, \quad w \in M_l.$$

Therefore,

$$I(w,a,b) = t\left[\|w\|_D^2 - \frac{a}{t}\|w^-\|^2 - \frac{b}{t}\|w^+\|^2\right] + (1-t)\|w\|_D^2 \geq (1-t)\|w\|_D^2.$$

Lemma 2 *If* $b > \gamma_l(a)$, *then there is an* $\varepsilon > 0$ *such that*

$$I(v,a,b) \leq -\varepsilon \|v\|_D^2, \quad v \in N_l. \tag{18}$$

Proof By the continuity of γ_l, there is a $t > 1$ such that $b/t > \gamma_l(a/t)$. Hence,

$$I(v,a/t,b/t) = \|v\|_D^2 - \frac{a}{t}\|v^-\|^2 - \frac{b}{t}\|v^+\|^2 \leq 0, \quad v \in N_l,$$

and

$$I(v,a,b) = t\left[\|v\|_D^2 - \frac{a}{t}\|v^-\|^2 - \frac{b}{t}\|v^+\|^2\right] + (1-t)\|v\|_D^2 \leq (1-t)\|v\|_D^2.$$

Lemma 3 *If*

$$t[f(x,t_1) - f(x,t_0)] \leq a(t^-)^2 + b(t^+)^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{19}$$

then

$$(G'(v+w_1)-G'(v+w_0),\, w) \geq 2I(w,a,b), \quad v, w_j \in D,\ w = w_1-w_0. \tag{20}$$

Lemma 4 *Suppose there are closed subspaces X,Y of a Banach space E such that* $E = X \oplus Y$, *and*

$$(G'(x+y_1)-G'(x+y_2),\, y_1-y_2) \geq m(\|y_1-y_2\|), \quad x \in X,\ y_1,\ y_2 \in Y, \tag{21}$$

for some function $m(t)$ *satisfying*

(a)

$$m(t) > 0, \quad t > 0,$$

(b) there is a $t_0 > 0$ *such that*

$$\int_0^{t_0} \frac{m(t)}{t}\, dt < \infty,$$

(c)

$$\frac{m(t)}{t} \to 0 \quad implies \quad t \to 0,$$

and

(d)

$$\frac{1}{M}\int_0^M \frac{m(t)}{t}\, dt \to \infty \quad as \quad M \to \infty.$$

Then there exists a continuous function $\varphi : X \to Y$ *satisfying*

(i) $G(x+\varphi(x)) = \min_{y\in Y} G(x+y)$.

(ii) The function $\tilde{G}(x) = G(x+\varphi(x)) : X \to \mathbb{R}$ *is of class* C^1 *and satisfies*

$$(\tilde{G}'(x), h) = (G'(x+\varphi(x)), h), \quad h \in X. \tag{22}$$

Lemma 5 *If* $f(x,t)$ *satisfies (19) and* $b < \Gamma_m(a)$, *then there is a continuous map* φ *from* $N_m \to M_m$ *such that*

$$J(v) \equiv G(v+\varphi(v)) = \min_{w\in M_m} G(v+w) \in C^1(N_m, \mathbb{R}), \quad v \in N_m, \tag{23}$$

and

$$J'(v) = G'(v+\varphi(v)), \quad v \in N_m. \tag{24}$$

Proof In view of Lemmas 1 and 3, we have

$$(G'(v + w_1) - G'(v + w_0), w) \geq \varepsilon \|w\|_D^2, \quad w \in M_m.$$

We can now apply Lemma 4 to arrive at the conclusion.

Lemma 6 *If, in addition,*

$$a_0(t^-)^2 + b_0(t^+)^2 \leq 2F(x,t), \quad |t| < \delta, \tag{25}$$

for some $\delta > 0$, *with* $a_0, b_0 < \lambda_{l+1}$, $b_0 > \mu_l(a_0)$, $l \leq m$, *then there are* $\varepsilon > 0$, $r > 0$ *such that*

$$J(v) \leq -\varepsilon \|v\|_D^2, \quad v \in N_l \cap B_r, \tag{26}$$

where

$$B_r = \{u \in D : \|u\|_D \leq r\}.$$

Proof Let q be any number satisfying

$$2 < q \leq 2n/(n-2T), \ 2T < n \tag{27}$$

$$2 < q < \infty, \ n \leq 2T. \tag{28}$$

It follows that there is a continuous map $\tau : N_l \to M_l$ such that

$$\tau(s\, v) = s\, \tau(v), \quad s \geq 0, \tag{29}$$

$$I(v + \tau(v), a_0, b_0) = \inf_{w \in M_l} I(v + w, a_0, b_0), \quad v \in N_l, \tag{30}$$

$$\|\tau(v)\|_D \leq C \|v\|_D, \quad v \in N_l. \tag{31}$$

Then, for $u = v + \tau(v)$, we have, by (2),

$$J(v) \leq G(u) \leq I(u, a_0, b_0) + \int_{|u|>\delta} [a_0(u^-)^2 + b_0(u^+)^2 - 2F(x,u)]\, dx$$

$$\leq F_{2l}(v, a_0, b_0) + C \int_{|u|>\delta} |u|^q \, dx \leq m_l(a_0, b_0)\|v\|_D^2 + o(\|v\|_D^2) \leq -\varepsilon \|v\|_D^2$$

for r sufficiently small.

Lemma 7 *Assume that*

$$a(t^-)^2 + b(t^+)^2 - W_1(x) \leq 2F(x,t), \quad |t| > K, \tag{32}$$

for some $K \geq 0$, *where* $a, b < \lambda_{m+1},\ b \geq \mu_m(a),\ l \leq m$, *and* $W_1 \in L^1(\Omega)$. *Then there is a* $K_1 < \infty$ *such that*

$$J(v) \leq K_1. \tag{33}$$

If $b > \mu_m(a)$, *then*

$$J(v) \longrightarrow -\infty \ as\ \|v\|_D \longrightarrow \infty. \tag{34}$$

Proof For $u = v + w,\ v \in N_m,\ w \in M_m$, we have

$$G(u) \leq I(u, a, b) + C \int_{|u|<K} |u|^q\, dx + \int_\Omega W_1(x)\, dx \leq I(u, a, b) + K'.$$

Thus,

$$\begin{aligned} J(v) &= \inf_{w \in M_m} G(v + w) \\ &\leq \inf_{w \in M_m} I(v + w, a, b) + K' \\ &= F_{2m}(v, a, b) + K' \\ &\leq m(a, b)\|v\|_D^2 + K'. \end{aligned}$$

If $b \geq \mu_m(a)$, then $m(a, b) \leq 0$. This proves (33). If $b > \mu_m(a)$, then $m(a, b) < 0$. This proves (34).

Lemma 8 *If* $l < m$ *and* $\lambda_l < a, b < \lambda_{m+1}$, *then there are continuous functions* $\xi : N_m \cap M_l \to N_l,\ \eta : N_m \cap M_l \to M_m$ *homogeneous of degree 1 and such that*

$$I(\xi(y)+\eta(y)+y, a, b) = \sup_{v \in N_l} \inf_{w \in M_m} I(v+w+y, a, b) = \inf_{w \in M_m} \sup_{v \in N_l} I(v+w+y, a, b) \tag{35}$$

for $y \in N_m \cap M_l$.

Proof Let

$$L_y(v, w) = I(v + w + y, a, b).$$

Then L_y is a strictly convex, lower semicontinuous functional in $w \in M_m$, and strictly concave and continuous in $v \in N_l$. Consequently, for each $y_0 \in N_m \cap M_l$, there are unique elements $v_0 = \xi(y_0) \in N_l,\ w_0 = \eta(y_0) \in M_m$ such that (35) holds, i.e., that

$$L_{y_0}(v, w_0) \leq L_{y_0}(v_0, w_0) \leq L_{y_0}(v_0, w), \quad v \in N_l,\ w \in M_m.$$

The functions ξ, η are clearly homogeneous of degree 1. To prove continuity, let $y_j \to y_0$ in $N_l \cap M_m$, and let $v_j = \xi(y_j),\ w_j = \eta(y_j)$. We note that the functions $v_j,\ w_j$ are bounded in D. Otherwise, it is easy to show that

$$I(v + w_j + y_j, a, b) \longrightarrow \infty \text{ as } j \longrightarrow \infty$$

for any $v \in N_l$, and

$$I(v_j + w + y_j, a, b) \longrightarrow -\infty \text{ as } j \longrightarrow \infty$$

for any $w \in M_m$. This would contradict (35). Thus there are renamed subsequences such that $v_j \to v_1,\ w_j \rightharpoonup w_1$ in D. Since

$$\begin{aligned} I(v + w_j + y_j, a, b) &\le I(v_j + w_j + y_j, a, b) \\ &\le I(v_j + w + y_j, a, b), \quad v \in N_l,\ w \in M_m, \end{aligned}$$

we have in the limit

$$\begin{aligned} I(v + w_1 + y_0, a, b) &\le I(v_1 + w_1 + y_0, a, b) \\ &\le I(v_1 + w + y_0, a, b), \quad v \in N_l,\ w \in M_m, \end{aligned}$$

showing that $v_1 = v_0,\ w_1 = w_0$. Since this is true for any subsequence, the result follows.

Lemma 9 *If*

$$2F(x,t) \le a_1(t^-)^2 + b_1(t^+)^2, \quad |t| \le \delta, \tag{36}$$

for some $\delta > 0$, with $a_1, b_1 > \lambda_l,\ b_1 < \nu_l(a_1),\ l < m$, then there are $\varepsilon > 0,\ r > 0$ such that

$$J(y + \xi(y)) \ge \varepsilon \|y\|_D^2, \quad y \in N_m \cap M_l \cap B_r. \tag{37}$$

Proof By Lemma 8

$$\inf_{w \in M_m} I(\xi(y) + y + w, a_1, b_1) = \inf_{w \in M_m} \sup_{v \in N_l} I(v + y + w, a_1, b_1) \tag{38}$$

for $y \in N_m \cap M_l$. Then, for $y \in (N_m \cap M_l \cap B_r) \setminus \{0\}$,

$$\begin{aligned} J(\xi(y) + y) &= G(\xi(y) + y + \varphi(\xi(y) + y)) \\ &\ge I(\xi(y) + y + \varphi(\xi(y) + y), a_1, b_1) - o(\|y\|_D^2) \\ &\ge \inf_{w \in M_m} I(\xi(y) + y + w, a_1, b_1) - o(\|y\|_D^2) \end{aligned}$$

$$
\begin{aligned}
&= \inf_{w\in M_m} \sup_{v\in N_l} I(v+y+w, a_1, b_1) - o(\|y\|_D^2)\\
&\geq \inf_{w\in M_m} M_l(a,b)\|y+w\|_D^2 - o(\|y\|_D^2)\\
&= M_l(a,b)\|y\|_D^2 - o(\|y\|_D^2)\\
&\geq \varepsilon\|y\|_D^2.
\end{aligned}
$$

Lemma 10 *Assume*

$$
t[f(x,t_1) - f(x,t_0)] \geq a(t^-)^2 + b(t^+)^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0. \tag{39}
$$

Then

$$
(G'(v_1+w) - G'(v_0+w), v) \leq 2I(v,a,b), \quad v_j, w \in D,\ v = v_1 - v_0. \tag{40}
$$

Proof We have

$$
(f(x,v_1+w) - f(x,v_0+w), v) \geq a\|v^-\|^2 + b\|v^+\|^2.
$$

Hence,

$$
\begin{aligned}
&(G'(v_1+w) - G'(v_0+w), v)/2\\
&\qquad = \|v\|_D^2 - (f(x,v_1+w) - f(x,v_0+w), v)\\
&\qquad \leq I(v,a,b).
\end{aligned}
$$

Lemma 11 *If $f(x,t)$ satisfies (39), and $b > \gamma_m(a)$, then there is a continuous map ψ from $M_m \to N_m$ such that*

$$
J(w) \equiv G(w+\psi(w)) = \max_{v\in N_m} G(v+w) \in C^1(M_m, \mathbb{R}), \quad w \in M_m, \tag{41}
$$

and

$$
J'(w) = G'(w+\psi(w)), \quad w \in M_m. \tag{42}
$$

Proof In view of Lemmas 2 and 10, we have

$$
(G'(v_1+w) - G'(v_0+w), v) \leq -\varepsilon\|v\|_D^2, \quad v \in N_m.
$$

We can now apply Lemma 4 to obtain the conclusion.

Lemma 12 *If, in addition,*

$$a_0(t^-)^2 + b_0(t^+)^2 \le 2F(x,t), \quad |t| < \delta, \tag{43}$$

for some $\delta > 0$, *with* $a_0, b_0 < \lambda_{l+1}$, $b_0 > \mu_l(a_0)$, $l > m$, *then there are* $\varepsilon > 0$, $r > 0$ *such that*

$$J(y + \eta(y)) \le -\varepsilon \|y\|_D^2, \quad y \in N_l \cap M_m \cap B_r. \tag{44}$$

Proof For $y \in M_m \cap N_l$, let $u = y + \eta(y) \in M_m$. By (2),

$$\begin{aligned} J(u) = G(u + \psi(u)) &\le I(u + \psi(u), a_0, b_0) + o(\|u\|_D^2) \\ &\le \sup_{v \in N_m} I(u + v, a_0, b_0) + o(\|u\|_D^2) \\ &= I(y + \eta(y) + \xi(y), a_0, b_0) + o(\|u\|_D^2) \\ &= \sup_{v \in N_m} \inf_{w \in M_l} I(y + v + w, a_0, b_0) + o(\|u\|_D^2) \\ &= \sup_{v \in N_m} F_{2l}(y + v, a_0, b_0) + o(\|u\|_D^2) \\ &\le \sup_{v \in N_m} m_l(a_0, b_0)\|y + v\|_D^2 + o(\|u\|_D^2) \\ &\le -\varepsilon \|y\|_D^2 \end{aligned}$$

for r sufficiently small.

Lemma 13 *If*

$$2F(x,t) \le a_1(t^-)^2 + b_1(t^+)^2, \quad |t| \le \delta, \tag{45}$$

for some $\delta > 0$, *with* $a_1, b_1 > \lambda_l$, $b_1 < \nu_l(a_1)$, $l > m$, *then there are* $\varepsilon > 0$, $r > 0$ *such that*

$$J(w) \ge \varepsilon \|w\|_D^2, \quad w \in M_l \cap B_r. \tag{46}$$

Proof There is a continuous map $\theta : M_l \to N_l$ such that

$$\theta(s\,w) = s\,\theta(w), \quad s \ge 0, \tag{47}$$

$$I(\theta(w) + w, a_1, b_1) = \sup_{v \in N_l} I(v + w, a_1, b_1), \quad w \in M_l. \tag{48}$$

Thus,

$$\begin{aligned} J(w) &\geq G(w+\theta(w), a_1, b_1) \\ &\geq I(w+\theta(w), a_1, b_1) - o(\|w\|_D^2) \\ &= \sup_{v \in N_l} I(v+w, a_1, b_1) - o(\|w\|_D^2) \\ &= F_{1l}(w, a_1, b_1) - o(\|w\|_D^2) \\ &\geq M_l(a_1, b_1)\|w\|_D^2 - o(\|w\|_D^2) \\ &\geq \varepsilon \|w\|_D^2 \end{aligned}$$

for r sufficiently small.

Lemma 14 *Assume that*

$$2F(x,t) \leq a(t^-)^2 + b(t^+)^2 + W_1(x), \quad |t| > K, \tag{49}$$

for some $K \geq 0$, *where* $a, b > \lambda_m$, $b \leq \nu_m(a)$, $l \geq m$, *and* $W_1 \in L^1(\Omega)$. *Then there is a* $K_1 < \infty$ *such that*

$$J(w) \geq -K_1, \quad w \in M_m. \tag{50}$$

If $b < \nu_m(a)$, *then*

$$J(w) \longrightarrow \infty \text{ as } \|w\|_D \longrightarrow \infty. \tag{51}$$

Proof For $u = v + w$, $v \in N_m$, $w \in M_m$, we have

$$G(u) \geq I(u,a,b) - C\int_{|u|<K} |u|^q \, dx - \int_\Omega W_1(x)dx \geq I(u,a,b) - K'.$$

Thus,

$$\begin{aligned} J(w) &= \sup_{v \in N_m} G(v+w) \\ &\geq \sup_{v \in N_m} I(v+w,a,b) - K' \\ &= F_{1m}(w,a,b) - K' \\ &\geq M_m(a,b)\|w\|_D^2 - K'. \end{aligned}$$

If $b \leq \nu_m(a)$, then $M_m(a,b) \geq 0$. This proves (50). If $b < \nu_m(a)$, then $M_m(a,b) > 0$. This proves (51).

Lemma 15 *If*

$$a_0(t^-)^2 + b_0(t^+)^2 \le 2F(x,t), \quad |t| < \delta, \tag{52}$$

for some $\delta > 0$, *with* $b_0 > \gamma_l(a_0)$, $l \le m$, *then there are* $\varepsilon > 0$, $r > 0$ *such that*

$$J(v) \le -\varepsilon \|v\|_D^2, \quad v \in N_l \cap B_r, \tag{53}$$

where

$$B_r = \{u \in D : \|u\|_D \le r\}.$$

Proof Let q be any number satisfying

$$\begin{aligned} 2 < q &\le 2n/(n-2T),\ 2T < n \\ 2 < q &< \infty,\ n \le 2T. \end{aligned}$$

By (2),

$$\begin{aligned} J(v) \le G(v) &\le I(v, a_0, b_0) + \int_{|v|>\delta} [a_0(v^-)^2 + b_0(v^+)^2 - 2F(x,v)]\, dx \\ &\le -\varepsilon \|v\|_D^2 + C \int_{|v|>\delta} |v|^q\, dx \\ &\le -\varepsilon \|v\|_D^2 + o(\|v\|_D^2) \\ &\le -\varepsilon \|v\|_D^2 \end{aligned}$$

for r sufficiently small.

Lemma 16 *If*

$$2F(x,t) \le a_1(t^-)^2 + b_1(t^+)^2, \quad |t| \le \delta, \tag{54}$$

for some $\delta > 0$, *with* $b_1 < \Gamma_l(a_1)$, $l < m$, *then there are* $\varepsilon > 0$, $r > 0$ *such that*

$$J(v) \ge \varepsilon \|v\|_D^2, \quad v \in N_m \cap M_l \cap B_r. \tag{55}$$

Proof Let $u = v + \varphi(v) \in M_l$. Then

$$\begin{aligned} J(v) = G(u) &\ge I(u, a_1, b_1) + \int_{|u|>\delta} [a_0(u^-)^2 + b_0(u^+)^2 - 2F(x,u)]\, dx \\ &\ge \varepsilon \|u\|_D^2 - C \int_{|u|>\delta} |u|^q\, dx \end{aligned}$$

$$\begin{aligned}
&\geq \varepsilon\|u\|_D^2 - o(\|u\|_D^2)\\
&\geq \varepsilon\|v\|_D^2 - o(\|v\|_D^2)\\
&\geq \varepsilon\|v\|_D^2
\end{aligned}$$

for r sufficiently small, since

$$\|v\|_D \leq \|u\|_D \leq C\|v\|_D.$$

Lemma 17 *If*

$$a_0(t^-)^2 + b_0(t^+)^2 \leq 2F(x,t), \quad |t| < \delta, \tag{56}$$

for some $\delta > 0$, *with* $b_0 > \gamma_l(a_0),\ l \geq m$, *then there are* $\varepsilon > 0,\ r > 0$ *such that*

$$J(w) \leq -\varepsilon\|w\|_D^2, \quad w \in N_l \cap M_m \cap B_r. \tag{57}$$

Proof For $w \in M_m \cap N_l$, let $u = w + \psi(w) \in N_l$. By (2),

$$\begin{aligned}
J(w) &= G(w + \psi(w)) = G(u)\\
&\leq I(u, a_0, b_0) + \int_{|u|>\delta} [a_0(v^-)^2 + b_0(u^+)^2 - 2F(x,u)]\, dx\\
&\leq -\varepsilon\|u\|_D^2 + C\int_{|u|>\delta} |u|^q\, dx\\
&\leq -\varepsilon\|u\|_D^2 + o(\|u\|_D^2)\\
&\leq -\varepsilon\|u\|_D^2
\end{aligned}$$

for r sufficiently small. Since

$$\|w\|_D \leq \|u\|_D \leq C\|w\|_D,$$

the result follows.

Lemma 18 *If*

$$2F(x,t) \leq a_1(t^-)^2 + b_1(t^+)^2, \quad |t| \leq \delta \tag{58}$$

for some $\delta > 0$, *with* $b_1 < \Gamma_l(a_1),\ l > m$, *then there are* $\varepsilon > 0,\ r > 0$ *such that*

$$J(w) \geq \varepsilon\|w\|_D^2, \quad w \in M_l \cap B_r. \tag{59}$$

Proof We have

$$\begin{aligned} G(w) &\geq I(w, a_1, b_1) + \int_{|w|>\delta} [a_0(u^-)^2 + b_0(w^+)^2 - 2F(x, w)]\, dx \\ &\geq \varepsilon \|w\|_D^2 - C \int_{|w|>\delta} |w|^q \, dx \\ &\geq \varepsilon \|w\|_D^2 - o(\|w\|_D^2) \\ &\geq \varepsilon \|w\|_D^2 - o(\|w\|_D^2) \\ &\geq \varepsilon \|w\|_D^2 \end{aligned}$$

for r sufficiently small. Since

$$J(w) = \sup_{v \in N_l} G(v + w) \geq G(w),$$

the result follows.

4 The Proofs

Proof of Theorem 4. By Lemma 5, it suffices to show that $J(v)$ has two nontrivial solutions. Now J is bounded from above by Lemma 7 and satisfies (PS) by (34). Moreover,

$$J(v) < 0, \quad v \in N_l \cap B_r \backslash \{0\}, \tag{60}$$

by Lemma 6, and

$$J(\xi(y) + y) > 0, \quad y \in N_m \cap M_l \cap B_r \setminus \{0\}, \tag{61}$$

by Lemma 9. Thus, J has a positive maximum on N_m. We can now apply Theorem 1 and Lemma 4 to obtain the desired conclusion.

Proof of Theorem 5. By Lemma 11, it suffices to show that $J(w)$ given by (41) has two nontrivial solutions. Now J is bounded from below by Lemma 14 and satisfies (PS) by (51). Moreover,

$$J(w + \eta(w)) < 0, \quad w \in N_l \cap M_m \cap B_r \backslash \{0\}, \tag{62}$$

by Lemma 12, and

$$J(w) > 0, \quad w \in M_l \cap B_r \setminus \{0\} \tag{63}$$

by Lemma 13. Thus, J has a negative minimum on M_m. We can now apply Theorem 1 and Lemma 4 to obtain the desired conclusion.

Proof of Theorem 6. With reference to Theorem 4, we note that, by Lemma 5, it suffices to show that $J(v)$ has two nontrivial solutions. Now J is bounded from above by Lemma 7 and satisfies (PS) by (34). Moreover,

$$J(v) < 0, \quad v \in N_l \cap B_r \backslash \{0\} \tag{64}$$

by Lemma 15, and

$$J(v) > 0, \quad v \in N_m \cap M_l \cap B_r \setminus \{0\} \tag{65}$$

by Lemma 16. Thus J has a positive maximum on N_m. We can now apply Theorem 1 and Lemma 4 to obtain the desired conclusion. With respect to Theorem 5, we note that by Lemma 11, it suffices to show that $J(w)$ given by (41) has two nontrivial solutions. Now J is bounded from below by Lemma 14 and satisfies (PS) by (51). Moreover,

$$J(w) < 0, \quad w \in N_l \cap M_m \cap B_r \backslash \{0\}, \tag{66}$$

by Lemma 17, and

$$J(w) > 0, \quad w \in M_l \cap B_r \setminus \{0\}, \tag{67}$$

by Lemma 18. Thus, J has a negative minimum on M_m. We can now apply Theorem 1 and Lemma 4 to obtain the desired conclusion.

5 Notes and Remarks

In his studies of semilinear elliptic problems with jumping nonlinearities, Các [2–7] proved the following.

Theorem 7 *Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$, with smooth boundary $\partial\Omega$. Let $0 < \lambda_0 < \lambda_1 < \cdots < \lambda_k < \cdots$ be the sequence of distinct eigenvalues of the eigenvalue problem*

$$-\Delta u = \lambda u \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{68}$$

Let $p(t)$ be a continuous function such that $p(0) = 0$ and

$$p(t)/t \longrightarrow a \quad \text{as} \quad t \longrightarrow -\infty$$

and

$$p(t)/t \longrightarrow b \quad \text{as} \quad t \longrightarrow +\infty.$$

Assume that for some $k \geq 1$, *we have* $a \in (\lambda_{k-1}, \lambda_k)$, $b \in (\lambda_k, \lambda_{k+1})$, *and the only solution of*

$$-\Delta u = bu^+ - au^- \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega \tag{69}$$

is $u \equiv 0$, *where* $u^\pm = \max[\pm u, 0]$. *Assume further that*

$$\frac{p(s) - p(t)}{s - t} \leq \nu < \lambda_{k+1}, \quad s, t \in \mathbb{R},\ s \neq t. \tag{70}$$

Assume also that $p'(0)$ *exists and satisfies* $p'(0) \in (\lambda_{j-1}, \lambda_j)$ *for some* $j \leq k$. *Then*

$$-\Delta u = p(u) \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega \tag{71}$$

has at least two nontrivial solutions.

This theorem generalizes the work of Gallouët and Kavian [9, 10] which required λ_k to be a simple eigenvalue and the left-hand side of (70) to be sandwiched in between λ_{k-1} and λ_{k+1} and bounded away from both of them. Các proves a counterpart of the theorem in which the inequalities are reversed.

In the present work we generalized this theorem and its reverse-inequality counterpart by not requiring $p(t)/t$ to converge to limits at either $\pm\infty$ or ± 0. Rather, we worked with the primitive

$$F(x, t) := \int_0^t f(x, s)ds$$

and bounded $2F(x, t)/t^2$ near $\pm\infty$ and ± 0 [we replaced $p(t)$ with a function $f(x, t)$ depending on x as well]. Our main assumptions were

$$t[f(x, t_1) - f(x, t_0)] \leq a(t^-)^2 + b(t^+)^2, \quad t_j \in \mathbb{R},\ t = t_1 - t_0, \tag{72}$$

$$a_0(t^-)^2 + b_0(t^+)^2 \leq 2F(x, t) \leq a_1(t^-)^2 + b_1(t^+)^2, \quad |t| < \delta, \tag{73}$$

for some $\delta > 0$,

$$a_2(t^-)^2 + b_2(t^+)^2 - W_1(x) \leq 2F(x, t), \quad |t| > K, \tag{74}$$

for some $K > 0$ and $W_1 \in L^1(\Omega)$, where the constants $a, a_0, a_1, a_2, b, b_0, b_1, b_2$ are suitably chosen (they include the cases considered by Các). The advantage of such inequalities is that they do not restrict the expression $2F(x, t)/t^2$ or $f(x, t)/t$ to any particular interval.

The results of this presentation come from [13] with changes in the proofs. Certain cases were considered in [11]. Local linking results similar to those of [1] can be found in [12]. Lemma 4 is due to Castro [8]. The quantities used in the theorems were introduced in [14] and [15].

References

1. H. Brezis, L. Nirenberg, Remarks on finding critical points. Comm. Pure Appl. Math. **44**, 939–964 (1991)
2. N.P. Các, On an elliptic boundary value problem at double resonance. J. Math. Anal. Appl. **132**, 473–483 (1988)
3. N.P. Các, On nontrivial solutions of an asymptotically linear Dirichlet problem. J. Differ. Equ. **75**(1), 103–117 (1988)
4. N.P. Các, On nontrivial solutions of a Dirichlet problem whose jumping nonlinearity crosses a multiple eigenvalue. J. Differ. Equ. **80**, 379–404 (1989)
5. N.P. Các, On the number of solutions of an elliptic boundary value problem with jumping nonlinearity. Nonlinear Anal. TMA **13**(3), 341–351 (1989)
6. N.P. Các, On a boundary value problem with nonsmooth jumping nonlinearity. J. Differ. Equ. **93**(2), 238–259 (1991)
7. N.P. Các, On a Dirichlet problem with jumping nonlinearity, in *Differential Equations*. Lecture Notes in Pure and Applied Mathematics, vol. 127 (Marcell Dekker, New York, 1991), pp. 63–67
8. A. Castro, Reduction methods via minimax, in *Differential Equations*. Lecture Notes in Mathematics, vol. 957 (Springer, Berlin, 1982), pp. 1–20
9. T. Gallouet, O. Kavian, Resultats d'existence et de non existence pour certains problemes demi lineaires a l'infini. Ann. Fac. Sci. Toulouse Math. **3**, 2011–246 (1981)
10. T. Gallouët, O. Kavian, Resonance for jumping non-linearities. Comm. PDE **7**, 325–342 (1982)
11. S. Li, M. Willem, Multiple solutions for asymptotically linear boundary value problems in which the nonlinearity crosses at least one eigenvalue. Nonlinear Differ. Equ. Appl. **5**, 479Ű490 (1998)
12. K. Perera, Homological local linking. Abstr. Appl. Anal. **3**, 181–189 (1998)
13. K. Perera, M. Schechter, Multiple nontrivial solutions of elliptic semilinear equations. Topol. Methods Nonlinear Anal. **16**(1), 1–15 (2000)
14. M. Schechter, The Fučík spectrum. Indiana Univ. Math. J. **43**, 1139–1157 (1994)
15. M. Schechter, Asymptotically linear elliptic boundary value problems. Monatsh. Math. **117**, 121–137 (1994)
16. M. Schechter, The intrinsic mountain pass. Pac. J. Math. **171**, 529–544 (1995)
17. M. Schechter, *Linking Methods in Critical Point Theory* (Birkhäuser, Boston, 1999)

Porous Medium Equation with Nonlocal Pressure



Diana Stan, Félix del Teso, and Juan Luis Vázquez

Dedicated to Profs. Haim Brezis and Louis Nirenberg with deep admiration

1 Introduction

The study of evolution equations associated to dissipative operators was intensely pursued in the second half of the twentieth century. It took the form of the abstract equation

$$\frac{du}{dt} = A(u) + f\,,$$

where u is a time function with values in a Hilbert or Banach space X and A is typically an unbounded linear or nonlinear operator with strict dissipative conditions, the simplest case being the Laplacian operator acting on $L^2(\Omega)$. Famous theorems were proved that widely extended what was known for the heat equation and then for parabolic equations. The first results dealt with the case where the underlying functional space X was a Hilbert space, [10, 11, 66], and then the theory applied to Banach spaces, [23, 31, 49]. The aim of the general theory was to construct the corresponding semigroups (for $f = 0$) or flows (for general f) with detailed properties [35, 43, 46, 51].

D. Stan
Basque Center for Applied Mathematics, Bilbao, Spain
e-mail: dstan@bcamath.org

F. del Teso
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: felix.delteso@ntnu.no

J. L. Vázquez (✉)
Department of Mathematics, Universidad Autónoma de Madrid, Madrid, Spain
e-mail: juanluis.vazquez@uam.es

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_12

(I) It was soon realized that a general theory was bound to be too rich in details and difficulties, and this led to concentrating the attention on particular equations with a relevant physical interest and significant novel properties, [9, 44]. One of these equations was the porous medium equation, PME for short,

$$\partial_t u = \Delta(u^m) = \nabla(mu^{m-1}\nabla u), \quad m > 1,$$

which had received much attention from the Russian school in the 1950s and 1960s, [1, 50], and was taken up in the abstract general setting in the 70–80s, [2, 3]. This equation is a relevant model for heat propagation with finite propagation speed. As a consequence of such property, interesting geometry occurs in the form free boundaries (interfaces), an issue that originated a great amount of mathematical analysis [12, 13, 19, 41, 42]. Work done for several decades led to considerable success and a rather complete theory was formulated, cf. [59]. The PME generates a nonlinear contraction semigroup in $X = L^1(\Omega)$, $\Omega \subset \mathbb{R}^N$ with definite regularity properties.

This led to extensions like the fast diffusion equation where $m < 1$, the p-Laplacian equation, $\partial_t u = \Delta_p(u) = \nabla(|\nabla u|^{p-2}\nabla u)$, $1 < p < \infty$, [29, 58], and several other equations in the areas of Nonlinear Diffusion and Reaction-Diffusion. Much attention was also given to the Stefan Problem, very important for its mathematics and its applications, [14, 15, 29, 40, 47]. For a general survey paper see [62].

(II) A decade ago there arose the interest in combining the PME nonlinear mechanism with nonlocal operators so as to take into account anomalous diffusion effects, the main examples of such operators being the so-called fractional Laplacian operators. This trend has been a very active area of research since then. The main problem to be treated was the Porous Medium Equation with fractional pressure formulated by Caffarelli and Vázquez as

$$\partial_t u = \nabla(u\nabla(-\Delta)^{-s}u), \tag{1}$$

for $0 < s < 1$, having in mind models in statistical mechanics, see [32, 33] that deal with the macroscopic evolution of interacting gas systems. In [16] existence of weak solutions is proved for initial data $u_0 \in L^1 \cap L^\infty$ via an approximation method that requires a suitable decay of the data. On the other hand, the 1D model was investigated by Biler et al. in [4] having an application to dislocation theory, [36].

Uniqueness is a key issue for this model. It holds for suitable solutions in 1D as proved in [4]. However, uniqueness of weak or better solutions is not known so far in several dimensions, except locally in time for good data (see [67] for Besov spaces). The lack of comparison principle is also an important issue to deal with.

A quite important, and rather surprising, feature of the equation is the property of finite speed of propagation. This property was not evident since fractional operators are known to imply very fast propagation. and could overrun the PME nonlinearity,

but the nonlinearity wins in this case. The property was proved [16] by comparison with special type of barrier functions, called by the authors true supersolutions. Hölder regularity of solutions is proved in [18, 20]. The large time asymptotic behavior of weak solutions of (1) is given by a unique fundamental solution constructed via an obstacle type problem (the proof is given in [17]). A suitable entropy function is constructed to prove the uniqueness of the self-similar solution and the asymptotic behavior. Refined asymptotics was done in [21]. Gradient flow methods are also an alternative to prove existence of solutions, see [45]. However, because of the lack of a uniqueness theory, the constructed weak solutions of [16] and gradient flow solutions ([45]) might be different.

(III) Extending model (1) to general exponents, as in the standard PME, is natural and motivates the model we study here:

$$\partial_t u = \nabla \cdot (u^{m-1} \nabla (-\Delta)^{-s} u) \qquad x \in \mathbb{R}^N,\ t > 0. \tag{M1}$$

The extension also agrees with the generality of the models proposed in [32, 34]. We are specially interested in better understanding well-posedness and velocity of propagation. It turns out that this equation has quite interesting properties, some of them are inherited from (1) (finite speed of propagation for $m \geq 2$) and therefore some techniques of the proofs can be successfully adapted, but many other different properties were discovered, like infinite speed of propagation for $m \in (1, 2)$. In what follows we describe this last model with its main properties that have been obtained so far: existence of solutions in the general setting of finite measure data (therefore this extends the result of [16] for (1)), the transition finite-to-infinite speed of propagation, uniqueness in dimension one, and so on. Moreover we prove new results like existence of selfsimilar solutions for $m \geq 2$ and $N = 1$, the asymptotic behavior for $m > 1$ and $N = 1$ and partial results on these topics in higher dimensions.

(IV) Equations with two nonlinearities of the form

$$\partial_t u = \nabla \cdot (u^{a-1} \nabla (-\Delta)^{-s} u^{b-1}) \tag{GM}$$

have also been considered in the literature. We refer to [55] for construction of self-similar solutions when $a > 1$, $b > 1$ and transformation formulas between self-similar solutions of (GM), (M1) and (FPME).

Dolbeault and Zhang [30] proved that for $a - 1 = \frac{1}{2}$, $b < \frac{3}{2}$, self-similar solutions are not optimal for the Gagliardo-Nirenberg-Sobolev inequalities, in strong contrast with usual standard fast diffusion equations based on non-fractional operators (see [27]). Their approach is mainly based on entropy methods.

Important related work is due to Biler, Imbert and Karch. In [5, 6] they considered the model

$$\partial_t u = \nabla \cdot (|u| \nabla (-\Delta)^{-s} |u|^{b-2} u), \tag{2}$$

corresponding to $a = 2$, $b > 1$ in (GM). They prove existence of a changing-sign weak solution and its main properties. Moreover, they find explicit self-similar solutions with compact support. In a later work [39] finite speed of propagation is established for general positive solutions.

2 Presentation of the Model

We consider the initial value problem

$$\begin{cases} \partial_t u = \nabla \cdot (u^{m-1} \nabla (-\Delta)^{-s} u) & \text{for } x \in \mathbb{R}^N,\ t > 0, \\ u(0,x) = u_0(x) & \text{for } x \in \mathbb{R}^N, \end{cases} \tag{M1}$$

for $u = u(x,t) \geq 0$, exponents $m > 1$, $0 < s < 1$, and space dimension $N \geq 1$.

This model has been studied in the series of papers [54–57] by the present authors. Many properties of the solutions were proved and there is of course work to be done. In this paper we report on the results obtained so far and also we make a step further in the theory of this model by describing the asymptotic behavior in dimension $N = 1$.

First, we introduce the notion of weak solution in the very general context of measures as initial data. It applies for all $m \in (1, +\infty)$ and $s \in (0,1)$. We denote by $\mathscr{M}^+(\mathbb{R}^N)$ the set of nonnegative Radon measures.

Definition 1 We say that $u \geq 0$ is a weak solution of Problem (M1) with initial data $\mu \in \mathscr{M}^+(\mathbb{R}^N)$ if: $u \in L^1_{\rm loc}(\mathbb{R}^N \times (0,T))$, $\nabla(-\Delta)^{-s} u \in L^1_{\rm loc}(\mathbb{R}^N \times (0,T))$, $u^{m-1}\nabla(-\Delta)^{-s} u \in L^1_{\rm loc}(\mathbb{R}^N \times (0,T))$, and

$$\int_0^T \int_{\mathbb{R}^N} u\phi_t \, dxdt - \int_0^T \int_{\mathbb{R}^N} u^{m-1} \nabla(-\Delta)^{-s} u \cdot \nabla\phi \, dxdt + \int_{\mathbb{R}^N} \phi(x,0) d\mu(x) = 0,$$

for all test functions $\phi \in C_c^1(\mathbb{R}^N \times [0,T))$.

Remark 1 Note that, if $\mu = u_0 \in L^1_{\rm loc}(\mathbb{R}^N)$, then $d\mu(x) = u_0(x)dx$ and

$$\int_{\mathbb{R}^N} \phi(x,0) d\mu(x) = \int_{\mathbb{R}^N} u_0(x)\phi(x,0)dx,$$

thus, the initial datum is taken in the usual sense (as initial trace). We will prove the existence of a constructed weak solution by an approximation method; we will call it constructed weak solution.

In what follows we will present the main results on Problem (M1). In Sect. 3 three different diffusion models are introduced and we show their relation to Problem (M1) and the consequences of this transformation on the qualitative properties of solutions to Problem (M1). Section 4 is devoted to the integrated version of (M1). In Sect. 5 we state general results on existence of solutions, velocity

of propagation. Uniqueness and asymptotic behavior are also stated, but only in dimension 1. Finally the proofs are given in Sect. 6.

3 Related Models and Transformations

The study of some properties of Problem (M1) is made more difficult because the comparison principle does not hold and no proof of uniqueness of weak solutions in dimension higher than one is known. This motivates us to search for a connection with other fractional diffusion models that allow a more friendly approach. We present here some models of fractional diffusion connected to Problem (M1) via useful transformations at the self-similar level.

3.1 The Fractional Porous Medium Equation

An alternative fractional version of the standard or local Porous Medium Equation, $w_t = \Delta w^q$, is given by the following equation

$$w_t + (-\Delta)^\sigma w^q = 0, \tag{FPME}$$

called the *Fractional Porous Equation* in the literature (we call the exponent q instead of the usual m for convenience in later comparisons). For $q = 1$ and $0 < \sigma < 1$ this leads to the linear Fractional Heat Equation, for which we refer to the survey [8] and also [63]. Note that (FPME) corresponds to the general model (GM) when $a = 1$, $b = q + 1$ and $s = 1 - \sigma$.

In recent years the theory for this model has been widely developed: the existence, uniqueness and continuous dependence of solutions of the Cauchy problem (FPME) for all $q > 0$ and $0 < \sigma < 1$ have been proved by De Pablo, Quirós, Rodríguez and Vázquez in [24, 25].

The (FPME) model inherits some of the properties of the classical PME. Using the Caffarelli-Silvestre extension method and the Bénilan-Brezis-Crandall functional semigroup approach, a weak energy solution is constructed, and $u \in C([0,\infty) : L^1(\mathbb{R}^N))$. Moreover, the set of solutions forms a semigroup of ordered contractions in $L^1(\mathbb{R}^N)$.

An important property of (FPME), which does not hold in its non-fractional version, is the *infinite speed of propagation*: assume $\sigma \in (0,1)$, $q > (q)_c = (N - 2\sigma)_+/N$. Then for non-negative initial data $u_0 \geq 0$ such that $\int_{\mathbb{R}^N} u_0(x)\,dx < \infty$, there exists a unique solution $w(x,t)$ of problem (FPME) satisfying $w(x,t) > 0$ for all $x \in \mathbb{R}^N$, $t > 0$. Moreover, there is conservation of mass $\int w(x,t)\,dx = \int w_0(x)\,dx$ for all $t > 0$. The solutions are C^α continuous as proved in [65]. More general diffusions were considered in [26]. We refer to the survey [61] for a complete description of the (FPME).

Asymptotic Behavior and Self-similarity for FPME The large time behaviour of such solutions is described by the self-similar solutions with finite mass (Barenblatt solutions) constructed in [60], which have the form

$$w(x,t) = t^{-N\beta_1}\phi_1(y), \quad y = x\,t^{-\beta_1},$$

where $\beta_1 = 1/(N(q-1)+2\sigma)$ and the profile function ϕ_1 satisfies the following equation

$$(-\Delta)^{\sigma}\phi_1^q = \beta_1 \nabla\cdot(y\,\phi_1). \tag{3}$$

These solutions are well defined for $q > (q)_c$, where β_1 is positive. The profile $\phi_1(y)$ is a smooth and positive radial function in $\mathbb{R}^N$, it is monotone decreasing in $r = |y|$ and has a polynomial decay rate as $|y| \to \infty$ depending on the exponent q.

3.2 Self-similar Solutions for Problem (M1) When $m < 2$

In a previous work [55] we have established three main types of self-similar solutions for model (M1) depending on the range of the parameter m, but always restricted to the range $m < 2$. The first type are functions that are positive for all times, while the second type are functions that extinguish in finite time, both separated by a transition type. We briefly present these solutions here with the purpose of relating model (M1) with its alternative (FPME). The rigorous computations and the derivation of the formulas are written in [55].

Self-similarity of First Type. Solutions that Exist for All Positive Times A self-similar solution $V(x,t)$ of the first type to equation (M1) conserving mass is given by

$$V(x,t) = t^{-\alpha_2}\phi_2(y), \quad y = x\,t^{-\beta_2}$$

with $\alpha_2 = N\beta_2$ and $\beta_2 = 1/(N(m-1)+2-2s)$, and with profile function ϕ_2 satisfying the equation

$$\nabla\cdot(\phi_2^{m-1}\,\nabla(-\Delta)^{-s}\phi_2) = -\beta_2\nabla\cdot(y\,\phi_2). \tag{4}$$

These solutions are considered in the range of parameters where $\beta_2 > 0$, that is, for $m > (N-2+2s)/N$.

Self-similarity of Second Type. Extinction in Finite Time These are solutions to Eq. (M1) with the self-similar form

$$V(x,t) = (T-t)^{\overline{\alpha}_2}\psi_2(y), \quad y = x(T-t)^{\overline{\beta}_2}.$$

where $\overline{\alpha}_2 = N\overline{\beta}_2$, $\overline{\beta}_2 = 1/(N(1-m)+2s-2)$. The profile ψ_2 satisfies the equation

$$\nabla \cdot (\psi_2^{m-1} \nabla(-\Delta)^{-s} \psi_2) = \nabla \cdot (y\psi_2). \tag{5}$$

Here $\overline{\beta}_2 = -\beta_2$, where β_2 is the self-similarity exponent of first type. We argue now in the range of parameters where $\overline{\beta}_2 > 0$, that is $m < (N-2+2s)/N$.

- **Self-similarity of Third Type. Eternal Solutions** For $m \to (N-2+2s)/N$ there is a class of self-similar solutions to Eq. (M1) conserving mass of the form

$$V(x,t) = e^{-c\,t} F(y), \quad y = x e^{-c\,t},$$

where $c > 0$ is a free parameter (exponential self-similarity, which usually plays a transition role) and F is a solution to the profile equation

$$\nabla \cdot (F^{m-1} \nabla(-\Delta)^{-s} F) = -c\nabla \cdot (yF). \tag{6}$$

Remark 2 Solutions of this type live backward and forward in time, they are eternal. Notice that in this borderline case $m \to (N-2+2s)/N$ we have $1/\beta_2 = 1/\overline{\beta}_2 \to 0$, and therefore self-similar solutions of the first and second type do not apply here.

- **The Transformation** In [55] we found an unexpected relationship that allows to transform the families of mass-conserving self-similar solutions of models (FPME) and (M1) into each other, if suitable parameter ranges are prescribed. Actually, there exists a precise correspondence between the profiles ϕ_1 and ϕ_2, ψ_2 or F, and the parameters q and m, as well as σ and s.

Theorem 1 *Let $q > \frac{N-2\sigma}{N}$, $s \in (0,1)$ and let $\phi_1 \geq 0$ be a solution to the profile equation (3). The following holds:*

(i) *If $q \in (\frac{N}{N+2\sigma}, \infty)$ then*

$$\phi_2(x) = (\beta_1/\beta_2)^{\frac{q}{1-q}} (\phi_1(x))^q$$

is a solution to the profile equation (4) if we put $m = \frac{2q-1}{q}$ and $s = 1-\sigma$. This corresponds to $m \in (\frac{N-2+2s}{N}, 2)$, $s \in (0,1)$.

(ii) *If $q \in (\frac{N-2\sigma}{N}, \frac{N}{N+2\sigma})$ then*

$$\psi_2(x) = (\beta_1/\beta_2)^{\frac{q}{1-q}} (\phi_1(x))^q$$

is a solution to the profile equation (5) if we put $m = \frac{2q-1}{q}$ and $s = 1-\sigma$. This corresponds to $m \in \left(\frac{N-4+4s}{N-2+2s}, \frac{N-2+2s}{N}\right)$, $s \in (0,1)$.

(iii) *If $q = \frac{N}{N+2\sigma}$ then*

$$F(x) = (\beta_1/c)^{\frac{N}{2\sigma}} \, (\phi_1(x))^{\frac{N}{N+2\sigma}}$$

is a solution to the profile equation (6) *if we put* $m = \frac{N-2+2s}{N}$ *and* $s = 1 - \sigma$.

This transformation is an algebraic relation that has important consequences. It provides a proof for the existence of self-similar solutions to Problem (M1) and their characterization. Note that $m < 2$.

Self-similar solutions of equation (FPME) can be also constructed for smaller values of q: these are very singular solutions that extinguish in finite time (see Vázquez [60]) or blow up in finite time (see Vázquez and Volzone [64]). These ones can be transformed in a similar manner to corresponding self-similar solutions of the model (M1). Rigorous proofs with complete computations can be found in [55]. Self-similar solutions do not have an explicit formula, except very particular cases of exponents $m = m(s)$ explicitly computed by Huang in [37].

3.3 Related Models for $m > 2$

In this section we will derive (in a formal way) a transformation formula between self-similar solutions of model (M1) and two new nonlocal problems. First, we will show that when $m > 2$, self-similar solutions of (M1) have a correspondence to self-similar solutions of

$$v_t + v^2(-\Delta)^{1-s} v^{\overline{m}} = 0, \quad x \in \mathbb{R}^N, \ t > 0. \tag{7}$$

with $\overline{m} = 1/(m-2)$. The algebraic change $v = 1/w$ maps solutions of problem (7) to solutions of

$$w_t - (-\Delta)^{1-s} w^{-\overline{m}} = 0.$$

which corresponds to solutions of (FPME) for negative exponents q (it is known this equation does not admit integrable solutions, cf. [7], therefore one must search for non-integrable solutions).

Self-similar solutions of Problem (7) are of the form

$$V(x,t) = t^{-a} \psi(y), \quad y = x\, t^b$$

with

$$a = bN, \quad b = \frac{1}{N(\overline{m}+1)+2(1-s)},$$

where ψ satisfies the profile equation

$$b(N\psi - y\nabla\psi) = \psi^2(-\Delta)^{1-s}\psi^{\overline{m}}. \tag{8}$$

Lemma 1 *Let $m > 2$ and let $\phi(x,t)$ a smooth solution to the profile equation (4). Let ψ and $\overline{m}$ defined by*

$$\phi = c\,\psi^{\overline{m}}, \quad \overline{m} = \frac{1}{m-2},$$

with $c = \left(\frac{\beta}{b}\right)^{1/(m-1)}$. Then ψ is a solution to the profile equation (8).

Proof The proof follows directly from the profile equation (4) . Indeed,

$$\begin{aligned}
\phi^{m-1}\,\nabla(-\Delta)^{-s}\phi &= -\beta y\,\phi \\
\nabla\cdot\nabla(-\Delta)^{-s}\phi &= -\beta\nabla\cdot(y\,\phi^{2-m}) \\
(-\Delta)^{1-s}\phi &= \beta(N\,\phi^{2-m} + (2-m)y\phi^{1-m}\nabla\phi) \\
\phi^{2(m-2)}(-\Delta)^{1-s}\phi &= \beta(N\,\phi^{m-2} - y\nabla\phi^{m-2}) \\
\psi^2(-\Delta)^{1-s}\psi^{\overline{m}} &= b(N\,\psi - y\nabla\psi).
\end{aligned}$$

Remark 3

(i) In Lemma 4 we will have prove the existence of a self-similar solution of model (M1) when $N = 1$ as a limit of the rescaled solutions. The transformation formula given by Lemma 1 would give the existence of a self-solution of model (M1) for $N \geq 1$ and $m \geq 2$ subject to a rigorous proof of existence of selfsimilar solutions of model (7).

(ii) Model (7) has not been studied in the literature yet. However, it seems natural to think that it will enjoy some good properties, such as comparison principle for viscosity solutions. It is a possible direction to follow in order to deal with the open problem of the asymptotic behavior of solutions of model (M1) for $m > 2$ and $N > 1$ where no uniqueness is known for (M1).

4 The Integrated Model in one Dimension for $m \in (1, \infty)$

We consider model (M1) in one space dimension

$$\partial_t u = \partial_x \cdot (u^{m-1}\partial_x(-\Delta)^{-s}u), \tag{9}$$

for $x \in \mathbb{R}$, $t > 0$ and $s \in (0, 1)$. We take compactly supported initial data $u_0 \geq 0$ such that $u_0 \in L^1_{\text{loc}}(\mathbb{R})$. The "integrated solution" v is defined by

$$v(x,t) = \int_{-\infty}^{x} u(y,t)\,dy \geq 0 \quad \text{for } t > 0,\ x \in \mathbb{R}. \tag{10}$$

Therefore $v_x = u$ and $v(x,t)$ will be a solution (in the viscosity sense) of the equation

$$\partial_t v = -|v_x|^{m-1}(-\Delta)^{\alpha} v, \tag{11}$$

with $\alpha = 1 - s$ and initial data

$$v(x,0) = v_0(x) := \int_{-\infty}^{x} u_0(x)\,dx \quad \text{for all } x \in \mathbb{R}. \tag{12}$$

Note that $v(x,t)$ is a non-decreasing function in the space variable x. Moreover, since $u(x,t)$ enjoys the property of conservation of mass, then $v(x,t)$ satisfies

$$\lim_{x\to-\infty} v(x,t) = 0, \qquad \lim_{x\to+\infty} v(x,t) = M$$

for all $t \geq 0$. We consider viscosity solutions $v(x,t)$ of (11)–(12) in the sense of Crandall-Lions.

Definition 2 Let v be a upper semi-continuous function (resp. lower semi-continuous function) in $\mathbb{R} \times (0,\infty)$.

(i) We say that v is a **viscosity sub-solution** (resp. **super-solution**) of Eq. (11) on $\mathbb{R} \times (0,\infty)$ if for any point (x_0, t_0) with $t_0 > 0$ and any $\tau \in (0, t_0)$ and any test function $\varphi \in C^2(\mathbb{R} \times (0,\infty)) \cap L^\infty(\mathbb{R} \times (0,\infty))$ such that $v - \varphi$ attains a global maximum (resp. minimum) at the point (x_0, t_0) on $Q_\tau = \mathbb{R} \times (t_0 - \tau, t_0]$ we have that

$$\partial_t \varphi(x_0,t_0) + |\varphi_x(x_0,t_0)|^{m-1}((-\Delta)^{\alpha}\varphi(\cdot,t_0))(x_0) \leq 0 \quad (\geq 0).$$

(ii) We say that v is a **viscosity sub-solution** (resp. **super-solution**) of the initial-value problem (11)–(12) on $\mathbb{R} \times (0,\infty)$ if it satisfies moreover at $t = 0$

$$v(x,0) \leq \limsup_{y\to x,\ t\to 0} v(y,t) \quad (\text{resp. } v(x,0) \geq \liminf_{y\to x,\ t\to 0} v(y,t)).$$

We say that $v \in C(\mathbb{R} \times (0,\infty))$ is a **viscosity solution** if v is a viscosity sub-solution and a viscosity super-solution on $\mathbb{R} \times (0,\infty)$.

Since Eq. (11) is invariant under translations, the test function φ in the above definition can be taken such that φ touches v from above in the sub-solution case, resp. φ touches v from below in the super-solution case.

Now we state the correspondence of solutions between the two models, together with the regularity estimates that u inherits from v.

Proposition 1 *Let $s \in (0,1)$ and $m \geq 1$. Let also u be a constructed weak solution for Problem (9) with initial data $u_0 \in \mathscr{M}_+(\mathbb{R})$. Then v defined by (10) is a viscosity solution for Problem (11)–(12) and $v \in C(\mathbb{R} \times (0,T))$. If additionally $u_0 \in L^\infty(\mathbb{R})$, then $v \in C(\mathbb{R} \times [0,T))$.*

Proof We start by proving the regularity of v. If $u_0 \in L^\infty(\mathbb{R})$ then $u \in L^\infty(\mathbb{R} \times (0,T))$. This implies that $v \in L^\infty([0,T] : \mathrm{Lip}(\mathbb{R}))$ since $v_x = u \in L^\infty([0,T] : L^\infty(\mathbb{R}))$. Moreover $v_t \in L^2([0,T] : L^2(B))$ by the second energy estimate for u. Then the continuity in time is shown by estimating the decay of the time shift: $|v(x_0,t_1) - v(x_0,t_0)| \leq K|t_1 - t_0|^{1/3}$ when t_1 and t_0 are close. This is done by combining the L^2 estimate on v_t and the already proved spatial regularity.

When $u_0 \notin L^\infty(\mathbb{R})$, we use the smoothing effect given in Theorem (2) to show that $u \in L^\infty(\mathbb{R}^N \times (\tau,\infty))$ for any $\tau > 0$. Consequently, we proceed as before but avoiding $t = 0$.

To show that v is in fact a viscosity solution, we consider the regularized problem

$$(v_\delta)_t = \delta\Delta(v_\delta) + |(v_\delta)_x|^{m-1}(-\Delta)^{1-s}v_\delta. \tag{13}$$

It is clear that $v_\delta = \int_{-\infty}^{x} u_\delta(y,t)dy$ where u_δ is the classical solution of

$$\partial_t u_\delta = \delta\Delta u_\delta + \partial_x \cdot (u_\delta^{m-1}\partial_x(-\Delta)^{-s}u_\delta),$$

Not that this is the same problem that in the last step of the proof of existence of u. Thus, up to a corresponding subsequence, we can pass to the limit in (13) and show that $v := \lim_{\delta\to 0} v_\delta$ is a viscosity solution of (11) since it is a limit of viscosity solutions.

The standard comparison principle for viscosity solutions holds true (see Chasseigne and Jakobsen [22]). We also refer to [4] for more details regarding properties of this integrated model when $m = 2$.

Proposition 2 (Comparison Principle) *Let $m \geq 1$ and $\alpha \in (0,1)$. Let w be a viscosity sub-solution and W be a viscosity super-solution of Eq. (11). If $w(x,0) \leq W(x,0)$, then $w \leq W$ in $\mathbb{R} \times (0,\infty)$.*

The following uniqueness result is also proved in [22]. This result is crucial to obtain uniqueness of weak solution in dimension one for (M1).

Proposition 3 *Let $m \geq 1$ and $\alpha \in (0,1)$. Then there exists a unique viscosity solution of Problem (11) with piecewise continuous initial data.*

5 Main Results for Model (M1)

5.1 Existence of Solutions

The most important contribution to the existence theory was done in [57] where we constructed a weak solution in the sense of Definition 1 in the general setting of initial data any $\mu \in \mathscr{M}^+(\mathbb{R}^N)$, the space of nonnegative Radon measures on $\mathbb{R}^N$ with *finite mass*. In particular, this includes the case of merely integrable data $u_0 \in L^1(\mathbb{R}^N)$.

Theorem 2 *Let $m \in (1, +\infty)$, $s \in (0, 1)$, $N \geq 1$ and $\mu \in \mathscr{M}^+(\mathbb{R}^N)$. Then there exists a nonnegative weak solution u of Problem (M1) and for all $\tau > 0$,*

$$u \in L^\infty((\tau, \infty) : L^1(\mathbb{R}^N)) \cap L^\infty(\mathbb{R}^N \times (\tau, \infty)) \cap L^\infty((0, \infty) : \mathscr{M}^+(\mathbb{R}^N)).$$

Moreover, u has the following properties:

1. ***(Conservation of mass)*** *For all $0 < t < T$ we have $\int_{\mathbb{R}^N} u(x,t)dx = \int_{\mathbb{R}^N} d\mu(x)$.*
2. *(L^∞ **estimate**) $0 < \tau < t < T$ we have $||u(\cdot,t)||_\infty \leq ||u(x,\tau)||_\infty$.*
3. *(L^p**-energy estimate**) For all $1 < p < \infty$ and $0 < \tau < t < T$ we have*

$$\int_{\mathbb{R}^N} u^p(x,t)dx + \frac{4p(p-1)}{(m+p-1)^2} \int_\tau^t \int_{\mathbb{R}^N} \left|(-\Delta)^{\frac{1-s}{2}} \left[u^{\frac{m+p-1}{2}}\right](x,s)\right|^2 dxds \leq \int_{\mathbb{R}^N} u^p(x,\tau)dx. \tag{14}$$

4. ***(Second energy estimate)*** *For all $0 < \tau < t < T$ we have*

$$\frac{1}{2} \int_{\mathbb{R}^N} \left|(-\Delta)^{-\frac{s}{2}} u(x,t)\right|^2 dx + \int_\tau^t \int_{\mathbb{R}^N} u^{m-1} \left|\nabla(-\Delta)^{-s} u(x,s)\right|^2 dxds \leq \frac{1}{2} \int_{\mathbb{R}^N} \left|(-\Delta)^{-\frac{s}{2}} u(x,\tau)\right|^2 dx. \tag{15}$$

5. ***(Smoothing effect)*** *For all $t > 0$, we have*

$$\|u(\cdot,t)\|_{L^\infty(\mathbb{R}^N)} \leq C_{N,s,m}\, t^{-\gamma}\, \mu(\mathbb{R}^N)^\delta$$

where $\gamma = \frac{N}{(m-1)N+2(1-s)} > 0$ and $\delta = \frac{2(1-s)}{(m-1)N+2(1-s)} > 0$.

Remark 4 If $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$, all the properties of Theorem (2) hold up to $\tau = 0$.

5.2 Uniqueness in Dimension $N = 1$

Uniqueness of weak solutions is proved in the one-dimensional case.

Theorem 3 *Let $m \in (1, +\infty)$, $s \in (0, 1)$, $N = 1$ and $\mu \in \mathscr{M}^+(\mathbb{R}^N)$. Then there exists a unique constructed weak solution to Problem* (M1).

The proof follows as consequence of Propositions 1 and 3.

5.3 Speed of Propagation

A very interesting property is the finite/infinite speed of propagation of the solution of Problem (M1) depending on the nonlinearity parameter m, as proved in [56, 57].

Theorem 4 *Let $m \in [2, \infty)$, $s \in (0, 1)$ and $N \geq 1$. Assume that $u_0 \in L^\infty(\mathbb{R}^N)$ has compact support and let u be the constructed weak solution of Problem* (M1) *given in Theorem 2. Then, $u(\cdot, t)$ has compact support for all $t > 0$, i.e., u has **finite speed of propagation**.*

Theorem 5 *Let $m \in (1, 2)$, $s \in (0, 1)$ and $N = 1$. Assume $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ and let u be the constructed weak solution of Problem* (M1). *Then, for any $t > 0$ and $R > 0$, the set $\mathscr{P}_{R,t} = \{x : |x| \geq R,\ u(x, t) > 0\}$ has positive measure (even if u_0 is compactly supported). This is a weak form of **infinite speed of propagation**. Moreover, if u_0 is radially symmetric and non-increasing in $|x|$, then $u(x, t) > 0$ for all $x \in \mathbb{R}$ and $t > 0$.*

5.4 Asymptotic Behavior

Once we know the uniqueness result of Theorem 3 and the existence of solutions for finite measure data, we can prove that there exists a unique fundamental solution to Problem (M1) and it describes the large time asymptotic behavior of a general class of solutions.

Theorem 6 (Asymptotic Behavior) *Let $m \in (1, \infty)$, $s \in (0, 1)$ and $N = 1$. Assume that $u_0 \in L^1(\mathbb{R})$ such that $\|u_0\|_{L^1(\mathbb{R})} = M$ and let u be the corresponding constructed weak solution of* (M1). *Then*

$$t^{\frac{N\left(1-\frac{1}{p}\right)}{(m-1)N+2-2s}} \|u(\cdot, t) - U_M(\cdot, t)\|_{L^p(\mathbb{R}^N)} \to 0 \quad \text{as} \quad t \to \infty$$

for any $p > 1$, where U_M is the unique self-similar solution of (M1) *with initial data $\mu = M\delta_0$.*

Notice that U_M can be transformed into a self-similar solution of (FPME) (for $m < 2$) or (7) (for $m > 2$) as explained in Sect. 3. In the first case $m < 2$ this transformation allows to obtain the main properties of U_M from the known properties of the Barenblatt solutions of the (FPME), which are derived in [60]. The precise decay for large $|x|$ of U_M is given in [55], Corollary 3.2.

6 Proofs of the Results

6.1 Sketch of the Proof of Existence

The proof of existence for general initial data given by an integrable measure is a long ride, with several nontrivial steps. In this section we will show the strategies of the proof, together with some details of the main ingredients of it.

We will first prove existence for $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ via a four steps approximation method by regularized versions of (M1). In this first part we will also obtain very useful energy estimates that ensure compactness, together with an $L^1 - L^\infty$ smoothing effect. Afterwards, using the smoothing effect we prove existence for initial data $\mu \in \mathscr{M}_+(\mathbb{R}^N)$ approximating μ by bounded integrable initial data.

6.1.1 Existence for $u_0 \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$

Formally, we consider an equivalent version of (M1) given by:

$$u_t = \nabla \cdot (u^{m-1} \nabla (-\Delta)^{-1} (-\Delta)^{1-s} u) \quad \text{in} \quad \mathbb{R}^N. \tag{16}$$

The idea is to consider a regularized version of (16) where all the problematic terms are approximated. More precisely, we add the vanishing viscosity term $\delta \Delta u$ to (16) that ensures good properties of regularity for the solution, we eliminate the degeneracy at the zero level sets by putting $u^{m-1} \sim (u + \mu)^{m-1}$ and we eliminate the singular character of the fractional Laplacian $(-\Delta)^{1-s}$ approximating it by the zero order pseudo-differential operator

$$\mathscr{L}_\varepsilon^{1-s}(u)(x) = C_{N,1-s} \int_{\mathbb{R}^N} \frac{u(x) - u(y)}{\left(|x-y|^2 + \varepsilon^2\right)^{\frac{N+2-2s}{2}}} dy. \tag{17}$$

Additionally, to ensure existence, we will restrict (16) to a bounded domain B_R. The approximated problem reads

$$\begin{cases} (U_1)_t = \delta\Delta U_1 + \nabla\cdot((U_1+\mu)^{m-1}\nabla(-\Delta)^{-1}\mathscr{L}_\varepsilon^{1-s}[U_1]) & \text{in } B_R\times(0,T), \\ U_1(x,0) = \widehat{u}_0(x) & \text{in } B_R, \\ U_1(x,t) = 0 & \text{in } B_R^c\times(0,T), \end{cases} \tag{$P_{\varepsilon\delta\mu R}$}$$

depending on the parameters $\varepsilon, \delta, \mu, R > 0$. We also consider $\widehat{u}_0$ to be a smooth approximation of u_0. We say that U_1 is a weak solution of ($P_{\varepsilon\delta\mu R}$) if

$$\int_0^T\int_{B_R} U_1(\phi_t - \delta\Delta\phi)dxdt - \int_0^T\int_{B_R}(U_1+\mu)^{m-1}\nabla(-\Delta)^{-1}\mathscr{L}_\epsilon^{1-s}[U_1]\cdot\nabla\phi dxdt + \int_{B_R}\widehat{u}_0(x)\phi(x,0)dx = 0$$

for smooth test functions $\phi(x,t)$ that vanish on the spatial boundary ∂B_R and for large t. Indeed, existence of smooth weak solutions is proved via mild solutions, i.e, U_1 is the fixed point of the following map:

$$\mathscr{T}(v)(x,t) = e^{\delta t\Delta}u_0(x) + \int_0^t \nabla e^{\delta(t-\tau)\Delta}\cdot G(v)(x,\tau)d\tau,$$

where $G(v) = (v+\mu)^{m-1}\nabla(-\Delta)^{-1}\mathscr{L}_\epsilon^s[v]$. The map,

$$\mathscr{T} : C((0,T) : L^1(B_R)\cap L^\infty(B_R)) \to C((0,T) : L^1(B_R)\cap L^\infty(B_R))$$

is well defined and it is also a contraction, thus, Banach contraction principle ensures existence of a fixed point. We refer to [6] for a very similar proof in a slightly different context.

Once existence and regularity of the approximated problem are obtained, we provide the solution with the following energy-type estimates, that will give compactness that allow to pass to the limit in all the approximation parameters.

- L^p **Energy Estimates for** $1 \le p < \infty$ For all $0 < t < T$ we have that:

$$\int_{B_R} U_1^p(x,t)dx + p(p-1)\int_0^t\int_{B_R}|(\mathscr{L}_\epsilon^{1-s})^{\frac{1}{2}}[\Psi(U_1)](x,s)|^2dxds + \frac{4(p-1)\delta}{p}\int_0^t\int_{B_R}\left|\nabla(U_1^{p/2})(x,s)\right|^2dxds \le \int_{B_R}u_0^p(x)dx \tag{18}$$

where $\Psi(z) = \int_0^z \zeta^{\frac{p-2}{2}}(\zeta+\mu)^{\frac{m-1}{2}}d\zeta$. We want to mention that a crucial step in the derivation of (18) relies on the generalized version of the Stroock-Varopoulos Inequality: Given $\psi : \mathbb{R}\to\mathbb{R}$ such that $\psi\in C^1(\mathbb{R})$, $\psi'\ge 0$ and Ψ such that $\psi' = (\Psi')^2$ we have

$$\int_{\mathbb{R}^N} \psi(w)\mathscr{L}_\epsilon^s[w]dx \geq \int_{\mathbb{R}^N} \left|(\mathscr{L}_\epsilon^s)^{\frac{1}{2}}[\Psi(w)]\right|^2 dx.$$

- **Second Energy Estimate** For all $0 < t < T$ we have that:

$$\frac{1}{2}\int_{B_R} \left|\left((-\Delta)^{-1}\mathscr{L}_\epsilon^{1-s}\right)^{\frac{1}{2}}[U_1(t)]\right|^2 dx + \int_0^t \int_{B_R} (U_1+\mu)^{m-1}\left|\nabla(-\Delta)^{-1}\mathscr{L}_\epsilon^{1-s}[U_1]\right|^2 dxdt + \delta\int_0^t \int_{B_R} \left|(\mathscr{L}_\epsilon^{1-s})^{\frac{1}{2}}[U_1]\right|^2 dxdt \leq \frac{1}{2}\int_{B_R} \left|\left((-\Delta)^{-1}\mathscr{L}_\epsilon^{1-s}\right)^{\frac{1}{2}}[u_0]\right|^2 dx.$$

- **(Decay of Total Mass)** For all $0<t<T$ we have $\int_{B_R} U_1(x,t)dx \leq \int_{B_R} u_0(x)dx$.
- **(L^∞-estimate)** For all $0 < t < T$ we have $||U_1(\cdot,t)||_\infty \leq ||u_0||_\infty$.

By combining these energy estimates we are able to apply some suitable parabolic compactness theorems to derive convergence of approximated solutions when the parameters of the approximations are passed to the limit step by step in the order:

$$(P_{\epsilon\delta\mu R}) \xrightarrow{\epsilon\to 0} (P_{\delta\mu R}) \xrightarrow{R\to\infty} (P_{\delta\mu}) \xrightarrow{\mu\to 0} (P_\delta) \xrightarrow{\epsilon\to 0} (P).$$

Remark 5

(i) Notice that the fractional operator is always defined by extending the function by 0 outside the ball B_R in the first two problems of the approximation $(P_{\epsilon\delta\mu R})$ and $(P_{\delta\mu R})$. This is a delicate aspect which needs to be properly justified. The functions U_1, U_2 are defined on a ball B_R and extended by 0 to $\mathbb{R}^N \setminus B_R$. We are able to do this extension since $U_1, U_2 \in H_0^1(B_R)$ by (18) therefore they have the right decay at the boundary ∂B_R that allows the extension by 0. This is also one of the reasons for which the term $\delta\Delta U$ in the approximating problems is the last one passing to the limit.

(ii) The term with δ coefficient in the L^p estimate (18) gives H^1 regularity, an essential information in using parabolic compactness criteria. Again, this motivates the $\delta \to 0$ limit to be the last one.

Passing to the Limit First limit is done as $\epsilon \to 0$ and is based on the compactness criteria of type Simon-Aubin-Lions [53] in the context of

$$H_{\epsilon_0}^{1-s}(B_R) \subset L^2(B_R) \subset H^{-1}(B_R),$$

where $H_{\epsilon_0}^{1-s}(B_R)$ is the space associated to (17), and thus the left hand side inclusion is compact. We conclude that the family of approximate solutions $\{U_1\}_{\epsilon>0}$ is relatively compact in $L^2(0,T : L^2(B_R))$ and we obtain that $(U_1)_{\epsilon,\delta,\mu,R} \to (U_2)_{\delta,\mu,R}$ as $\epsilon \to 0$ in $L^2(0,T : L^2(B_R))$, up to subsequences.

As usual, the limit U_2 is identified to be a weak solution of a limit problem, in this case $(P_{\delta\mu R})$. Moreover, U_2 will satisfy the corresponding energy estimates which are proved by passing to the limit as $\epsilon \to 0$ the estimates for U_1. The following two limits $R \to \infty$ and $\mu \to 0$ are similar using the same type of compactness criteria of Simon.

The novelty appears in the last limit as $\delta \to 0$ where the regularity given by H_1 term with δ coefficient is lost. Here we need to use a different compactness criteria due to Rakotoson and Temam [52] which does not ask for such strong regularity assumptions as before. We conclude that the solution U_4 of (P_δ) satisfies

$$U_4 \to u \text{ as } \delta \to 0 \text{ in } L^2_{\text{loc}}(\mathbb{R}^N \times (0,T)).$$

In the end we prove that u is a weak solution to Problem (M1) and it satisfies the corresponding energy estimates. We call this *u constructed weak solution* since there is no uniqueness theory available in $\mathbb{R}^N$. For $N = 1$ we prove in Sect. 5.2 that uniqueness holds in the class of weak solutions and therefore the u we have constructed is indeed the weak solution to Problem (M1).

An L^p-L^∞ smoothing effect is proved by combining L^p energy estimate (14) with the Nash-Gagliardo-Nirenberg inequality (See Theorem 7.4 in [57]) for the function $u^{\frac{m+p+1}{2}}$. More precisely we get

$$\|u(\cdot,t)\|_{L^\infty(\mathbb{R}^N)} \le C_{N,s,m,p}\, t^{-\gamma_p} \|u_0\|^{\delta_p}_{L^p(\mathbb{R}^N)} \quad \text{for all} \quad t > 0, \tag{19}$$

where $\gamma_p = \frac{N}{(m-1)N+2p(1-s)}$, $\delta_p = \frac{2p(1-s)}{(m-1)N+2p(1-s)}$.

6.1.2 Existence for Initial Data in $\mathscr{M}_+(\mathbb{R}^N)$

The existence of a solution for measure data is done via an approximating problem with data $(u_0)_n \in L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ where $(u_0)_n \to \mu$ and it conserves the mass $\|(u_0)_n\|_{L^1(\mathbb{R}^N)} = \mu(\mathbb{R}^N)$. More precisely, let u_n be the solution to Problem (M1) with data

$$(u_0)_n(x) := \int_{\mathbb{R}^N} \rho_n(x-z)\, d\mu(z).$$

We use the smoothing effect (19) for $L^1(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ initial data in the particular case $p = 1$. Then, as in the previous section u_n satisfies the energy estimates plus the smoothing effect:

$$\|u_n(\cdot,t)\|_{L^\infty(\mathbb{R}^N)} \le \|u_n(\cdot,\tau)\|_{L^\infty(\mathbb{R}^N)} \le C_{N,s,m}\, \tau^{-\gamma} \|(u_0)_n\|^{\delta}_{L^1(\mathbb{R}^N)} = C_{N,s,m}\, \tau^{-\gamma} \mu(\mathbb{R}^N)^{\delta},$$

where $\gamma = \frac{N}{(m-1)N+2(1-s)}$, $\delta = \frac{2(1-s)}{(m-1)N+2(1-s)}$. Note that the bound does not depend on the approximation parameter n. In a similar way as before, we derive compactness estimates and apply the Rakotoson-Temam criteria [52] in order to obtain a limit as $n \to \infty$ away from $t = 0$

$$u_n \longrightarrow u^\tau \quad \text{as} \quad n \to \infty \quad \text{in} \quad L^2_{\text{loc}}(\mathbb{R}^N \times (\tau, T)).$$

We also show that the initial data is recovered. Basically, the second energy estimate given by (15) allows us to prove that for any test function ϕ we have that

$$\left| \int_0^\tau \int_{\mathbb{R}^N} u_n^{m-1} \nabla (-\Delta)^{-s} u_n \nabla \phi dx dt \right| \leq \Lambda(\tau)$$

for some modulus of continuity that only depends on ϕ, μ and s. Thus,

$$\begin{aligned} \left| \int_{\mathbb{R}^N} (u_n(\tau) - (u_0)_n)\phi dx \right| &= \left| \int_0^\tau \int_{\mathbb{R}^N} \partial_t u_n \phi \, dx dt \right| \\ &= \left| \int_0^\tau \int_{\mathbb{R}^N} u_n^{m-1} \nabla (-\Delta)^{-s} u_n \nabla \phi \, dx dt \right| \leq \Lambda(\tau). \end{aligned}$$

A standard diagonal argument in τ and n completes the proof of existence for measure data.

- **Conservation of Mass** is proved by using the previous estimate with the sequence of cutoff type test functions $\phi_R(x) = \phi(x/R)$ with $0 \leq \phi \leq 1$ and $\phi_1(x) = 1$ for $|x| \leq 1$ and such that $\|\nabla \phi_R\|_{L^\infty(\mathbb{R}^N)} = O(R^{-1})$.

6.2 Sketch of the Proof of Speed of Propagation

The proof requires delicate barrier arguments since Problem (M1) is proved to have a lack of comparison principle. We refer to [16] for an explicit example of this fact.

6.2.1 Finite Speed of Propagation for $m \in [2, \infty)$

However, a special kind of super solutions (so-called true super-solutions), are of particular interest. We can show that, comparing any solution with a true super-solution, no contact point between them is possible. This will be enough to show the property of finite speed of propagation.

Without loss of generality, we assume that $0 \le u_0 \le 1$ (thus, $0 \le u(x,t) \le 1$) and consider the parabola-like function

$$U(x,t) = ((Ct - (|x| - b))_+)^2.$$

where $b > 0$ is such that $u_0(x) < U(x,0) =: U_0(x)$ for all $x \in B_b(0)$ and C is a suitable constant to be chosen later.

We argue by contradiction at a possible first contact point (x_c, t_c) between u and U. The fact that such a first contact point happens for $t > 0$ and $x \neq \infty$ is justified by regularization. We also exclude the extreme case where the contact is made at the boundary of the support of U given by $|x_f(t_c)| := b + Ct_c$ (see Lemma 7.2 in [56]). Then, there exists $h > 0$ such that $b + Ct_c - |x_c| = h > 0$. At (x_c, t_c), we have that $u = U$, $\nabla(u - U) = 0$, $\Delta(u - U) \le 0$, $(u - U)_t \ge 0$, that is

$$u(x_c, t_c) = h^2, \quad u_r = -2h, \quad \Delta u \le 2N, \quad u_t \ge 2Ch,$$

where $r = |x|$ denotes the radial coordinate. We also have the following estimates on $\nabla p := \nabla(-\Delta)^{-s}$ for $0 < s < 1/2$ (see Theorem 5.1. of [16]):

$$-p_r(|x_c|, t_c) \le K_1 + K_2 h^{1+2s} + K_3 h, \quad \Delta p(|x_c|, t_c) \le K_4$$

for some $K_1, K_2, K_3, K_4 \ge 0$. We now use the expanded form of Problem (M1) given by $u_t = (m-1)u^{m-2}\nabla u \cdot \nabla p + u^{m-1}\Delta p$, we get the inequality

$$\begin{aligned} C &\le (m-1)h^{2m-4}\left(-p_r(|x_c|, t_c) + \frac{h}{2}\Delta p(|x_c|, t_c)\right) \\ &\le (m-1)h^{2m-4}\left(K_1 + K_2 h^{1+2s} + (K_3 + \frac{K_4}{2})h\right), \end{aligned}$$

which leads to a contradiction choosing $C = C(s, N)$ large enough.

When $1/2 \le s < 1$, an improved version on the estimate of p_r leads to similar result, but this time $C = C(t)$. This is again enough to prove the property of finite speed of propagation, but this time, we do not have a quantitative estimate on the growth of the support.

One can easily see that the term h^{2m-4} in the last estimate needs $m \ge 2$ to create a contradiction. In fact $m = 2$ is show to be the critical exponent, as we show in the following section.

6.2.2 Infinite Speed of Propagation

In dimension $N = 1$, we have already established a duality between constructed weak solutions of (M1)

$$\partial_t u = \nabla \cdot (u^{m-1}\nabla(-\Delta)^{-s}u)$$

and viscosity solutions of the "integrated problem"

$$\partial_t v = -|v_x|^{m-1}(-\Delta)^{\alpha} v, \tag{20}$$

where $v(x,t) = \int_{-\infty}^{x} u(x,t)dx$ and $\alpha = 1 - s$. It will be enough to consider the initial data given by

$$v_0(x) \geq H_{x_0}(x) = \begin{cases} 0, & x < x_0, \\ 1, & x > x_0. \end{cases} \tag{21}$$

Indeed, (20) has suitable comparison principles for viscosity solutions. See Propositions 8.5 and 8.6 in [56] for a standard comparison principle and a parabolic type comparison principle respectively. At this point, we need to find a subsolution $\Phi = \Phi(x,t)$ of (20) such that $\Phi(x,0) \leq v_0(x)$ and $\Phi(x,t) > 0$ for any $t > 0$ and $|x|$ arbitrary large. See Figs. 1 and 2 for a graphic version of the proof.

We need to find this kind of subsolution. First, consider a function $G = G(x)$ such that G is compactly supported in the interval $(-x_0, \infty)$, $G(x) \leq C_1$ for all $x \in \mathbb{R}$ and $(-\Delta)^s G(x) \leq -C_2|x|^{-(1+2s)}$ for all $x < x_0$ (see Lemma 9.1 in [56]

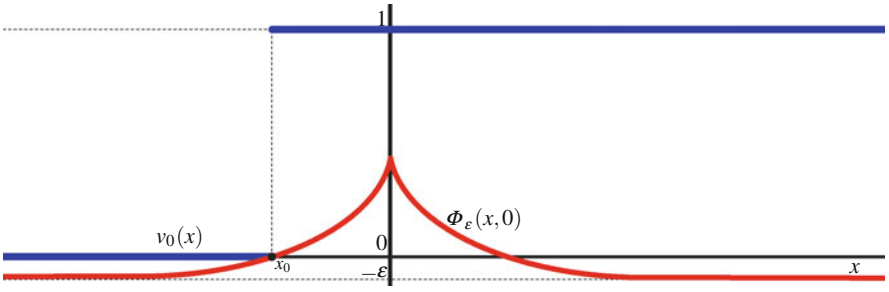


Fig. 1 Comparison with the barrier at time $t = 0$

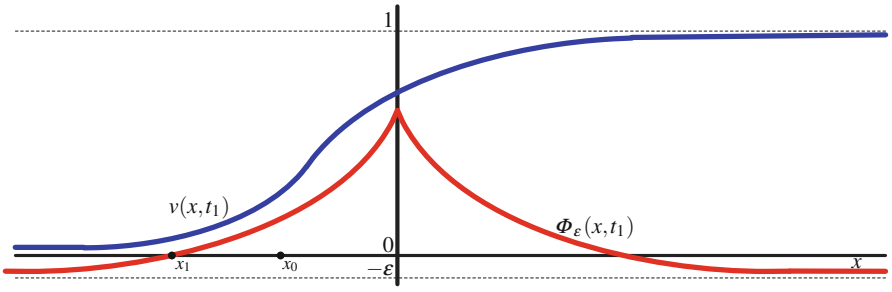


Fig. 2 Comparison with the barrier at time $t > 0$

for the existence of such a function G). Now, given any $\tau, \xi, \epsilon > 0$, we can find $C_2 = C_2(N, s, \alpha, \tau)$ such that the function

$$\Phi_\epsilon(x,t) = (t+\tau)^{b\gamma}\left((|x|+\xi)^{-\gamma} + G(x)\right) - \epsilon, \quad t \geq 0,\ x \in \mathbb{R}.$$

satisfies

$$(\Phi_\epsilon)_t + |(\Phi_\epsilon)_x|^{m-1}(-\Delta)^\alpha \Phi_\epsilon \leq 0 \quad \text{for } x < x_0,\ t > 0$$

for $\gamma = \dfrac{m+2\alpha}{2-m}$ and $b = \frac{1}{m-1+2\alpha}$. The parameters γ and b are found in the study of self-similar solutions of (20).

The main tool to finish the proof is given by the following parabolic comparison principle proved in [56]:

Proposition 4 *Let $m > 1$, $\alpha \in (0,1)$ and $N = 1$. Let v be a viscosity solution of Problem (11)–(12). Let $\Phi : \mathbb{R} \times [0,\infty) \to \mathbb{R}$ such that $\Phi \in C^2(\Omega \times (0,T))$. Assume that*

- *$\Phi_t + |\Phi_x|^{m-1}(-\Delta)^\alpha \Phi < 0$ for $x \in \Omega$, $t \in [0,T]$;*
- *$\Phi(x,0) < v(x,0)$ for all $x \in \mathbb{R}$ (comparison at initial time);*
- *$\Phi(x,t) < v(x,t)$ for all $x \in \mathbb{R} \setminus \Omega$ and $t \in (0,T)$ (comparison on the parabolic boundary).*

Then $\Phi(x,t) \leq v(x,t)$ for all $x \in \mathbb{R}$, $t \in (0,T)$.

At this point we need to show that Φ_ϵ can be compared at initial time and also on the parabolic boundary.

The initial data (21) naturally impose the following conditions on Φ_ϵ:

$$\xi > x_0 + \epsilon^{-\frac{1}{\gamma}}.$$

that ensures that $\Phi_\epsilon(x_0,0) < v_0(x_0)$. Now let $k_1 := \min\{v(x,t) :\ x \geq x_0,\ 0 < t \leq T\} > 0$ (we recall that $v \in C([0,T] : C(\mathbb{R}))$ and $v_0(x_0) = 1$). The condition $\Phi_\epsilon(x,t) < v(x,t)$ for all $x \geq x_0,\ t \in [0,T]$ requires

$$(t+1)^{b\gamma}(\xi^{-\gamma} + C_1) < k_1.$$

The maximum value of $t = T$ for which this inequality holds is

$$T < \left(\frac{k_1}{\xi^{-\gamma} + C_1}\right)^{1/b\gamma} - 1.$$

Thus, in order to have $T > 0$ we require $\xi > (k_1 - C_1)^{-\frac{1}{\gamma}}$. The remaining parameter C_1 from assumption (G2) is chosen here such that: $C_1 < k_1$. By Proposition 4 we obtain the desired comparison:

$$v(x,t) \geq \Phi_\epsilon(x,t) \quad \text{for all } (x,t) \in Q_T.$$

Now, let $x_1 < x_0 < 0$ and $t_1 \in (0, T)$ be arbitrary. It is now straightforward to show that for

$$\epsilon < \left[\frac{(t_1+1)^b}{(k_1 - C_1)^{-\frac{1}{\gamma}} - x_1} \right]^{\gamma}.$$

we have that $\Phi_\epsilon(x_1, t_1) > 0$ and thus, by comparison $v(x_1, t_1) > 0$. In this way, we have proves the following result:

Theorem 7 (Infinite Speed of Propagation for v) *Assume that $u_0 \in L^\infty(\mathbb{R}^N)$ is nonnegative and compactly supported. Let v be the solution of Problem (11)–(12). Then $0 < v(x,t) < M$ for all $t > 0$ and $x \in \mathbb{R}$.*

The result for u follows immediately. We have proved that $v(x,t)$ is positive for every $t > 0$ and $x \in \mathbb{R}$, thus u has accumulated mass at every (x,t). This fact ensures that for every time $t > 0$ there exists an $x \in \mathbb{R}$ arbitrary far from the origin such that $u(x,t) > 0$. Moreover, when u_0 is radially symmetric and non-increasing in $|x|$ then u inherits these properties, ensuring that u can not take zero values.

Remark 6

(i) This method is working only in one dimension since we use the integrated function. It is an open issue the proof of infinite speed of propagation in dimension $N \geq 2$. New methods should be employed and, at least for particular cases of data, one can see a possible direction to continue: for instance radial data will produce radial solutions and then one could search for a suitable transformation between (M1) and a $1 - D$ problem.

(ii) Infinite speed of propagation holds for $m < 2$ for any self-similar solution as a consequence of the transformation formula from the previous section. Moreover their properties are imported from the alternative model (FPME).

6.3 Proof of the Asymptotic Behavior

Here we provide the proof of the asymptotic behavior in dimension $N = 1$ using a four step method. This will be a new contribution to the study of Problem (M1). The result can only be presented in dimension 1 due to the lack of uniqueness for (M1). However, we will present intermediate steps valid in $\mathbb{R}^N$, and the reader could trivially see that the asymptotic behaviour result for general $N \geq 1$ would follow from a result of uniqueness of solutions with Dirac delta type initial data.

6.3.1 Existence of a Rescaled Solution

Lemma 2 *Let $m \in (1, +\infty)$, $s \in (0, 1)$ and $N \geq 1$. Assume that $u_0 \in L^1(\mathbb{R}^N)$ and let u be the constructed weak solution of (M1) given by Theorem 2. Then, for any $\lambda > 0$ the rescaled function*

$$u_\lambda(x, t) = \lambda^N u(\lambda x, \lambda^b t),$$

with $b = (m-1)N + 2 - 2s$, is a weak solution of

$$\begin{cases} \partial_t u_\lambda = \nabla \cdot (u_\lambda^{m-1} \nabla (-\Delta)^{-s} u_\lambda) \text{ for } x \in \mathbb{R}^N,\ t > 0, \\ u_\lambda(0, x) = \lambda^N u_0(\lambda x) \qquad\qquad \text{for } x \in \mathbb{R}^N. \end{cases} \tag{22}$$

Moreover, u_λ has the following properties:

1. ***(Conservation of mass)** For all $0 < t < T$ we have $\int_{\mathbb{R}^N} u_\lambda(x, t)dx = \int_{\mathbb{R}^N} u_0(x)dx$.*
2. *(**L^p energy estimate**) For all $1 < p < \infty$ and $0 < \tau < t < T$ we have*

$$\int_{\mathbb{R}^N} u_\lambda^p(x, t)dx + \frac{4p(p-1)}{(m+p-1)^2} \int_\tau^t \int_{\mathbb{R}^N} \left| (-\Delta)^{\frac{1-s}{2}} \left[u_\lambda^{\frac{m+p-1}{2}} \right](x, s) \right|^2 dxds \leq \int_{\mathbb{R}^N} u_\lambda^p(x, \tau)dx.$$

3. ***(Second energy estimate)** For all $0 < \tau < t < T$ we have*

$$\frac{1}{2} \int_{\mathbb{R}^N} \left| (-\Delta)^{-\frac{s}{2}} u_\lambda(x, t) \right|^2 dx + \int_\tau^t \int_{\mathbb{R}^N} u^{m-1} \left| \nabla (-\Delta)^{-s} u_\lambda(x, s) \right|^2 dxds \leq \frac{1}{2} \int_{\mathbb{R}^N} \left| (-\Delta)^{-\frac{s}{2}} u_\lambda(x, \tau) \right|^2 dx.$$

4. ***(Smoothing effect)** For all $t > 0$,*

$$\|u_\lambda(\cdot, t)\|_{L^\infty(\mathbb{R}^N)} \leq C_{N,s,m}\, t^{-\gamma} \|u_0\|_{L^1(\mathbb{R}^N)}^\delta$$

where $\gamma = \frac{N}{(m-1)N+2(1-s)} > 0$ and $\delta = \frac{2(1-s)}{(m-1)N+2(1-s)} > 0$.

Note that estimates 2 and 3 in Lemma 2 are not uniform in λ up to $\tau = 0$ since the hypothesis $u_0 \in L^1(\mathbb{R}^N)$ is not enough to find a uniform bound for the right hand side term. Note that u_λ only belongs to $L^1(\mathbb{R}^N)$ and $u_\lambda(x, 0)$ will converge to $\|u_0\|_{L^1(\mathbb{R}^N)}\delta_0$ as $\lambda \to \infty$. However, for any $\tau > 0$, the smoothing effect ensures that u_λ is bounded uniformly in λ and then, the right hand side terms of estimates 2 and 3 can be bounded by the terms involving only L^1 norm of u_0. This kind of uniform estimates are very useful and will be given in more details later.

Proof

I) u_λ is a weak solution of (22). Note that given any test function $\phi \in C_c^\infty(\mathbb{R}^N \times (0,T))$, we can define $\psi \in C_c^\infty(\mathbb{R}^N \times (0,\lambda^b T))$ such that $\phi(x,t) = \psi(\lambda x, \lambda^b t)$. Then, the first term in the weak formulation reads

$$\begin{aligned}\int_0^T \int_{\mathbb{R}^N} u_\lambda(x,t)\phi_t(x,t)\,dxdt &= \lambda^N \int_0^T \int_{\mathbb{R}^N} u(\lambda x,\lambda^b t)\phi_t(x,t)\,dxdt \\ &= \lambda^{N+b} \int_0^T \int_{\mathbb{R}^N} u(\lambda x,\lambda^b t)\psi_t(\lambda x,\lambda^b t)\,dxdt \\ &= \int_0^{\lambda^b T} \int_{\mathbb{R}^N} u(y,s)\psi_s(y,s)\,dyds.\end{aligned}$$

The second term is as follows:

$$\begin{aligned}&\int_0^T \int_{\mathbb{R}^N} u_\lambda^{m-1}(x,t)\nabla(-\Delta)^{-s}u_\lambda(x,t)\cdot\nabla\phi(x,t)\,dxdt \\ &= \lambda^{N(m-1)+N} \int_0^T \int_{\mathbb{R}^N} u^{m-1}(\lambda x,\lambda^b t)\nabla(-\Delta)^{-s}[u(\lambda\cdot,\lambda^b t)] \\ &\quad\cdot\nabla\phi(x,t)\,dxdt \\ &= \lambda^{Nm+1-2s} \int_0^T \int_{\mathbb{R}^N} u^{m-1}(\lambda x,\lambda^b t)\nabla(-\Delta)^{-s}u(\lambda x,\lambda^b t) \\ &\quad\cdot\nabla\phi(x,t)\,dxdt \\ &= \lambda^{Nm+2-2s} \int_0^T \int_{\mathbb{R}^N} u^{m-1}(\lambda x,\lambda^b t)\nabla(-\Delta)^{-s}u(\lambda x,\lambda^b t) \\ &\quad\cdot\nabla\psi(\lambda x,\lambda^b t)\,dxdt \\ &= \lambda^{N(m-1)+2-2s-b} \int_0^{\lambda^b T} \int_{\mathbb{R}^N} u^{m-1}(y,s)\nabla(-\Delta)^{-s}u(y,s) \\ &\quad\cdot\nabla\psi(y,s)\,dyds \\ &= \int_0^{\lambda^b T} \int_{\mathbb{R}^N} u^{m-1}(y,s)\nabla(-\Delta)^{-s}u(y,s)\cdot\nabla\psi(y,s)\,dyds.\end{aligned}$$

Finally the initial condition is reformulated as

$$\int_{\mathbb{R}^N} u_\lambda(x,0)\phi(x,0)dx = \lambda^N \int_{\mathbb{R}^N} u_0(\lambda x)\phi(x,0)dx = \int_{\mathbb{R}^N} u_0(y)\psi(y,0)dy,$$

which concludes the proof of (I).

II) u_λ has conservation of mass independent of λ. Since u preserves the mass, we have that

$$\int_{\mathbb{R}^N} u_\lambda(x,t)dx = \lambda^N \int_{\mathbb{R}^N} u(\lambda x, \lambda^b t)dx = \int_{\mathbb{R}^N} u(y, \lambda^b t)dx = \int_{\mathbb{R}^N} u_0(x)dx.$$

III) Energy estimates. The energy estimates are obtained by similar scaling arguments using the energy estimates available for u.

IV) u_λ has smoothing effect uniform in λ. Since u_λ is a weak solution of (M1) we can use the smoothing effect of Theorem 2 together with the result of conservation of mass independent of λ to get

$$\begin{aligned}\|u_\lambda(\cdot,t)\|_{L^\infty(\mathbb{R}^N)} &\le C_{N,s,m}\, t^{-\gamma} \|u_\lambda(x,0)\|^\delta_{L^1(\mathbb{R}^N)} \\ &= C_{N,s,m}\, t^{-\gamma} \|u_0\|^\delta_{L^1(\mathbb{R}^N)} \quad \text{for all } t > 0.\end{aligned}$$

6.3.2 Convergence of the Rescaled Solution as $\lambda \to \infty$

Lemma 3 *Let $m \in (1, +\infty)$, $s \in (0,1)$ and $N \ge 1$. Assume that $u_0 \in L^1(\mathbb{R}^N)$ such that $\|u_0\|_{L^1(\mathbb{R}^N)} = M$. Let also u_λ defined as in Lemma 2. Then, for any $0 < t_1 < t_2 < \infty$, there exists a function $U_M \in L^1(\mathbb{R}^N \times (0,\infty))$ and a subsequence $\{\lambda_j\}_{j=1}^\infty$ such that*

$$u_{\lambda_j} \to U_M \quad \text{as} \quad \lambda_j \to \infty \quad \text{in} \quad L^p(\mathbb{R}^N \times [t_1, t_2]) \quad \text{for} \quad 1 \le p < \infty, \tag{23}$$

where U_M is a weak solution of (M1) with measure initial data $U_M(x,0) = M\delta_0$ and it satisfies the properties 1–5 of Theorem 2. Moreover, in dimension $N = 1$, the full sequence u_λ converge in the sense of (23).

Proof **I. Existence of a Limit** Estimates 1–3 from Lemma 2 are enough to follow the same proof of Theorem 5.2 in [57]. We get that (up to a subsequence),

$$u_{\lambda_j} \to U_M \quad \text{as} \quad \lambda_j \to \infty \quad \text{in} \quad L^2_{\text{loc}}(\mathbb{R}^N \times [t_1, t_2])$$

where U_M is a weak solution of (M1) with initial data $U_M(x,0) = M\delta_0$ with all the desired properties. Now we need to have a uniform control of the tails of the solutions in order to be able to pass from local convergence in L^2 to global convergence in any L^p. Moreover, in dimension $N = 1$, uniqueness of weak solutions ensures that the full sequence u_λ converges.

II. Tail Control Let $\phi \in C^\infty(\mathbb{R}^N)$ be a nondecreasing function such that $\phi(x) = 0$ if $|x| < 1$ and $\phi(x) = 1$ if $|x| > 2$. Now we take $\phi_R(x) := \phi(x/R)$ as test function (after an approximation argument) to get

$$\int_{\mathbb{R}^N} u_\lambda(x,t)\phi_R(x)dx$$
$$\leq \int_{\mathbb{R}^N} u_\lambda(x,0)\phi_R(x)dx$$
$$- \int_{\mathbb{R}^N}\int_0^t u_\lambda^{m-1}(x,t)\nabla(-\Delta)^{-s}u_\lambda(x,t)\cdot\nabla\phi_R(x)dx = I + II.$$

First, we note that since $u_\lambda \geq 0$, then

$$\int_{\mathbb{R}^N} u_\lambda(x,t)\phi_R(x)dx = \int_{|x|>R} u_\lambda(x,t)\phi_R(x)dx \geq \int_{|x|>2R} u_\lambda(x,t)dx.$$

On the other hand, for $\lambda > 1$ we have

$$\int_{\mathbb{R}^N} u_\lambda(x,0)\phi_R(x)dx \leq \lambda^N \int_{|x|>R} u_0(\lambda x)dx = \int_{|y|>\lambda R} u_0(y)dy \leq \int_{|y|>R} u_0(y)dy$$

and the last term clearly goes to zero as $R \to \infty$ since $u_0 \in L^1(\mathbb{R}^N)$. We also have, as in part III in the proof of Theorem 5.2 in [57], that

$$|II| \leq \|\nabla\phi_R\|_\infty \int_0^t \int_{\mathbb{R}^N} u_\lambda^{m-1}(x,t)|\nabla(-\Delta)^{-s}u_\lambda(x,t)|dxdt \leq \Lambda(t)/R$$

where Λ is a locally bounded function in t. Combining the above estimates, we conclude that $\int_{|x|>2R} u_\lambda(x,\tau)dx \to 0$ as $R \to \infty$ for all $\tau \in (0,t)$. Passing to the limit, the same estimate is inherited by U_M. A similar tail control argument has been used by one of the authors in [38] for a fractional diffusion-convection equation.

III. Convergence in $L^p(\mathbb{R}^N \times [t_1,t_2])$. First, we prove L^1 convergence. From step II in this proof, for any $\epsilon > 0$ we can choose R large enough such that

$$I_{B_R^c} = \int_{t_1}^{t_2}\int_{B_R^c} |u_{\lambda_j}(x,t) - U_M(x,t)|dxdt$$
$$= \int_{t_1}^{t_2}\int_{B_R^c} |u_{\lambda_j}(x,t)|dxdt + \int_{t_1}^{t_2}\int_{B_R^c} |U_M(x,t)|dxdt < \epsilon/2.$$

On the other hand,

$$I_{B_R} = \int_{t_1}^{t_2}\int_{B_R} |u_{\lambda_j}(x,t) - U_M(x,t)|dxdt$$
$$\leq |B_R|(t_2-t_1)\left(\int_{t_1}^{t_2}\int_{B_R} |u_{\lambda_j}(x,t) - U_M(x,t)|^2 dxdt\right)^{\frac{1}{2}}.$$

Since $u_{\lambda_j} \to U_M$ as $\lambda_j \to \infty$ in $L^2_{\text{loc}}(\mathbb{R}^N \times [t_1, t_2])$, we can now choose λ_j big enough such that $I_{B_R} < \epsilon/2$. In this way,

$$\|u_{\lambda_j} - U_M\|_{L^1(\mathbb{R}^N \times [t_1,t_2])} \le I_{B_R} + I_{B_R^c} < \epsilon$$

which concludes the proof of L^1 convergence. By the smoothing effect, both u_λ and U_M are uniformly bounded outside $t = 0$, for any $R > 0$ we have that

$$\begin{aligned}\|u_{\lambda_j} - U_M\|^p_{L^p(\mathbb{R}^N \times [t_1,t_2])} &\le \|u_{\lambda_j} - U_M\|^{p-1}_{L^\infty(\mathbb{R}^N \times [t_1,t_2])} \|u_{\lambda_j} - U_M\|_{L^1(\mathbb{R}^N \times [t_1,t_2])} \\ &\le C \|u_{\lambda_j} - U_M\|_{L^1(\mathbb{R}^N \times [t_1,t_2])}\end{aligned}$$

which again converges to 0 as $\lambda_j \to \infty$.

6.3.3 Self-similarity of the Limit Solution

Lemma 4 (Existence of Self-similar Solution) *Let $m \in (1, +\infty)$, $s \in (0, 1)$, $N = 1$. The solution U_M constructed in Lemma 3 is a selfsimilar solution of the form*

$$U_M(x, t) = t^{-\alpha} \phi(x t^{-\beta}),$$

for a certain function $\phi \in \mathbb{R}^N \to \mathbb{R}$ and $\alpha = N\beta$ with $\beta = 1/(N(m-1)+2-2s)$.

Proof First we note that for $b = N(m-1) + 2 - 2s$, U_M is invariant under the following scaling

$$\begin{aligned}\lambda_0^N U_M(\lambda_0 x, \lambda_0^b t) &= \lim_{\lambda \to \infty} \lambda_0^N u_\lambda(\lambda_0 x, \lambda_0^b t) = \lim_{\lambda \to \infty} (\lambda_0 \lambda)^N u(\lambda_0 \lambda x, (\lambda_0 \lambda)^b t) \\ &= \lim_{\lambda \lambda_0 \to \infty} u_{\lambda_0 \lambda}(x, t) = U_M(x, t). \end{aligned} \tag{24}$$

Thus, since $b\beta = 1$ and then, we choose $\lambda_0 = t^{-\frac{1}{b}} = t^{-\beta}$ to get

$$\begin{aligned}U_M(x, t) &= \lambda_0^N U_M(\lambda_0 x, \lambda_0^b t) = t^{-\frac{N}{b}} U_M(x t^{-\frac{1}{b}}, 1) \\ &= t^{-\frac{N}{b}} U_M(x t^{-\frac{1}{b}}, 1) = t^{-\alpha} U_M(x t^{-\beta}, 1).\end{aligned}$$

Remark 7 Note that, in identity (24), the fact the full sequence u_λ converges plays a crucial role. If we do not have this property, we cannot ensure that the sequence in $\{\lambda_0 \lambda_j\}_{j=1}^\infty$ gives a convergent $u_{\lambda_0 \lambda_j}$.

6.3.4 Proof of Theorem 6

We have that

$$\|u_\lambda - U_M\|_{L^p(\mathbb{R}^N \times [t_1,t_2])} \to 0 \quad \text{as} \quad \lambda \to \infty,$$

which in particular implies

$$\|u_\lambda(\cdot,t) - U_M(\cdot,t)\|_{L^p(\mathbb{R}^N)} \to 0 \quad \text{as} \quad \lambda \to \infty \quad \text{for a.e} \quad t \in [t_1,t_2].$$

Without loos of generality, assume the above limit holds for $t = 1$. Then, choosing $\tau = \lambda^b = \lambda^{1/\beta}$,

$$\begin{aligned}\|u_\lambda(\cdot,1) - U_M(\cdot,1)\|_{L^p(\mathbb{R}^N)} &= \lambda^N \|u(\lambda\cdot,\lambda^b) - U_M(\lambda\cdot,\lambda^b)\|_{L^p(\mathbb{R}^N)} \\ &= \lambda^{N-\frac{N}{p}} \|u(\cdot,\lambda^b) - U_M(\cdot,\lambda^b)\|_{L^p(\mathbb{R}^N)} \\ &= \tau^{N(1-\frac{1}{p})\beta} \|u(\cdot,\tau) - U_M(\cdot,\tau)\|_{L^p(\mathbb{R}^N)}.\end{aligned}$$

Since $\tau \to \infty$ as $\lambda \to \infty$, we conclude that

$$\tau^{\frac{N(1-\frac{1}{p})}{(m-1)N+2-2s}} \|u(\cdot,\tau) - U_M(\cdot,\tau)\|_{L^p(\mathbb{R}^N)} \to 0 \quad \text{as} \quad \tau \to \infty.$$

7 Comments and Open Problems

- As a summary, we establish the theory of existence of suitable weak solutions of problem (M1) and settle the question of finite vs infinite speed of propagation depending on whether $m \geq 2$ or $m < 2$. We also settle the asymptotic behavior in one dimension by means of an integrated version of the problem.
- The questions of uniqueness in several dimensions are widely open and ought to be addressed. Once this result is available, the existence of selfsimilar solutions together with the asymptotic behaviour would follow with the techniques showed in this paper.

  If $m \in (1,2)$, $N \geq 1$ we have uniqueness by the 1-to-1 correspondence of self-similar solutions between (M1) and (FPME), and the last ones are known to be unique. However, we need a uniqueness result regarding general initial data in order to be able to prove the asymptotic behaviour in dimension higher than one.

  If $m \in (1,\infty)$ and $N = 1$, the solutions (not only self-similar ones) are unique since there exists a 1-to-1 correspondence with viscosity solutions of the integrated problem, which are known to be unique.
- Another pending issue is continuity of weak solutions. In the case $m = 2$ Hölder continuity is proved in [17, 20].

- Recently, the problem posed in a bounded domain was considered in [48] for dimension $N \geq 1$. Further work is to be done on that issue.
- Satisfying numerical experiments have been performed, see [57] for some numerical experiments using ideas of [28]. A systematic and rigorous numerical analysis is still pending.

Acknowledgements This work was partially supported by Spanish Project MTM2014-52240-P. D. Stan is partially supported by the MEC-Juan de la Cierva postdoctoral fellowship number FJCI-2015-25797, by the ERCEA Advanced Grant 2014 669689—HADE, by the MINECO project MTM2014-53850-P, by Basque Government project IT-641-13 and also by the Basque Government through the BERC 2014–2017 program, and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323. F.d.Teso is partially supported by the Toppforsk (research excellence) project Waves and Nonlinear Phenomena (WaNP), grant no. 250070 from the Research Council of Norway and by the ERCIM "Alain Bensoussan" Fellowship programme.

References

1. G.I. Barenblatt, On some unsteady motions of a liquid or a gas in a porous medium. Prikl. Mat. Mekh. **16**(1), 67–78 (1952) (in Russian)
2. P. Bénilan, Equations d'évolution dans un espace de Banach quelconque et applications, Ph.D. Thesis, University of Orsay, 1972 (in French)
3. Ph. Bénilan, H. Brezis, M.G. Crandall, A semilinear equation in $L^1(R^N)$. Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **2**, 523–555 (1975)
4. P. Biler, G. Karch, R. Monneau, Nonlinear diffusion of dislocation density and self-similar solutions. Commun. Math. Phys. **294**, 145–168 (2010)
5. P. Biler, C. Imbert, G. Karch, Barenblatt profiles for a nonlocal porous medium equation. C. R. Math. Acad. Sci. Paris. **349**, 641–645 (2011)
6. P. Biler, C. Imbert, G. Karch. The nonlocal porous medium equation: barenblatt profiles and other weak solutions. Arch. Ration. Mech. Anal. **215**, 497–529 (2015)
7. M. Bonforte, A. Segatti, J.L. Vázquez, Non-existence and instantaneous extinction of solutions for singular nonlinear fractional diffusion equations. Calc. Var. PDEs **55**, 55–68 (2016)
8. M. Bonforte, Y. Sire, J.L. Vázquez, Optimal existence and uniqueness theory for the fractional heat equation. Nonlinear Anal. **153**, 142–168 (2017)
9. H. Brezis, Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations, in *Proceedings of Symposium Nonlinear Functional Analysis, Madison (1971)*, Contribution to Nonlinear Functional Analysis (Academic, New York, 1971), pp. 101–156
10. H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert* (North-Holland, Amsterdam, 1973)
11. H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext (Springer, New York, 2011)
12. L.A. Caffarelli, A. Friedman, Continuity of the density of a gas flow in a porous medium. Trans. Am. Math. Soc. **252**, 99–113 (1979)
13. L.A. Caffarelli, A. Friedman, Regularity of the free boundary of a gas flow in an n-dimensional porous medium. Indiana Univ. Math. J. **29**, 361–391 (1980)
14. L.A. Caffarelli, L.C. Evans, Continuity of the temperature in the two-phase Stefan problem. Arch. Ration. Mech. Anal. **81**(3), 199–220 (1983)
15. L.A. Caffarelli, S. Salsa, *A Geometric Approach to Free Boundary Problems* (American Mathematical Society, Providence, 2005)

16. L.A. Caffarelli, J.L. Vázquez, Nonlinear porous medium flow with fractional potential pressure. Arch. Ration. Mech. Anal. **202**(2), 537–565 (2011)
17. L.A. Caffarelli, J.L. Vázquez, Asymptotic behaviour of a porous medium equation with fractional diffusion. Discrete Contin. Dyn. Syst. A **29**(4), 1393–1404 (2011)
18. L.A. Caffarelli, J.L. Vázquez, Regularity of solutions of the fractional porous medium flow with exponent 1/2. St. Petersburg Math. J. **27**(3), 437–460 (2016)
19. L.A. Caffarelli, J.L. Vázquez, N.I. Wolanski, Lipschitz continuity of solutions and interfaces of the N-dimensional porous medium equation. Indiana Univ. Math. J. **36**, 373–401 (1987)
20. L.A. Caffarelli, F. Soria, J.L. Vázquez, Regularity of solutions of the fractional porous medium flow. J. Eur. Math. Soc. **15**(5), 1701–1746 (2013)
21. J.A. Carrillo, Y. Huang, M.C. Santos, J.L. Vázquez, Exponential convergence towards stationary states for the 1D porous medium equation with fractional pressure. J. Differ. Equ. **258**, 736–763 (2015)
22. E. Chasseigne, E.R. Jakobsen, On nonlocal quasilinear equations and their local limits. J. Differ. Equ. **262**(6), 3759–3804 (2017)
23. M.G. Crandall, T.M. Liggett, Generation of semi-groups of nonlinear transformations on general Banach spaces. Am. J. Math. **93**, 265–298 (1971)
24. A. de Pablo, F. Quirós, A. Rodríguez, J.L. Vázquez, A fractional porous medium equation. Adv. Math. **226**(2), 1378–1409 (2011)
25. A. de Pablo, F. Quirós, A. Rodríguez, J.L. Vázquez, A general fractional porous medium equation. Comm. Pure Appl. Math. **65**(9), 1242–1284 (2012)
26. A. de Pablo, F. Quirós, A. Rodríguez, J.L. Vázquez. Classical solutions for a logarithmic fractional diffusion equation. J. Math. Pures Appl. (9) **101**(6), 901–924 (2014)
27. M. Del Pino, J. Dolbeault, Best constants for Gagliardo-Nirenberg inequalities and applications to nonlinear diffusions. J. Math. Pures Appl. **81**(9), 847–875 (2002)
28. F. del Teso, Finite difference method for a fractional porous medium equation. Calcolo **51**(4), 615–638 (2014)
29. E. DiBenedetto, *Degenerate Parabolic Equations* (Springer, Berlin, 1993)
30. J. Dolbeault, A. Zhang, Flows and functional inequalities for fractional operators. Appl. Anal. **96**(9), 1547–1560 (2017)
31. L.C. Evans, Applications of nonlinear semigroup theory to certain partial differential equations, in *Nonlinear Evolution Equations*, ed. by M.G. Crandall (Academic, New York, 1978), pp. 163–188
32. G. Giacomin, J.L. Lebowitz, Phase segregation dynamics in particle systems with long range interaction I. Macroscopic limits. J. Stat. Phys. **87**, 37–61 (1997)
33. G. Giacomin, J.L. Lebowitz, Phase segregation dynamics in particle systems with long range interaction II. Interface motion. SIAM J. Appl. Math. **58**, 1707–29 (1998)
34. G. Giacomin, J.L. Lebowitz, R. Marra., Macroscopic evolution of particle systems with short and long-range interactions. Nonlinearity **13**(6), 2143–2162 (2000)
35. J.A. Goldstein, *Semigroups of Linear Operators and Applications*. Oxford Mathematical Monographs (The Clarendon Press, Oxford University Press, New York, 1985)
36. A.K. Head., Dislocation group dynamics II. Similarity solutions of the continuum approximation. Phil. Mag. **26**, 65–72 (1972)
37. Y. Huang, Explicit Barenblatt profiles for fractional porous medium equations. Bull. Lond. Math. Soc. **46**, 857–869 (2014)
38. L. Ignat, D. Stan, Asymptotic behavior of solutions to fractional diffusion convection equations. J. Lond. Math. Soc. **97**(2), 258–281 (2018). https://doi.org/10.1112/jlms.12110
39. C. Imbert, Finite speed of propagation for a non-local porous medium equation. Colloq. Math. **143**(2), 149–157 (2016)
40. S. Kamenomostskaya (Kamin), On the Stefan problem. Mat. Sbornik **53**, 489–514 (1961)
41. C. Kienzler, H. Koch, J.L. Vázquez, Flatness implies smoothness for solutions of the porous medium equation. Calc. Var. **57**(1), 18 (2018)

42. H. Koch, Non-Euclidean singular integrals and the porous medium equation, University of Heidelberg, Habilitation Thesis, 1999. http://www.iwr.uniheidelberg.de/groups/amj/koch.html
43. O.A. Ladyzhenskaya, *Attractors for Semigroups and Evolution Equations. Lezioni Lincee*. Lincei Lectures (Cambridge University Press, Cambridge, 1991)
44. J.-L. Lions, E. Magenes, *Problèmes aux limites non homogènes et applications* (French). Travaux et Recherches Mathématiques, No. 17, 18, 20, vols. 1, 2, 3 (Dunod, Paris, 1968–1970) (2017)
45. S. Lisini, E. Mainini, A. Segatti, A gradient flow approach to the porous medium equation with fractional pressure. Arch. Ration. Mech. Anal. **227**, 567 (2018). https://doi.org/10.1007/s00205-017-1168-2
46. L. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*. Modern Birkhäuser Classics (Birkhäuser/Springer Basel AG, Basel, 1995)
47. A.M. Meirmanov, *The Stefan Problem*. de Gruyter Expositions in Mathematics, 3 (Walter de Gruyter & Co., Berlin, 1992) (translated from the Russian)
48. Q.H. Nguyen, J.L. Vázquez, Porous medium equation with nonlocal pressure in a bounded domain. Preprint (2017). arXiv:1708.00660
49. L. Nirenberg. Ordinary differential equations in Banach spaces. (Original in Italian, C.I.M.E., 1963). Reprinted in "Abstract differential equations", 123–170. C.I.M.E. Summer School 29 (Springer, Heidelberg, 2011)***
50. O.A. Oleinik, A.S. Kalashnikov, Y.-I. Chzou, The Cauchy problem and boundary problems for equations of the type of unsteady filtration. Izv. Akad. Nauk SSR Ser. Math. **22**, 667–704 (1958)
51. A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Applied Mathematical Sciences, vol. 44 (Springer, New York, 1983)
52. J.M. Rakotoson, R. Temam, An optimal compactness theorem and application to elliptic-parabolic systems. Appl. Math. Lett. **14**(3), 303–306 (2001)
53. J. Simon, Compact sets in the space $L^p(0, T; B)$. Ann. Mat. Pura Appl. **146**, 65–96 (1987)
54. D. Stan, F. del Teso, J.L. Vázquez. Finite and infinite speed of propagation for porous medium equations with fractional pressure. C. R. Math. Acad. Sci. Paris **352**(2), 123–128 (2014)
55. D. Stan, F. del Teso, J.L. Vázquez, Transformations of self-similar solutions for porous medium equations of fractional type. Nonlinear Anal. **119**, 62–73 (2015)
56. D. Stan, F. del Teso, J.L. Vázquez, Finite and infinite speed of propagation for porous medium equations with nonlocal pressure. J. Differ. Equ. **260**(2), 1154–1199 (2016)
57. D. Stan, F. del Teso, J.L. Vázquez, Existence of weak solutions for a general porous medium equation with nonlocal pressure (2017). arXiv:1609.05139
58. J.L. Vázquez, *Smoothing and Decay Estimates for Nonlinear Diffusion Equations*. Oxford Lecture Series in Mathematics and Its Applications, vol. 33 (Oxford University Press, Oxford, 2006)
59. J.L. Vázquez, *The Porous Medium Equation. Mathematical Theory*. Oxford Mathematical Monographs (Oxford University Press, Oxford, 2007)
60. J.L. Vázquez, Barenblatt solutions and asymptotic behaviour for a nonlinear fractional heat equation of porous medium type. J. Eur. Math. Soc. (JEMS) **16**(4), 769–803 (2014)
61. J.L. Vázquez, Recent progress in the theory of nonlinear diffusion with fractional Laplacian operators, nonlinear elliptic and parabolic differential equations. Discrete Contin. Dyn. Syst. **S7**(4), 857–885 (2014)
62. J.L. Vázquez, The mathematical theories of diffusion: nonlinear and fractional diffusion, in *Nonlocal and Nonlinear Diffusions and Interactions: New Methods and Directions*, ed. by M. Bonforte, G. Grillo. Lecture Notes in Mathematics, vol. 2186. Fond. CIME/CIME Found. Subser. (Springer, Cham, 2017), pp. 205–278
63. J.L. Vázquez, Asymptotic behaviour for the fractional heat equation in the Euclidean space, in *Complex Variables and Elliptic Equations. Special Volume in Honor of Vladimir I. Smirnov's 130th Anniversary*. Published online in November 2017

64. J.L. Vázquez, B. Volzone, Optimal estimates for fractional fast diffusion equations. J. Math. Pures Appl. (9) **103**(2), 535–556 (2015)
65. J.L. Vázquez, A. de Pablo, F. Quirós, A. Rodríguez, Classical solutions and higher regularity for nonlinear fractional diffusion equations. J. Eur. Math. Soc. **19**(7), 1949–1975 (2017)
66. K. Yosida, *Functional Analysis*. Die Grundlehren der Mathematischen Wissenschaften (Springer, Berlin, 1965)
67. X. Zhou, W. Xiao, J. Chen. Fractional porous medium and mean field equations in Besov spaces. Electron. J. Differ. Equ. **199**, 14 (2014)

Nondiscrete Lassonde-Revalski Principle and Dependent Choice



Mihai Turinici

1 Introduction

Let (M, d) be a complete metric space; and $\varphi \in \mathscr{F}(M, R \cup \{\infty\})$ be (M, d)-*regular*; that is,

(a01) φ is proper ($\mathrm{Dom}(\varphi) := \{x \in M; \varphi(x) < \infty\} \neq \emptyset$)
(a02) φ is bounded below ($\inf \varphi(M) > -\infty$)
(a03) φ is d-lsc ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \stackrel{d}{\longrightarrow} x$).

[Here, for each couple A, B of nonempty sets, $\mathscr{F}(A, B)$ denotes the class of all functions from A to B; when $A = B$, we write $\mathscr{F}(A)$ in place of $\mathscr{F}(A, A)$]. The following 1974 statement in Ekeland [15] (referred to as Ekeland's variational principle; in short: EVP) is our starting point.

Theorem 1 *Let the precise conditions hold. Then, for each element $u \in \mathrm{Dom}(\varphi)$, there exists some other element $v = v(u) \in \mathrm{Dom}(\varphi)$, with*

(11-a) $d(u, v) \leq \varphi(u) - \varphi(v)$ *(hence* $\varphi(u) \geq \varphi(v)$*)*
(11-b) $d(v, x) > \varphi(v) - \varphi(x)$, *for all* $x \in M \setminus \{v\}$.

This principle found some basic applications to control and optimization, generalized differential calculus, critical point theory, and global analysis; see the 1979 Ekeland's paper [16] for a survey of these. As a consequence, many extensions of (EVP) were proposed. For example, the *(abstract) order* one starts from the fact that, with respect to the Brøndsted quasi-order [7]

(a04) $(x, y \in M)$: $x \leq y$ iff $d(x, y) + \varphi(y) \leq \varphi(x)$,

M. Turinici (✉)
"A. Myller" Mathematical Seminar, "A. I. Cuza" University, Iaşi, Romania
e-mail: mturi@uaic.ro

© Springer International Publishing AG, part of Springer Nature 2018

T. M. Rassias (ed.), *Current Research in Nonlinear Analysis*, Springer Optimization and Its Applications 135, https://doi.org/10.1007/978-3-319-89800-1_13

the element $v \in M$ appearing in the second conclusion above is *maximal*; so that, (EVP) is nothing but a denumerable variant of the Zorn-Bourbaki maximality principle; its precise formulation is just the 1976 ordering principle due to Brezis and Browder [6] (in short: BB). Further, the *dimensional* way of extension refers to the ambient space (R) of $\varphi(M)$ being substituted by a (topological or not) vector space; an account of the results in this area is to be found in the 2003 monograph by Goepfert et al. [17, Ch 3]. Finally, the *metrical* one consists in conditions imposed to the ambient metric over M being relaxed; the basic result in this direction was obtained in 1996 by Kada et al. [19].

Now, the natural question arising here is that of all these extensions for (EVP) (including (BB)) being or not effective. Some partial answers were stated in Turinici [31]; see also Bao and Khanh [1]. According to these, the dimensional and metrical extensions of (EVP) are obtainable from either (EVP) or (BB), via straightforward techniques. Concerning the question of (BB) (and its subsequent extensions) being reducible to (EVP), the basic tool for solving it is the Dependent Choice Principle (in short: DC) due—independently—to Bernays [2] and Tarski [28]. Precisely, note that, by the developments in Cârjă et al. [9, Ch 2, Sect 2.1]

$$(\mathrm{DC}) \Longrightarrow (\mathrm{BB}) \Longrightarrow (\mathrm{EVP});$$

and (cf. Brunner [8]), (EVP) $\Longrightarrow$ (DC). Hence, any maximal/variational result—(MP) say—with

$$\mathrm{DC} \Longrightarrow \mathrm{MP} \Longrightarrow \mathrm{EVP}$$

is logically equivalent with both (DC) and (EVP); see Turinici [32] for details. Note that, this is the case with many sequential extensions of (EVP) and/or (BB); in particular (cf. Turinici [33]), the conclusion is retainable for a technical 1987 generalization of (EVP) referred to as the Smooth Borwein-Preiss Variational Principle [4] (in short: BP). Now—especially from a practical perspective—this last principle found a lot of applications in Variational Analysis; so that, it was the subject of numerous extensions. For example, a basic contribution in the area is the 1993 Generic Smooth Variational Principle due to Deville, Godefroy and Zizler [13] (in short: DGZ). Further, in their 2005 paper, Lassonde and Revalski [21] established that (DGZ)—and many other statements of this type—may be viewed as a particular case of some fragmentability result involving sequences of multivalued maps (referred to as: Lassonde-Revalski Principle; in short: LR). Taking the preceding observations into account, it is legitimate to ask whether the obtained results are genuine (logical) extensions of (EVP). As we shall see, the answer to this is *negative*, in the larger sense: the Semi-Bump Deville-Godefroy-Zizler Generic Smooth Variational Principle (DGZ-sb) (including (DGZ)) and Nondiscrete Lassonde-Revalski Principle; (LR-ndiscr) (including (LR)) enter in the "logical" segment between (DC) and (EVP); i.e.

$$(\mathrm{DC}) \Longrightarrow (\text{LR-ndiscr}) \Longrightarrow (\text{DGZ-sb}) \Longrightarrow (\mathrm{EVP}) \Longrightarrow (\mathrm{DC});$$

whence: both (LR-ndiscr) and (DGZ-sb) are equivalent with (DC)—or, equivalently: with (EVP). The steps of this investigation process are to be described as follows. In Sect. 2, we show that (DC) is equivalent with Baire Category Theorem over complete metric spaces (in short: BC); hence, in particular, (BC) includes the Denumerable Axiom of Choice (in short: AC(N)). Further, in Sects. 3 and 4, the obtained facts are used to get the inclusion chain

$$\text{(DC) implies (LR-ndiscr)} \implies \text{(LR-ndiscr-m)} \implies \text{(LR-VP-m)};$$

here, (LR-ndiscr-m) is the *metrical* version of (LR-ndiscr) and (LR-VP-m) denotes its associated *variational principle*. The objective of Sect. 5 is to establish that

$$\text{(LR-VP-m)} \implies \text{(DGZ); hence (by the above) (BC)} \implies \text{(DGZ)};$$

an explicit proof of this last inclusion is also provided, for completeness reasons. Finally, in Sects. 6 and 7, we prove that (DGZ) $\implies$ (EVP) and (EVP) $\implies$ (DC), respectively; so, we close the circle between all these. Further aspects, involving the 2000 Pseudometric Variational Principle in Li and Shi [22] or the 2001 Perturbed Minimization Principle due to Deville and Ghoussoub [12] will be discussed elsewhere.

2 (DC) $\Longleftrightarrow$ (BC)

Throughout this exposition, the axiomatic system to be used is Zermelo-Fraenkel's (abbreviated: ZF), as described in Cohen [10, Ch 2, Sect 3]. The notations and basic facts about these are standard. Some important ones are given below.

(A) Let X be a nonempty set. By a *relation* over it, we mean any (nonempty) part $\mathscr{R}$ of $X \times X$; in this case, $(X, \mathscr{R})$ is called a *relational structure*. As usually, we may regard $\mathscr{R}$ as a mapping from X to 2^X (=the class of all subsets in X). Precisely, for each $x \in X$, denote

$$X(x, \mathscr{R}) = \{y \in X;\ x\mathscr{R}y\} \text{ (the } \textit{section} \text{ of } \mathscr{R} \text{ through } x);$$

then, the mapping in question is

$$\mathscr{R}(x) = X(x, \mathscr{R}),\ x \in X.$$

Call $\mathscr{R}$, *proper* when $\mathscr{R}(x) \neq \emptyset$, for all $x \in X$; note that, in such a case, $\mathscr{R}$ appears as a mapping between X and $(2)^X$ (=the class of all nonempty parts in X). This will be also referred to as: $(X, \mathscr{R})$ is a *proper* relational structure; note that a basic example of such object is

$$\mathscr{I} = \{(x, x);\ x \in X\} \text{ (the } \textit{identity} \text{ relation).}$$

Given the relations $\mathscr{R}$, $\mathscr{S}$ over X, define their *product* $\mathscr{R} \circ \mathscr{S}$ as

$(x, z) \in \mathscr{R} \circ \mathscr{S}$, if there exists $y \in X$ with $(x, y) \in \mathscr{R}$, $(y, z) \in \mathscr{S}$.

Also, for each relation $\mathscr{R}$ over X, denote

$\mathscr{R}^{-1} = \{(x, y) \in X \times X;\ (y, x) \in \mathscr{R}\}$ (the *inverse* of $\mathscr{R}$).

Finally, given the relations $\mathscr{R}$ and $\mathscr{S}$ on X, let us say that $\mathscr{R}$ is *coarser* than $\mathscr{S}$ (or, equivalently: $\mathscr{S}$ is *finer* than $\mathscr{R}$), provided

$\mathscr{R} \subseteq \mathscr{S}$; i.e.: $x \mathscr{R} y$ implies $x \mathscr{S} y$.

Given a relation $\mathscr{R}$ on X, the following properties are to be discussed here:

(P1) $\mathscr{R}$ is *reflexive*: $\mathscr{I} \subseteq \mathscr{R}$
(P2) $\mathscr{R}$ is *irreflexive*: $\mathscr{R} \cap \mathscr{I} = \emptyset$
(P3) $\mathscr{R}$ is *transitive*: $\mathscr{R} \circ \mathscr{R} \subseteq \mathscr{R}$
(P4) $\mathscr{R}$ is *symmetric*: $\mathscr{R}^{-1} = \mathscr{R}$
(P5) $\mathscr{R}$ is *antisymmetric*: $\mathscr{R}^{-1} \cap \mathscr{R} \subseteq \mathscr{I}$.

This yields the classes of relations to be used; the following ones are important for our developments:

(C0) $\mathscr{R}$ is *trivial* (i.e.: $\mathscr{R} = X \times X$)
(C1) $\mathscr{R}$ is a *(partial) order* (reflexive, transitive, antisymmetric)
(C2) $\mathscr{R}$ is a *strict order* (irreflexive and transitive)
(C3) $\mathscr{R}$ is a *quasi-order* (reflexive and transitive)
(C4) $\mathscr{R}$ is an *equivalence* (reflexive, transitive, symmetric).

A basic ordered structure is $(N, \leq)$; here, $N = \{0, 1, \ldots\}$ is the set of *natural numbers* and $(\leq)$ is defined as

$m \leq n$ iff $m + p = n$, for some $p \in N$.

For each $n \in N(1, \leq)$, let $N(n, >) := \{0, \ldots, n-1\}$ stand for the *initial interval* (in N) induced by n. Any set P with $P \sim N$ (in the sense: there exists a bijection from P to N) will be referred to as *effectively denumerable*; also denoted as: $\mathrm{card}(P) = \aleph_0$. In addition, given some natural number $n \geq 1$, any set Q with $Q \sim N(n, >)$ will be said to be *n-finite*; and we write this: $\mathrm{card}(Q) = n$; when n is generic here, we say that Q is *finite*. Finally, the (nonempty) set Y is called (at most) *denumerable* iff it is either effectively denumerable or finite.

Let X be a nonempty set. By a *sequence* in X, we mean any mapping $x : N \to X$. For simplicity reasons, it will be denoted as $(x(n);\ n \geq 0)$, or $(x_n;\ n \geq 0)$; moreover, when no confusion can arise, we further simplify this notation as $(x(n))$ or (x_n), respectively. Given such an object (x_n), any sequence $(y_n := x_{i(n)};\ n \geq 0)$ with

$(i(n);\ n \geq 0)$ is *divergent* $(i(n) \to \infty$ as $n \to \infty)$

will be referred to as a *subsequence* of $(x_n;\ n \geq 0)$.

Remember that, an outstanding part of (ZF) is the *Axiom of Choice* (abbreviated: AC); which, in a convenient manner, may be written as

(AC) For each nonempty set M, there exists a (selective) function $f : (2)^M \to M$ (in the sense: $f(Y) \in Y$, $\forall Y \in (2)^M$).

There are many logical equivalents of (AC); see, for instance, Moore [24, Appendix 2]. Sometimes, when the ambient set M is endowed with denumerable type structures, this general choice procedure may be replaced with a weaker form of it, called: *Dependent Choice Principle*. Some preliminaries are needed. Let M be a nonempty set. For each natural number $k \geq 1$, call the map $F : N(k, >) \to M$, a *k-sequence*; if $k \geq 1$ is generic, we talk about a *finite* sequence. The following result, referred to as the *Finite Dependent Choice property* (in short: (DC-fin)) is available in the *strongly reduced* Zermelo-Fraenkel system (ZF-AC). Given $a \in M$, let us say that the k-sequence $F : N(k, >) \to M$ (where $k \geq 2$) is $(a, \mathscr{R})$-*iterative* provided

$$F(0) = a \text{ and } F(i+1) \in \mathscr{R}(F(i)), \text{ for all } i \in N(k-1, >).$$

Lemma 1 *Let the relational structure $(M, \mathscr{R})$ be proper. Then, for each $k \geq 2$, the following property holds:*

(P(k)) for each $a \in M$, there exists an $(a, \mathscr{R})$-iterative k-sequence.

Proof Clearly, $(P(2))$ is true; just take $b \in \mathscr{R}(a)$ and define $F : N(2, >) \to M$ as: $F(0) = a$, $F(1) = b$. Assume that (for some $k \geq 2$),

$(P(h))$ is true, for $2 \leq h \leq k$;

we claim that $(P(k+1))$ is also true. For, let $F : N(k, >) \to M$ be an $(a, \mathscr{R})$-iterative k-sequence, assured by hypothesis. As $\mathscr{R}$ is proper, $\mathscr{R}(F(k-1))$ is nonempty; let u be some element of it. The map $G : N(k+1, >) \to M$ introduced as

$$G(i) = F(i), i \in N(k, >); G(k) = u$$

is an $(a, \mathscr{R})$-iterative $(k+1)$-sequence; and then, we are done.

Now, it is natural to see what happens when k "tends to infinity". At a first glance, the following *Dependent Choice Principle* (in short: DC) is obtainable in (ZF-AC) from this "limit" process. Given $a \in M$, let us say that the sequence $(x_n; n \geq 0)$ in M is $(a; \mathscr{R})$-*iterative*, provided

$$x_0 = a; x_{n+1} \in \mathscr{R}(x_n), \forall n.$$

Proposition 1 *Let the relational structure $(M, \mathscr{R})$ be proper. Then, for each $a \in M$ there exists at least one $(a, \mathscr{R})$-iterative sequence in M.*

Formally, the "argument" involved here consists in the possibility of constructing an (infinite) sequence via finite sequences. This, ultimately, cannot be done under the precise context [so, the limit process in question does not work in (ZF-AC)]; whence, (DC) is not obtainable from the axioms of our strongly reduced system. On the other hand, this principle—proposed, independently, by Bernays [2] and Tarski [28]—is deductible from (AC), but not conversely; cf. Wolk [35]. Moreover, by the developments in Moskhovakis [25, Ch 8], and Schechter [27, Ch 6], the *reduced system* (ZF-AC+DC) it large enough so as to cover the "usual" mathematics; see also Moore [24, Appendix 2, Table 4].

(B) Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of relations on M. Given $a \in M$, let us say that the sequence $(x_n; n \geq 0)$ in M is $(a; (\mathscr{R}_n; n \geq 0))$-*iterative*, provided

$$x_0 = a; x_{n+1} \in \mathscr{R}_n(x_n), \forall n.$$

The following "Diagonal Dependent Choice" principle (in short: DDC) is to be taken into consideration.

Proposition 2 *Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of proper relations on M. Then, for each $a \in M$ there exists at least one $(a; (\mathscr{R}_n; n \geq 0))$-iterative sequence in M.*

Clearly, (DDC) includes (DC); to which it reduces when $(\mathscr{R}_n; n \geq 0)$ is constant. The reciprocal of this is also true. In fact, letting the premises of (DDC) hold, put $P = N \times M$; and let $\mathscr{S}$ be the relation over P introduced as

$$\mathscr{S}(i, x) = \{i + 1\} \times \mathscr{R}_i(x), \ (i, x) \in P.$$

It will suffice applying (DC) to $(P, \mathscr{S})$ and $b := (0, a) \in P$ to get the conclusion in our statement; we do not give details.

Summing up, (DDC) is provable in (ZF-AC+DC). This is valid as well for its variant, referred to as: the "Selected Dependent Choice" principle (in short: SDC).

Proposition 3 *Let the map $F : N \to (2)^M$ and the relation $\mathscr{R}$ over M fulfill*

$$(\forall n \geq 0): \mathscr{R}(x) \cap F(n + 1) \neq \emptyset, \ \forall x \in F(n).$$

Then, for each $a \in F(0)$ there exists a sequence $(x(n); n \geq 0)$ in M with

$$x(0) = a, \ x(n) \in F(n), \ x(n + 1) \in \mathscr{R}(x(n)), \ \forall n.$$

As before, (SDC) $\Longrightarrow$ (DC) ($\Longleftrightarrow$ (DDC)); just take $F(n) = M$, $n \geq 0$. But, the reciprocal is also true, in the sense: (DDC) $\Longrightarrow$ (SDC). This follows from

Proof (Proposition 3) Let the premises of (SDC) be true. Define a sequence of relations $(\mathscr{R}_n; n \geq 0)$ over M as: for each $n \geq 0$,

$$\mathscr{R}_n(x) = \mathscr{R}(x) \cap F(n + 1), \text{ if } x \in F(n),$$
$$\mathscr{R}_n(x) = \{x\}, \text{ otherwise } (x \in M \setminus F(n)).$$

Clearly, $\mathscr{R}_n$ is proper, for all $n \geq 0$. So, by (DDC), it follows that, given the starting $a \in F(0)$, there exists an $(a, (\mathscr{R}_n; n \geq 0))$-iterative sequence $(x(n); n \geq 0)$ in M. Combining with the very definition above, it follows that conclusion in the statement is holding.

(C) In particular, when $\mathscr{R} = M \times M$, the regularity condition imposed in (SDC) holds. The corresponding variant of the underlying statement is just

(AC(N))=the Denumerable Axiom of Choice.

Precisely, we have

Proposition 4 *Let $F : N \to (2)^M$ be a function. Then, for each $a \in F(0)$ there exists a function $f : N \to M$ with $f(0) = a$ and $f(n) \in F(n)$, $\forall n$.*

Note that, as a consequence of the above facts,

(DC) $\Longrightarrow$ (AC(N)) in the strongly reduced system (ZF-AC).

A direct verification of this is obtainable by taking $A = N \times M$ and introducing the relation $\mathscr{R}$ over it, according to:

$$\mathscr{R}(n, x) = \{n + 1\} \times F(n + 1),\ n \in N, x \in M;$$

we do not give details. The reciprocal of the written inclusion is not true; see Moskhovakis [25, Ch 8, Sect 8.25] for details.

(D) As a useful application of these developments, we show that the Baire Category Theorem (in short: BC) is provable in (ZF-AC+DC).

Let (X, d) be a metric space. For each $x \in X$, $\varepsilon > 0$, denote

$$X(x, \varepsilon) = \{y \in X; d(x, y) < \varepsilon\},\ X[x, \varepsilon] = \{y \in X; d(x, y) \leq \varepsilon\}$$

(the *open/closed sphere* with center x and radius ε). Let $Y \in 2^X$ be arbitrary fixed. We say that $y \in X$ is *interior* to Y when

$X(y, \varepsilon) \subseteq Y$, for some $\varepsilon = \varepsilon(y) > 0$;
or, equivalently:
$X[y, \varepsilon] \subseteq Y$, for some $\varepsilon = \varepsilon(y) > 0$;

the set of all such points will be denoted as $\mathrm{int}(Y)$ (the *interior* of Y). Also, let us say that $z \in X$ is *adherent* to Y when

$X(z, \varepsilon) \cap Y \neq \emptyset$, for each $\varepsilon > 0$;
or, equivalently:
$X[z, \varepsilon] \cap Y \neq \emptyset$, for each $\varepsilon > 0$.

the set of all such points will be denoted $\mathrm{cl}(Y)$ (the *closure* of Y). Call Y, *open* (resp., *closed*) provided $Y = \mathrm{int}(Y)$ (resp., $Y = \mathrm{cl}(Y)$). In particular, any open (resp., closed) sphere in X is an open (resp., closed) subset of X. Clearly, the following *basic duality formula* holds

(dual-1) (for each $Y \in 2^X$):
$\mathrm{int}(X \setminus Y) = X \setminus \mathrm{cl}(Y),\ \mathrm{cl}(X \setminus Y) = X \setminus \mathrm{int}(Y)$;

see, for instance, Bourbaki [5, Ch 1, Sect 1.6]. Finally, given $Y \in (2)^X$, denote

$\mathrm{diam}(Y) = \sup\{d(u, v); u, v \in Y\}$ (the *diameter* of Y);

clearly, for any such part, we have

$\mathrm{diam}(Y) = \mathrm{diam}(Z)$, where $Z = \mathrm{cl}(Y)$.

Now, call the ambient metric space (X, d), *complete* when each d-Cauchy sequence is d-convergent. To give a characterization of this concept, call the sequence of (nonempty) closed parts $(F_n; n \geq 0)$,

(i) *descending* when $F_i \supseteq F_j$ for $i \leq j$;
(ii) *0-descending* if, in addition, $\mathrm{diam}(F_n) \to 0$.

We say that (X, d) has the *Cantor property* when

each 0-descending sequence of nonempty closed sets in X has a nonempty intersection.

The following *Cantor Intersection Theorem* (in short: CI) is useful for us.

Proposition 5 *The following (generic type) properties are valid, in the class of metric structures:*

(25-1) Cantor property $\Longrightarrow$ *completeness, in (ZF-AC)*
(25-2) completeness $\Longrightarrow$ *Cantor property, in (ZF-AC+DC).*

Proof

(i) Assume that the metric space (X, d) has the Cantor property; and let $(x_n; n \geq 0)$ be a d-Cauchy sequence in X:

$\lim_n \operatorname{diam}(Y_n) = 0$, where $(Y_n = \{x_k; k \geq n\}; n \geq 0)$.

By the properties of diameter function (see above), this yields

$\lim_n \operatorname{diam}(F_n) = 0$, where $(F_n = \operatorname{cl}(Y_n); n \geq 0)$.

So, taking into account the Cantor property, we must have

$\cap\{F_n; n \geq 0\} = \{z\}$, for a (uniquely determined) $z \in X$.

And this, along with

$$Y_n \subseteq F_n \subseteq X[z, \operatorname{diam}(F_n)],\ n \geq 0$$

tells us that $\lim_n (x_n) = z$; so that, (X, d) is complete.

(ii) Assume that (X, d) is a complete metric space; and let $(F_n; n \geq 0)$ be a 0-descending sequence of nonempty closed sets in X. By the Denumerable Axiom of Choice (AC(N)), we may construct a sequence $(x_n; n \geq 0)$ in X, according to

$x_n \in F_n$, for each $n \geq 0$.

From the choice of our sequence $(F_n; n \geq 0)$, one has

$\lim_n \operatorname{diam}(Y_n) = 0$, where $(Y_n = \{x_k; k \geq n\}; n \geq 0)$;

which tells us that (x_n) is a d-Cauchy sequence in X. But then, by completeness,

$\lim_n x_n = z$, for some (uniquely determined) $z \in X$.

This in turn yields, in a direct way,

$$\cap\{F_n; n \geq 0\} = \{z\};$$

so that, the Cantor property follows.

(E) Having these precise, call $Y \in 2^X$, *nowhere dense* when $\mathrm{int}(\mathrm{cl}(Y)) = \emptyset$; and *dense*, provided $\mathrm{cl}(Y) = X$; note that—as a consequence of some related fact—the *derived* duality formula holds

(dual-2) (for each $Y \in 2^X$):
(Y=closed, nowhere dense) $\Longleftrightarrow$ ($X \setminus Y$=open, dense).

Finally, let us say that $Y \in 2^X$ is of $\mathscr{F}_\sigma$ (resp., $\mathscr{G}_\delta$) type, when

Y may be represented as a denumerable union (resp., intersection) of closed (resp., open) subsets in X.

The following Baire Category Theorem (in short: BC) is available.

Proposition 6 *Let (X, d) be a complete metric space. Then*

(26-1) If (F_n) is a sequence of (proper) nowhere dense closed subsets of X, then $F = \cup\{F_n; n \geq 0\}$ is a (proper) nowhere dense $\mathscr{F}_\sigma$ type subset of X
(26-2) If (G_n) is a sequence of (nonempty) dense open subsets of X, then $G = \cap\{G_n; n \geq 0\}$ is a (nonempty) dense $\mathscr{G}_\delta$ type subset of X.

An appropriate proof of this may be found in Schechter [27, Ch 20, Sect 20.16]; for completeness reasons, we shall provide it, with some modifications.

Proof (Proposition 6) Clearly, the properties (26-1) and (26-2) are equivalent, via preceding duality formula; so, it will suffice verifying that, e.g., (26-2) holds. Without loss, one may assume that $G_0 = X$; otherwise, we simply replace $(G_0, G_1, \ldots)$ by $(X, G_0, G_1, \ldots)$. Denote $M := X \times R^0_+$, where $R^0_+ :=]0, \infty[$; and, for each $n \geq 0$, put

$M_n = \{(x, \varepsilon) \in M; X[x, \varepsilon] \subseteq G_n\}$; hence, $M_0 = M$.

We introduce a relation $\mathscr{R}$ over M as:

$(x, \varepsilon)\mathscr{R}(y, \delta)$ iff $X[x, \varepsilon] \supseteq X[y, \delta]$, $\varepsilon \geq 2\delta$.

Fix $n \geq 0$ and take some $(x, \varepsilon) \in M_n$; i.e., $X[x, \varepsilon] \subseteq G_n$. As $\mathrm{cl}(G_{n+1}) = X$, we have that $X(x, \varepsilon) \cap G_{n+1}$ is a nonempty open set in X; hence, there exists $(y, \delta) \in M$ in such a way that

$X(y, 2\delta) \subseteq X(x, \varepsilon) \cap G_{n+1}$, $\delta \leq \varepsilon/2$;

which yields

$X[y, \delta] \subseteq X[x, \varepsilon] \cap G_{n+1}$, $\delta \leq \varepsilon/2$;
hence, in particular, $(x, \varepsilon)\mathscr{R}(y, \delta)$ and $(y, \delta) \in M_{n+1}$.

By the Selected Dependent Choice Principle (SDC), it follows that for each starting $(x_0, \varepsilon_0) \in M (= M_0)$ there exists a sequence $((x_n, \varepsilon_n); n \geq 0)$ in M, with the properties (for each n)

(26-3) $(x_n, \varepsilon_n) \in M_n$ (that is: $X[x_n, \varepsilon_n] \subseteq G_n$)
(26-4) $(x_n, \varepsilon_n)\mathscr{R}(x_{n+1}, \varepsilon_{n+1})$
(i.e.: $X[x_n, \varepsilon_n] \supseteq X[x_{n+1}, \varepsilon_{n+1}]$, $\varepsilon_n \geq 2\varepsilon_{n+1}$).

By the latter of these, one derives—via Cantor Intersection Theorem—that there must be a unique $z \in X$ with $\cap\{X[x_n, \varepsilon_n]; n \geq 0\} = \{z\}$; so, combining with the former, $z \in G$. Since, on the other hand, $z \in X[x_0, \varepsilon_0]$, we have $X[x_0, \varepsilon_0] \cap G \neq \emptyset$. This, by the arbitrariness of (x_0, ε_0), gives $\mathrm{cl}(G) = X$; and completes the argument.

(F) By the developments above, it follows that

$$\text{(DC)} \implies \text{(BC), in (ZF-AC).}$$

The remarkable fact to be added is that the converse inclusion is valid too.

Proposition 7 *We have [in (ZF-AC)]:*

(27-1) (BC) $\Longrightarrow$ *(DC); hence, (BC)* $\Longleftrightarrow$ *(DC)*
(27-2) (BC) $\Longrightarrow$ *(AC(N)) (but not conversely).*

The original argument may be found in Blair [3]. For completeness reasons, we provide however a verification of it, with certain modifications.

Proof (Proposition 7) The argument will be divided into several steps.

Part 1 Let P be a nonempty set; and d stand for the *discrete* metric over it:

$(s, t \in P)$: $[d(s, t) = 0$, if $s = t]$ and $[d(s, t) = 1$, if $s \neq t]$;

(Note that d is even an *ultrametric* on P, in the sense

$d(x, z) \leq \max\{d(x, y), d(y, z)\}$, $\forall x, y, z \in X$;

but this is not essential for us). Let $\mathscr{S}(P) := \mathscr{F}(N, P)$ stands for the class of all sequences in P. Fix some $a \in P$; and put $X = \{x \in \mathscr{S}(P); x(0) = a\}$. Define a mapping $d_\infty : X \times X \to R_+$ as

$d_\infty(x, y) = \sum_n 2^{-n} d(x(n), y(n))$, for $x = (x(n))$, $y = (y(n))$ in X.

It is not hard to see that d_∞ appears as a metric on X. The natural question to be discussed here involves its completeness property. In this direction, we have

Lemma 2 *Under the above conventions, the metric space* (X, d_∞) *is complete: each* d_∞*-Cauchy sequence in* X *is* d_∞*-convergent.*

Proof Let $(x^n; n \geq 0)$ be a sequence in X; it may be written as

$$(x^n = (x^n(0), x^n(1), \ldots) = (a, x^n(1), \ldots); n \geq 0).$$

Assume that $(x^n; n \geq 0)$ is d_∞-Cauchy; this may be also characterized as:

$\forall \varepsilon > 0$: $C(\varepsilon) := \{n \in N; n \leq p \leq q \implies d_\infty(x^p, x^q) < \varepsilon\} \neq \emptyset$.

As a consequence, the map $\varepsilon \mapsto C(\varepsilon)$ is increasing on R_+^0, in the sense:

$\varepsilon_* < \varepsilon^*$ implies $C(\varepsilon_*) \subseteq C(\varepsilon^*)$;

so that, the map $\varepsilon \mapsto \Gamma(\varepsilon) := \min[C(\varepsilon)]$ is decreasing on R_+^0:

$\varepsilon_* < \varepsilon^*$ implies $\Gamma(\varepsilon_*) \geq \Gamma(\varepsilon^*)$.

Let $(\varepsilon_n; n \geq 0)$ be a strictly descending sequence in R^0_+ with

$\varepsilon_n < 2^{-n}, \forall n$ (hence: $\varepsilon_n \to 0$ as $n \to \infty$).

Denote for simplicity

$m(k) = \Gamma(\varepsilon_k), n(k) = m(k) + k, \ k \geq 0.$

By the properties above, the map $k \mapsto m(k)$ is increasing; hence, the map $k \mapsto n(k)$ is strictly increasing. For the moment, it is clear that

$x^{n(0)}(0) = x^p(0) = a, \forall p \geq n(0).$

Further, by the very definition of these maps,

$n(1) \leq p \leq q \Longrightarrow d_\infty(x^p, x^q) < \varepsilon_1.$

Combining with the definition of d_∞, gives

$2^{-1} d(x^p(1), x^q(1)) < \varepsilon_1$, if $n(1) \leq p \leq q$;

so that (as $\varepsilon_1 < 2^{-1}$),

$x^{n(1)}(1) = x^p(1)$, for all $p \geq n(1)$.

The procedure may continue indefinitely; and gives us

$$x^{n(i)}(i) = x^p(i), \quad \text{for all } p \geq n(i) \text{ and all } i \geq 0.$$

Let $y = (y(i); i \geq 0)$ be the "diagonal" sequence $(y(i) = x^{n(i)}(i); i \geq 0)$; clearly, it is an element of X. We claim that our initial sequence $(x^n; n \geq 0)$ is convergent (modulo d_∞) towards y; i.e.,

$d_\infty(x^n, y) \to 0$, as $n \to \infty$.

In fact, let $\varepsilon > 0$ be arbitrary fixed; and $h = h(\varepsilon) \geq 0$ be such that

$2^{-h} < \varepsilon$ (hence: $2^{-j} < \varepsilon, \forall j \geq h$).

For each $n \geq n(h)$ we have (by the above properties)

$$d(x^n, y) = \sum_{i \leq h} 2^{-i} d(x^n(i), x^{n(i)}(i)) + \sum_{i > h} 2^{-i} d(x^n(i), x^{n(i)}(i)) =$$
$$\sum_{i > h} 2^{-i} d(x^n(i), x^{n(i)}(i)) \leq \sum_{i > h} 2^{-i} = 2^{-h} < \varepsilon;$$

and, from this, we are done.

Part 2 Let M be a nonempty set; and $\mathscr{R}$ be a proper relation over M:

(p-rel) $\forall a \in M, \exists b \in M$: $(a, b) \in \mathscr{R}$.

Fix some $a \in M$; and put $X = \{x \in \mathscr{S}(M); x(0) = a\}$. Let the metric $d_\infty(.,.)$ over X be introduced as before; remember that

(X, d_∞) is complete:
each d_∞-Cauchy sequence in X is a d_∞-convergent one.

For each $x \in X$, define the relation $\mathscr{H}(x) : N \to 2^N$ and the associated functions $H(x) : N \to N \cup \{\infty\}$, $H[x] : N \to N \cup \{\infty\}$, according to

$\mathscr{H}(x)(i) = \{j \in N(i, <); (x(i), x(j)) \in \mathscr{R}\}, i \geq 0$
$H(x)(i) = \min \mathscr{H}(x)(i), i \geq 0$
$H[x](n) = \max\{H(x)(i), i \leq n\}, n \geq 0;$

here, as usual, $\min(\emptyset) = \infty$. Clearly, the following progressive type relations hold

(prog-1) $i < H(x)(i)$, for each $i \geq 0$
(prog-2) $n < H[x](n)$, for each $n \geq 0$.

Some other useful completions of these are being collected in

Lemma 3 *Let $n \geq 0$ and $x \in X$ be such that*

$k := H[x](n) < \infty$*; hence, $n < k$ and ($H(x)(i) \leq k$, for $0 \leq i \leq n$).*

Then,

(p1) for each $i \leq n (< k)$, there exists $j \leq k$ with $i < j$, $(x(i), x(j)) \in \mathscr{R}$
(p2) if $y \in X$ fulfills ($y(h) = x(h)$, whenever $0 \leq h \leq k$), then
($H(y)(i) = H(x)(i)$, for $0 \leq i \leq n$); so that, $H[y](n) = H[x](n)$.

Proof

(i) Evident, by letting $j = H(x)(i)$ and combining with a progressive type relation we just listed.

(ii) Let $y \in X$ be taken as before; and let $i \in N(n, \geq)$ be arbitrary fixed. By the preceding step, we have with $p := H(x)(i)$ (hence, $p \leq k$)

$$i < p \text{ and } (x(i), x(p)) \in \mathscr{R};$$

or, equivalently (by our working hypothesis)

$$i < p \text{ and } (y(i), y(p)) \in \mathscr{R};$$

whence $q := H(y)(i) \leq p$. This, again by our working hypothesis, gives

$$i < q \text{ and } (x(i), x(q)) \in \mathscr{R};$$

wherefrom $p = H(x)(i) \leq q$. Combining these, yields $H(x)(i) = H(y)(i)$; so that [by the arbitrariness of $i \in N(n, \geq)$], it results that $H[y](n) = H[x](n)$.

We may now pass to the basic part of our developments. To each $n \geq 0$, let us associate a subset A_n of X as

$A_n = \{x \in X; H(x)(i) < \infty, \text{ for all } i \leq n\};$
or, equivalently: $A_n = \{x \in X; H[x](n) < \infty\}$.

Note that, for the moment, the alternative of

$A_n = \emptyset$ for some $n \geq 0$

cannot be avoided. This, however, cannot hold; as results from

Lemma 4 *The subset A_n is nonempty in X, for each $n \geq 0$.*

Proof Let $n \geq 0$ be arbitrary fixed. By the Finite Dependent Choice property (DC-fin), there exists a $(n+2)$-sequence $F : N(n+2, >)$, in such a way that

$$F(0) = a \text{ and } F(k)\mathscr{R}F(k+1), \text{ for } k \leq n.$$

The sequence $x \in \mathscr{S}(M)$ introduced as

$$x(i) = F(i), i \leq n; x(i) = F(n+1), i > n$$

belongs to X and

$$H(x)(i) = i + 1 \leq n + 1, \forall i \leq n; \text{ whence, } H[x](n) = n + 1;$$

and this tells us that $x \in A_n$.

Concerning the remaining properties of the family $(A_n; n \geq 0)$, a basic one is expressed by

Lemma 5 *Let $n \geq 0$ and $x \in A_n$ be arbitrary fixed; hence,*

$$k := H[x](n) < \infty; \text{ so that: } n < k \text{ and } (H(x)(i) \leq k, \text{ for all } i \leq n).$$

Then, each $y \in X$ with

$$d_\infty(x, y) = \textstyle\sum_i 2^{-i} d(x(i), y(i)) < 2^{-k}$$

fulfills $y \in A_n$. In particular, A_n is open in X (for each $n \geq 0$).

Proof By the meaning of our notation (and a previous auxiliary fact)

$$\forall i \leq n, \exists j \leq k: i < j \text{ and } (x(i), x(j)) \in \mathscr{R}.$$

On the other hand, by the imposed condition,

$$x(h) = y(h), \text{ whenever } 0 \leq h \leq k.$$

Combining these yields (see above)

$$H(y)(i) = H(x)(i), \forall i \leq n; \text{ wherefrom } H[y](n) = H[x](n)(< \infty);$$

so that $y \in A_n$.

Finally, as another basic property of $(A_n; n \geq 0)$, we have

Lemma 6 *The subset A_n is dense in X, for each $n \geq 0$.*

Proof Let us consider the logical property

(Dense(n)) A_n is dense in X (modulo d_∞).

We show, by induction, that

(Dense(n)) is true, for all $n \geq 0$.

(I) First, we are verifying the starting property:

A_0 is dense in X (modulo d_∞).

Let the sequence $z \in X$ and the number $\varepsilon > 0$ be arbitrary fixed. Further, let the rank $m \geq 0$ be such that $2^{-m} < \varepsilon$. By hypothesis, there exists $b \in M$ such that $(z(0), b) \in \mathscr{R}$. Define the sequence $w \in X$ according to

$$w(n) = z(n), n \leq m; w(n) = b, n > m.$$

As a consequence of these facts,

(i) $H(w)(0) = H[w](0) \leq m + 1$; hence, $w \in A_0$
(ii) $d_\infty(z, w) \leq \sum_{i>m} 2^{-i} = 2^{-m} < \varepsilon$;

and our assertion follows.

(II) Further, we pass to the inductive step:

(A_j is dense in X for $0 \leq j \leq m$) implies A_{m+1} is dense in X.

Let the sequence $z \in X$ be given, as well as the number $\varepsilon > 0$ with $\varepsilon < 2^{-m}$. As A_m is dense in X,

$$d_\infty(z, u) < \varepsilon/2, \text{ for some } u \in A_m.$$

On the other hand, as $u \in A_m$,

$$p := H[u](m) < \infty; \text{ whence, } m < p \text{ and } (H(u)(j) \leq p, \forall j \leq m).$$

Fix some index q according to

$$q \geq p + 1, 2^{-q} < \varepsilon/2;$$

and choose (by the properness of $\mathscr{R}$) some $c \in M$ with $(u(m + 1), c) \in \mathscr{R}$. Finally, define the sequence $w \in X$, according to

$$w(n) = u(n), n \leq q; w(n) = c, n > q.$$

By this very convention (and $p = H[u](m) < q$),

$$H(w)(j) = H(u)(j) \leq p < q + 1, \forall j \leq m; H(w)(m + 1) \leq q + 1;$$
wherefrom $H[w](m + 1) \leq q + 1 (< \infty)$;

which tells us that $w \in A_{m+1}$. Moreover (combining with the choice of q)

$$d_\infty(u, w) \leq \sum_{j>q} 2^{-j} = 2^{-q} < \varepsilon/2;$$

wherefrom

$$d_\infty(z, w) \leq d_\infty(z, u) + d_\infty(u, w) < \varepsilon.$$

Putting these together, our assertion is proved.

Having these precise, let us continue our argument, by passing to

Part 3 Since all sets in the sequence $(A_n; n \geq 0)$ are (nonempty) open dense, it results—via Baire Category Theorem (BC)—that $A_\infty := \cap\{A_n; n \geq 0\}$ is a dense $\mathscr{G}_\delta$ type subset of X. Fix some sequence $y = (y(0), y(1), \ldots)$ in A_∞; note that, by this very choice,

$H(y)(i) < \infty$, for each $i \geq 0$; whence: $H[y](n) < \infty$, for each $n \geq 0$.

Define a strictly ascending rank sequence $(k(n); n \geq 0)$, according to the iterative type construction

$k(0) = 0, k(n+1) = H(y)(k(n)), n \geq 0.$

The subsequence $(z(n) := y(k(n)); n \geq 0)$ of $(y(n); n \geq 0)$ is an element of X, with

$$z(n) \mathscr{R} z(n+1), \ \forall n;$$

so that, $(z(n); n \geq 0)$ is $(a, \mathscr{R})$-iterative. This ends our argument.

Note, finally, that in a *separable* complete metric space, the Baire Category theorem does not require any choice techniques; cf. Oxtoby [26, Supplem. Notes Ch 9]. Further aspects may be found in Schechter [27, Ch 20, Sect 20.16].

3 (BC) Implies (LR-ndiscr-m)

Let Z be a nonempty set. Remember that, by a *topology* over Z we mean any (nonempty) family of subsets $\mathscr{T} \subseteq 2^Z$, fulfilling

(top-1) $\mathscr{A}$=arbitrary part of $\mathscr{T} \Longrightarrow \cup\mathscr{A} \in \mathscr{T}$
(top-2) $\mathscr{A}$=finite part of $\mathscr{T} \Longrightarrow \cap\mathscr{A} \in \mathscr{T}$;

in this case, $(Z, \mathscr{T})$ will be referred to as a *topological space*. Given any $z \in Z$, any subset $U \in \mathscr{T}$ with $U \ni z$ will be referred to as a *open neighborhood* of z; the class of all these will be denoted as $\mathscr{V}^0(z)$. Also, by a *neighborhood* of z we mean any subset V of Z with

$V \supseteq W$, for some $W \in \mathscr{V}^0(z)$;

the family of all such subsets will be denoted as $\mathscr{V}(z)$. In this case, call the topological space $(Z, \mathscr{T})$, *Hausdorff separated* if

$(a, b \in Z, a \neq b) \Longrightarrow (\exists A \in \mathscr{V}(a), \exists B \in \mathscr{V}(b))$: $A \cap B = \emptyset$.

Note that such a condition is fulfilled whenever the ambient topology $\mathscr{T}$ is generated by a *metric* $\Delta(.,.)$ over Z.

Let $Y \in 2^Z$ be arbitrary fixed. We say that $y \in X$ is *interior* to Y, when

$V \subseteq Y$, for some $V \in \mathscr{V}(y)$;

the set of all such points will be denoted as $\text{int}(Y)$ (the *interior* of Y). Also, let us say that $z \in X$ is *adherent* to Y, when

$$V \cap Y \neq \emptyset, \text{ for each } V \in \mathscr{V}(z);$$

the set of all such points will be denoted $\text{cl}(Y)$ (the *closure* of Y). Call Y, *open* (resp., *closed*) provided $Y = \text{int}(Y)$ (resp., $Y = \text{cl}(Y)$); note that the *basic* duality formula from the metric setting continues to hold here. Finally, call $Y \in 2^Z$, *nowhere dense* when $\text{int}(\text{cl}(Y)) = \emptyset$; and *dense*, provided $\text{cl}(Y) = X$. As before, the *derived* duality formula holds too in this extended setting; we do not give details.

Having these precise, we may now pass to the formulation of our problem. Let $(Z, \mathscr{T})$ be a Hausdorff separated topological space. Denote for simplicity

$$\mathscr{T}^0 = \mathscr{T} \setminus \{\emptyset\} = \mathscr{T} \cap (2)^Z;$$

this is nothing else than the class of all nonempty open subsets of Z. Further, let (X, d) be a metric space (in the usual sense). Remember that, for each $Y \in (2)^X$, we denoted

$$\text{diam}(Y) = \sup\{d(u, v); u, v \in Y\} \text{ (the } \textit{diameter} \text{ of } Y).$$

Let $T : R_+^0 \times Z \to (2)^X$ be a (nonempty valued) map; it may be also viewed as a family $(T(\beta); \beta > 0)$ of (nonempty valued) maps in $\mathscr{F}(Z, (2)^X)$, according to

$$T(\beta)(z) = T(\beta, z), \beta > 0, z \in Z.$$

Call this family, *ascending* provided

$$\beta_1 \leq \beta_2 \Longrightarrow T(\beta_1)(z) \subseteq T(\beta_2)(z), \forall z \in Z;$$
or, equivalently:
$$\beta_1 \leq \beta_2 \Longrightarrow T(\beta_1)(U) \subseteq T(\beta_2)(U), \forall U \in (2)^Z.$$

Given such a family $(T(\beta); \beta > 0)$ over $\mathscr{F}(Z, (2)^X)$, define the *extremal* mapping $T(0+) : Z \to 2^X$, as

$$T(0+)(z) = \cap\{T(\beta)(z); \beta > 0\}, z \in Z.$$

Denote, for simplicity

$$\text{Dom}(T(0+)) = \{z \in Z; T(0+)(z) \neq \emptyset\} \text{ (the } \textit{domain} \text{ of } T(0+)).$$

Note that the alternative

$$T(0+) \text{ is not proper: } \text{Dom}(T(0+)) = \emptyset$$

cannot be avoided; but, even if

$$T(0+) \text{ is proper: } \text{Dom}(T(0+)) \neq \emptyset$$

we must emphasize that, given $z \in \text{Dom}(T(0+))$, the extremal representation of $T(0+)(z)$ relative to $(T(\beta)(z); \beta > 0)$ cannot be viewed as a limit one (from a topological perspective). So, we may ask under which (extra) regularity conditions about our family $(T(\beta)(.); \beta > 0)$ is this possible. As we shall see, the following general conditions are needed:

(c01) $(Z, \mathcal{T})$ is a *Baire space*; i.e.: ($\mathcal{T}$ is Hausdorff separated and) the intersection of a countable family of open dense subsets in Z is dense
(c02) (X, d) is a complete metric space
(c03) $(T(\beta); \beta > 0)$ is ascending (see above) and nonempty closed valued; i.e.: $T(\beta)(z) \in (2c)^X$, for all $\beta > 0, z \in Z$.

Here, we denoted for simplicity

$(2c)^X$=the class of all nonempty closed parts in X.

Concerning the specific conditions to be added, some preliminaries are in order. Define, for each $\varepsilon > 0$

$$\mathcal{A}(\varepsilon) = \{(U, \beta) \in \mathcal{T}^0 \times R^0_+; \operatorname{diam}(T(\beta)(U)) < \varepsilon\};$$

note that $\varepsilon \mapsto \mathcal{A}(\varepsilon)$ is a multivalued map from R^0_+ to $\mathcal{T}^0 \times R^0_+$, with (possibly empty values and) the *hereditary* property

(for each $(U, \beta), (U^*, \beta^*) \in \mathcal{T}^0 \times R^0_+$):
$(U, \beta) \in \mathcal{A}(\varepsilon) \Longrightarrow (U^*, \beta^*) \in \mathcal{A}(\varepsilon)$, whenever $U \supseteq U^*, \beta \geq \beta^*$.

Let also $\varepsilon \mapsto \Pi_1(\mathcal{A}(\varepsilon))$ and $\varepsilon \mapsto \Pi_2(\mathcal{A}(\varepsilon))$ stand for the projection maps

$$\Pi_1(\mathcal{A}(\varepsilon)) = \{U \in \mathcal{T}^0; (U, \beta) \in \mathcal{A}(\varepsilon), \text{for some } \beta \in R^0_+\}, \varepsilon > 0$$
$$\Pi_2(\mathcal{A}(\varepsilon)) = \{\beta \in R^0_+; (U, \beta) \in \mathcal{A}(\varepsilon), \text{for some } U \in \mathcal{T}^0\}, \varepsilon > 0.$$

Now, the announced specific condition upon our family may be written as

(c04) $(T(\beta); \beta > 0)$ is *d-fragmentable*:
$\forall U \in \mathcal{T}^0, \forall \varepsilon > 0, \exists V \in \Pi_1(\mathcal{A}(\varepsilon))$: $U \supseteq V$.

Finally, let us introduce the generalized Hausdorff-Pompeiu pseudo-distance over $(2)^X$ as: for each $A, B \in (2)^X$,

$$\Gamma(A, B) = sup\{d(x, B); x \in A\},$$
where $d(x, B) = \inf\{d(x, b); b \in B\}$, for all such x and B.

Note that its symmetric counterpart

$$\Theta(A, B) = \max\{\Gamma(A, B), \Gamma(B, A)\}, A, B \in (2)^X$$

is a generalized distance (in the Luxemburg-Jung sense [23], [18]) over $(2c)^X$; see Kuratowski [20, Ch 2, Sect 15] for details.

The following result (referred to as: Nondiscrete Lassonde-Revalski Principle; in short: (LR-ndiscr)) is a basic tool for our developments.

Theorem 2 *Let $(T(\beta) : Z \to (2c)^X; \beta > 0)$ be an ascending family of nonempty closed valued maps from Z to X, where $(V, \mathcal{T})$ is a (Hausdorff separated) Baire space and (X, d) is a complete metric space. In addition, suppose that $(T(\beta); \beta > 0)$ is d-fragmentable. Then, there exists a dense $\mathcal{G}_\delta$ type subset $W \in (2)^Z$, with*

(31-a) $\operatorname{Dom}(T(0+))$ *is W-including (*$\operatorname{Dom}(T(0+)) \supseteq W$*)*

such that, for each $z \in W$ [hence, $z \in \operatorname{Dom}(T(0+))$],

(31-b) $T(0+)(z)$ *is a (nonempty) singleton*
(31-c) for each $\varepsilon > 0$*, there exists* $(U, \beta) \in \mathscr{A}(\varepsilon)$*, such that*
$U \in \mathscr{V}^0(z)$ *and* $T(0+)(z) \in T(\beta)(z) \subseteq T(\beta)(U) \subseteq X(T(0+)(z), \varepsilon)$
(31-d) $\Theta(T(\beta)(z), T(0+)(z)) \to 0$ *as* $\beta \to 0+$.

Technically speaking, this result is to be obtained by means of certain choice procedures. The most general one is (AC) (= Axiom of Choice). But, as we shall see, (AC(N)) (= Denumerable Axiom of Choice) will suffice for the proposed argument to work. This will follow from

Proposition 8 *We have*

(31-1) $((DC) \Longrightarrow)$ *(AC(N))* $\Longrightarrow$ *(LR-ndiscr), in (ZF-AC);*

or, in other words,

(31-2) (LR-ndiscr) is deductible in (ZF-AC+DC).

Proof Let $(\varepsilon_n; n \geq 0)$ be a strictly descending sequence in R_+^0 with $\lim_n \varepsilon_n = 0$. (The natural choice is $(\varepsilon_n = 2^{-n}; n \geq 0)$; but, this is not the only possible one).

Step 1 Denote, for simplicity

$$W_k = \cup \Pi_1(\mathscr{A}(\varepsilon_k)), k \geq 0.$$

Clearly, W_k is (nonempty) open in Z (for each $k \geq 0$), as a union of (nonempty) open parts of Z. Moreover,

W_k is dense in Z, for each $k \geq 0$.

In fact, let $z \in Z$ be arbitrary fixed; and $U \in \mathscr{V}^0(z)$ be an open neighborhood of it. As $(T(\beta); \beta > 0)$ is d-fragmentable, there exists $V \in \Pi_1(\mathscr{A}(\varepsilon_k))$ (hence, $V \subseteq W_k$), with $U \supseteq V$. Consequently, $\emptyset \neq V \subseteq W_k \cap U$; wherefrom, $z \in \mathrm{cl}(W_k)$.

Step 2 Further, denote $W = \cap\{W_k; k \geq 0\}$; this (by the Baire space condition) is a dense $\mathscr{G}_\delta$ type subset of Z. We claim that conclusions of (LR-ndiscr) are necessarily holding under such a choice.

In fact, let $z \in W$ be arbitrary fixed. By definition,

$$z \in W_k = \cup \Pi_1(\mathscr{A}(\varepsilon_k)), \text{ for each } k \geq 0;$$

so that, the multivalued map

$$\mathscr{F}_z(k) = \{(U, \beta) \in \mathscr{A}(\varepsilon_k);\ U \ni z\}, k \geq 0$$

is nonempty valued. By the Denumerable Axiom of Choice (AC(N)), there exists a sequence $((U(k), \beta(k)); k \geq 0)$ in $\mathscr{T}^0 \times R_+^0$, such that

$$(U(k), \beta(k)) \in \mathscr{F}_z(k),\ \forall k \geq 0;$$

that is

$$U(k) \in \mathscr{V}^0(z), \mathrm{diam}[T(\beta(k))(U(k))] < \varepsilon_k,\ k \geq 0.$$

Define a new sequence $(\gamma(k); k \geq 0)$ in R^0_+ as

$$\gamma(k) = 2^{-k} \min\{\beta(0), \ldots, \beta(k)\}, k \geq 0.$$

Clearly, $(\gamma(k); k \geq 0)$ is strictly descending in R^0_+ with

$$(\gamma(k) \leq \beta(k), \forall k \geq 0) \text{ and } \lim_k(\gamma(k)) = 0;$$

moreover, by the ascending property of the family $(T(\beta); \beta > 0)$, we must have

$$U(k) \in \mathscr{V}^0(z), \operatorname{diam}[T(\gamma(k))(U(k))] < \varepsilon_k, \ k \geq 0.$$

In particular, this tells us that

$$\operatorname{diam}[T(\gamma(k))(z)] \leq \operatorname{diam}[T(\gamma(k))(U(k))] < \varepsilon_k, \ k \geq 0;$$

which, by the Cantor Intersection Theorem, assures us that

$$T(0+)(z) = \cap\{T(\gamma(k))(z); k \geq 0\}$$

is a nonempty singleton; hence, the first conclusion in the statement follows. On the other hand, let $\varepsilon > 0$ be arbitrary fixed; and $k \geq 0$ be such that

$$\varepsilon_k < \varepsilon; \text{ hence, } \varepsilon_h < \varepsilon, \text{ for all } h \geq k.$$

In view of

$$T(0+)(z) \in T(\gamma(k))(U(k)), \ \operatorname{diam}[T(\gamma(k))(U(k))] < \varepsilon_k < \varepsilon,$$

we therefore get

$$T(\gamma(k))(U(k)) \subseteq X(T(0+)(z), \varepsilon_k) \subseteq X(T(0+)(z), \varepsilon);$$

and this, along with

$$T(0+)(z) \in T(\gamma(k))(z) \subseteq T(\gamma(k))(U(k)),$$

shows that our second conclusion in the statement is holding with $[U := U(k), \beta := \gamma(k)]$. Finally, the last conclusion is clear, by the very definition of Hausdorff-Pompeiu (generalized) distance.

Technically speaking, the Baire space condition is comprehensive enough so as to include some nice particular cases of (Hausdorff separated) topological spaces; see, for instance, Willard [34, Ch 7, Sect 25]. However (cf. our previous developments), the most natural circumstance when $(Z, \mathscr{T})$ is a Baire space is that of our ambient topology $\mathscr{T}$ being induced by a complete metric $\Delta(.,.)$ over Z. This yields the following practical statement, (referred to as: Nondiscrete Lassonde-Revalski Metrical Principle; in short: (LR-ndiscr-m)).

Theorem 3 *Let $(T(\beta) : Z \to (2c)^X; \beta > 0)$ be an ascending family of nonempty closed valued maps from Z to X, where (Z, Δ) and (X, d) are complete metric spaces. In addition, suppose that $(T(\beta); \beta > 0)$ is d-fragmentable. Then, there exists a dense $\mathscr{G}_\delta$ type subset $W \in (2)^Z$, with*

(32-a) $\mathrm{Dom}(T(0+))$ *is W-including (*$\mathrm{Dom}(T(0+)) \supseteq W$*)*

such that, for each $z \in W$ [hence, $z \in \mathrm{Dom}(T(0+))$],

(32-b) $T(0+)(z)$ is a (nonempty) singleton
(32-c) for each $\varepsilon > 0$, there exists $(U, \beta) \in \mathscr{A}(\varepsilon)$, such that
$U \in \mathscr{V}^0(z)$ and $T(0+)(z) \in T(\beta)(z) \subseteq T(\beta)(U) \subseteq X(T(0+)(z), \varepsilon)$
(32-d) $\Theta(T(\beta)(z), T(0+)(z)) \to 0$ as $\beta \to 0+$.

Formally, this result may be viewed as a particular case of the preceding one. However, the Baire property of the complete metric space (V, Δ) is no longer assured under (AC(N)); but (see above) it is retainable under the Baire Category theorem (BC). Summing up, the following inclusion type result is valid.

Proposition 9 *We have*

(32-1) ((DC) $\Longleftrightarrow$) (BC) $\Longrightarrow$ (LR-ndiscr-m), in (ZF-AC);

or, in other words,

(32-2) (LR-ndiscr-m) is deductible in (ZF-AC+DC).

Proof Let $(\varepsilon_n; n \geq 0)$ be a strictly descending sequence in R_+^0 with $\lim_n \varepsilon_n = 0$.

Part 1 Denote, for simplicity

$$W_k = \cup \Pi_1(\mathscr{A}(\varepsilon_k)), k \geq 0.$$

Clearly, W_k is (nonempty) open (for each $k \geq 0$), as a union of (nonempty) open parts of Z. Moreover (see above), W_k is dense in Z, for each $k \geq 0$.

Part 2 Further, denote $W = \cap\{W_k; k \geq 0\}$; this—by the Baire Category theorem (BC) applied to the complete metric space (Z, Δ)—is a dense $\mathscr{G}_\delta$ type subset of Z.

Finally, by our previous developments, conclusions of (LR-ndiscr-m) are necessarily holding under such a choice.

4 (LR-ndiscr-m) Implies (LR-VP-m)

Let (M, d) be a complete metric space; and the function $\psi \in \mathscr{F}(M, R \cup \{\infty\})$ be (M, d)-*regular*; i.e.,

(reg-1) ψ is proper ($\mathrm{Dom}(\psi) := \{x \in M; \psi(x) < \infty\} \neq \emptyset$)
(reg-2) ψ is bounded below ($\inf[\psi(M)] > -\infty$)
(reg-3) ψ is d-lsc ($\liminf_n \psi(x_n) \geq \psi(x)$, whenever $x_n \xrightarrow{d} x$).

Denote, for each $\beta \geq 0$

$$\text{ulev}(\psi; M; \beta) = \{x \in M; \psi(x) \leq \inf \psi(M) + \beta\};$$

this will be referred to as the *β-upper level set* of ψ with respect to M; note that, $\text{ulev}(\psi; M; 0)$ is nothing else than the set of *global minimum points* of ψ over M. Let us say that ψ is (M, d)-*admissible*, when the optimization problem attached to these data

$$(M, d; \psi): \ \inf\{\psi(x); x \in M\}$$

is *well posed*, in the sense

(wp-1) $\text{ulev}(\psi; M; 0)$ is a (nonempty) singleton, $\{x_\infty\}$
(wp-2) $\lim_{\beta \to 0+} \Theta(\text{ulev}(\psi; M; \beta), x_\infty) = 0$; hence, any minimizing sequence (x_n) in M (in the sense: $\psi(x_n) \to \inf \psi(M) = \psi(x_\infty)$) fulfills $x_n \xrightarrow{d} x_\infty$.

In what follows, the Nondiscrete Lassonde-Revalski Metrical Principle (LR-ndiscr-m) is used to get a generic type description of (M, d)-admissible functions in $\mathscr{F}(M, R \cup \{\infty\})$. To this end, some preliminaries are needed.

(A) Let (X, d) be a complete metric space. Denote

$\mathscr{C}(X)$=the (real) vector space of all continuous $f \in \mathscr{F}(X, R)$.

With the *supremum* generalized norm

$$||f||_\infty = \sup\{|f(x)|; x \in X\}, \ \ f \in \mathscr{C}(X),$$

$\mathscr{C}(X)$ becomes a generalized Banach space; equivalently, this means: $(\mathscr{C}(X), D_\infty)$ is a complete generalized metric space; where

$$D_\infty(f, g) = ||f - g||_\infty, \ \ f, g \in \mathscr{C}(X) \text{ (the } \textit{supremum} \text{ generalized metric).}$$

Further, take a linear subspace $\mathscr{A}$ of $\mathscr{C}(X)$. Let $||.||_{\mathscr{A}}$ be a (standard) norm over $\mathscr{A}$; and $D_{\mathscr{A}}$ stand for its associated (standard) metric

$$D_{\mathscr{A}}(f, g) = ||f - g||_{\mathscr{A}}, \ f, g \in \mathscr{A}.$$

The following regularity conditions are accepted

(d01) (completeness):
$(\mathscr{A}, ||.||_{\mathscr{A}})$ is a Banach space; or, equivalently: $(\mathscr{A}, D_{\mathscr{A}})$ is complete
(d02) (subordination):
there exists $\mu > 0$, with $(\mu||g||_{\mathscr{A}} \geq ||g||_\infty, \forall g \in \mathscr{A})$.

A basic consequence of the second one is the following. Denote

$\mathscr{C}_b(X)$=the linear subspace of all bounded $f \in \mathscr{C}(X)$.

Note that

$(\mathscr{C}_b(X), ||.||_\infty)$ is a (standard) Banach space;
hence, $(\mathscr{C}_b(X), D_\infty)$ is a (standard) complete metric space.

By the quoted condition, it results that

$\mathscr{A}$ is a subset of $\mathscr{C}_b(X)$; i.e.; each $g \in \mathscr{A}$ is bounded over X.

Now, let in the following $\varphi : X \to R \cup \{\infty\}$ be (X, d)-*regular*; i.e. (see above)

(reg-1) φ is proper ($\mathrm{Dom}(\varphi) := \{x \in X; \varphi(x) < \infty\} \neq \emptyset$)
(reg-2) φ is bounded below ($\inf \varphi(X) > -\infty$)
(reg-3) φ is d-lsc: ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \stackrel{d}{\longrightarrow} x$).

The following hereditary type result is useful in the sequel.

Proposition 10 *Under these conventions,*

$\varphi + g$ *is* (X, d)*-regular (see above), for each* $g \in \mathscr{A}$.

Proof

(j) As $g \in \mathscr{A}$ is bounded (on X),

$$\mathrm{Dom}(\varphi + g) = \mathrm{Dom}(\varphi) \neq \emptyset,$$
$$\inf(\varphi + g)(X) \geq \inf \varphi(X) - ||g||_\infty > -\infty;$$

hence, (reg-1) and (reg-2) are holding for $\varphi + g$.

(jj) As g is d-continuous on X, the sum function $\varphi + g$ is d-lsc (on X); so that, it fulfills (reg-3) as well.

Remember that, for each (X, d)-regular function $\psi : X \to R \cup \{\infty\}$ and each $\beta \geq 0$, we introduced the convention

$$\mathrm{ulev}(\psi; X; \beta) = \{x \in X; \psi(x) \leq \inf \psi(X) + \beta\};$$

this will be referred to as the β-*upper level set* of ψ with respect to X; moreover, $\mathrm{ulev}(\psi; X; 0)$ is nothing else than the set of global minimum points of ψ over X. Returning to our initial setting, the following auxiliary fact involving the upper level sets attached to the family $(\varphi + g; g \in \mathscr{A})$ will be in effect for us.

Proposition 11 *Let* $k, h \in \mathscr{A}$ *and* $\delta > 0$ *be such that* $||k - h||_{\mathscr{A}} \leq \delta$. *Then, for each* $\lambda \geq 0$, *we have*

$$\mathrm{ulev}(\varphi + k; X; \lambda) \subseteq \mathrm{ulev}(\varphi + h; X; \lambda + 2\delta\mu).$$

Proof By the subordination property, we have for the moment

$$||k - h||_\infty \leq \mu ||k - h||_{\mathscr{A}} \leq \delta\mu;$$

wherefrom, by definition

$$|k(x) - h(x)| \leq \delta\mu, \ \forall x \in X.$$

This yields the evaluations, for each $x \in X$

$$\varphi(x) + k(x) \leq \varphi(x) + h(x) + \delta\mu$$
$$\varphi(x) + h(x) \leq \varphi(x) + k(x) + \delta\mu;$$

wherefrom, passing to infimum

$$\inf(\varphi + k)(X) \leq \inf(\varphi + h)(X) + \delta\mu$$
$$\inf(\varphi + h)(X) \leq \inf(\varphi + k)(X) + \delta\mu.$$

Combining these facts yields, for each $x \in \text{ulev}(\varphi + k; X; \lambda)$

$$(\varphi + h)(x) \leq (\varphi + k)(x) + \delta\mu \leq \inf(\varphi + k)(X) + \lambda + \delta\mu$$
$$\inf(\varphi + k)(X) \leq \inf(\varphi + h)(X) + \delta\mu;$$

wherefrom

$$(\varphi + h)(x) \leq \inf(\varphi + h)(X) + \lambda + 2\delta\mu.$$

This proves that $x \in \text{ulev}(\varphi + h; X; \lambda + 2\delta\mu)$; and concludes the argument.

We may now pass to the effective part of our developments. Let $\varphi : X \to R \cup \{\infty\}$ be (X, d)-regular. Define a mapping $T : R^0_+ \times \mathscr{A} \to (2)^X$ as

$$T(\beta, g) = \text{ulev}(\varphi + g; X; \beta),\ \beta \in R^0_+,\ g \in \mathscr{A}.$$

As precise, this may be viewed as a family of maps $(T(\beta) : \mathscr{A} \to (2)^X; \beta > 0)$, according to

$$T(\beta)(g) = T(\beta, g),\ \beta \in R^0_+,\ g \in \mathscr{A}.$$

From the developments above, $\varphi + g$ is d-lsc over X; so that, $T(\beta)$ is nonempty closed valued, for each $\beta > 0$. On the other hand, the (partial) map $\beta \mapsto T(\beta)$ is increasing, in the sense

$$T(\beta_1)(g) \subseteq T(\beta_2)(g), \text{ whenever } 0 < \beta_1 \leq \beta_2,\ g \in \mathscr{A}.$$

Finally, the limiting map $T(0+) : \mathscr{A} \to 2^X$ attached to this family (see above) has the representation

$$T(0+)(g) = \cap\{T(\beta)(g); \beta > 0\} = \text{ulev}(\varphi + g; X; 0),\ g \in \mathscr{A}.$$

Clearly, $\text{Dom}(T(0+))$ is just the set of all $g \in \mathscr{A}$ where the optimization problem

$$(X; d; \varphi + g):\ \inf\{(\varphi + g)(x); x \in X\}$$

has global minimum solutions. Note that the alternative

$$T(0+) \text{ is not proper: } \text{Dom}(T(0+)) = \emptyset$$

cannot be avoided; but, even if

$$T(0+) \text{ is proper: } \text{Dom}(T(0+)) \neq \emptyset$$

we must emphasize that, given $g \in \text{Dom}(T(0+))$, the extremal representation of $T(0+)(g)$ relative to $(T(\beta)(g); \beta > 0)$ cannot be viewed as a limit one (from a topological perspective). So, we may ask under which (extra) regularity

conditions about our family $(T(\beta)(.);\ \beta > 0)$ is this possible. The needed conditions are just the ones encountered in our previous developments. Precisely: whenever $(T(\beta);\ \beta > 0)$ is d-fragmentable, the Nondiscrete Lassonde-Revalski Metrical Principle (LR-ndiscr-m) is applicable to these data; and the desired conclusion follows.

(B) Passing to sufficient conditions under which the underlying property holds, we start with

(d03) X is d-unbounded: $\mathrm{diam}(X) = \infty$.

For example, this holds when X is a linear space and its ambient metric $d(.,.)$ is generated by a norm; we do not give details.

Under this framework, put for $x \in X$, $\varepsilon > 0$

$$X^c(x,\varepsilon) = X \setminus X(x,\varepsilon),\ X^c[x,\varepsilon] = X \setminus X[x,\varepsilon];$$

these are just the complements of the open/closed sphere centered at x with radius ε. Note that, by the unboundedness condition,

$$X^c(x,\varepsilon) \neq \emptyset \text{ and } X^c[x,\varepsilon] \neq \emptyset, \text{ for each } \varepsilon > 0;$$

since the verification is immediate, we omit it.

Proposition 12 *Let the general completeness, subordination and unboundedness conditions above be accepted; and the extra requirements below be considered*

(extra-1) for each (X,d)-regular function $\psi : X \to R \cup \{\infty\}$
and each $\varepsilon > 0$, there exist $h \in \mathscr{A}$ and $\alpha > 0$ such that
$||h||_{\mathscr{A}} \leq \varepsilon$ and $\mathrm{diam}[\mathrm{ulev}(\psi + h; X; \alpha)] < \varepsilon$
(extra-2) for each $\varepsilon > 0$, there exists $\delta > 0$, such that:
for each $y \in X$, there exists $g \in \mathscr{A}$, with the properties
$||g||_{\mathscr{A}} = 1$ and $g(y) + \delta \leq \inf g(X^c(y,\varepsilon))$.

Then, the following are valid:

(43-1) (extra-2) implies (extra-1)
(43-2) if (extra-1) holds, then $(T(\beta);\ \beta > 0)$ is d-fragmentable.

The initial argument may be found in Lassonde and Revalski [21]. For completeness reasons, we provide its proof [which differs in part from the original one].

Proof (Proposition 12) There are two steps to be passed.

(i) Suppose that (extra-2) holds; and fix some (X,d)-regular function $\psi : X \to R \cup \{\infty\}$, as well as some $\varepsilon > 0$. Put $\eta := \varepsilon/3$; and let $\delta > 0$ be the number associated to it by (extra-2). Denote $\alpha := \eta\delta/2$; and pick $y \in X$ so that

$$\psi(y) < \inf \psi(X) + \alpha.$$

Further, let $g \in \mathscr{A}$ be the function attached to (η, δ, y) via (extra-2); and put $h = \eta g$. We claim that the couple (h, α) is the desired one for (extra-1). In fact, let $x \in \mathrm{ulev}(\psi + h; X; \alpha)$ be arbitrary fixed. By definition (and the above relations)

$$\inf \psi(X) + h(x) \leq (\psi + h)(x) \leq \inf(\psi + h)(X) + \alpha \leq$$
$$(\psi + h)(y) + \alpha < \inf \psi(X) + h(y) + 2\alpha.$$

This, along with $-\infty < \inf \psi(X) < \infty$, gives

$$h(x) < h(y) + 2\alpha \leq h(y) + \eta\delta; \text{ hence, } g(x) < g(y) + \delta;$$

so that, by the choice of g [in (extra-2)]

$$g(x) < \inf g(X^c(y, \eta)); \text{ wherefrom } x \in X(y, \eta).$$

This (by the arbitrariness of x in $\text{ulev}(\psi + h; X; \alpha)$), tells us that

$$\text{ulev}(\psi + h; X; \alpha) \subseteq X(y, \eta); \text{ whence } \text{diam}[\text{ulev}(\psi + h; X; \alpha)] \leq 2\eta < \varepsilon.$$

Finally, $h = \eta g$ is an element of $\mathscr{A}$, with $||h||_{\mathscr{A}} = \eta < \varepsilon$. Putting these together, it follows that (extra-1) holds.

(ii) Suppose that (extra-1) holds; and let the nonempty subset U of $\mathscr{A}$ as well as the number $\varepsilon > 0$ be arbitrary fixed. Fix some $f \in U$, $\gamma > 0$ according to

$$U_0 := \mathscr{A}(f, \gamma) \subseteq U; \text{ where } \mathscr{A}(f, \gamma) = \{h \in \mathscr{A}; ||f - h||_{\mathscr{A}} < \gamma\};$$

without loss, one may work with U_0 in place of U. Further, put $\eta := (1/2)\min\{\varepsilon, \gamma\}$. By (extra-1) there exists, for the (X, d)-regular function $\psi + f$ and the number $\eta > 0$, some element $g \in \mathscr{A}$ and some number $\alpha > 0$ with the properties

$$||g||_{\mathscr{A}} \leq \eta \text{ and } \text{diam}[\text{ulev}(\psi + f + g; X; \alpha)] < \eta.$$

Let the number $\delta > 0$ be taken as:

$$\delta < \min\{\gamma/2, \alpha/4\mu\} \text{ (where } \mu > 0 \text{ is given by (d02))};$$

and put $V_0 = \mathscr{A}(f + g, \delta)$. By the triangular property, one gets (via $\eta \leq \gamma/2$)

$$V_0 \subseteq \mathscr{A}(f, \delta + \eta) \subseteq \mathscr{A}(f, \gamma) = U_0.$$

Moreover, in view of

$$\alpha/2 + 2\delta\mu < \alpha/2 + \alpha/2 = \alpha,$$

one derives (by an auxiliary fact above)

$$\text{ulev}(\varphi + k; X; \alpha/2) \subseteq \text{ulev}(\varphi + f + g; X; \alpha), \ \forall k \in V_0;$$

so that (by the imposed notations)

$$T(\alpha/2)(V_0) \subseteq \text{ulev}(\varphi + f + g; X; \alpha).$$

Taking a preceding relation into account, yields (by the monotone property)

$$(V_0 \subseteq U_0 \text{ and}) \ \mathrm{diam}[T(\beta)(V_0)] < \eta < \varepsilon, \ \text{whenever } 0 < \beta \leq \alpha/2;$$

so that, $(T(\beta);\ \beta > 0)$ is d-fragmentable. The proof is thereby complete.

As a consequence of this, the following generic statement (referred to as: Metrical Lassonde-Revalski Variational Principle; in short: (LR-VP-m)) is available. Let (X, d) be a complete metric space fulfilling (d03). Further, let $(\mathscr{A}, ||.||_{\mathscr{A}})$ be a subspace of the generalized normed space $(\mathscr{C}(X), ||.||_\infty)$, with the properties (d01) and (d02). Finally, let the function $\varphi : X \to R \cup \{\infty\}$ be (X, d)-regular.

Theorem 4 *Suppose (in addition) that one of the extra conditions (extra-1) or (extra-2) is holding. Then, there exists a dense $\mathscr{G}_\delta$ type subset $\mathscr{B}$ of $\mathscr{A}$ such that for each $g \in \mathscr{B}$, the perturbed function $\varphi + g$ is (X, d)-admissible; i.e.*

(41-a) $\mathrm{ulev}(\varphi + g; X; 0)$ *is a (nonempty) singleton,* $\{x_\infty(g)\}$
(41-b) $\lim_{\beta \to 0+} \Theta(\mathrm{ulev}(\varphi + g; X; \beta), x_\infty(g)) = 0$*; so, any $(\varphi + g)$-minimizing sequence (x_n) in X [i.e.:* $(\varphi + g)(x_n) \to \inf(\varphi + g)(X) = (\varphi + g)(x_\infty(g))$*] fulfills* $x_n \overset{d}{\longrightarrow} x_\infty(g)$.

Proof Denote, for simplicity,

$$T(\beta, g) = \mathrm{ulev}(\varphi + g; X; \beta),\ \beta > 0,\ g \in \mathscr{A}.$$

This yields a family of nonempty closed valued maps $(T(\beta) : \mathscr{A} \to (2c)^X;\ \beta > 0)$, according to

$$T(\beta)(g) = T(\beta, g),\ \beta > 0,\ g \in \mathscr{A}.$$

In addition, by this very definition, $(T(\beta);\ \beta > 0)$ is increasing:

$$T(\beta_1)(g) \subseteq T(\beta_2)(g), \text{ whenever } 0 < \beta_1 \leq \beta_2,\ g \in \mathscr{A}.$$

Denote the limiting map attached to this family as

$$T(0+)(g) = \cap\{T(\beta)(g);\ \beta > 0\} = \mathrm{ulev}(\varphi; X; 0).$$

From the auxiliary statement above, $(T(\beta);\ \beta > 0)$ is d-fragmentable. So, by the Nondiscrete Lassonde-Revalski Metrical Principle (LR-ndiscr-m), there exists a dense $\mathscr{G}_\delta$ type subset $\mathscr{B}$ of $\mathscr{A}$, with

(41-c) $\mathrm{Dom}(T(0+))$ is $\mathscr{B}$-including $(\mathrm{Dom}(T(0+)) \supseteq \mathscr{B})$

such that, for each $g \in \mathscr{B}$ [hence, $g \in \mathrm{Dom}(T(0+))$], the following conclusions hold:

(41-d) $T(0+)(g)$ is a singleton $\{x_\infty(g)\}$
(41-e) $\lim_{\beta \to 0+} \Theta(T(\beta)(g), x_\infty(g)) = 0$.

But, these are just the conclusions in the statement.

Finally, note that conclusion of (LR-VP-m) is no longer retainable if the (complete) subspace $\mathscr{A}$ of $\mathscr{C}(X)$ is to be substituted by a *(closed) convex cone* $\mathscr{A}_+$ of the same (endowed with usual mild properties). This, ultimately, shows that a quasi-order version of (LR-VP-m) under the lines in Turinici [29] is not available here; further aspects will be provided elsewhere.

5 (LR-VP-m) Implies (DGZ)

In the following, we show that the above established Metrical Lassonde-Revalski Variational Principle (LR-VP-m) may be used to get (a certain bump version of) the Generic Smooth Variational Principle in Deville et al. [13].

(A) Let $(X, ||.||)$ be a Banach space; hence, if d stands for the associated metric

$$d(x, y) = ||x - y||, x, y \in X,$$

we have that

(X, d) is complete and X is d-unbounded.

Denote

$\mathscr{C}(X)$=the (real) vector space of all continuous $f \in \mathscr{F}(X, R)$.

Remember that, with the *supremum* generalized norm

$$||f||_\infty = \sup\{|f(x)|; x \in X\}, \ f \in \mathscr{C}(X),$$

$\mathscr{C}(X)$ becomes a generalized Banach space; equivalently, this means: $(\mathscr{C}(X), D_\infty)$ is a complete generalized metric space; where

$$D_\infty(f, g) = ||f - g||_\infty, \ f, g \in \mathscr{C}(X) \text{ (the } \textit{supremum} \text{ generalized metric).}$$

Further, take a linear subspace $\mathscr{A}$ of $\mathscr{C}(X)$. Let $||.||_{\mathscr{A}}$ be a certain norm over $\mathscr{A}$; and $D_{\mathscr{A}}$ stand for its associated metric

$$D_{\mathscr{A}}(f, g) = ||f - g||_{\mathscr{A}}, f, g \in \mathscr{A}.$$

The basic regularity conditions to be accepted are our old ones:

(e01) (completeness):
$(\mathscr{A}, ||.||_{\mathscr{A}})$ is a Banach space; or, equivalently: $(\mathscr{A}, D_{\mathscr{A}})$ is complete
(e02) (subordination):
there exists $\mu > 0$, with $(\mu||g||_{\mathscr{A}} \geq ||g||_\infty, \forall g \in \mathscr{A})$.

As precise, a basic consequence of the second one is the following. Denote

$\mathscr{C}_b(X)$=the linear subspace of all bounded $f \in \mathscr{C}(X)$.

Note that

$(\mathscr{C}_b(X), ||.||_\infty)$ is a (standard) Banach space;
hence, $(\mathscr{C}_b(X), D_\infty)$ is a (standard) complete metric space.

By the quoted condition, it results that

$\mathscr{A}$ is a subset of $\mathscr{C}_b(X)$; i.e.: each $g \in \mathscr{A}$ is bounded over X.

To state the remaining (specific) conditions, we need a convention. For each $\lambda \in R_+^0$, $u \in X$, define the *affine transformations* of $\mathscr{C}(X)$

$$J(\lambda, u)g(x) = g(\lambda x + u),\ x \in X,\ g \in \mathscr{C}(X).$$

Let us now accept the compatible type requirements:

(e03) (functional invariance)
$\forall \lambda \in R^0, \forall u \in X$: $J(\lambda, u)(\mathscr{A}) \subseteq \mathscr{A}$,
(e04) (norm conservation)
$\forall u \in X, \forall g \in \mathscr{A}$: $||J(1, u)g||_{\mathscr{A}} = ||g||_{\mathscr{A}}$.

For the last condition to be stated, some preliminaries are needed. Define, for each function $b \in \mathscr{C}(X)$

$$K(b) = b^{-1}(R \setminus \{0\}),\ \mathrm{supp}(b) = \mathrm{cl}(K(b));$$

these will be referred to as the *semi-support* and *support* of b, respectively.

The following simple properties of these concepts will be useful for us.

Proposition 13 *Let $b \in \mathscr{C}(X)$ be a function; and, given $\lambda > 0$, $u \in X$, put*

$$c := J(\lambda, u)b \text{ (i.e.: } c(x) = b(\lambda x + u),\ x \in X).$$

Then,

(51-1) $K(c) = (1/\lambda)(K(b) - u)$
(51-2) $\mathrm{supp}(c) = (1/\lambda)(\mathrm{supp}(b) - u)$.

Proof

(i) Let $x \in K(c)$ be arbitrary fixed. By definition, this means

$$c(x) = b(\lambda x + u) \neq 0;\ \text{i.e., } \lambda x + u \in K(b);$$

and, from this $x \in (1/\lambda)(K(b) - u)$. Conversely, let $y \in (1/\lambda)(K(b) - u)$ be arbitrary fixed. This gives

$$\lambda y + u \in K(b);\ \text{i.e., } c(y) = b(\lambda y + u) \neq 0;$$

which tells us that $y \in K(c)$.

(ii) Evident, by the properties of closure operator attached to $(X, ||.||)$.

(B) Call $b \in \mathscr{C}_b(X)$, a *bump function* if

$K(b)$ is nonempty bounded;
or, equivalently:
$\mathrm{supp}(b) = \mathrm{cl}(K(b))$ is nonempty bounded.

Now, the last condition to be added reads:

(e05) (bump-function):
there exists a bump function $b \in \mathscr{A}$.

When the requirements (e01)–(e04) and (e05) hold, we say that $(\mathscr{A}, ||.||_{\mathscr{A}})$ is an *admissible* subspace of $\mathscr{C}_b(X)$.

Finally, let us remember that the function $\psi : X \to R \cup \{\infty\}$ is called (X, d)-*regular*, when

(reg-1) ψ is proper ($\mathrm{Dom}(\psi) := \{x \in X;\ \psi(x) < \infty\} \neq \emptyset$)
(reg-2) ψ is bounded below ($\inf \psi(X) > -\infty$)
(reg-3) ψ is d-lsc: ($\liminf_n \psi(x_n) \geq \psi(x)$, whenever $x_n \xrightarrow{d} x$).

Denote, for each $\beta \geq 0$

$$\mathrm{ulev}(\psi; X; \beta) = \{x \in X;\ \psi(x) \leq \inf \psi(X) + \beta\};$$

this will be referred to as the β-*upper level set* of ψ with respect to X; note that, $\mathrm{ulev}(\psi; X; 0)$ is nothing else than the set of global minimum points of ψ over X. We say that ψ is (X, d)-*admissible*, when the attached optimization problem

$$(X, d; \psi):\ \inf\{\psi(x);\ x \in X\}$$

is *well posed*, in the sense

(wp-1) $\mathrm{ulev}(\psi; X; 0)$ is a (nonempty) singleton, $\{x_\infty\}$
(wp-2) $\lim_{\beta \to 0+} \Theta(\mathrm{ulev}(\psi; X; \beta), x_\infty) = 0$; hence (in particular) any minimizing sequence (x_n) in X (in the sense: $\psi(x_n) \to \inf \psi(X) = \psi(x_\infty)$) fulfills $x_n \xrightarrow{d} x_\infty$.

Under these facts, the following generic type statement (referred to as: the Bump Deville-Godefroy-Zizler Variational Principle (in short: (DGZ-b)) is available. Let $(X, ||.||)$ be a Banach space; and $(\mathscr{A}, ||.||_{\mathscr{A}})$ be be an admissible subspace of $\mathscr{C}_b(X)$.

Theorem 5 *Let the function $\varphi : X \to R \cup \{\infty\}$ be (X, d)-regular. Then, there exists a dense $\mathscr{G}_\delta$ type subset $\mathscr{B}$ of $\mathscr{A}$, such that: for each $g \in \mathscr{B}$, the perturbed function $\varphi + g$ is (X, d)-admissible; i.e.*

(51-a) $\mathrm{ulev}(\varphi + g; X; 0)$ *is a (nonempty) singleton,* $\{x_\infty(g)\}$
(51-b) $\lim_{\beta \to 0+} \Theta(\mathrm{ulev}(\varphi+g; X; \beta), x_\infty(g)) = 0$*; so, any minimizing sequence* (x_n) *in X [i.e.:* $(\varphi + g)(x_n) \to \inf(\varphi + g)(X) = (\varphi + g)(x_\infty(g))$*]*
fulfills $x_n \xrightarrow{d} x_\infty(g)$.

Proof We show that the Metrical Lassonde-Revalski Variational Principle (LR-VP-m) is applicable here; precisely, that one of the extra conditions below holds:

(extra-1) for each (X, d)-regular function $\psi : X \to R \cup \{\infty\}$
and each $\varepsilon > 0$, there exist $h \in \mathscr{A}$ and $\alpha > 0$ such that
$||h||_{\mathscr{A}} \leq \varepsilon$ and $\mathrm{diam}[\mathrm{ulev}(\psi + h; X; \alpha)] < \varepsilon$
(extra-2) for each $\varepsilon > 0$, there exists $\delta > 0$, such that: for each $y \in X$, there exists $g \in \mathscr{A}$, with $||g||_{\mathscr{A}} = 1$ and $g(y) + \delta \leq \inf g(X^c(y, \varepsilon))$.

In fact, the latter of these is more appropriate here. For, let $b \in \mathscr{A}$ be the bump function assured by (e05); hence,

$K(b) := b^{-1}(R \setminus \{0\})$ is nonempty bounded;
or, equivalently:
$\text{supp}(b) := \text{cl}(K(b))$ is nonempty bounded.

This, by definition, gives (passing to $b_1 := -b \in \mathscr{A}$ if needed)

$K(b) \subseteq X[0, \lambda]$, for some $\lambda > 0$
$b(u) > 0$, for at least one $u \in K(b)$.

Passing to the bump function (in $\mathscr{A}$) $b_2 = J(1, u)b$, it results that (cf. a previous auxiliary fact)

$$K(b_2) = K(b) - u \subseteq X[0, 2\lambda] \text{ and } b_2(0) = b(u) > 0.$$

Further, passing to the bump function (in $\mathscr{A}$) $b_3 = J(4\lambda, 0)b_2$, we get

$$K(b_3) = (1/4\lambda)K(b_2) \subseteq X[0, 1/2], \text{ and } b_3(0) = b_2(0) > 0.$$

Finally, taking the bump function (in $\mathscr{A}$)

$$b_4 = (1/\nu)b_3, \text{ where } \nu := b_3(0),$$

we have the evaluations

$$K(b_4) = K(b_3) \subseteq X[0, 1/2] \text{ and } b_4(0) = (1/\nu)b_3(0) = 1.$$

Put $c = b_4$; we show that it is essential for solving our question. In fact, let $\varepsilon > 0$ be arbitrary fixed. The bump function (in $\mathscr{A}$) $c_1 := J(1/\varepsilon, 0)c$ fulfills

$$K(c_1) = \varepsilon K(c) \subseteq X[0, \varepsilon/2], \text{ and } c_1(0) = c(0) = 1.$$

Let $y \in X$ be arbitrary fixed. The bump function (in $\mathscr{A}$) $c_2 = J(1, -y)c_1$ has the properties (cf. the choice of $\mathscr{A}$)

$$K(c_2) = K(c_1) + y \subseteq X[y, \varepsilon/2],\ ||c_2||_{\mathscr{A}} = ||c_1||_{\mathscr{A}},\ c_2(y) = c_1(0) = 1.$$

Finally, the bump function (in $\mathscr{A}$)

$$c_3 = -\beta c_2, \text{ where } \beta = 1/||c_2||_{\mathscr{A}}$$

has the properties

$K(c_3) = K(c_2) \subseteq X[y, \varepsilon/2]$, $||c_3||_{\mathscr{A}} = 1$, $c_3(y) + \beta = 0$;
whence, $\text{supp}(c_3) \subseteq X[y, \varepsilon/2] \subseteq X(y, \varepsilon)$.

This tells us that the desired condition holds with $g = c_3$, $\delta = \beta$. And then, by the underlying variational principle, we are done.

(C) In the following, an (extended) semi-bump counterpart of this principle is to be considered. Remember that we introduced the notations

$$X^c(x, \varepsilon) = X \setminus X(x, \varepsilon),\ X^c[x, \varepsilon] = X \setminus X[x, \varepsilon],\ x \in X, \varepsilon > 0;$$

these are just complements of the open/closed sphere centered at x with radius ε. Note that, in our normed context, these subsets are always nonempty; further properties of them are contained in

Proposition 14 *Under these conventions, we have, for* $\lambda, \varepsilon > 0$ *and* $u, y \in X$

(52-1) $\lambda X^c(0,\varepsilon) + u \subseteq X^c(0, \lambda\varepsilon - \|u\|)$, *whenever* $\lambda\varepsilon - ||u|| > 0$
(52-2) $X^c(y,\varepsilon) - y = X^c(0,\varepsilon)$.

Proof

(i) Let $x \in \lambda X^c(0,\varepsilon) + u$ be arbitrary fixed. By definition, this gives

$(1/\lambda)(x-u) \in X^c(0,\varepsilon)$; hence $(1/\lambda)||x-u|| \geq \varepsilon$.

But then (from the triangular property),

$||x|| \geq \lambda\varepsilon - ||u|| > 0$; that is: $x \in X^c(0, \lambda\varepsilon - \|u\|)$.

(ii) Evident.

Now, given the bounded function $g : X \to R$, denote

$\liminf_{x\to\infty} g(x) = \sup_{\varepsilon>0} \inf g(X^c(0,\varepsilon))$,
$\limsup_{x\to\infty} g(x) = \inf_{\varepsilon>0} \sup g(X^c(0,\varepsilon))$.

Note that, from the boundedness property,

$$-\infty < \liminf_{x\to\infty} g(x) \leq \limsup_{x\to\infty} g(x) < \infty.$$

Generally, these (finite) numbers may be distinct; if equality holds, we denote

$\lim_{x\to\infty} g(x)$=the common value of underlying limits.

A natural question is that of these functional operations being invariant with respect to the affine transformations $J(.,.)$. The answer to this is positive; as results from

Proposition 15 *Under these conventions, we have, for* $\lambda > 0$, $u \in X$, $g \in \mathscr{C}_b(X)$

(53-1) $\liminf_{x\to\infty} J(\lambda,u)g(x) = \liminf_{x\to\infty} g(x)$
(53-2) $\limsup_{x\to\infty} J(\lambda,u)g(x) = \limsup_{x\to\infty} g(x)$
(53-3) $\lim_{x\to\infty} J(\lambda,u)g(x) = \lim_{x\to\infty} g(x)$,
whenever the right member exists (in R).

Proof It will suffice verifying the first part. So, let $\lambda > 0$ and $u \in X$ be fixed in the sequel. Define the function $\delta : R_+^0 \to R$ as

$\delta(\varepsilon) = \lambda\varepsilon - ||u||$, $\varepsilon > 0$.

Further, given $g \in \mathscr{C}_b(X)$, denote

$h = J(\lambda,u)g$; i.e.: $h(x) = g(\lambda x + u)$, $x \in X$.

Let $\varepsilon > (1/\lambda)||u||$ be arbitrary fixed. By the auxiliary facts above,

$$h(X^c(0,\varepsilon)) = g(\lambda X^c(0,\varepsilon) + u) \subseteq g(X^c(0,\delta(\varepsilon)));$$

wherefrom,

$\inf h(X^c(0,\varepsilon)) \geq \inf g(X^c(0,\delta(\varepsilon))), \varepsilon > (1/\lambda)||u||.$

Passing to supremum over all such ε, one gets

$\liminf_{x\to\infty} h(x) \geq \liminf_{x\to\infty} g(x).$

The reciprocal inequality is also true, in view of

$g = J(\mu, v)h$, where $\mu := 1/\lambda$, $v = -(1/\lambda)u$;

and then, all is clear.

Under the precise facts, let us say that $b \in \mathscr{C}_b(X)$ is a *semi-bump* function, if

$\limsup_{x\to\infty} b(x) < \sup b(X).$

Note that, in view of

$\sup b(X^c(0,\varepsilon)) \leq \sup b(X), \varepsilon > 0,$

we have

$$\limsup_{x\to\infty} b(x) \leq \sup b(X),\ b \in \mathscr{C}_b(X);$$

so, this relation is "almost" fulfilled by every element $b \in \mathscr{C}_b(X)$.

Denote the class of all these functions as $\mathscr{C}_b^s(X)$. Some basic properties of it are being collected in

Proposition 16 *The class $\mathscr{C}_b^s(X)$ is invariant to affine transformations:*

(54-1) $b \in \mathscr{C}_b^s(X) \Longrightarrow J(\lambda, u)b \in \mathscr{C}_b^s(X), \forall \lambda \in R_+^0, \forall u \in X.$

In addition, we have

(54-2) $b \in \mathscr{C}_b^s(X) \Longrightarrow \mu b + \theta \in \mathscr{C}_b^s(X), \forall \mu \in R_+^0, \forall \theta \in R.$

The verification is immediate, by the auxiliary facts above; so, we do not give further details.

To see the effectiveness of this notion, the following facts are useful. Call $b \in \mathscr{C}_b(X)$, a *strong semi-bump* function if

$K(b) := b^{-1}(R \setminus \{0\})$ (the *semi-support* of b) is nonempty,
and $\lim_{x\to\infty} b(x) = 0$.

The following inclusion holds (over $\mathscr{C}_b(X)$):

$$b \text{ is strong semi-bump } \Longrightarrow b \text{ or } (-b) \text{ is semi-bump.}$$

But, the reciprocal is not in general true: just take $X = R$ (endowed with the usual norm) and the function

$b(t) = 1 + |t|/(1 + t^2), t \in R.$

As another example, remember that $b \in \mathscr{C}_b(X)$ is called a *bump function* if

$K(b) := b^{-1}(R \setminus \{0\})$ is nonempty bounded; or, equivalently:
$\mathrm{supp}(b) := \mathrm{cl}(K(b))$ (the *support* of b) is nonempty bounded.

Clearly, we have the inclusion (over $\mathscr{C}_b(X)$):

$$b \text{ is bump} \implies b \text{ is strong semi-bump.}$$

The reciprocal is not in general true: just take $X = R$ (endowed with the usual norm) and the function

$b(t) = |t|/(1 + t^2), t \in R.$

Now, the last conditions to be added reads:

(e06) there exists a semi-bump function $b \in \mathscr{A}$.

When the requirements (e01)–(e04) and (e06) hold, we say that $(\mathscr{A}, ||.||_{\mathscr{A}})$ is a *semi-admissible* subspace of $\mathscr{C}_b(X)$.

Some useful properties of this last class are described in

Proposition 17 *Assume that $(\mathscr{A}, ||.||_{\mathscr{A}})$ is a semi-admissible subspace of $\mathscr{C}_b(X)$; and let $\varepsilon > 0$, $\alpha > 0$ be arbitrary fixed. Then,*

(55-1) there exists a semi-bump function $b \in \mathscr{A}$, with
$||b||_{\mathscr{A}} < \varepsilon$, $\sup b(X^c(0, \alpha)) < b(0)$
(55-2) for each $y \in X$, the semi-bump (in $\mathscr{A}$) function $e := J(1, -y)b$
fulfills $||e||_{\mathscr{A}} = ||b||_{\mathscr{A}} < \varepsilon$, $\sup e(X^c(y, \alpha)) = \sup b(X^c(0, \alpha))$, $e(y) = b(0)$;
hence, $\sup e(X^c(y, \alpha)) < e(y)$.

Proof By hypothesis, there are semi-bump functions in $\mathscr{A}$; let $a(.)$ be one of these:

$\limsup_{x \to \infty} a(x) < \sup a(X).$

From the very definition of our involved concepts,

$\sup a(X^c(0, \beta)) < a(u)$, for some $\beta > 0, u \in X$.

Passing to the semi-bump (in $\mathscr{A}$) function $a_1 := J(1, u)a$, this becomes

$\sup a_1(X^c(0, \delta)) = \sup a(X^c(0, \delta) - u)) \leq \sup a(X^c(0, \beta))$
$< a(u) = a_1(0)$; where $\delta := ||u|| + \beta$.

As a consequence, the semi-bump (in $\mathscr{A}$) function

$a_2 := J(\gamma, 0)a_1$, where $\gamma := \delta/\alpha$,

fulfills the property

$$\sup a_2(X^c(0,\alpha)) = \sup a_1(\gamma X^c(0,\alpha)) \le \sup a_1(X^c(0,\delta)) < a_1(0) = a_2(0).$$

It will suffice then passing to the semi-bump (in $\mathscr{A}$) function

$$b = (\varepsilon/2\nu)a_2, \text{ where } \nu := ||a_2||_{\mathscr{A}} > 0,$$

to get the first conclusion. The second one follows at once by the semi-admissible condition, and the preceding auxiliary facts.

Finally, let $\psi : X \to R \cup \{\infty\}$ be a (X,d)-*regular* function; i.e.:

(reg-1) φ is proper ($\mathrm{Dom}(\varphi) := \{x \in X; \varphi(x) < \infty\} \neq \emptyset$)
(reg-2) φ is bounded below ($\inf \varphi(X) > -\infty$)
(reg-3) φ is d-lsc ($\liminf_n \varphi(x_n) \ge \varphi(x)$, whenever $x_n \xrightarrow{d} x$).

Denote, for each $\beta \ge 0$

$$\mathrm{ulev}(\psi; X; \beta) = \{x \in X; \psi(x) \le \inf \psi(X) + \beta\};$$

this will be referred to as the β-*upper level set* of ψ with respect to X; note that, $\mathrm{ulev}(\psi; X; 0)$ is nothing else than the set of global minimum points of ψ over X. Remember that ψ is called (X,d)-*admissible*, when the optimization problem attached to these data

$$(X, d; \psi):\ \inf\{\psi(x); x \in X\}$$

is *well posed*, in the sense:

(wp-1) $\mathrm{ulev}(\psi; X; 0)$ is a (nonempty) singleton, $\{x_\infty\}$
(wp-2) any minimizing sequence (x_n) in X
(i.e.: $\psi(x_n) \to \inf \psi(X) = \psi(x_\infty)$) fulfills $x_n \xrightarrow{d} x_\infty$.

A related concept of this type may be introduced as follows. Let us say that ψ is (X,d)-*sequentially-admissible*, when

(sa-1) there exists $(x_n; n \ge 0)$ in $\mathrm{Dom}(\psi)$ and $x_\infty \in \mathrm{Dom}(\psi)$ with
$x_n \xrightarrow{d} x_\infty$ and $\limsup_n \psi(x_n) \le \inf \psi(X \setminus \{x_\infty\})$
(sa-2) if $(y_n; n \ge 0)$ in $\mathrm{Dom}(\psi)$ fulfills $\psi(y_n) \to \psi(x_\infty)$, then $y_n \xrightarrow{d} x_\infty$.

The connection between these admissible concepts is described as follows.

Proposition 18 *For each (X,d)-regular function ψ, we have in (ZF-AC+DC),*

$$(X,d)\text{-sequentially-admissible} \iff (X,d)\text{-admissible.}$$

Precisely, the following conclusions hold:

(56-1) if ψ is (X,d)-sequentially-admissible, then the point x_∞ from (sa-1) is the unique element of $\mathrm{ulev}(\psi; X; 0)$*; so that, ψ is (X,d)-admissible*
(56-2) if ψ is (X,d)-admissible, then (sa-1) holds, with x_∞ taken as the unique element in $\mathrm{ulev}(\psi; X; 0)$*; so that, ψ is (X,d)-sequentially-admissible.*

Proof

(I) Suppose that ψ is (X, d)-sequentially-admissible; hence, (sa-1)+(sa-2) hold. From the former of these, one gets (as ψ is d-lsc)

$$\psi(x_\infty) \leq \liminf_n \psi(x_n) \leq \limsup_n \psi(x_n) \leq \inf \psi(X \setminus \{x_\infty\});$$

wherefrom $x_\infty \in \text{ulev}(\psi; X; 0)$. Moreover, from the latter of these, one gets

$$y \in \text{Dom}(\psi), \psi(y) = \psi(x_\infty) \Longrightarrow y = x_\infty;$$

and this yields $\text{ulev}(\psi; X; 0) = \{x_\infty\}$. Combining again with (sa-2), one derives that ψ is (X, d)-admissible.

(II) Suppose that ψ is (X, d)-admissible. By (wp-1), we formally have

$$\inf\psi(X) = \psi(x_\infty), \text{ for a uniquely determined } x_\infty \in X.$$

From the Denumerable Axiom of Choice (AC(N)) (deductible in (ZF-AC+DC); see above), we may pick some sequence (x_n) in X with

$$x_n \in \text{ulev}(\psi; X; 2^{-n}), n \geq 0; \text{ whence, } \lim_n \psi(x_n) = \inf \psi(X) = \psi(x_\infty).$$

Note that, as a direct consequence of the second equality,

$$\psi(x_\infty)(= \inf \psi(X)) \leq \inf \psi(X \setminus \{x_\infty\});$$

and this, along with the first equality, gives

$$\lim_n \psi(x_n) \leq \inf \psi(X \setminus \{x_\infty\});$$

wherefrom (sa-1) holds. On the other hand, (sa-2) holds as well, in view of (wp-2); so, combining these, one derives that ψ is (X, d)-sequentially-admissible. The proof is thereby complete.

Finally, letting the (X, d)-regular function function $\psi : X \to R \cup \{\infty\}$ (see above), denote for each $\alpha > 0$

$$\text{amin}(\psi; X; \alpha) = \{y \in X; \psi(y) < \inf \psi(X^c(y, \alpha))\};$$

each point y of this set will be referred to as *(spherically) almost minimum* for ψ over X. Note that the alternative

$$\text{amin}(\psi; X; \alpha) = \emptyset, \text{ for certain } \alpha > 0$$

cannot be avoided. On the other hand, if $y \in X$ is a global minimum point for ψ, then we must have

$$\psi(y) = \inf \psi(X) \leq \inf \psi(X^c(y, \alpha)), \text{ for each } \alpha > 0;$$

hence, y is "very close" to be (spherically) almost minimum for ψ over X.

Under these facts, the following generic type statement, comparable with the one in Deville et al. [13] (referred to as: the Semi-Bump Deville-Godefroy-Zizler

Variational Principle (in short: (DGZ-sb)) is available. Let $(X, ||.||)$ be a Banach space; and $(\mathscr{A}, ||.||_{\mathscr{A}})$ be a semi-admissible Banach subspace of $\mathscr{C}_b(X)$.

Theorem 6 *Let (in addition) $\varphi : X \to R \cup \{\infty\}$ be a (X, d)-regular function. There exists then, in (ZF-AC+DC), a subset $\mathscr{B}$ of $\mathscr{A}$, with*

(52-a) $\mathscr{B}$ is of $\mathscr{G}_\delta$ type and dense (modulo $D_{\mathscr{A}}$)
(52-b) for each $g \in \mathscr{B}$, $\varphi + g$ is (X, d)-sequentially-admissible
(or, equivalently: (X, d)-admissible).

Proof There are two steps to be passed.

Step 1 Denote, for simplicity

$$U(\alpha; \varphi) := \{g \in \mathscr{A}; \operatorname{amin}(\varphi + g; X; \alpha) \neq \emptyset\}, \alpha > 0.$$

The basic properties of these functional sets are concentrated in

Proposition 19 *The following conclusions hold, in the strongly reduced system (ZF-AC), for each $\alpha > 0$:*

(57-1) $U(\alpha; \varphi)$ is dense (modulo $D_{\mathscr{A}}$) in $\mathscr{A}$
(57-2) $U(\alpha; \varphi)$ is open (modulo $D_{\mathscr{A}}$) in $\mathscr{A}$.

Proof (Proposition 19) There are two parts to be verified.

(I) (Denseness). Let $\varepsilon > 0$ be arbitrary fixed. By a preceding result, there exists a semi-bump function $b(.)$ in $\mathscr{A}$, with the property

$$||b||_{\mathscr{A}} < \varepsilon, \sup b(X^c(0, \alpha)) < b(0).$$

Let $g \in \mathscr{A}$ be arbitrary fixed. Since $\varphi + g$ is (X, d)-regular, there exists $y \in X$ with

$$(\varphi + g)(y) < \inf(\varphi + g)(X) + b(0) - \sup b(X^c(0, \alpha)).$$

Passing to the semi-bump function $e := J(1, -y)b$, one has (by an auxiliary fact)

$$||e||_{\mathscr{A}} = ||b||_{\mathscr{A}}, e(y) = b(0), \sup e(X^c(y, \alpha)) = \sup b(X^c(0, \alpha));$$

so, replacing in the above relation,

$$(\varphi + g)(y) < \inf(\varphi + g)(X) + e(y) - \sup e(X^c(y, \alpha)).$$

Put $h = g - e$; we claim that it is our desired element. In fact, $g = h + e$ gives $\|g - h\|_{\mathscr{A}} = \|e\|_{\mathscr{A}} < \varepsilon$. Moreover, as $\varphi + g = \varphi + h + e$, one gets (via direct calculations)

$$\inf(\varphi + g)(X) \leq \inf(\varphi + g)(X^c(y, \alpha)) \leq$$
$$\inf(\varphi + h)(X^c(y, \alpha)) + \sup e(X^c(y, \alpha)).$$

On the other hand, the preceding relation (involving $e(.)$) gives

$$(\varphi + h)(y) = (\varphi + g)(y) - e(y) < \inf(\varphi + g)(X) - \sup e(X^c(y, \alpha)).$$

Combining these relations yields

$$(\varphi + h)(y) < \inf(\varphi + h)(X^c(y, \alpha)); \text{ hence, } h \in U(\alpha; \varphi).$$

(II) (Openness) Fix $g \in U(\alpha; \varphi)$; hence, there exists $y \in X$ so as

$(\varphi + g)(y) < \inf(\varphi + g)(X^c(y, \alpha))$;

note that, in particular, this yields

$$-\infty < \inf(\varphi + g)(X) \leq (\varphi + g)(y) < \infty; \text{ whence, } y \in \mathrm{Dom}(\varphi).$$

The case $X^c(y, \alpha) \cap \mathrm{Dom}(\varphi) = \emptyset$ yields (via $\mathscr{A} \subseteq \mathscr{C}_b(X)$)

$$(\varphi + h)(y) < \infty = \inf(\varphi + h)(X^c(y, \alpha)), \text{ for all } h \in \mathscr{A};$$

wherefrom $U(\alpha; \varphi) = \mathscr{A}$; so, without loss, one may assume that

$X^c(y, \alpha) \cap \mathrm{Dom}(\varphi) \neq \emptyset$; hence $-\infty < \inf(\varphi + g)(X^c(y, \alpha)) < \infty$.

Take the number $\delta > 0$ according to:

$(\varphi + g)(y) + 2\delta < \inf(\varphi + g)(X^c(y, \alpha))$;

and let $h \in \mathscr{A}$ be such that $||g - h||_{\mathscr{A}} < \delta/\mu$. By the subordination hypothesis

$||g - h||_\infty < \delta$; wherefrom $|g(x) - h(x)| < \delta$, for all $x \in X$.

As a consequence, we have (via $y \in \mathrm{Dom}(\varphi)$)

$$(\varphi + h)(y) < (\varphi + g)(y) + \delta < \inf(\varphi + g)(X^c(y, \alpha)) - \delta,$$
$$\inf(\varphi + g)(X^c(y, \alpha)) \leq \inf(\varphi + h)(X^c(y, \alpha)) + \delta.$$

Combining these gives

$$(\varphi + h)(y) < \inf(\varphi + h)(X^c(y, \alpha));$$

or, in other words, $h \in U(\alpha; \varphi)$.

Step 2 Fix a strictly descending sequence $(\gamma_n; n \geq 0)$ in R^0_+ with $\gamma_n \to 0$. [For example, one may take $\gamma_n = 2^{-n}$, $n \geq 0$]. By the previous part, each member of the sequence $(U(\gamma_n; \varphi); n \geq 0)$ is dense and open (modulo $D_{\mathscr{A}}$) in $\mathscr{A}$. Since $(\mathscr{A}, D_{\mathscr{A}})$ is complete, an application of the [deductible from (DC)] Baire Category Theorem (BC) assures us that $\mathscr{B} := \cap\{U(\gamma_n; \varphi); n \geq 0\}$ is a (nonempty) dense (modulo $D_{\mathscr{A}}$) $\mathscr{G}_\delta$ type subset of $\mathscr{A}$. We now claim that, for each $g \in \mathscr{B}$, the associated function $\varphi + g$ is (X, d)-sequentially-admissible (or, equivalently: (X, d)-admissible). In fact, as $[g \in U(\gamma_n; \varphi)$, for all $n]$, we have by definition that

$\mathrm{amin}(\varphi + g; X; \gamma_n)$ is nonempty, for each $n \geq 0$.

By the Denumerable Axiom of Choice (AC(N))—deductible, as already established, from (DC)—there exists then a sequence $(x_n; n \geq 0)$ in X such that, for each $n \geq 0$,

$x_n \in \text{amin}(\varphi + g; X; \gamma_n)$; i.e.: $(\varphi + g)(x_n) < \inf(\varphi + g)(X^c(x_n, \gamma_n))$;

note that, necessarily, $\{x_n; n \geq 0\} \subseteq \text{Dom}(\varphi)$. We claim that

$$||x_n - x_p|| \leq \gamma_n, \quad \text{for } n < p.$$

For, otherwise, there exists a couple (n, p) with

$n < p, ||x_n - x_p|| > \gamma_n (> \gamma_p)$;

and this, by the very definition of our sequence, gives

$$(\varphi + g)(x_n) < (\varphi + g)(x_p), \quad (\varphi + g)(x_p) < (\varphi + g)(x_n);$$

a contradiction; hence the claim. As a consequence of this, $(x_n; n \geq 0)$ is a d-Cauchy sequence; so (from the completeness of X)

$x_n \xrightarrow{d} x_\infty$ as $n \to \infty$, for some (uniquely determined) $x_\infty \in X$.

We claim that $(x_n; n \geq 0)$ and x_∞ are the desired data so as to verify the (X, d)-sequentially-admissible property for $\varphi + g$.

(sa-1) Let $x \in X \setminus \{x_\infty\}$ be arbitrary fixed. As

$||x_n - x|| - \gamma_n \to ||x_\infty - x|| > 0$ as $n \to \infty$,

there must be some rank $p = p(x) \geq 0$ such that

$$||x_n - x|| > \gamma_n \text{ (hence : } x \in X^c(x_n, \gamma_n)), \text{ for all } n \geq p.$$

By the very definition of these points,

$$(\varphi + g)(x_n) < \inf(\varphi + g)(X^c(x_n, \gamma_n)) \leq (\varphi + g)(x), \quad \forall n \geq p;$$

so that

$$\limsup_n (\varphi + g)(x_n) \leq (\varphi + g)(x).$$

This, by the arbitrariness of $x \in X \setminus \{x_\infty\}$, gives

$$\limsup_n (\varphi + g)(x_n) \leq \inf(\varphi + g)(X \setminus \{x_\infty\});$$

wherefrom, the first half of (X, d)-sequentially-admissible condition holds (with respect to $\varphi + g$). For, as a consequence of this, one gets (by the d-lsc property of φ)

$$(\varphi+g)(x_\infty) \le \liminf_n (\varphi+g)(x_n) \le \limsup_n (\varphi+g)(x_n) \le \inf(\varphi+g)(X\setminus\{x_\infty\});$$

and then

$(\varphi+g)(x_\infty) = \inf(\varphi+g)(X) < \infty$; i.e.: $x_\infty \in \mathrm{Dom}(\varphi)$ is a global minimum for $\varphi+g$.

(sa-2) Let (y_n) be a sequence in $\mathrm{Dom}(\varphi)$ with

$$(\varphi+g)(y_n) \to (\varphi+g)(x_\infty) = \inf(\varphi+g)(X).$$

Assume that (y_n) does not converge to x_∞. Passing to a subsequence if necessary, one may assume that, for some $\varepsilon > 0$,

$$||y_n - x_\infty|| \ge 2\varepsilon, \forall n; \quad ||x_p - x_\infty|| \le \varepsilon, \quad \forall p;$$

wherefrom (by the triangle inequality)

$$||y_n - x_p|| \ge ||y_n - x_\infty|| - ||x_p - x_\infty|| \ge \varepsilon, \quad \forall n, \forall p.$$

As a consequence, there exists a rank $p \ge 0$ such that

$||y_n - x_p|| > \gamma_p$ (whence, $y_n \in X^c(x_p, \gamma_p)$), for all n.

This yields (by the definition of x_p)

$$\inf(\varphi+g)(X) \le (\varphi+g)(x_p) < \inf(\varphi+g)(X^c(x_p,\gamma_p)) \le (\varphi+g)(y_n), \forall n;$$

wherefrom (passing to limit as $n \to \infty$)

$$\inf(\varphi+g)(X) < \inf(\varphi+g)(X^c(x_p,\gamma_p)) \le \inf(\varphi+g)(X);$$

contradiction. Hence, $y_n \xrightarrow{d} x_\infty$; and the proof is complete.

(D) Let again $(X, ||.||)$ be a Banach space; and $(\mathscr{A}, ||.||_{\mathscr{A}})$ be a semi-admissible Banach subspace of $\mathscr{C}_b(X)$. The following "plane" version of the obtained statement (referred to as: Plane Semi-Bump Deville-Godefroy-Zizler Variational Principle (in short: (DGZ-psb)) is now available.

Theorem 7 *Let $(\varphi_i; i \ge 0)$ be a sequence of (X,d)-regular functions. There exists then, in (ZF-AC+DC), a subset $\mathscr{B}$ of $\mathscr{A}$, with*

(53-a) $\mathscr{B}$ is of $\mathscr{G}_\delta$ type and dense (modulo $D_{\mathscr{A}}$)
(53-b) for each $i \ge 0$ and $g \in \mathscr{B}$, $\varphi_i + g$ is (X,d)-sequentially-admissible (or, equivalently: (X,d)-admissible).

Proof (Sketch) Let $(\gamma_n; n \ge 0)$ be a strictly descending sequence in R^0_+ with $\lim_n \gamma_n = 0$. By the argument of (DGZ-sb) we just exposed, it follows that

$U(\gamma_n; \varphi_i)$ is dense and open in $\mathscr{A}$(modulo $D_{\mathscr{A}}$), for each $n \geq 0, i \geq 0$;

so that, by the Baire Category Theorem (BC), one gets

$\mathscr{B} := \cap\{U(\gamma_n; \varphi_i); n \geq 0, i \geq 0\}$ is of $\mathscr{G}_\delta$ type and dense in $\mathscr{A}$ (modulo $D_{\mathscr{A}}$).

Moreover, by the same argument of (DGZ-sb), we have that, for each $g \in \mathscr{B}$,

$$\varphi_i + g \text{ is } (X, d)\text{-sequentially-admissible, for all each } i \geq 0.$$

This gives us all desired facts.

Formally, (DGZ-psb) is an extension of (DGZ-sb); to which it reduces when the sequence $(\varphi_i; i \geq 0)$ is constant. But, as we shall see, the reciprocal inclusion holds too; hence, these statements are in fact equivalent to each other. Note, finally, that a similar plane construction is obtainable by starting from the Bump Deville-Godefroy-Zizler Variational Principle (DGZ-b); we do not give details.

6 (DGZ-b) Implies (EVP)

Now, as already precise in the introductory section, a complete ontological status of the obtained result is to be reached as long as (DGZ-b) includes (EVP). A positive answer is available; but, it will necessitate some preliminary facts.

(A) Let $(X, ||.||)$ be a Banach space; hence, if d stands for the associated metric

$d(x, y) = ||x - y||, x, y \in X,$

we have that (X, d) is complete. Denote

$\mathscr{C}(X)$=the (real) vector space of all continuous $f \in \mathscr{F}(X, R)$.

Remember that, with the *supremum* generalized norm

$||f||_\infty = \sup\{|f(x)|; x \in X\}, \ f \in \mathscr{C}(X),$

$\mathscr{C}(X)$ becomes a generalized Banach space; equivalently, this means: $(\mathscr{C}(X), D_\infty)$ is a complete metric space; where

$D_\infty(f, g) = ||f - g||_\infty, \ f, g \in \mathscr{C}(X)$ (the *supremum* generalized metric).

Further, let us introduce the functional set

$\mathscr{C}_b(X)$=the linear subspace of all bounded $f \in \mathscr{C}(X)$.

It is not hard to see that

$(\mathscr{C}_b(X), ||.||_\infty)$ is a (standard) Banach space;
hence, $(\mathscr{C}_b(X), D_\infty)$ is a (standard) complete metric space.

Having these precise, let K be a (nonempty) d-closed subset of X; clearly, (K, d) is complete too. Remember that $\varphi : K \to R \cup \{\infty\}$ is (K, d)-*regular*, when

(reg-1) φ is proper ($\mathrm{Dom}(\varphi) := \{x \in K; \varphi(x) < \infty\} \neq \emptyset$)
(reg-2) φ is bounded below ($\inf \varphi(K) > -\infty$)
(reg-3) φ is d-lsc ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \xrightarrow{d} x$).

The following statement (referred to as: the *normed* version of (EVP); and denoted as: (EVP-n)) is our starting point in these investigations.

Theorem 8 *Let the function $\varphi : K \to R \cup \{\infty\}$ be (K, d)-regular. Then, for each $u \in \mathrm{Dom}(\varphi)$ there exists $v \in \mathrm{Dom}(\varphi)$ in such a way that*

(61-a) $d(u, v) \leq \varphi(u) - \varphi(v)$ *(hence* $\varphi(u) \geq \varphi(v)$*)*
(61-b) $d(v, x) > \varphi(v) - \varphi(x)$, *for all* $x \in K \setminus \{v\}$.

The normed context we just used suggests us that an application of the preceding (normed) smooth variational principles must be possible here. A confirmation of this fact is contained in

Proposition 20 *We have in the strongly reduced system (ZF-AC),*

(DGZ-b) $\Longrightarrow$ *(EVP-n); hence (LR-VP-n)* $\Longrightarrow$ *(EVP-n).*

Proof Let $(\leq)$ stand for the Brøndsted quasi-order [7]

$(x, y \in K)$: $x \leq y$ iff $d(x, y) + \varphi(y) \leq \varphi(x)$.

Denote, for simplicity

$K_u := K(u, \leq)$ (hence, $K_u \subseteq \mathrm{Dom}(\varphi)$).

Clearly, $(\leq)$ is an order on $\mathrm{Dom}(\varphi)$; hence, all the more on K_u; in addition (as φ is d-lsc on K), we must have

K_u is d-closed in K; so that, K_u is d-closed in X as well.

Define a new function $\psi : X \to R \cup \{\infty\}$, as

$\psi(x) = \varphi(x)$, if $x \in K_u$; $\psi(x) = \infty$, if $x \in X \setminus K_u$.

Clearly, ψ is proper and bounded from below (over X); because

$\mathrm{Dom}(\psi) = K_u \neq \emptyset$ (in view of $u \in K_u \subseteq \mathrm{Dom}(\varphi)$)
$\inf \psi(X) = \inf \psi(K_u) = \inf \varphi(K_u) \geq \inf \varphi(K) > -\infty$.

Moreover, ψ is lsc (on X); since K_u is closed; hence, summing up, ψ is (X, d)-regular. Further, let the generalized Banach space $(\mathscr{C}(X), ||.||_\infty)$ and the standard Banach space $(\mathscr{C}_b(X), ||.||_\infty)$ be introduced as above; and $\mathscr{A}$ stand for the subspace of $\mathscr{C}_b(X)$ consisting of all bounded d-Lipschitz (hence, d-continuous) functions $g : X \to R$, endowed with the norm

$$||g||_{\mathscr{A}} = ||g||_\infty + ||g||_L \text{ (hence } ||g||_{\mathscr{A}} \geq ||g||_\infty), \quad g \in \mathscr{A};$$

where

$$||g||_L = \sup\{|g(x) - g(y)|/||x - y||; x, y \in X, x \neq y\}, \quad g \in \mathscr{A}.$$

Clearly, the regularity conditions (e01)–(e04) hold for $\mathscr{A}$; moreover, the function

$$b(x) = d(x, X^c(0,1))/[1 + d(x, X^c(0,1))], \quad x \in X$$

is an element of $\mathscr{A}$, with

supp$(b) = X[0,1]$ (=the closed unit sphere in X);

hence, b is a bump function over X and (e05) holds too. Summing up, $(\mathscr{A}, ||.||_{\mathscr{A}})$ is an admissible subspace of $\mathscr{C}_b(X)$; wherefrom, (DGZ-b) applies to $(X, ||.||)$, $(\mathscr{A}, ||.||_{\mathscr{A}})$ and ψ. As a consequence, there exists a dense $\mathscr{G}_\delta$ type subset $\mathscr{B}$ of $\mathscr{A}$, such that:

for each $g \in \mathscr{B}$, the perturbed function $\psi + g$ is (X, d)-admissible.

In particular, letting $\mathscr{A}(0,1)$ stand for the open unit sphere in $\mathscr{A}$, we must have $\mathscr{B} \cap \mathscr{A}(0,1) \neq \emptyset$; so that, there must be some $g \in \mathscr{A}$ with

(61-1) $||g||_{\mathscr{A}} < 1$ (hence $||g||_\infty, ||g||_L < 1$)
(61-2) $\psi + g$ is (X, d)-admissible (see above).

By the very definition of this concept, the perturbed function $\psi + g$ admits a global minimum point $x_\infty \in X$; i.e.

$$\psi(x) + g(x) \geq \psi(x_\infty) + g(x_\infty), \forall x \in X.$$

Taking $x = u$ in this relation, we derive $x_\infty \in \text{Dom}(\psi) = K_u$; which tells us (by the very definition of the underlying subset) that (61-a) holds (with x_∞ in place of v). Further, assume by contradiction that (61-b) would be false: there must be some $x \in K \setminus \{x_\infty\}$, with

$$||x_\infty - x|| \leq \varphi(x_\infty) - \varphi(x).$$

We then have $x_\infty \leq x$; so, combining with $u \leq x_\infty$, one gets $u \leq x \in K$; whence, $x \in K_u = \text{Dom}(\psi)$; which allows us to write the preceding relation as

$$||x - x_\infty|| + \psi(x) - \psi(x_\infty) \leq 0.$$

On the other hand, by the choice of g, we have

$$|g(x_\infty) - g(x)| \leq ||g||_L \cdot ||x_\infty - x|| < ||x_\infty - x||;$$

so, combining with our global minimum point relation (and $x \in \text{Dom}(\psi)$)

$$\psi(x) - \psi(x_\infty) + ||x - x_\infty|| > \psi(x) - \psi(x_\infty) + g(x) - g(x_\infty) \geq 0.$$

This, however, contradicts the preceding relation; and proves our claim.

(B) The obtained normed type variational principle (EVP-n) is just an intermediary step towards the final inclusion (DGZ-b) $\Longrightarrow$ (EVP). For technical reasons, it would be useful to complete it with another intermediary step of metrical nature.

Let (M, d) be a complete metric space with

(f01) M is d-bounded (i.e.: diam$(M) < \infty$);

and $\varphi : M \to R$ be a function with

(f02) φ is (bilaterally) bounded:
$-\infty < \varphi_* := \inf \varphi(M) \leq \varphi^* := \sup \varphi(M) < \infty$
(f03) φ is d-lsc: $\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \xrightarrow{d} x$.

Clearly, φ is proper; because $\mathrm{Dom}(\varphi) = M$; whence, φ is (M, d)-regular (see above). The following "bounded" version of (EVP) (referred to as: (EVP-b)) is an useful tool for the question we are dealing with.

Theorem 9 *Let (M, d) and φ be taken as before. Then, for each $u \in M$ there exists some $v \in M$ with the properties*

(62-a) $d(u, v) \leq \varphi(u) - \varphi(v)$ *(hence* $\varphi(u) \geq \varphi(v)$*)*
(62-b) $d(v, x) > \varphi(v) - \varphi(x)$, *for all* $x \in M \setminus \{v\}$.

Formally, (EVP-b) is deductible from (EVP). But, the reciprocal inclusion is also true; as results from

Proposition 21 *We have in the strongly reduced system (ZF-AC),*

$$\textit{(EVP-b)} \implies \textit{(EVP); hence (EVP-b)} \iff \textit{(EVP).}$$

Proof Let the premises of (EVP) be assured; precisely, let (X, d) be a complete metric space and $\varphi : X \to R \cup \{\infty\}$ be (X, d)-regular. Further, take some point $u \in \mathrm{Dom}(\varphi)$. Let $(\leq)$ stand for the Brøndsted quasi-order [7]

$(x, y \in X)$: $x \leq y$ iff $d(x, y) + \varphi(y) \leq \varphi(x)$;

remember that it is an ordering on $\mathrm{Dom}(\varphi)$. Denote $M := X(u, \leq)$; it is a closed (hence complete) subset of X. Let again φ stand for the restriction of φ on M; it is (bilaterally) bounded, since

$-\infty < \varphi_* \leq \varphi(x) \leq \varphi(u) < \infty, \forall x \in M.$

This, along with

$d(u, x) \leq \varphi(u) - \varphi(x) \leq \varphi(u) - \varphi_*, \forall x \in M,$

tells us that M is a d-bounded part of X. Summing up, (EVP-b) is applicable to (M, d) and φ. It gives us, for the starting $u \in M$, some $v \in M$ with the properties

(62-1) $u \leq v$ (i.e.: $d(u, v) \leq \varphi(u) - \varphi(v)$)
(62-2) $v \leq x$ is false (i.e.: $d(v, x) > \varphi(v) - \varphi(x)$), for all $x \in M \setminus \{v\}$.

The former of these is just the first conclusion (11-a) of (EVP); so, it remains to establish that the latter of these yields the second conclusion (11-b) of (EVP). Suppose that this would be false:

$v \leq x$ (i.e.: $d(v, x) \leq \varphi(v) - \varphi(x)$), for some $x \in X \setminus \{v\}$.

Combining with $u \leq v$ yields $u \leq x$; whence, $x \in M \setminus \{v\}$; and this, along with the preceding relations, yields a contradiction to (62-2); hence the claim.

(C) We are now in position to give an appropriate answer to the posed question. As already noted, (EVP-n) is but a particular case of (EVP); hence (see above) it is deductible from (EVP-b). The converse implication is discussed in

Proposition 22 *We have, in the strongly reduced system (ZF-AC),*

(63-1) (EVP-n) $\Longrightarrow$ *(EVP-b)* $\Longrightarrow$ *(EVP)*
(63-2) (EVP-n), (EVP-b) and (EVP) are equivalent.

Proof

(i) Let the premises of (EVP-b) be assured; i.e.: take a bounded complete metric space (M, d) and a (bilaterally) bounded d-lsc function $\varphi : M \to R$. Let $X = \mathscr{C}_b(M)$ stand for the linear space of all continuous bounded over M functions. The supremum norm

$$||y||_\infty = \sup\{|y(t)|; t \in M\}, y \in X$$

is complete, in the sense: its associated metric

$$d_\infty(x, y) = ||x - y||_\infty, x, y \in X$$

is complete upon X. For each $z \in M$, let $T(z)$ stand for the real function (defined over M)

$$T(z)(t) = d(z, t), t \in M \text{ (hence, } T(z) \in X = \mathscr{C}_b(M)).$$

The canonical map $z \mapsto T(z)$ fulfills

$$d(z, w) = d_\infty(T(z), T(w)), \forall z, w \in M;$$

so, it is an isometry between M and its image $K := T(M)$ in X. As a consequence,

(63-3) K is bounded and closed (hence complete) in X (modulo d_∞).

In fact, assume that the sequence $(y_n = T(x_n); n \geq 0)$ in K fulfills

$$\lim_n d_\infty(y_n, y) = 0, \text{ for some } y \in X.$$

As (y_n) is d_∞-Cauchy, so is (x_n) (by the isometry relation); wherefrom, there must be some $x \in M$, with

$$\lim_n d(x_n, x) = 0; \text{ or, equivalently: } \lim_n d_\infty(T(x_n), T(x)) = 0.$$

This (by the uniqueness of our limit) yields $y = T(x) \in K$; hence the claim. Define the associated function (from K to R) as: for each $y \in K$,

$$\psi(y) = \varphi(T^{-1}(y)); \text{ i.e.: } \psi(y) = \varphi(x), \text{ where } T(x) = y.$$

By the imposed conditions,

(63-4) ψ is (bilaterally) bounded over K
(63-5) ψ is d_∞-lsc over K (in the above described sense).

The former of these is clear, by the corresponding property of φ. For the latter one, we may proceed as follows. Assume that

$$\psi(y_n) \leq \tau, \forall n, \text{ for some } \tau \in R;$$

where $(y_n = T(x_n); n \geq 0)$ is a sequence in K with $\lim_n d_\infty(y_n, y) = 0$, for some $y = T(x)$ in K (see above). By definition, this gives

$$(\varphi(x_n) \leq \tau, \forall n), \text{ and } x_n \xrightarrow{d} x;$$

hence (as φ is d-lsc) $\varphi(x) \leq \tau$; (i.e.: $\psi(y) \leq \tau$); and our assertion follows.

Summing up, (EVP-n) applies to $(X, ||.||)$, K and ψ; so, given the starting point $p = T(u)$ of K there exists another one $q = T(v)$ of K with the properties

(63-6) $d_\infty(p, q) \leq \psi(p) - \psi(q)$,
(63-7) $d_\infty(q, y) > \psi(q) - \psi(y), \forall y \in K \setminus \{q\}$.

But then, (61-a) and (61-b) are clear, by the properties of canonical map T (and the definition of ψ); wherefrom, our first conclusion follows.

(ii) Clearly, (EVP) $\Longrightarrow$ (EVP-n); and this establishes our second conclusion.

Finally, note that, as (EVP-b) $\Longrightarrow$ (EVP), we must have (EVP-n) $\Longrightarrow$ (EVP). A direct argument for this is available, under the lines in Daneš [11] and Turinici [30]; we do not give details.

7 (EVP) Implies (DC)

By the developments above, we have [in the strongly reduced Zermelo-Fraenkel system (ZF-AC)] the inclusions

$$((\text{DC}) \Longleftrightarrow) \text{ (BC)} \Longrightarrow \text{(LR-ndiscr)} \Longrightarrow \text{(LR-ndiscr-m)}$$

$$\text{(LR-ndiscr-m)} \Longrightarrow \text{(LR-VP-m)} \Longrightarrow \text{(DGZ-b)}$$

$$\text{(BC)} \Longrightarrow \text{(DGZ-psb)} \Longrightarrow \text{(DGZ-sb)} \Longrightarrow \text{(DGZ-b)}$$

$$\text{(DGZ-b)} \Longrightarrow \text{(EVP-n)} \Longrightarrow \text{(EVP-b)} \Longrightarrow \text{(EVP)}.$$

So, it is natural asking whether these inclusion chains may be reversed. At a first glance, a negative answer is highly expectable; because (DC) is "too general" with respect to (EVP). However, the situation is exactly opposite; i.e.: (EVP) includes (DC); and then, we closed the circle between all such principles. An early result of this type was provided in 1987 by Brunner [8]; for a different answer to the same, we refer to the 1999 paper by Dodu and Morillon [14]. It is our aim in the following to show that a further extension of this last result is possible, in the sense: (DC)

is deductible from a certain Lipschitz bounded countable version of (EVP). Some other aspects occasioned by these developments are also discussed.

Let $(X, \leq)$ be a partially ordered structure. Remember that $z \in X$ is $(\leq)$-*maximal*, if $z \leq w \in X$ implies $z = w$; the class of all these will be denoted as $\max(X, \leq)$. In this case, we say that $(\leq)$ is a *Zorn order* when

$\max(X, \leq)$ is (nonempty and) cofinal in X
(for each $u \in X$, there exists $v \in \max(X, \leq)$ with $u \leq v$).

In particular, when $d(.,.)$ is a *(standard) metric* on X and $\varphi : X \to R_+$ is some function, a good example of partial order on X is that introduced by the convention

$$x \leq_{(d,\varphi)} y \text{ iff } d(x, y) \leq \varphi(x) - \varphi(y);$$

referred to as the *Brøndsted order [7]* attached to the couple (d, φ). Further, let us say that φ is *d-Lipschitz*, provided

$$|\varphi(x) - \varphi(y)| \leq Ld(x, y), \forall x, y \in X, \text{ for some } L > 0;$$

note that, any such function is uniformly continuous on X.

The following stronger variant of (EVP) enters in our discussion.

Theorem 10 *Let the metric space (X, d) and the function $\varphi : X \to R_+$ satisfy*

(g01) X is d-bounded and d-complete
(g02) φ is d-Lipschitz (hence, bounded)
(g03) $\varphi(X)$ is (at most) countable.

Then, $(\leq_{(d,\varphi)})$ is a Zorn order.

We call this, the *Lipschitz bounded countable* version of (EVP) (in short: (EVP-Lbc)). By the above developments, we thus have

$$\text{(DC)} \implies \text{(EVP)} \implies \text{(EVP-Lbc)}.$$

The remarkable fact to be added is that this last principle yields (DC); so—as precise—it completes the circle between all these.

Proposition 23 *We have, in the strongly reduced system (ZF-AC)*

(EVP-Lbc) $\implies$ (DC).

As a consequence of this,

(71-1) the variational principles (BC), (LR-ndiscr), (LR-ndiscr-m), (LR-VP-m), (DGZ-psb), (DGZ-sb), (DGZ-b), (EVP-n), (EVP-b) are equivalent with both (DC) and (EVP); hence, mutually equivalent.
(71-2) any maximal/variational principle (VP) with (DC) $\implies$ (VP) $\implies$ (EVP) is equivalent with both (DC) and (EVP).

The proof of this result may be found in Turinici [33]. However, for completeness reasons, we shall provide the argument, with certain modifications.

Proof There are several steps to be followed.

Part 0 Let M be a nonempty set; and $\mathscr{R}$ be a proper relation over it. Fix $a \in M$; and take some other point α, that does not belong to M. Put $P = M \cup \{\alpha\}$; and let $d(.,.)$ stand for the discrete metric on P:

$$d(s,t) = 0, \text{ if } s = t;\ d(s,t) = 1, \text{ if } s \neq t.$$

(In fact, $d(.,.)$ is even an *ultrametric* on P; but this is not essential for us).

Part 1 Let $\mathscr{S}(P)$ stand for the class of all sequences $x = (x(n); n \geq 0)$ with elements in P. Denote $X = \{x \in \mathscr{S}(P); x(0) = a\}$; and let us introduce the map

$$d_\infty(x, y) = \sum_n 2^{-n} d(x(n), y(n)),\ \text{for } x = (x(n)) \text{ and } y = (y(n)) \text{ in } X.$$

It is not hard to see that d_∞ acts as a (standard) metric on X; moreover,

$$d_\infty(x, y) \leq \sum_{n \geq 1} 2^{-n} = 1, \forall x, y \in X;\ \text{whence } X \text{ is } d_\infty\text{-bounded.}$$

A natural question to be discussed here is the completeness property. In this direction, we have

Proposition 24 *Under the above conventions, the metric space (X, d_∞) is complete: each d_∞-Cauchy sequence in X is d_∞-convergent.*

Proof A verification of this was already performed in a previous place; however, for technical reasons, we shall remember its basic lines.

Let $(x^n; n \geq 0)$ be a sequence in X; it may be written as

$$(x^n = (x^n(0), x^n(1), \ldots) = (a, x^n(1), \ldots); n \geq 0).$$

Assume that $(x^n; n \geq 0)$ is d_∞-Cauchy; this may be also characterized as:

$$\forall \varepsilon > 0\text{: } C(\varepsilon) := \{n \in N; n \leq p \leq q \Longrightarrow d_\infty(x^p, x^q) < \varepsilon\} \neq \emptyset.$$

(I) As a consequence, the map $\varepsilon \mapsto C(\varepsilon)$ is increasing on R^0_+, in the sense:

$$\varepsilon_* < \varepsilon^* \text{ implies } C(\varepsilon_*) \subseteq C(\varepsilon^*);$$

so that, the map $\varepsilon \mapsto \Gamma(\varepsilon) := \min[C(\varepsilon)]$ is decreasing on R^0_+:

$$\varepsilon_* < \varepsilon^* \text{ implies } \Gamma(\varepsilon_*) \geq \Gamma(\varepsilon^*).$$

(II) Let $(\varepsilon_n; n \geq 0)$ be a strictly descending sequence in R^0_+ with

$$\varepsilon_n < 2^{-n}, \forall n\ \text{(hence: } \varepsilon_n \to 0 \text{ as } n \to \infty).$$

Denote for simplicity

$$m(k) = \Gamma(\varepsilon_k), n(k) = m(k) + k,\ k \geq 0.$$

By the properties above, the map $k \mapsto m(k)$ is increasing; hence, the map $k \mapsto n(k)$ is strictly increasing. Moreover, the following evaluation is valid

$$x^{n(i)}(i) = x^p(i),\ \text{ for all } p \geq n(i) \text{ and all } i \geq 0.$$

(III) Let $y = (y(i); i \geq 0)$ be the "diagonal" sequence $(y(i) = x^{n(i)}(i); i \geq 0)$; clearly, it is an element of X. The initial sequence $(x^n; n \geq 0)$ is convergent (modulo d_∞) to y; i.e.,

$$d_\infty(x^n, y) \to 0, \text{ as } n \to \infty;$$

and, from this, we are done.

Part 2 Let Y stand for the class of all sequences $x = (x(n); n \geq 0)$ in X, with

$$(\forall n):\ x(n), x(n+1) \in M \Longrightarrow x(n) \mathscr{R} x(n+1).$$

Note that $Y \neq \emptyset$; for, given $b \in \mathscr{R}(a)$, the sequence $y = (y(n); n \geq 0)$ in X introduced as below

$$(y(0) = a, y(1) = b; y(n) = \alpha, n \geq 2)$$

is an element of it.

Proposition 25 *The subset Y is d_∞-closed; hence, d_∞-complete as well.*

Proof Let $(x^n := (x^n(0) = a, x^n(1), \ldots); n \geq 0)$ be a sequence in Y, and $y = (y(n); n \geq 0)$ be an element of X with

$x^n \to y$ (modulo d_∞); that is,
$d_\infty(x^n, y) := \sum_i 2^{-i} d(x^n(i), y(i)) \to 0$, as $n \to \infty$.

Note that, as a direct consequence of this,

$$x^n(i) \xrightarrow{d} y(i) \text{ as } n \to \infty,\ \forall i \geq 0.$$

Further, as d_∞=metric, $(x^n; n \geq 0)$ is d_∞-Cauchy; so, by a preceding completeness statement, there exists a strictly ascending sequence of ranks $(n(i); i \geq 0)$, with

$$(\forall i \geq 0):\ \ x^{n(i)}(i) = x^p(i),\ \ \forall p \geq n(i).$$

In this case, the d_∞-limit $y = (y(n); n \geq 0)$ of our sequence must have the form

$$y(i) = x^{n(i)}(i), \text{ for all } i \geq 0.$$

We now claim that the representation of Y gives us the desired conclusion: $y \in Y$. In fact, let $i \geq 0$ be such that $y(i), y(i+1) \in M$. By the previous relations,

$$y(i) = x^{n(i)}(i) = x^{n(i+1)}(i) \in M;\ \ y(i+1) = x^{n(i+1)}(i+1) \in M.$$

This, along with $x^{n(i+1)} \in Y$, yields

$$x^{n(i+1)}(i) \mathscr{R} x^{n(i+1)}(i+1);\ \text{that is,}\ \ y(i) \mathscr{R} y(i+1).$$

The argument is thereby complete.

Part 3 Now, let us note that, conclusion of our statement is equivalent with $Y \cap \mathscr{S}(M) \neq \emptyset$. For, taking some sequence $y = (y(n); n \geq 0)$ in this intersection, we have $y(n), y(n+1) \in M$, $\forall n$; so that, by definition, $y(n) \mathscr{R} y(n+1)$, $\forall n$; whence, $(y(n); n \geq 0)$ is $(a, \mathscr{R})$-iterative. Assume by contradiction that this is not true:

$Y \cap \mathscr{S}(M) = \emptyset$; i.e.: for each $y = (y(n); n \geq 0) \in Y$,
there exists some $k = k(y) \geq 1$, such that $y(k) = \alpha$.

As a consequence, the functions below are well defined:

$$g(y) = \min\{k \geq 1; y(k) = \alpha\}, \varphi(y) = 2^{2-g(y)}, y \in Y.$$

Some basic properties of these are described in

Proposition 26 *The following are valid:*

(74-1) *the functions g, φ are continuous on Y; precisely,*

$$\forall y \in Y, \exists \beta = \beta(y) > 0 : z \in Y, \ d_\infty(z, y) < \beta \Longrightarrow g(z) = g(y), \ \varphi(z) = \varphi(y)$$

(74-2) *the function φ is d_∞-Lipschitz, in the sense:*

$$|\varphi(x) - \varphi(y)| \leq 4 d_\infty(x, y), \quad \forall x, y \in Y$$

(74-3) *$g(Y)$ is countable; hence, so is $\varphi(Y)$.*

Proof

(i) Fix $y = (y(n); n \geq 0) \in Y$, and put $r = g(y)$; we therefore have

$$r \geq 1, y(r) = \alpha, y(k) \in M, \forall k \in N(r, >).$$

Take some $\beta \in]0, 2^{-r}[$; and let $z = (z(n); n \geq 0) \in Y$ be such that $d_\infty(y, z) < \beta$. By the definition of our metric,

$$2^{-k} d(y(k), z(k)) < \beta < 2^{-r}, \quad \forall k \in N(r, \geq);$$

and this yields

$$z(k) = y(k), \ \forall k \in N(r, \geq).$$

In particular, we must have

$$(z(k) \in M, \quad \forall k \in N(r, >)) \text{ and } z(r) = \alpha;$$

so that $g(z) = r = g(y)$ (whence, $\varphi(z) = \varphi(y)$).

(ii) Let $x = (x(n); n \geq 0)$ and $y = (y(n); n \geq 0)$ be two points in Y. Denote, for simplicity $r = g(x)$, $s = g(y)$. If $r = s$, all is clear; so, it remains the opposite case $r \neq s$; without loss, one may assume that $r < s$. As a consequence,

$$x = (x(0) = a, \ldots, x(r-1), \alpha, \ldots, x(s-1), x(s), \ldots),$$
$$y = (y(0) = a, \ldots, y(r-1), y(r), \ldots y(s-1), \alpha, \ldots).$$

In particular, $y(r) \in M$; hence $y(r) \neq \alpha$; and then,

$$d_\infty(x, y) \geq 2^{-r} \geq 2^{-r} - 2^{-s} = |2^{-r} - 2^{-s}|.$$

This gives the conclusion we need.

(iii) Evident.

Part 4 We show that, under the introduced conventions,

for each $v \in Y$ there exists $y \in Y \setminus \{v\}$ such that $d_\infty(v, y) \leq \varphi(v) - \varphi(y)$;

or, in other words: each element of Y is non-maximal with respect to the Brøndsted ordering $(\leq)$ attached to d_∞ and φ:

$(z, w \in Y)$: $z \leq w$ iff $d_\infty(z, w) \leq \varphi(z) - \varphi(w)$.

In fact, let $v = (v(n); n \geq 0)$ be its representation. Put $g(v) = r$; hence

$$r \geq 1;\ v(0), \ldots, v(r-1) \in M;\ v(r) = \alpha.$$

Take the sequence $y = (y(n); n \geq 0)$ in $X \setminus \{v\}$ according to

$y(k) = v(k), \forall k \in N(r, >)$; $y(h) = \alpha, \forall h \in N(r+1, <)$;
$y(r), y(r+1) \in M$; $y(i) \mathscr{R} y(i+1),\ \forall i \in \{r-1, r\}$.

(The last relation is possible, by the Finite Dependent Choice property). We claim that, necessarily,

$y \in Y$; hence, $y \in Y \setminus \{v\}$.

The assertion is evident in case of $r = 1$. For the case of $r \geq 2$, it will suffice noting that, by the definition of Y, one gets the relations

$v(i) \mathscr{R} v(i+1)$, whenever $i \leq r-2$;

and this proves our claim. Note that, as a direct consequence of these facts, we must have $g(y) = r + 2$. Now, the desired relation above becomes

$$d_\infty(v, y) \leq 2^{2-r} - 2^{-r} = 3 \cdot 2^{-r}.$$

According to the representation of $y \in Y \setminus \{v\}$, this means

$$\sum_{i \geq r} 2^{-i} d(v(i), y(i)) \leq 3 \cdot 2^{-r}.$$

But then, the last relation is clear, in view of

$$\sum_{i\geq r} 2^{-i} d(v(i), y(i)) \leq \sum_{i\geq r} 2^{-i} = 2^{1-r} < 3 \cdot 2^{-r}.$$

Part 5 We may now pass to the final part of our argument. By the above facts, (EVP-Lbc) is applicable to the metric space (Y, d_∞) and the function $\varphi : Y \to R_+$ (introduced as before). Hence, the associated Brøndsted order $(\leq)$ (see above) is a Zorn one. As a consequence, there exists, for the starting point (in Y)

$$u = (u(n); n \geq 0): \ u(0) = a, u(n) = \alpha, \forall n \geq 1,$$

some other point $v = (v(n); n \geq 0)$ in Y with

(zorn-1) $u \leq v$: $d_\infty(u, v) \leq \varphi(u) - \varphi(v)$
(zorn-2) v is $(\leq)$-maximal: $d_\infty(v, y) > \varphi(v) - \varphi(y), \forall y \in Y \setminus \{v\}$.

This, however, contradicts the preceding step; and shows that $Y \cap \mathscr{S}(M) \neq \emptyset$. But then (by the very definition of Y) there must be some sequence $y = (y(n); n \geq 0)$ in M, with

$$y(0) = a \text{ and } y(n) \mathscr{R} y(n+1), \ \forall n.$$

The proof is complete.

In particular, when the boundedness and Lipschitz properties are ignored, this result is just the one in Dodu and Morillon [14]. Further aspects may be found in the paper by Turinici [32].

Summing up, all variational principles in this exposition (derived from (DC)) are nothing but logical equivalents of (EVP). So, it is natural to ask whether the remaining (sequential) ones—including the Smooth Variational Principle in Deville and Ghoussoub [12]—are endowed as well with such a property. The answer to this is affirmative; further aspects will be delineated elsewhere.

References

1. T.Q. Bao, P.Q. Khanh, Are several recent generalizations of Ekeland's variational principle more general than the original principle? Acta Math. Vietnam. **28**, 345–350 (2003)
2. P. Bernays, A system of axiomatic set theory: Part III. Infinity and enumerability analysis. J. Symb. Log. **7**, 65–89 (1942)
3. C.E. Blair, The Baire category theorem implies the principle of dependent choice. Bull. Acad. Pol. Sci. (Sér. Math. Astronom. Phys.) **10**, 933–934 (1977)
4. J.M. Borwein, D. Preiss, A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions. Trans. Am. Math. Soc. **303**, 517–527 (1987)
5. N. Bourbaki, *General Topology*, chaps. 1–4 (Springer, Berlin, 1989)
6. H. Brezis, F.E. Browder, A general principle on ordered sets in nonlinear functional analysis. Adv. Math. **21**, 355–364 (1976)
7. A. Brøndsted, Fixed points and partial orders. Proc. Am. Math. Soc. **60**, 365–366 (1976)
8. N. Brunner, Topologische Maximalprinzipien. Zeitschr. Math. Logik Grundl. Math. **33**, 135–139 (1987)

9. O. Cârjă, M. Necula, I. I. Vrabie, *Viability, Invariance and Applications*. North Holland Mathematics Studies, vol. 207 (Elsevier, Amsterdam, 2007)
10. P.J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966)
11. J. Daneš, Equivalence of some geometric and related results of nonlinear functional analysis. Comment. Math. Univ. Carol. **26**, 443–454 (1985)
12. R. Deville, N. Ghoussoub, Perturbed minimization principles and applications, in Handbook of the Geometry of Banach Spaces, ed. by W.B. Johnson, J. Lindenstrauss, vol. I, chap. 10 (Elsevier, Amsterdam, 2001), pp. 399–435
13. R. Deville, G. Godefroy, V. Zizler, A smooth variational principle with applications to Hamilton-Jacobi equations in infinite dimensions. J. Funct. Anal. **111**, 197–212 (1993)
14. J. Dodu, M. Morillon, The Hahn-Banach property and the axiom of choice. Math. Log. Q. **45**, 299–314 (1999)
15. I. Ekeland, On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)
16. I. Ekeland, Nonconvex minimization problems. Bull. Am. Math. Soc. (New Series) **1**, 443–474 (1979)
17. A. Goepfert, H. Riahi, C. Tammer, C. Zălinescu, *Variational Methods in Partially Ordered Spaces*. Canadian Mathematical Society Books in Mathematics, vol. 17 (Springer, New York, 2003)
18. C.F.K. Jung, On generalized complete metric spaces. Bull. Am. Math. Soc. **75**, 113–116 (1969)
19. O. Kada, T. Suzuki, W. Takahashi, Nonconvex minimization theorems and fixed point theorems in complete metric spaces. Math. Jpn. **44**, 381–391 (1996)
20. C. Kuratowski, *Topologie (vol. I)*, Mathematical Monographs, vol. 20 (Polish Scientific Publishers, Warsaw, 1958)
21. M. Lassonde, J.P. Revalski, Fragmentability of sequences of set-valued mappings with applications to variational principles. Proc. Am. Math. Soc. **133**, 2637–2646 (2005)
22. Y. Li, S. Shi, A generalization of Ekeland's ε-variational principle and its Borwein-Preiss smooth variant. J. Math. Anal. Appl. **246**, 308–319 (2000)
23. W.A.J. Luxemburg, On the convergence of successive approximations in the theory of ordinary differential equations (II). Indag. Math. **20**, 540–546 (1958)
24. G.H. Moore, *Zermelo's Axiom of Choice: its Origin, Development and Influence* (Springer, New York, 1982)
25. Y. Moskhovakis, *Notes on Set Theory* (Springer, New York, 2006)
26. J.C. Oxtoby, *Measure and Category* (Springer, New York, 1980)
27. E. Schechter, *Handbook of Analysis and its Foundation* (Academic Press, New York, 1997)
28. A. Tarski, Axiomatic and algebraic aspects of two theorems on sums of cardinals. Fundam. Math. **35**, 79–104 (1948)
29. M. Turinici, A monotone version of the variational Ekeland's principle. An. Şt. Univ. "A. I. Cuza" Iaşi (S. I-a: Mat) **36**, 329–352 (1990)
30. M. Turinici, An order drop theorem. Le Matematiche **61**, 213–230 (2006)
31. M. Turinici, Variational statements on KST-metric structures. An. Şt. Univ. "Ovidius" Constanţa (Mat.) **17**, 231–246 (2009)
32. M. Turinici, *Sequential Maximality Principles*, in Mathematics Without Boundaries, ed. by T.M. Rassias, P.M. Pardalos (Springer, New York, 2014), pp. 515–548
33. M. Turinici, *Variational Principles in Gauge Spaces*, in Optimization in Science and Engineering, ed. by T.M. Rassias, C.A. Floudas, S. Butenko (Springer, New York, 2014), pp. 503–542
34. S. Willard, *General Topology* (Addison-Wesley, Don Mills, 1970)
35. E.S. Wolk, On the principle of dependent choices and some forms of Zorn's lemma. Can. Math. Bull. **26**, 365–367 (1983)